La raccolta di:

Matematica

il giornalino degli OpenDays

...notizie, giochi e pillole di matematica

Dipartimento di Matematica

Introduzione

Questo volume è una raccolta degli articoli comparsi nei *Giornalini degli Open days*, prodotti due volte l'anno in occasione degli eventi di orientamento promossi dal Dipartimento di Matematica dell'Università di Pisa e rivolti agli studenti delle scuole superiori. Il volume è "in progress", in quanto verrà via via aggiornato con ulteriori articoli, in occasione della pubblicazione dei nuovi numeri del giornalino. Potete trovare le versioni elettroniche complete dei vari giornalini fin qui prodotti nella sezione Orientamento del sito del Dipartimento di Matematica, all'indirizzo `https://www.dm.unipi.it/webnew/it/orientamento/il-giornalino-degli-open-days`. Potete inoltre contattare gli studenti counseling all'indirizzo mail *counseling.dm.unipi@gmail.com*.
Buona lettura!

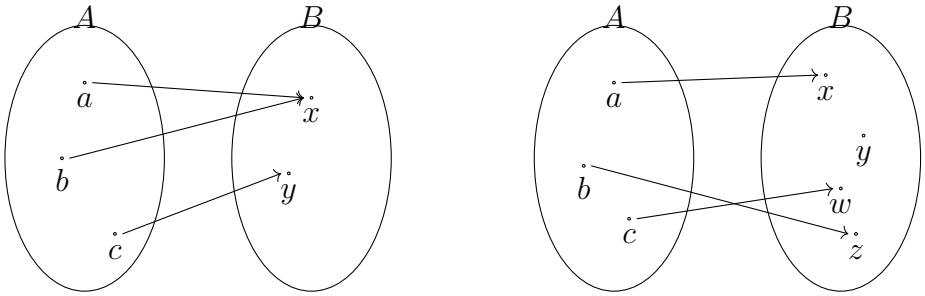
Indice

1 Contare gli Infiniti

Dario Villanis Ziani, n.0, Febbraio 2015

Tutti noi, fin dalle scuole elementari, siamo in grado di contare un numero finito di oggetti, e di dire se un insieme abbia più elementi di un altro: ad esempio, l'insieme dei giorni in un anno ha più elementi dell'insieme dei mesi in un anno, l'insieme dei giocatori in una partita di calcio ha più elementi dell'insieme degli arbitri nella stessa partita, e così via. In linguaggio matematico il numero di elementi di un insieme A si dice *cardinalità* e si denota con $|A|$ (o talvolta $\#A$). Così ad esempio l'insieme dei giorni in un anno ha cardinalità 365 (se non bisestile), l'insieme dei mesi in un anno ha cardinalità 12, eccetera. Ma quando si tratta di passare ad insiemi infiniti, le cose si fanno più complicate. Quanti infiniti esistono? Gli infiniti sono tutti uguali, tutti diversi, alcuni uguali ed alcuni diversi? Ad esempio quanti sono i numeri naturali? Sono tanti quanti i numeri interi? Sebbene la risposta possa sembrare semplice, non è affatto così; certamente non possiamo più contare direttamente gli elementi. Però è lecito chiedersi quanto siano "grandi" gli insiemi, pertanto dobbiamo tentare di definire il concetto di cardinalità in altro modo.

Dovrebbe essere già noto al lettore il concetto di funzione, che comunque richiamiamo brevemente. Siano A e B due insiemi; una *funzione* associa ad ogni elemento dell'insieme A *uno e un solo* elemento dell'insieme B: si dice funzione *definita su A a valori in B* e si denota $f\colon A \longrightarrow B$. Niente vieta che a due diversi elementi di A sia associato lo stesso elemento dell'insieme B, si pensi ad esempio alla funzione elevamento al quadrato definita su tutti i numeri reali: il quadrato di 3 è uguale al quadrato di -3. Non è neanche detto che un qualunque elemento di B possa essere raggiunto mediante la funzione a partire da un elemento di A: pensando ancora alla funzione $f(x) = x^2$ con $A = B = \mathbb{R}$, un numero negativo non è quadrato di alcun numero reale. Diciamo che una funzione è *iniettiva* se, dati due diversi elementi x ed x' di A, si ha $f(x) \neq f(x')$. Una funzione invece è *suriettiva* se per ogni elemento y di B esiste *almeno* un elemento x di A tale che $f(x) = y$. Infine, una funzione è *biunivoca* se per ogni elemento y di B esiste *esattamente* un elemento x di A tale che $f(x) = y$. Nelle due figure che seguono presentiamo due esempi grafici: la figura a sinistra rappresenta una funzione suriettiva ma non iniettiva tra A e B; la figura a destra invece rappresenta una funzione iniettiva ma non suriettiva tra tali due insiemi.



Le funzioni sono il concetto che ci permetterà di estendere in modo naturale la definizione di cardinalità ad insiemi infiniti. Diciamo che un insieme A ha cardinalità minore o uguale a quella di un altro insieme B, e scriviamo $|A| \leq |B|$, se esiste una funzione iniettiva $f : A \longrightarrow B$. Se inoltre esiste una funzione biunivoca da A a B, allora essi hanno la stessa cardinalità, e scriviamo $|A| = |B|$; se infine esiste una funzione iniettiva da A a B ma non è possibile trovare una funzione biunivoca tra essi, allora la cardinalità di A è strettamente minore di quella di B, e scriviamo $|A| < |B|$. Vedremo tra poco che avere un'inclusione stretta tra due insiemi (ossia avere un insieme propriamente incluso in un altro, come i numeri pari ed i numeri naturali) non implica necessariamente che le loro cardinalità siano diverse; anzi, in molti casi non è così.
Intanto vediamo che queste definizioni si accordano con il caso finito di cui abbiamo parlato poco fa. Prendiamo ad esempio l'insieme dei primi dodici numeri naturali $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ (attenzione: se includiamo lo zero nei numeri naturali, i primi n numeri naturali sono $0, 1, \ldots, n-1$) e vediamo che è possibile costruire una funzione biunivoca tra questo e l'insieme dei mesi dell'anno: è facile, basta associare a Gennaio il numero 0, a Febbraio il numero 1, a Marzo il numero 2 è così via, fino a Dicembre che deve essere associato al numero 11. Che la funzione così definita sia biunivoca è piuttosto facile da verificare. Ovviamente potevamo definire in modo completamente diverso la funzione, ad esempio associando a Gennaio il numero 3, a Febbraio il numero 11 e così via, facendo attenzione ad associare ad ogni mese un numero non utilizzato in precedenza (altrimenti non avremmo più avuto una funzione iniettiva). Se proviamo a costruire una funzione definita sull'insieme dei mesi a valori nell'insieme dei giorni di un anno, non possiamo certamente ottenere una funzione biunivoca: in effetti ci sono meno mesi che giorni in un anno. È facile verificare che il confronto tra cardinalità ottenuto mediante le funzioni si accorda bene con il caso finito, e pertanto estende l'intuizione del "contare" gli elementi. Vogliamo porre l'accento sul fatto che costruire una funzione da un insieme ad un altro non ci dice direttamente *quanti* elementi abbia uno specifico insieme, ma piuttosto se ne abbia più o meno di un altro: come già detto si tratta di un *confronto* tra cardinalità.

Siamo dunque pronti per affrontare la questione delle cardinalità infinite. Cominciamo dai numeri naturali. Il matematico tedesco Georg Cantor (1845-1918) ha

chiamato la cardinalità dell'insieme dei numeri naturali $\aleph_0$ ($\aleph$ si legge "alef", ed è la prima lettera dell'alfabeto ebraico). Intanto, possiamo facilmente vedere che $\aleph_0$ è maggiore di qualsiasi numero naturale; se ad esempio prendiamo un insieme con 288 elementi, è facile costruire una funzione iniettiva da tale insieme nell'insieme dei numeri naturali; è invece impossibile costruire una funzione biunivoca tra questi due insiemi, pertanto $288 < \aleph_0$; lo stesso ragionamento funziona con 7, 2.056.786 e con qualsiasi altro numero naturale.

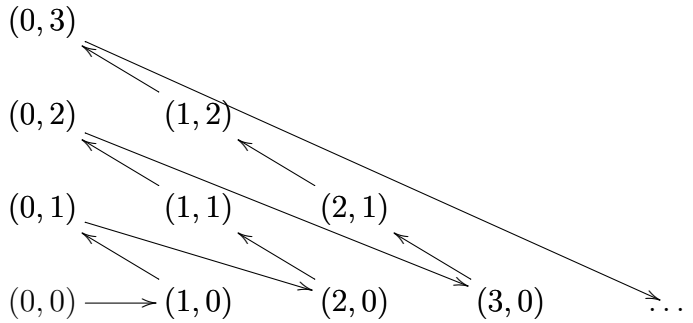
Si tratta adesso di vedere se, prendendo altri insiemi infiniti, le loro cardinalità siano uguali o meno a quella dei numeri naturali. Un insieme di cardinalità uguale a quella dei numeri naturali si dice *numerabile*. Prima di procedere però dobbiamo fare due osservazioni che sembrano banali ma che in realtà nel caso infinito non lo sono affatto. In effetti la prima di queste è un vero e proprio teorema, detto teorema di *Cantor-Bernstein*, il quale ci assicura che la relazione tra cardinalità è simmetrica: se $|X| \leq |Y|$ e se $|Y| \leq |X|$, allora $|X| = |Y|$; in altre parole, se esiste una funzione iniettiva da X ad Y e se esiste una funzione iniettiva da Y a X, allora esiste una funzione biunivoca tra X e Y. La seconda osservazione, decisamente più semplice e che il lettore può provare a dimostrare, assicura che se $|X| \leq |Y|$ e se $|Y| \leq |Z|$, allora $|X| \leq |Z|$; in altre parole, se esiste una funzione iniettiva da X ad Y e se esiste un'altra funzione iniettiva da Y a Z, allora esiste anche una funzione iniettiva da X a Z.

Vediamo adesso, con un racconto molto famoso, cosa può succedere di strano quando consideriamo le cardinalità infinite. L'*albergo di Hilbert* è un albergo molto particolare: ha $\aleph_0$ stanze, tante quante i numeri naturali. In un periodo particolarmente affollato, l'albergo è completamente pieno. Un giorno però arriva un nuovo ospite, che chiede se vi siano stanze libere. "Nessun problema!" risponde il gestore della struttura, facendo spostare i propri ospiti nella stanza successiva a quella che occupano: ossia fa spostare l'ospite della stanza n nella stanza $n + 1$. Così facendo, tutti i vecchi ospiti hanno trovato una nuova sistemazione, ed il nuovo arrivato può prendere possesso della stanza numero 0. Il giorno dopo arrivano k nuovi ospiti, diciamo $k = 100$ per semplicità, ma il ragionamento che stiamo per fare funziona con qualsiasi numero. Anch'essi chiedono se vi sia disponibilità di stanze, ed il gestore, analogamente a quanto fatto il giorno precedente, fa spostare gli ospiti, stavolta dalla stanza n alla stanza $n + 100$; in questo modo le stanze da 0 a 99 vengono liberate ed i nuovi arrivati possono essere accolti. Il giorno successivo arriva una gita molto affollata. "Quanti siete?" chiede il gestore. "Infiniti! Precisamente, numerabili!" risponde la guida. Senza farsi prendere dallo sconforto, il proprietario predispone un nuovo spostamento di stanza: l'occupante della stanza numero n si deve spostare nella stanza numero $2n$. In questo modo tutti i nuovi arrivati possono prendere possesso delle camere con numero dispari.
Questo racconto mostra intanto che l'unione tra l'insieme dei numeri naturali ed

un qualsiasi singoletto (insieme formato da un solo elemento) è ancora numerabile; ad esempio $|\mathbb{N} \cup \{\pi\}| = |\mathbb{N}|$. Ma non solo, in realtà resta numerabile anche l'unione dell'insieme dei numeri naturali ed un qualsiasi insieme finito, cioè $|\mathbb{N} \cup \{a_1, \ldots, a_k\}| = |\mathbb{N}|$ per ogni numero naturale k; ossia esiste una funzione biunivoca tra l'insieme dei numeri naturali e l'insieme dei numeri naturali più un numero finito di elementi. Ma c'è di più: abbiamo anche mostrato che possiamo posizionare infiniti (purchè numerabili) nuovi ospiti nell'albergo sebbene sia già pieno, ossia che $|\mathbb{N} \cup \mathbb{N}| = |\mathbb{N}|$. In realtà possiamo spingerci ancora oltre. Consideriamo l'insieme $\mathbb{N} \times \mathbb{N}$, ossia il prodotto cartesiano di due copie dei numeri naturali; in parole povere, l'insieme $\mathbb{N} \times \mathbb{N}$ è costituito da *coppie* di numeri naturali, ossia i suoi elementi sono della forma $(1, 1)$, $(2, 7)$, $(256, 14)$ e così via. Vogliamo provare che la cardinalità di $\mathbb{N} \times \mathbb{N}$ è uguale alla cardinalità di $\mathbb{N}$. Dobbiamo pertanto costruire una funzione biunivoca tra $\mathbb{N}$ e $\mathbb{N} \times \mathbb{N}$, ed il compito non è facile come i precedenti, ma è comunque alla nostra portata. Posizioniamo gli elementi di $\mathbb{N} \times \mathbb{N}$ su un piano:

$$\begin{array}{cccccc}
 & \vdots & & & \iddots & \\
(0,3) & \vdots & & & & \\
(0,2) & (1,2) & (2,2) & (3,2) & \ldots & \\
(0,1) & (1,1) & (2,1) & (3,1) & \ldots & \\
(0,0) & (1,0) & (2,0) & (3,0) & \ldots &
\end{array}$$

completare essa servono infiniti passi. Allo stesso modo non riusciamo a costruire una funzione biunivoca iniziando dagli elementi della prima colonna. debbano contare soltanto finiti elementi, come ad esempio il cosiddetto *argomento diagonale di Cantor*, che suggerisce di procedere nel seguente modo: Il cosiddetto *argomento diagonale di Cantor* suggerisce di procedere nel modo suggerito dalla seguente figura:



ossia consideriamo prima gli elementi la cui somma delle componenti sia 0 (il solo $(0, 0)$); una volta esauriti questi, consideriamo gli elementi la cui somma delle componenti sia 1, ordinati a partire da quello con prima componente più grande (quindi $(1, 0)$ e $(0, 1)$); poi passiamo agli elementi la cui somma delle componenti sia 2, ordinati come i precedenti (quindi $(2, 0)$, $(1, 1)$, $(0, 2)$); e così via. In questo modo

ad ogni passo dobbiamo contare solo un numero finito di elementi, prima di passare alla diagonale successiva. Si può facilmente verificare che questo procedimento costruisce una funzione biunivoca tra $\mathbb{N}$ e $\mathbb{N} \times \mathbb{N}$. ‘e come se nell'albergo di Hilbert fosse arrivata una quantità numerabile di gite ciascuna composta da una quantità numerabile di persone: c'è posto anche per loro, basta numerare questi nuovi ospiti con una coppia di numeri naturali, dove la prima componente indica il numero della gita alla quale appartengono e la seconda componente indica il partecipante; ad esempio, l'ospite numero $(5, 14)$ è il quattordicesimo partecipante della quinta gita.
Dunque, abbiamo appena visto che l'insieme dei numeri naturali ha la stessa cardinalità dell'insieme delle coppie di numeri naturali, cosa che in effetti sembra piuttosto strana, e che mostra come con le cardinalità infinite possano accadere cose che nel caso finito sono escluse. Nascosta nel racconto della storia dell'albergo di Hilbert, c'è anche la dimostrazione di un altro fatto, ossia che l'insieme dei numeri pari ha la stessa cardinalità dell'insieme dei numeri naturali, sebbene il primo insieme sia strettamente incluso nel secondo: c'è infatti una funzione biunivoca tra questi due insiemi, data da $n \mapsto 2n$.

E i numeri interi (ossia l'insieme $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$) quanti sono? Intuitivamente verrebbe da pensare che siano più dei numeri naturali, ma ormai abbiamo capito che l'intuito che abbiamo sviluppato nel caso finito può non funzionare con insiemi infiniti. Ed in effetti, anche l'insieme dei numeri interi ha la stessa cardinalità dell'insieme dei numeri naturali. Proviamo quindi a costruire una funzione biunivoca tra questi due insiemi. L'idea è quella di associare ai numeri naturali pari gli interi positivi, ed ai numeri naturali dispari gli interi negativi, nel seguente modo: se n è pari, gli associamo il numero $\frac{n}{2}$; se n è dispari, gli associamo il numero $-\frac{n+1}{2}$. Ossia

$$0 \mapsto 0, \qquad 1 \mapsto -1, \qquad 2 \mapsto 1, \qquad 3 \mapsto -2$$

eccetera; anche questa funzione è biunivoca.

Possiamo andare oltre, e dimostrare che i numeri razionali sono tanti quanti i numeri naturali. Iniziamo con il provare che l'insieme dei razionali positivi è numerabile. Per far ciò, ci viene in soccorso quanto detto per il prodotto cartesiano $\mathbb{N} \times \mathbb{N}$, ossia l'insieme delle coppie di numeri naturali: possiamo leggere le due componenti di tale coppia come, rispettivamente, numeratore e denominatore di una frazione (a patto di considerare, al posto dei numeri naturali, i numeri naturali positivi, per evitare eventi spiacevoli come uno zero al denominatore: questo comunque non cambia la sostanza, essendo i numeri naturali tanti quanti i numeri naturali positivi); ad esempio l‘elemento $(2, 1)$ corrisponde alla frazione $\frac{2}{1}$, l‘elemento $(4, 7)$ corrisponde alla frazione $\frac{4}{7}$ e così via. In realtà questi non sono proprio i numeri razionali, ma sono qualcosa in più, perché diverse coppie

possono rappresentare la stessa frazione; in effetti però i numeri razionali si possono ottenere dalle coppie di numeri naturali considerando come equivalenti due coppie che danno luogo ad una stessa frazione, e prendendo quindi un unico rappresentante per ogni famiglia di coppie equivalenti: ad esempio, dato che le coppie $(3, 2)$, $(6, 4)$, $(18, 12)$ ed in generale $(3n, 2n)$ danno luogo tutte alla stessa frazione $\frac{3}{2}$, prendiamo $(3, 2)$ come rappresentante di questa famiglia (numerabile) di coppie equivalenti. Pertanto, l'insieme di queste coppie di rappresentanti dei numeri razionali positivi risulta un sottoinsieme di $\mathbb{N} \times \mathbb{N}$, che sappiamo avere la stessa cardinalità di $\mathbb{N}$, ed è facile dimostrare che un sottoinsieme infinito di un insieme numerabile è ancora numerabile (esercizio per il lettore!). D'altronde, esiste una facile funzione iniettiva definita su $\mathbb{N}$ a valori in $\mathbb{N} \times \mathbb{N}$, pertanto il teorema di Cantor-Bernstein permette di concludere che l'insieme dei razionali positivi ha la stessa cardinalità dell'insieme dei naturali. Passare poi all'insieme di tutti i razionali è facile: è sufficiente provare, in modo analogo a quanto fatto per i numeri naturali e per i numeri interi, che i numeri razionali positivi sono tanti quanti tutti i numeri razionali; oppure basta considerare coppie di numeri in cui la prima componente possa essere negativa, ossia elementi di $\mathbb{Z} \times \mathbb{N}$.

A questo punto la domanda sorge spontanea: ma allora tutti gli infiniti sono uguali? La risposta è no, ed anzi possiamo "salire" di cardinalità in modo piuttosto semplice, ossia considerando l'insieme delle parti. Dato un insieme A, l'*insieme delle parti* di A è l'insieme dei suoi sottoinsiemi e si denota con $\mathcal{P}(A)$. Facciamo degli esempi: l'insieme vuoto, pertanto l'insieme delle parti di A ha esattamente un elemento, l'insieme vuoto, ossia $\mathcal{P}(A) = \{\emptyset\}$ (si faccia attenzione a quanto appena detto: l'insieme vuoto e l'insieme costituito dal solo insieme vuoto sono diversi!); pertanto, se A ha cardinalità zero, il suo insieme delle parti ha cardinalità 1. se A ha cardinalità 1, ad esempio $A = \{a\}$, allora il suo insieme delle parti è $\{\emptyset, \{a\}\}$, in quanto sia l'insieme vuoto che A stesso sono sottoinsiemi di A; pertanto in questo caso $\mathcal{P}(A)$ ha cardinalità 2. Se consideriamo un insieme con due elementi, diciamo $A = \{a, b\}$, allora $\mathcal{P}(A) = \{\emptyset, \{a\}, \{b\}, \{a, b\}\}$. Proseguendo in questo modo si intuisce che se $|A| = n$, allora $|\mathcal{P}(A)| = 2^n$ (lo sapreste dimostrare?). Passando agli insiemi infiniti, si può dimostrare (ma non lo faremo qui) che la cardinalità dell'insieme delle parti è strettamente maggiore della cardinalità dell'insieme di partenza. Dunque, nel caso dei numeri naturali, avremo che $|\mathbb{N}| < |\mathcal{P}(\mathbb{N})|$, dove, utilizzando la stessa notazione vista per il caso finito, $2^{\aleph_0}$ denota la cardinalità di $\mathcal{P}(\mathbb{N})$.

Adesso vogliamo provare che $|\mathbb{N}| < |\mathbb{R}|$, mediante il cosiddetto *secondo argomento diagonale di Cantor*. In realtà dimostriamo la non numerabilità dell'intervallo $\left[0, \frac{1}{2}\right]$, da cui seguirà facilmente la non numerabilità di $\mathbb{R}$. Tale dimostrazione procede per assurdo: ossia supponiamo la numerabilità di $\left[0, \frac{1}{2}\right]$ per poi trovare una contraddizione. Se dunque $\left[0, \frac{1}{2}\right]$ è numerabile, possiamo mettere in corrispondenza

biunivoca tale intervallo con i numeri naturali, e dunque possiamo enumerare tutti i suoi elementi: $r_1, r_2, r_3, \ldots$. Scriviamo ciascuno di questi numeri in forma decimale, indicando con $a_k^{(n)}$ la k-esima cifra decimale dell'n-esimo numero:

$$\begin{aligned} r_1 &= 0, a_1^{(1)} a_2^{(1)} a_3^{(1)} a_4^{(1)} \ldots \\ r_2 &= 0, a_1^{(2)} a_2^{(2)} a_3^{(2)} a_4^{(2)} \ldots \\ r_3 &= 0, a_1^{(3)} a_2^{(3)} a_3^{(3)} a_4^{(3)} \ldots \\ r_4 &= 0, a_1^{(4)} a_2^{(4)} a_3^{(4)} a_4^{(4)} \ldots \\ &\vdots \end{aligned}$$

In più assumiamo che in ciascuno sviluppo decimale non compaia mai solo la cifra nove da un certo punto in poi: questo per un problema di duplice rappresentazione di uno stesso numero reale, in quanto ad esempio $0, 3\bar{9}$ e $0, 4$ sono due diverse rappresentazioni dello stesso numero. Fatto ciò, definiamo una nuova successione di numeri naturali:

$$b_n = \begin{cases} 1, \text{ se } a_n^{(n)} = 0 \\ 0, \text{ se } a_n^{(n)} \neq 0 \end{cases}$$

Stiamo così definendo i numeri b_n a partire dagli elementi diagonali, che nella seguente figura sono posti in grassetto:

$$\begin{aligned} r_1 &= 0, \mathbf{a_1^{(1)}} a_2^{(1)} a_3^{(1)} a_4^{(1)} \ldots \\ r_2 &= 0, a_1^{(2)} \mathbf{a_2^{(2)}} a_3^{(2)} a_4^{(2)} \ldots \\ r_3 &= 0, a_1^{(3)} a_2^{(3)} \mathbf{a_3^{(3)}} a_4^{(3)} \ldots \\ r_4 &= 0, a_1^{(4)} a_2^{(4)} a_3^{(4)} \mathbf{a_4^{(4)}} \ldots \\ &\vdots \end{aligned}$$

Se $a_1^{(1)} = 0$, poniamo $b_1 = 1$, altrimenti $b_1 = 0$; poi, se $a_2^{(2)} = 0$, poniamo $b_2 = 1$, altrimenti $b_2 = 0$, e così via. In questo modo possiamo costruire un nuovo numero reale

$$r = 0, b_1 b_2 b_3 b_4 \ldots$$

Adesso osserviamo che $b_1 \neq a_1^{(1)}$ per costruzione, pertanto $r \neq r_1$; analogamente, $b_2 \neq a_2^{(2)}$, dunque $r \neq r_2$. Per ogni n si ha che $b_n \neq a_n^{(n)}$, dunque il numero r è diverso da ogni r_n, pur appartenendo all'intervallo $\left[0, \frac{1}{2}\right]$. Ma avevamo iniziato supponendo che i numeri $r_1, r_2, r_3, \ldots$ esaurissero tutto l'intervallo $\left[0, \frac{1}{2}\right]$, pertanto siamo giunti ad una contraddizione. Rileggendo attentamente la nostra dimostrazione (invitiamo il lettore a farlo) possiamo renderci conto che abbiamo in realtà provato che non

esiste una funzione suriettiva da $\mathbb{N}$ a $\left[0, \frac{1}{2}\right]$, pertanto non può esistere una funzione biunivoca tra essi. Questo prova che l'intervallo $\left[0, \frac{1}{2}\right]$ è non numerabile. In particolare, allora, non può esistere neanche una funzione suriettiva da $\mathbb{N}$ ad $\mathbb{R}$ (se infatti esistesse potremmo facilmente trovare, a partire da essa, una funzione suriettiva tra $\mathbb{N}$ e $\left[0, \frac{1}{2}\right]$), pertanto anche $\mathbb{R}$ è non numerabile.
In verità risulta $|\mathcal{P}(\mathbb{N})| = |\mathbb{R}|$. Quindi l'insieme numerico dei reali ci fornisce un esempio concreto di come anche le cardinalità infinite possano aumentare.

dell'insieme dei numeri reali è pari alla cardinalità dell'intervallo $[0, 1]$, ed anzi essa è pari alla cardinalità di un qualsiasi intervallo limitato $[a, b]$. Come ultima cosa, vogliamo mostrare che la cardinalità dell'insieme dei numeri reali è pari alla cardinalità dell'intervallo $[0, 1]$, ed anzi è pari alla cardinalità di un qualsiasi intervallo limitato $[a, b]$. Anche qui dobbiamo trovare una funzione biunivoca tra l'insieme dei numeri reali e l'intervallo $[0, 1]$. La risposta ci è data da una funzione che dovrebbe essere ben nota, la funzione arcotangente; infatti essa è definita su tutti i numeri reali, e come immagine ha un intervallo limitato: questo intervallo non è proprio $[0, 1]$ ma è $\left]-\frac{\pi}{2}, \frac{\pi}{2}\right[$. Per aggiustare le cose, dobbiamo applicare una traslazione alla funzione, in modo che l'immagine venga a coincidere con l'intervallo $]0, \pi[$, e successivamente una contrazione dell'ampiezza di tale intervallo, per portarla a $]0, 1[$. Consideriamo pertanto la funzione

$$f(x) = \frac{1}{\pi}\left(\arctan(x) + \frac{\pi}{2}\right)$$

Per quanto appena spiegato, essa è biunivoca come funzione definita sui numeri reali a valori in $]0, 1[$. Restano fuori gli elementi estremi dell'intervallo, ma aggiungere due singoli elementi non altera la cardinalità di un insieme infinito, quindi la cardinalità di $]0, 1[$ è pari alla cardinalità di $[0, 1]$. In definitiva, i numeri reali sono tanti quanti i numeri compresi tra 0 e 1. Un argomento analogo mostra che i numeri reali sono tanti quanti i numeri compresi tra a e b per ogni $a < b$.

In breve, per tirare le somme di quanto abbiamo mostrato riguardo alla teoria delle cardinalità, possiamo riassumere queste pagine nel seguente schema:

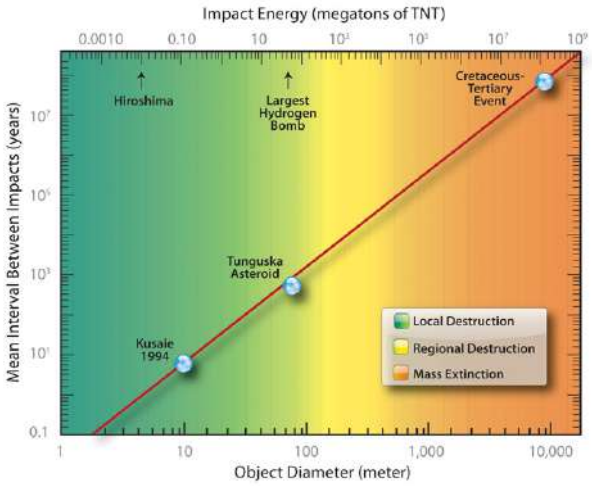
$$|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}| < |\mathbb{R}| = |\mathcal{P}(\mathbb{N})| < |\mathcal{P}(\mathcal{P}(\mathbb{N}))| < |\mathcal{P}(\mathcal{P}(\mathcal{P}(\mathbb{N})))| < \dots$$

Dario Villanis Ziani
Laureato Triennale in Matematica

2 Matematica Vs Armageddon

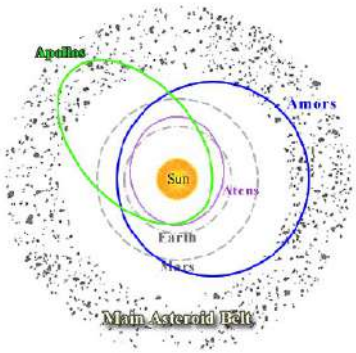
Giacomo Tommei, n.0, Febbraio 2015

A cosa serve la Matematica? Se le sono chiesto in tanti (...e molti altri continueranno a farlo), la risposta è complessa, e poche righe sarebbero inutili. Possiamo affermare, senza possibilità di smentita, che la Matematica pervade ogni cosa ed è necessaria in molti campi e in diversi aspetti della nostra vita. Un contesto particolare è la protezione del nostro pianeta da pericoli provenienti dallo spazio (non pensate subito ad un'invasione aliena...). Detto in altri termini, la Matematica ci può aiutare a non fare la fine dei dinosauri. Questi rettili giganteschi hanno dominato la Terra da 230 a 65 milioni di anni fa, quando improvvisamente si sono estinti. L'ipotesi più accreditata per la causa dell'estinzione (ipotesi di Alvarez, 1980) è quella di un impatto asteroidale o cometario che ha fortemente alterato l'ecosistema in cui vivevano i dinosauri. Nel 1990 è stato identificato il cratere di Chicxulub, sulla costa dello Yucatan, in Messico, che ha i requisiti per soddisfare l'ipotesi di Alvarez: questo cratere ha un diametro stimato di circa 180 chilometri, e il suo ritrovamento è stato possibile solo grazie a tecniche moderne in quanto interamente sommerso dal mare. Gli asteroidi da tenere sotto controllo non sono solo quelli che portano ad un'estinzione di massa, ce ne sono molti altri, più piccoli, che possono creare danni considerevoli. Nell'estate del 1908 un oggetto, probabilmente un asteroide roccioso con un diametro stimato tra i 60 e i 190 m, esplose ad un'altitudine tra i 6 e i 10 Km vicino a Tunguska, in Siberia. Nell'esplosione furono rilasciati circa 15 Megatoni di energia bruciando svariati ettari di tundra siberiana. Il 15 Febbraio 2013 a Chelyabinsk, in Russia, un oggetto di natura asteroidale (di circa 18 m di diametro) è esploso nei cieli a 23 Km di altezza provocando più di 1000 feriti e molti milioni di euro di danni (su YouTube si trovano diversi filmati). Pensate se un simile oggetto (o, peggio, un oggetto di tipo Tunguska) invece di esplodere e frantumarsi finendo la sua corsa in un lago ghiacciato, avesse impattato su una città o su una zona popolosa, oppure in mare creando forti tsunami. È questo il motivo che spinge gli scienziati a tenere sotto controllo gli incontri ravvicinati di asteroidi. Naturalmente non si deve creare allarmismo, consultando le frequenze degli impatti e le loro conseguenze in termini di energia rilasciata (vedi Figura) si può notare che un evento tipo Tunguska si verifica in media ogni 100 anni, mentre un evento catastrofico come quello che ha portato all'estinzione dei dinosauri ogni 100 milioni di anni. Quindi, da una parte la statistica ci può far stare tranquilli, ma dall'altra la comunità scientifica non può ignorare il problema, che, come vedremo nei successivi paragrafi, non è così semplice da risolvere.



I nostri vicini scomodi

La parola asteroide significa "come una stella", anche se questi corpi minori del Sistema Solare non emettono luce propria, ma sono visibili solo perché riflettono la luce solare. Le dimensioni degli asteroidi variano notevolmente: si va dalle centinaia di chilometri in diametro (*Cerere*, il più grande ed il primo ad essere scoperto, misura 913 Km in diametro) ai pochi metri. La massa totale di tutti gli asteroidi è inferiore a quella della Luna. Vi sono asteroidi in diverse posizioni del Sistema Solare e se ne conoscono più di 500000. La maggioranza di essi orbita nella fascia principale (detta anche cintura principale, *Main Belt*) tra Marte e Giove. I pianetini che maggiormente ci interessano, però, sono quelli che arrivano in un intorno della Terra e sperimentano incontri ravvicinati con il nostro pianeta: tali asteroidi sono chiamati NEAs (*Near Earth Asteroids*) e hanno la proprietà di avere una distanza minima dal Sole (perielio) inferiore a 1.3 Unità Astronomiche (le orbite di questi oggetti sono ellissi con il Sole in uno dei fuochi secondo la prima legge di Keplero). Fu con la scoperta di *433 Eros* (1898) che si accertò l'esistenza di tali asteroidi e al momento (Gennaio 2015) se ne contano 12051. L'attenzione per questi oggetti aumentò notevolmente con l'avvento dell'era spaziale e, in particolare, con le missioni lunari. Queste missioni misero in luce la natura da impatto dei crateri sulla Luna costringendo la comunità scientifica a porsi il problema degli impatti sulla Terra. Il monitoraggio di impatti (impact monitoring) è quindi una "scienza" giovane, basti pensare che nel 1998, anno in cui uscirono ben due film americani sul tema (Armageddon e Deep Impact), ancora non esistevano algoritmi per il calcolo della probabilità d'impatto.



Ci serve la Matematica!

Capire se un asteroide può in un futuro (prossimo o lontano) impattare con il nostro pianeta è un problema difficile che può essere suddiviso in tre passi.

1. Guardare il cielo e scoprire gli oggetti. Più osservazioni si hanno (purché siano di qualità), maggiore è l'accuratezza con la quale si può conoscere l'orbita dell'oggetto.

2. Calcolare, a partire dalle osservazioni, le orbite degli oggetti scoperti (**determinazione orbitale**).

3. Capire, con le informazioni a disposizione (osservazioni e orbite), se in futuro gli oggetti potranno impattare con il nostro pianeta, calcolando una probabilità d'impatto (**impact monitoring**).

Il passo 1) è compito degli astronomi di tutto il mondo, mentre per i passi 2) e 3) la Matematica è essenziale. La determinazione dell'orbita di un corpo celeste è un processo costituito da due fasi: (I) costruzione di un'orbita preliminare (chiamata anche orbita nominale) a partire da un numero minimo di osservazioni; (II) perfezionamento dell'orbita con metodi correttivi grazie ad un numero più consistente di osservazioni. Tradizionalmente esistono due metodi per il calcolo di orbite preliminari, il metodo di Laplace (matematico, fisico e astronomo francese, 1749-1827) ed il metodo di Gauss (matematico, fisico e astronomo tedesco, 1777-1855). In entrambi i metodi si cerca di determinare un'orbita kepleriana determinata dall'attrazione gravitazionale agente tra il corpo e il Sole. La base di partenza è data da tre osservazioni, ognuna delle quali è costituita da una terna (t, α, δ) in cui t rappresenta l'istante dell'osservazione, α l'ascensione retta e δ la declinazione del corpo (due angoli che servono a individuare l'oggetto sulla sfera celeste, come un punto sulla superficie terrestre è individuato da longitudine e latitudine). Ciò che non conosciamo è la distanza dell'oggetto, ed è proprio tale quantità che

deve essere calcolata. Si tratterà, nella risoluzione del problema, di trovare le radici di un polinomio di ottavo grado (peraltro tale polinomio risulterà simile in entrambi i metodi), e scoprire quali di queste possono essere accettate e quali invece devono essere scartate. Dovremo tener conto del fatto che le osservazioni iniziali contengono un errore di misura in nessun modo eliminabile; inoltre sarà necessario procedere ad ulteriori approssimazioni, utilizzando metodi numerici dei quali non è sempre assicurata la convergenza. La soluzione del problema non è quindi di per sé garantita (anzi ci sono casi in cui è impossibile trovarla). Nonostante questo, i metodi in questione sono ancora molto attuali e permettono spesso di risolvere il problema della costruzione di un'orbita preliminare. Naturalmente esistono oggi altri metodi per il calcolo di un'orbita preliminare, alcuni dei quali sviluppati dal Gruppo di Meccanica Celeste dell'Università di Pisa pochi anni fa. Con l'aggiunta di nuove osservazioni l'orbita preliminare può essere migliorata con il metodo dei minimi quadrati, inventato da Gauss, che permise di ritrovare l'asteroide *Cerere* (scoperto da Giuseppe Piazzi il primo giorno dell'anno 1801) un anno circa dopo la scoperta. Ogni volta che calcoliamo un'orbita dobbiamo tener conto della sua incertezza, dovuta alla propagazione degli errori di misura negli algoritmi. Quindi se vogliamo scoprire possibili impatti di un certo oggetto nel futuro dobbiamo, non solo propagare la sua orbita nominale, ma anche l'incertezza associata. Ed è qui che il problema si complica a causa della natura caotica delle orbite.

Breve storia del caos

Un sistema caotico ha le seguenti caratteristiche:

(a) l'evoluzione su tempi lunghi non è predicibile e simula un processo stocastico, ovvero una variazione casuale del sistema;

(b) due sistemi con condizioni molto vicine possono avere un futuro radicalmente diverso;

(c) le orbite degli elementi che compongono il sistema restano generalmente confinate, ovvero il sistema non evolve verso l'infinito.

Contrariamente a quanto si tenda a pensare comunemente dire che un sistema è caotico non vuole dire che sia instabile, ma piuttosto che sia non predicibile: la nostra conoscenza della sua evoluzione ha un orizzonte temporale limitato che dipende dal sistema in considerazione. La Teoria del Caos è una disciplina relativamente giovane che, nella seconda metà del Novecento ha fatto breccia nell'immaginario collettivo anche grazie a numerosi romanzi e film. È famoso il personaggio di Ian Malcolm, il matematico del libro (e dell'omonimo film, interpretato magistralmente da Jeff Goldblum) Jurassic Park, il quale illustra il caos parlando del famoso effetto

farfalla, enunciato per la prima volta dal matematico statunitense Edward Norton Lorenz (1917 - 2008): una farfalla sbatte le ali a Pechino e a Central Park, invece di essere soleggiato, piove. Lorenz fu colui che riscoprì il caos negli anni Sessanta del secolo scorso, ma il primo a riscontrare fenomeni caotici fu un brillante matematico francese Henri Poincaré (1854 - 1912). Nel 1885 il re Oscar II di Svezia, per celebrare il suo sessantesimo compleanno, decise di offrire un premio di 2500 corone a chiunque fosse stato in grado di descrivere matematicamente il moto di un certo numero di corpi soggetti alla forza di attrazione gravitazionale, quella descritta da Newton un paio di secoli prima. L'intento era chiaro: sfruttare la potenza della Matematica per predire il futuro, ma soprattutto la stabilità del Sistema Solare. Il problema, noto oggi come problema degli N-corpi gravitazionale, è difficilissimo perché appartiene a quella classe di problemi matematici detti non integrabili, ovvero la soluzione non può essere espressa mediante un algoritmo che include quadrature (calcolo di integrali) e funzioni implicite. Nonostante il problema fosse così difficile, il premio fu assegnato a Poincaré che studiò il moto di tre corpi, dimostrando che la soluzione non è esprimibile in forma esplicita. Nel lavoro che gli valse il premio enunciò inoltre un risultato di stabilità sul moto di questi tre corpi, a cui era arrivato arrotondando differenze molto piccole nelle posizioni dei corpi, pensando che ciò non avrebbe influito sul risultato finale. Solo dopo la consegna del lavoro, si accorse di aver commesso un grave errore: a differenza di quanto aveva creduto inizialmente, un piccolo cambiamento nelle condizioni iniziali portava ad orbite completamente diverse. Si affrettò quindi a contattare l'editore per far cessare la stampa del suo lavoro e si prodigò per recuperare e distruggere tutte le copie già stampate. Si vocifera che dovette ricomprarne una buona quantità spendendo più delle 2500 corone del premio. Ma dopo l'errore, cosa sarebbe successo al premio? Sembrava che stesse per scoppiare uno scandalo e invece, come spesso succede quando si fa ricerca, un errore clamoroso aveva spianato la strada ad un risultato sensazionale: Poincaré aveva scoperto il caos. I tempi, però, non erano maturi per quella scoperta. Con l'inizio del Novecento le attenzioni della comunità scientifica (e quindi anche di Poincaré) si spostarono verso nuove teorie fisiche, la teoria della relatività (ristretta e generale) di Albert Einstein (1879-1955) e la meccanica quantistica di Max Planck (1858-1947), che avrebbero influenzato tutto il secolo. Toccò allora a Lorenz riscoprire il caos durante alcune simulazioni di moti turbolenti nell'atmosfera. Si cominciavano ad usare elaboratori elettronici i quali lavorano in aritmetica finita, troncando i valori numerici durante i calcoli. Lorenz si accorse che, trascurando le cifre troncate dall'elaboratore nelle condizioni iniziali utilizzate per una nuova propagazione, si arrivava a risultati drasticamente diversi: i moti erano impredicibili. I fenomeni caotici nel Sistema Solare sono ben visibili, l'intero Sistema Solare può considerarsi caotico su tempi scala lunghi, non si possono infatti fare previsioni attendibili oltre i 100 milioni di anni. Ci sono

inoltre dei meccanismi, come le collisioni, gli incontri ravvicinati, le risonanze e l'effetto Yarkovsky, che amplificano il caos e che sono particolarmente importanti per capire la dinamica e l'origine degli asteroidi potenzialmente pericolosi.

(a) Collisioni. Le collisioni tra i numerosi asteroidi che abitano nella fascia principale rappresentano un evento consueto, le cui conseguenze dipendono dalle condizioni e dalle caratteristiche dei corpi. I principali fenomeni che possono verificarsi in questo caso sono due: l'accrescimento e la frammentazione. La prima ha luogo quando due corpi collidono a una velocità relativa sufficientemente scarsa e porta alla fusione dei due corpi. La frammentazione avviene invece a velocità relative maggiori, o se la composizione chimica dei due corpi è più fragile. In questo caso, i corpi si frammentano dando vita a diversi frammenti più piccoli. Il destino dei frammenti prodotti da una collisione può essere diverso: possono diventare singoli asteroidi oppure dar luogo a una delle tante famiglie dinamiche di asteroidi.

(b) Incontri ravvicinati. Il Sole, con la sua grande massa, costituisce l'attrattore principale di tutti i corpi del Sistema Solare; per questo le orbite possono essere calcolate, in prima approssimazione, applicando le leggi di Keplero per il moto a due corpi. La principale eccezione a questa legge si ha quando il corpo passa vicino ad un pianeta. In questo caso c'è una zona dove l'attrazione esercitata dal pianeta è più forte di quella degli altri corpi, incluso il Sole. Questa zona si chiama sfera di influenza del pianeta e al suo interno l'orbita originale del corpo cambia drasticamente diventando indipendente da quella originaria, anche se molto sensibile alle condizioni iniziali, cioè allo stato di moto prima dell'entrata nella sfera d'influenza. Per questa ragione gli incontri ravvicinati sono visti come amplificatori del caos.

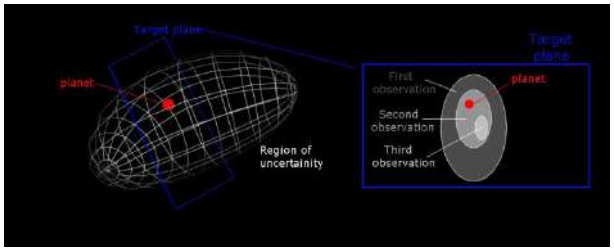
(c) Risonanze. La risonanza è un fenomeno efficiente nel rendere caotiche le orbite su tempi scala molto lunghi. Ma cos'è una risonanza? Consideriamo un asteroide con un'orbita che incrocia quella di Giove. Come è noto, l'attrazione newtoniana tra due corpi, è inversamente proporzionale alla loro distanza, per cui quando i due corpi sono vicini, l'asteroide subisce la massima attrazione da parte del pianeta. Se i periodi di Giove e dell'asteroide non sono proporzionali, i due corpi si ritroveranno saltuariamente vicini e l'effetto generale della reciproca attrazione sarà nullo. Invece l'interazione tra Giove e l'asteroide torna ad avere importanza quando le orbite dei due corpi hanno dei periodi risonanti. In altre parole, quando il periodo dell'asteroide è una frazione di quello di Giove, i due si trovano vicini periodicamente e la perturbazione si amplifica rendendo l'orbita dell'asteroide instabile.

(d) Effetto Yarkovsky. Yarkovsky era un ingegnere russo che aveva proposto, più di un secolo fa, una teoria sui cambiamenti delle orbite dei corpi vaganti nello spazio interplanetario, innescati dalle variazioni di temperatura superficiale fra l'emisfero diurno e quello notturno. Dato che un corpo solido tanto più caldo tanto più emette radiazione infrarossa, e questa emissione produce una piccola forza di rinculo (un po' come i gas emessi dall'ugello di un razzo), Yarkovsky propose che un piccolo corpo roccioso orbitante e rotante sul proprio asse avrebbe avuto la sua orbita lentamente modificata a causa del riscaldamento asimmetrico della superficie prodotto dalla radiazione solare. Yarkovsky e il suo effetto furono quasi dimenticati fino a tempi recenti, quando calcoli accurati, basati sulle proprietà termiche delle meteoriti e delle rocce lunari hanno mostrato che, per gli asteroidi di diametro fino a 20 km, l'effetto Yarkovsky altera lentamente, ma in misura non trascurabile i semiassi maggiori delle orbite. In una predizione d'impatto con orizzonte lontano nel futuro è necessario includere tale effetto nel modello.

Impact monitoring

Quando un asteroide viene scoperto, come abbiamo già evidenziato, non si sa niente circa l'orbita reale dell'oggetto. Vi è un insieme di possibili orbite, tutte compatibili con le osservazioni, che formano una regione di confidenza nello spazio delle orbite (tale spazio ha dimensione 6, servono infatti sei numeri per individuare univocamente posizione e velocità dell'oggetto). Possiamo descrivere questo fatto pensando ad uno sciame di asteroidi virtuali (*VAs, Virtual Asteroids*), con orbite diverse, ma molto vicine, e tutte compatibili con le osservazioni. La verità dell'asteroide è divisa tra tutti quelli virtuali, nel senso che solo uno è reale, ma non si sa quale. Per rendere le cose semplici è possibile pensare che tutti gli asteroidi virtuali abbiano la stessa probabilità di essere quello reale, ma in realtà si utilizza una distribuzione di probabilità più complicata (Gaussiana). Poiché la regione di confidenza contiene un continuo di orbite, ogni asteroide virtuale è il rappresentante di una piccola porzione di spazio. C'è da notare che l'orbita nominale, soluzione del fit ai minimi quadrati dei residui osservativi, è solo uno dei tanti asteroidi virtuali, senza nessun altro specifico significato. Nel caso che un asteroide sia Earth-crossing, è possibile che esistano uno, o più asteroidi virtuali associati ad esso, per i quali è ammissibile una collisione. Quindi può esistere una piccola regione connessa piena di orbite collisionali, la quale definisce un impattore virtuale (*VI, Virtual Impactor*). Lo scopo dell'impact monitoring è individuare impattori virtuali ed assegnare a ciascuno una probabilità d'impatto. Per raggiungere lo scopo dobbiamo prendere ciascun asteroide virtuale e propagare la sua orbita nel futuro (di solito 100 anni) registrando eventuali incontri ravvicinati e impatti con il nostro pianeta. Per la propagazione dobbiamo naturalmente campionare con

un numero finito di VAs la regione di confidenza. Questo viene fatto utilizzando un sottospazio unidimensionale dello spazio degli elementi orbitali (una curva) che chiamiamo Linea Delle Variazioni (LOV, Line Of Variations). Su questa curva vengono presi un certo numero di punti che rappresentano l'insieme dei VAs da propagare. Per la ricerca di impattori è necessario poi utilizzare un piano bersaglio, un piano passante per il centro della Terra e perpendicolare alla velocità relativa imperturbata dell'asteroide, sul quale viene segnata la sezione della Terra e l'intersezione con le orbite dei VAs. Se un asteroide virtuale si trova dentro la sezione della Terra siamo in presenza di un impattore virtuale e possiamo procedere al calcolo della probabilità utilizzando una proiezione della regione di confidenza 6-dimensionale sul piano bersaglio che rappresenta l'incertezza associata all'impattore virtuale. Gli algoritmi qui descritti sono implementati in un software chiamato CLOMON2 (la prima versione CLOMON risale al 1998), creato dal Gruppo di Meccanica Celeste dell'Università di Pisa, che dal 2002 si occupa di calcolare orbite e probabilità d'impatto degli asteroidi potenzialmente pericolosi: gli output sono riassunti in una Risk Page pubblicata sul sito web NEODyS. Un sistema analogo, Sentinel, si trova al Jet Propulsion Laboratory (JPL) di Pasadena, California. I due gruppi di ricerca (UniPI e JPL) sono in continuo contatto per la verifica ed il confronto dei risultati.



Se dovesse capitare di trovare una probabilità d'impatto uguale a 1, cosa si fa? I fattori da prendere in considerazione sono essenzialmente due: dimensioni dell'oggetto e tempo rimanente all'impatto. È chiaro che, maggiore è il tempo all'impatto, più possibilità si hanno di trovare una soluzione, che potrebbe anche consistere nell'organizzare una missione spaziale per la deflessione dell'asteroide. Se l'impatto è imminente (pochi giorni o addirittura poche ore) possiamo fare poco, ma gli oggetti in questi casi sono piccoli e spesso si sgretolano all'ingresso in atmosfera. Ad ogni modo il 'cosa fare in caso di previsione di impatto' è un argomento molto complesso e ancora estremamente dibattuto sia a livello scientifico che politico.

dott. Giacomo Tommei,
Ricercatore presso il Dipartimento di Matematica, Università di Pisa

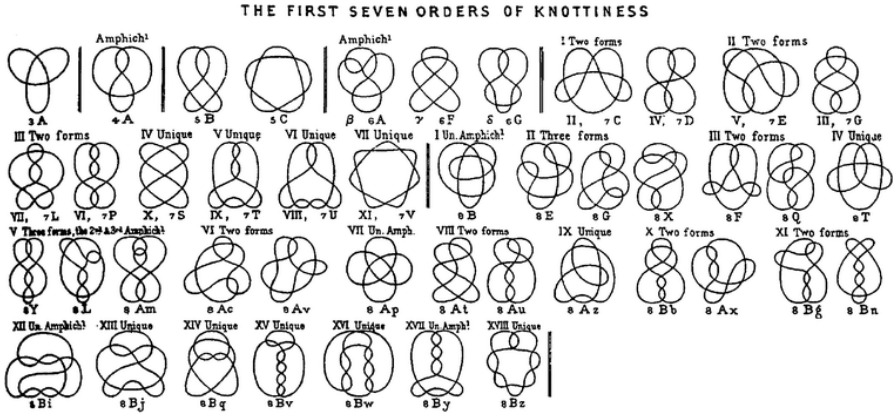
3 Nodi e Colorazioni

Agnese Barbensi, n.1, Settembre 2015

Sapreste dire quali dei lacci in figura sono annodati e quali sono sciolti?

Come ci ricordano certi rompicapo con le corde o alcuni trucchi da mago, la risposta non è sempre ovvia. La *Teoria dei Nodi* è la branca della matematica che si propone di rispondere a quesiti come questo.

Nella seconda metà del 1800 il fisico P.G. Tait, motivato dalla teoria di Lord Kelvin secondo la quale gli atomi sarebbero stati vortici annodati nell'etere, cercò di creare una classificazione completa dei nodi, ordinati per complicazione crescente: la prima *tavola di nodi*.



Sebbene pochi anni dopo la teoria di Lord Kelvin venne confutata, il tentativo di Tait motivò i matematici a chiedersi come fosse possibile distinguere tra di loro due nodi non equivalenti, cioè due nodi per cui l'unico modo di trasformare l'uno nell'altro sia "tagliare" il laccio. La Teoria dei Nodi divenne un argomento molto studiato all'interno della nascente branca della matematica, la *Topologia*.

La Teoria dei Nodi è ad oggi un settore di ricerca molto attivo. Oltre all'enorme interesse teorico, ha moltissime applicazioni, che spaziano dalla Fisica

Teorica alla Biologia. Dagli anni '90 è stata applicata allo studio del DNA (*acido desossiribonucleico*).

Infatti, la struttura a doppia elica del DNA risulta a volte ulteriormente annodata.

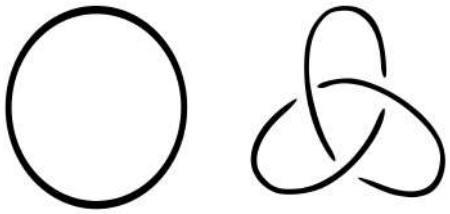
Un enzima, detto *Topoisomerasi* "snoda" le strisce durante i processi di trascrizione e duplicazione. Le conoscenze provenienti dalla Teoria dei Nodi aiutano i biologi a comprendere meglio questi processi di snodamento, dando ad esempio delle stime sul tempo che occorre agli enzimi per completare i loro compiti.

Ma che cos'è un *nodo* in Matematica?

Definizione. *Un nodo è una curva chiusa, intrecciata nello spazio.*
Due nodi sono equivalenti *(cioè sono lo stesso nodo) se è possibile deformare l'uno nell'altro nello spazio con continuità (in altre parole, senza creare autointersezioni e senza* tagliare *il laccio).*

Dire che un nodo è equivalente al nodo *banale* (cioè quello sciolto) vuol dire che è possibile scioglierlo senza creare autointersezioni e tagliare il laccio.

I nodi vengono rappresentati attraverso i loro *diagrammi*, cioè il disegno che si ottiene proiettandoli su un piano e segnalando graficamente i sottopassaggi ad ogni incrocio. In figura, a destra, un diagramma per il nodo Trifoglio, a sinistra uno per il nodo banale.



Come potete osservare usando una corda chiusa su se stessa (ad esempio un elastico), ci sono tantissimi (*infiniti!*) diagrammi diversi che rappresentano lo stesso nodo!

Nel 1927 il matematico Kurt Reidemeister ha dimostrato che due diagrammi rappresentano lo stesso nodo se e solo se è possibile ottenere l'uno dall'altro tramite una sequenza finita delle tre mosse in figura.

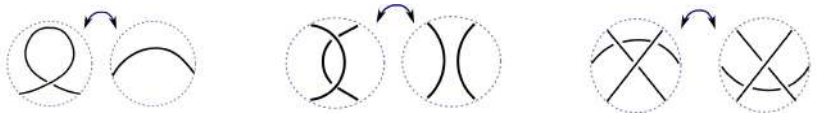


Figura 1: Ognuna delle tre mosse consiste nel lasciare invariato il diagramma fuori dal cerchio, e sostituire la parte interna al cerchio con la figura corrispondente.

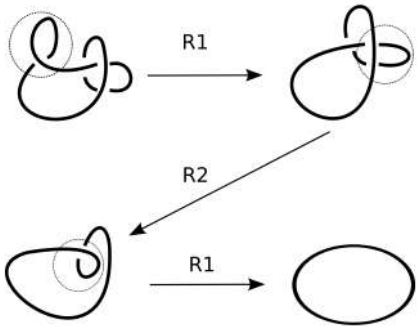


Figura 2: Un esempio: il diagramma rappresenta il nodo banale. Posso infatti "scioglierlo" con una sequenza di 3 mosse di Reidemeister.

Questo risultato, seppur di fondamentale importanza, non è sufficiente per capire se un nodo è banale, o se due diagrammi effettivamente rappresentano lo stesso nodo! Infatti, sebbene si sappia che diagrammi equivalenti differiscono per un numero finito di mosse di Reidemeister, il teorema non dice niente su quante, e soprattutto quali, mosse servano.

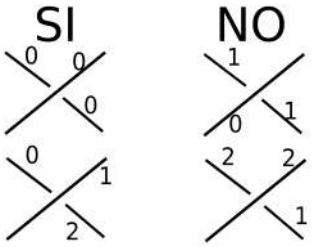
Queste considerazioni hanno portato i matematici a cercare altri modi per distinguere i nodi tra di loro. Nella pagine successive introdurremo uno strumento semplice ma funzionale allo scopo. Ad esempio saremo in grado di dimostrare che il nodo trifoglio non è banale.

Diagrammi e Colorazioni

Un *arco* di un diagramma è una porzione della proiezione del nodo delimitata da due *sottopassaggi:*

Scegliete i vostri 3 colori preferiti; con *colorazione* di un diagramma intendiamo l'assegnazione di un colore ad ogni arco, in modo che ad ogni incrocio valga questa regola:

- tutti gli archi hanno lo stesso colore, **oppure**
- tutti gli archi hanno colori diversi



Per motivi grafici indicheremo i colori con delle etichette agli archi date dai numeri $0, 1$ e 2.

La colorazione ottenuta assegnando ad ogni arco lo stesso colore è detta *banale*. Ogni diagramma ammette ovviamente almeno le 3 colorazioni banali.
Un diagramma si dice 3-*colorabile* se ammette una colorazione non banale. Il diagramma del nodo banale nella terza figura non è 3-*colorabile*.

Il motivo per cui questa costruzione è interessante sta nel seguente risultato dovuto a Ralph H. Fox nel 1961:

Teorema 1. *Diagrammi di nodi equivalenti ammettono lo stesso numero di 3-*colorazioni

Questo in particolare ci dice che se un nodo ammette un diagramma 3-*colorabile*, questo non può essere equivalente al nodo banale.

Il numero di 3-colorazioni è un esempio semplice di quelli che in matematica vengono chiamati *invarianti* di nodi, ossia oggetti (numeri, polinomi etc) associati ai diagrammi, che sono uguali per diagrammi che rappresentano lo stesso nodo.

Dimostrazione. Grazie al Teorema di Redeimeister sappiamo che due diagrammi rappresentano lo stesso nodo se e solo se differiscono per sequenze di mosse dei tre tipi di cui sopra.
Per dimostrare il Teorema di Fox basterà allora verificare la tesi per diagrammi che differiscono per una delle tre mosse.
Considereremo per ogni mossa di Reidemeister i due diagrammi L e L^* che coincidono ovunque fuori da un cerchio e differiscono per tale mossa all'interno, e mostreremo che esiste una bigezione tra le colorazioni di L e L^*, cioè una funzione iniettiva e surgettiva. In particolare il numero di colorazioni di L e L^* è lo stesso.

Dunque, ad ogni colorazione di L vogliamo associare una colorazione di L^*. Supponiamo quindi assegnata una colorazione di L. Siccome L e L^* coincidono

fuori dai cerchi, possiamo costruire la colorazione corrispondente per L^* mantenendo i colori assegnati agli archi di L fuori dal disco, provando poi (se possibile) a estenderla in modo coerente nella parte in cui i diagrammi differiscono.

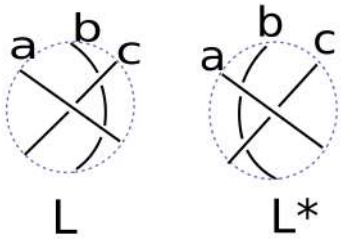
Vediamo il primo caso: per ogni scelta di colore $a \in \{0, 1, 2\}$ da dare all'arco del diagramma L ne abbiamo una sola in L^* che rispetti la regola all'incrocio (si veda la figura seguente).

Nel secondo, scegliamo una colorazione di L. Questa sarà data (all'interno del cerchio) da due colori $a, b \in \{0, 1, 2\}$.

Se $a = b$, l'unico modo di completare la colorazione in L^* è porre $* = a = b$.
Analogamente, se $a \neq b$, l'unico modo di completare la colorazione in L^* è porre $* = c$, dove $c \in \{0, 1, 2\}$ è l'unico dei tre colori diverso sia da a che da b.
In ognuna delle situazioni, abbiamo la bigezione cercata.

Il caso della terza mossa è ugualmente semplice, ma necessita di diversi passaggi.

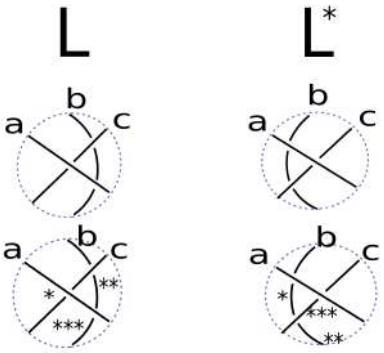
Supponiamo che ai capi degli archi siano assegnati i colori a, b e c in L. In L^* dovrò allora avere la stessa situazione, come in figura.



Se $a = b = c$, in entrambi L e L^* posso estendere le colorazioni dentro ai dischi solo nel modo banale, cioè colorando tutti gli archi con dello stesso colore.

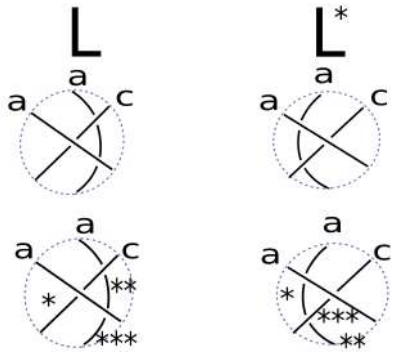
Supponiamo adesso che a, b, c siano a due a due distinti, e guardiamo L. Dato che $b \neq c$, per completare la colorazione necessariamente dovrò porre $** = a$.

Analogamente, da $a \neq c$ ottengo $* = b$. L'unica scelta possibile per completare rimane assegnare $*** = a$.
Con lo stesso ragionamento applicato ad L^* vediamo che anche in questo caso ho un unico modo di estendere la colorazione, ottenuto ponendo $* = c$, $*** = b$ e $** = a$.

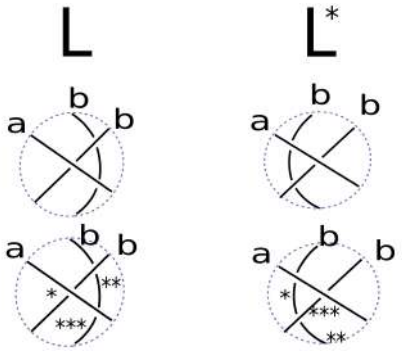


Rimangono adesso da verificare i casi in cui due dei colori sono uguali, e il terzo è distinto.

Supponiamo adesso che i primi due archi siano colorati uguali, e il terzo sia diverso, come in figura.



In L siamo obbligati a porre $** = b = *$, e quindi necessariamente $*** = c$. Allo stesso modo, in L^*, l'unico modo di rispettare la regola è imporre $* = a$ e $*** = b$, da cui $** = c$.



Se è il primo arco ad essere distinto dagli altri due, la situazione è simile:

In L dobbiamo infatti necessariamente avere $** = b$ e $* = c$, da cui $* * * = c$. Analogamente, in L^*, dobbiamo necessariamente porre $* = c = * * *$ da cui $** = c$.

Lasciamo per esercizio la (facile) verifica dell'ultimo caso, quello in cui sono il primo e il terzo arco ad avere lo stesso colore.

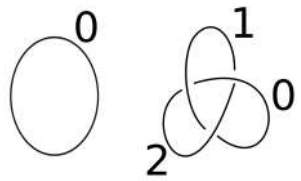
Abbiamo quindi mostrato che, se due diagrammi corrispondono allo stesso nodo (sono equivalenti), ogni 3-colorazione di L corrisponde esattamente a una di L^* (la bigezione cercata!), quindi il numero totale di colorazioni possibili è lo stesso per i due diagrammi, e il teorema è dimostrato.

□

Come possiamo facilmente intuire, è impossibile trasformare un nodo trifoglio in un laccio snodato! Ma solo grazie al Teorema appena dimostrato, possiamo esserne matematicamente certi:

Corollario. *Il Nodo Trifoglio non è banale.*

Dimostrazione. Il nodo trifoglio ammette un diagramma 3-colorabile, e non può quindi essere equivalente al nodo banale.



□

Non sempre la risposta giusta è quella intuitiva! Sapreste per esempio dire se il nodo in figura è banale? (Spoiler: la risposta è affermativa, ma le mosse di Reidemeister che lo sciolgono sono particolarmente difficili da trovare...)

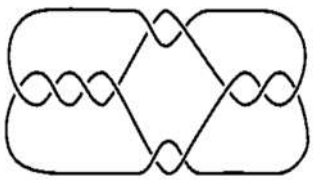


Figura 3: Diagramma

Agnese Barbensi,
Laureata Triennale in Matematica

4 Automi Cellulari e Biologia dei Tumori

Franco Flandoli, n.2, Febbraio 2016

Matematica ed oncologia

L'incontro con la malattia ed il dolore è tra le esperienze più importanti della vita e per questo molte persone elaborano il desiderio di contribuire alla cura o all'ideazione di cure per le malattie. La strada maestra senza dubbio è diventare medico ed operare direttamente nel settore sanitario ma, pensando anche al versante dello sviluppo di nuove cure, lo studio della chimica e della biologia, o di discipline più recenti come la bio-ingegneria, può portare su quella strada. Persino la fisica ha sviluppato una specializzazione che si avvicina alla medicina: la figura professionale del fisico medico.

La matematica invece, per quanto sia a fondamento della fisica, della chimica e dell'ingegneria, ha un rapporto meno diretto con le scienze della vita ed in particolare con la medicina. La connessione più ovvia è attraverso la statistica: tutti sentiamo esporre in forma di percentuali alcuni fatti relative alla salute dell'uomo, percentuali ottenute tramite indagini statistiche - è recente ad esempio l'affermazione che il consumo di carni trasformate - salumi ecc. - aumenta del 18% il rischio relativo di contrarre un tumore.

La matematica però ha tra le sue abilità quella di sviluppare modelli. La fisica è basata su questo. I modelli possono adattarsi a situazioni anche distanti dalla fisica, come ad esempio la finanza e l'economia. Allora perché non cercare di sviluppare modelli anche su temi specifici di medicina?

Sostanzialmente la resistenza a questa direzione di ricerca sta nella complessità dei fenomeni coinvolti. La fisica ha saputo scomporre i problemi in blocchi elementari, isolati dalla complessità dei fenomeni usuali; la matematica ha poi permesso di descrivere accuratamente il singolo tassello e tornare successivamente a costruzioni più complesse. In medicina, o nella vicina biologia, questo ancora non è accaduto.

Pensiamo ad esempio ad un tessuto umano: esso è composto da una miriade di cellule ed altre strutture. Anche un gas, studiato dalla fisica, è composto da una miriade di atomi o molecole. Ma gli atomi sono oggetti relativamente semplici, tutti uguali se della stessa materia, hanno gli uni con gli altri dei rapporti (interazioni) piuttosto elementari, mentre ogni cellula è un organismo di complessità eccezionale, compie una miriade di operazioni diverse, interagisce in modo estremamente complesso con le altre cellule. Ogni semplice modello matematico del comportamento di un tessuto non potrà tener conto di questa complessità. Anche un atomo di un gas internamente ha una sua complessità ma minore e di solito essa non interviene sul comportamento del gas nel suo complesso così fortemente quanto invece succede per l'esempio del tessuto e delle cellule.

Sembra quindi una strada senza speranza o per lo meno lontana ma credo che non ci si debba perdere d'animo e si debbano invece compiere con fiducia dei primi passi, ben sapendo che la strada è molto più lunga rispetto ad altre discipline. D'altra parte, se si potesse contribuire anche solo in minima parte alla cura, ad esempio, dei tumori, o delle malattie di tipo circolatorio, il contributo al bene dell'umanità sarebbe enorme.

Descrivere matematicamente la crescita di un tumore

Che forma può avere un modello matematico in oncologia? Focalizziamo su pochi aspetti, per arrivare rapidamente a qualche risultato. Poniamo la nostra attenzione sulla *crescita di una massa tumorale*.

Premessa su alcuni elementi di biologia cellulare dei tumori

Prima di sviluppare la matematica, è bene ricordare alcuni fatti essenziali di biologia dei tumori. All'origine c'è una prima cellula normale di un tessuto, di solito epiteliale (pelle, superfici interne al corpo), che ha subito un certo numero di modificazioni genetiche, piuttosto precise. Le cellule epiteliali, essendo soggette ad un continuo ricambio ed essendo maggiormente esposte ad agenti esterni come radiazioni, fumo, inquinamento, eccessivo utilizzo di certi cibi, hanno maggior probabilità di sviluppare generiche modificazioni genetiche; si trasformano in cellule tumorali se le modificazioni sono di tipo specifico. La caratteristica principale della nuova cellula modificata è quella di poter attivare la divisione cellulare liberamente, senza bisogno di stimoli precisi come avviene per le cellule sane, e di trasmettere questa caratteristica alle due cellule che si generano dalla sua divisione cellulare. Nulla può allora arrestare questo meccanismo proliferativo: da una cellula se ne formano due in breve tempo (la durata del ciclo cellulare varia molto da cellula a cellula; per molte però dura circa 16 ore; per semplificare, nel seguito supponiamo che serva un giorno per una duplicazione), poi da ciascuna delle due se ne formano altre due e così via. Se il processo fosse così rigido e preciso, supponendo per semplicità che serva un giorno per una duplicazione cellulare, dopo 10 giorni da una cellula ne avremmo $2^{10} = 1024$, e dopo un solo mese $2^{30} \sim 10^9$. Questo numero, 10^9, è l'ordine di grandezza giusto del numero di cellule in un tumore maturo, di quelli percepibili con analisi cliniche. Però in genere, per fortuna, serve più di un mese per arrivarci, perché le cellule tumorali rallentano il loro ritmo di duplicazione quando sono eccessivamente circondate da altre cellule tumorali, a causa della carenza di ossigeno e nutrimenti che si crea per l'eccessivo affollamento. Continuano a duplicare secondo la norma solo le cellule della parte più esterna del tumore, mentre quasi si fermano quelle più interne.

Nel frattempo, tra le modificazioni genetiche tipiche delle cellule tumorali, c'è quella che permette il movimento. Tutte le cellule hanno una certa capacità di movimento che però di solito è impedita dall'adesione tra le membrane cellulari: due cellule in contatto attivano, nella parte di membrana con cui sono in contatto, delle proteine che, come fossero dei gancetti, le tengono unite. La loro capacità di movimento si restringe quindi a poche mosse di assestamento, ad esempio necessarie in fase di proliferazione, per lasciar spazio alle nuove nate. Ma le cellule tumorali, magari non sin dall'inizio ma dopo un po' di maturazione della popolazione cellulare, sviluppano attraverso ulteriori trasformazioni genetiche la capacità di sciogliere quel legame tra membrane ed essere quindi libere di muoversi nell'ambiente circostante. Tale ambiente, che chiameremo genericamente *matrice extracellulare*, è composto da altre cellule, varie sostanze chimiche, varie proteine, strutture filamentose e colloidali che conferiscono una certa stabilità, vasi sanguigni ed altro. Lentamente ed in modo erratico, una cellula tumorale può spostarsi nella matrice extracellulare, di solito con l'obiettivo di raggiungere zone più ricche di ossigeno e nutrimento ed in particolare avvicinandosi così ai vasi sanguigni, da cui viene diffuso l'ossigeno. Il contatto, purtroppo, tra cellule tumorali e vasi sanguigni è la maggior causa di pericolosità dei tumori perché, se il loro effetto sul corpo vivente fosse solo di aumentare di dimensione nella regione in cui è nata la prima cellula tumorale, intanto provocherebbero un danno relativo, solo in quella regione, poi sarebbero dopo un po' facilmente identificabili e contrastabili chirurgicamente; il corpo stesso, per le ragioni di difficoltà di approvvigionamento di ossigeno alla zona interna del tumore, porrebbe un freno alla crescita. Invece, grazie al movimento delle cellule tumorali che le porta a contatto coi vasi sanguigni - e grazie ad un altro fenomeno simmetrico, l'angiogenesi, di cui però ora non parleremo per non allargare troppo il discorso - le cellule possono entrare nel circolo ematico e raggiungere altre parti del corpo vivente, sviluppando colonie in molti punti e impedendo così di contrastarne efficacemente la crescita. Non che questi viaggi siano una passeggiata: si pensi che mediamente una sola cellula tumorale su 10.000 sopravvive al viaggio nel sangue e riesce a impiantarsi in un altro punto del corpo; ma i numeri di cellule in gioco sono elevatissimi, come già detto sopra.

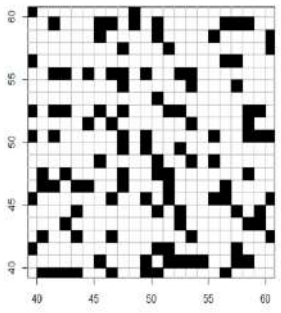
Gli automi cellulari

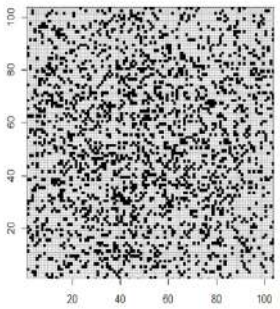
Descriviamo ora un modello matematico molto versatile che si può adattare a varie situazioni della realtà. La coincidenza di nome, "cellulare", non deve far pensare che sia stato ideato specificamente per la descrizione delle cellule tumorali.

Si pensi ad una scacchiera, magari immensa, in teoria anche infinita, magari tridimensionale - come è in realtà qualsiasi tessuto di un corpo; nel seguito, per varie ragioni, la supporremo bidimensionale, proprio come una scacchiera usuale. Ogni casella ha un colore non predeterminato ma variabile, secondo le regole del

gioco che verranno specificate. Pensiamo al caso più semplice, in cui le caselle possono essere solo bianche e nere, ma nulla impedisce di sviluppare giochi più complicati con varie colorazioni.

Supponiamo che "nero" significhi "occupato", mentre "bianco" sia "vuoto". La scacchiera con certe caselle bianche ed altre nere ci mostra quali zone del piano sono occupate e quali libere. Si deve pensare anche ad una visione dall'alto, da lontano, con lo zoom, su una grandissima scacchiera: emergeranno delle masse nere su fondo bianco, composte di una miriade di caselle occupate. Ecco due figure che illustrano un dettaglio della scacchiera vista da vicino ed una visione più da lontano:





Nel nostro esempio dei tumori, si può pensare che per "occupato" intendiamo "occupato da una cellula tumorale" invece che da cellule sane o altre strutture della matrice extracellulare. La visione da vicino ci mostra il dettaglio delle singole cellule, come sono disposte le une rispetto alle altre; la visione da lontano mostra la massa tumorale nel suo complesso, rispetto allo spazio "vuoto" (non occupato dal tumore).

Ciò di cui abbiamo parlato per ora sono le cosiddette *configurazioni* del sistema. Una configurazione è la specificazione di quali caselle sono nere e quali bianche. Se ci sono in tutto N caselle nella scacchiera, le configurazioni possibili sono 2^N: numerando le caselle e percorrendole una dopo l'altra, la prima casella potrebbe trovarsi in due condizioni - bianco o nero - e data la sua condizione, la seconda casella potrebbe trovarsi anch'essa in due e così via (2 per 2 per 2 ..., N volte). L'insieme delle configurazioni possibili è molto ampio; si pensi che una piccola scacchiera quadrata avente ogni lato di lunghezza 100 (come nella maggior parte delle figure mostrate qui), ha un totale di $N = 100 \times 100$ caselle; ma allora il numero di configurazioni possibili sarà $2^{100\times100}$, numero nemmeno immaginabile nei termini dei numeri usualmente accessibili. In questo senso, pur essendo l'automa cellulare un *modello discreto*, esso può approssimare abbastanza bene la complessità di un *modello continuo*.

Ora dobbiamo parlare della dinamica di un automa cellulare, dobbiamo specificare le mosse con cui esso modifica la propria configurazione. In teoria, si potrebbero specificare mosse "globali", che toccano contemporaneamente tutte le caselle della scacchiera, ma tali mosse avrebbero una complessità troppo elevata, metterebbero in gioco i numeri troppo grossi visti sopra. Conviene allora specificare "mosse locali". Con questo si intende che viene identificata una casella, con qualche criterio - ad esempio del tutto a caso - e viene prescritto cosa avviene ad essa ed eventualmente a quelle vicine. Una mossa globale, sempre che abbia senso o sia utile pensarla, è in un certo senso il risultato di una miriade di mosse locali successive.

Così facendo, agendo cioè per mosse locali, da un lato possiamo più facilmente prescrivere delle regole che ricalchino la realtà che vogliamo descrivere, dall'altro abbiamo uno schema più facile da implementare con un codice numerico al calcolatore. Un'ultima osservazione generale: nei modelli usuali di automi cellulari il "caso" viene immesso nell'algoritmo a vari livelli. Abbiamo già accennato al fatto che la scelta della casella in cui effettuare una mossa locale è spesso effettuata a caso, dando ad ogni casella la stessa probabiità di essere selezionata; oltre a questo, viene spesso scelto a caso il tempo che deve intercorrere tra una mossa e l'altra; scelta la casella, nell'operare la mossa viene spesso inserito il caso per scegliere tra varie mosse possibili (in questo caso pesando opportunamente le varie mosse in modo da far accadere più spesso quelle più frequenti).

Un esempio di automa cellulare per la crescita di un tumore

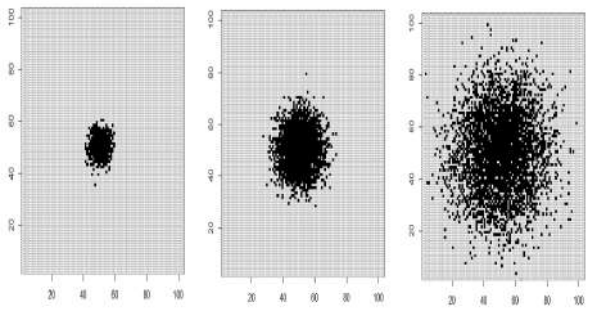
Immaginiamo quindi di prescrivere come condizione iniziale della nostra dinamica cellulare la configurazione formata da una casella occupata e tutte le altre libere - è il momento iniziale in cui si è formata la prima cellula tumorale. Prescriviamo poi che il sistema scelga a caso tra due possibilità, movimento o proliferazione, magari con probabilità non uguali, $1 - p$ e p. Se l'algoritmo sceglie il movimento, poi deve scegliere una a caso delle 4 caselle circostanti per spostare lì la cellula tumorale. Se sceglie la proliferazione, deve analogamente scegliere a caso una delle 4 caselle circostanti per generare lì una delle due nuove cellule; l'altra nuova generata, occuperà la posizione del genitore.

Dopo la prima duplicazione, abbiamo due caselle occupate, due cellule tumorali. Il sistema sceglie a caso tra le due: questa è la fase iniziale casuale della scelta della casella in cui operare la mossa locale. Non si sceglie a caso tra tutte le caselle ma solo tra quelle occupate, perché non serve fare mosse nelle caselle vuote, in questo modello molto semplificato - si possono invece immaginare situazioni più complicate in cui invece degli spazi vuoti ci sono spazi occupati da varie componenti della matrice extracellulare, che quindi sono coinvolte nella dinamica.

Scelta una delle due, l'algoritmo opera come in precedenza: decide a caso se operare un movimento o una proliferazione e poi sceglie la casella limitrofa in cui muoversi o proliferare. L'unico accorgimento è che la scelta della casella limitrofa è limitata a quelle libere, perché non tutte lo sono. A questo punto l'algoritmo procede nello stesso modo, con le stesse regole, anche quando la scacchiera contiene molte cellule tumorali: ad ogni passo viene scelta una cellula tumorale, viene deciso se muoverla o proliferare, viene scelta una casella limitrofa in ciò avviene.

Può accadere che non ci siano caselle limitrofe libere: in questo caso la mossa viene cancellata e si ricomincia scegliendo a caso una casella. In questo modo, automaticamente, sia il movimento sia soprattutto la proliferazione vengono inibite nelle zone troppo piene, rispettando una caratteristica reale.

Le seguenti tre figure mostrano uno stadio iniziale, uno più avanzato in cui il tumore ha ancora un carattere prevalentemente proliferativo (qui il parametro p è quasi uguale ad 1) ed uno ancora più avanzato in cui viene maggiormente espresso il *fenotipo* diffusivo (qui il parametro p è quasi uguale ad 0):



L'algoritmo ora descritto a parole è suscettibile di una ben precisa formulazione matematica tramite il concetto di *catena di Markov* e la *teoria dei processi stocastici*. Il calcolo delle probabilità permette di formalizzare le varie mosse casuali descritte sopra e descrivere l'intera procedura tramite uno schema iterativo casuale. La teoria matematica consente ad esempio di esaminare rigorosamente cosa accade quando si prende una scacchiera infinita e la si guarda con lo zoom: si studia così il *limite al continuo*, in cui al posto del reticolo della scacchiera discreta c'è il piano ed al posto di una configurazione discreta di caselle nere c'è una *densità* continua di cellule tumorali. Mentre la crescita nel tempo del tumore a livello discreto è basata su uno schema iterativo, nel limite continuo fa uso delle cosiddette equazioni differenziali alle derivate parziali, in modo simile a come si descrivono in fisica vari fenomeni che cambiano nel tempo ed hanno variazioni spaziali. Tramite opportuni codici di calcolo numerico, queste dinamiche vengono poi simulate al calcolatore. Facendo varie simulazioni, anche al variare dei diversi parametri simili

a p, si possono esaminare diversi comportamenti e fare stime ad esempio sul tempo necessario al tumore per raggiungere i vasi sanguigni.

L'utilizzo medico di questo tipo di modelli è ancora allo stato embrionale, come dicevamo all'inizio. La strada è lunga ed incerta. Ma la prospettiva di poter contribuire alla ricerca ed applicazione medica in un futuro non troppo lontano è sufficiente a motivare ogni sforzo possibile per rendere sempre più realistici questi modelli.

Franco Flandoli,
Professore Ordinario presso il Dipartimento di Matematica, Università di Pisa

5 La Matematica dell'Importanza

Federico Poloni, n.2, Febbraio 2016

Il problema del ranking In questo articolo, vogliamo descrivere un problema che ha un'importanza e un'utilità pratica sempre maggiori negli ultimi anni caratterizzati dalla crescita di internet e dalla disponibilità sempre maggiore di dati da analizzare.

Supponete di fare una ricerca su internet per trovare siti che parlano dell'ultimo film della Disney. Ci sono diversi risultati pertinenti: per esempio, la pagina ufficiale del film, le recensioni sui giornali, e le pagine dei fan che pubblicano opinioni, immagini e commenti. Quando digitiamo il titolo del film, i motori di ricerca moderni riescono con un'efficacia sorprendente a distinguere quali pagine sono più importanti per noi (per esempio, il sito ufficiale) e quali sono secondarie (per esempio, le pagine dei fan). Vi siete mai chiesti come fanno?

Per gli argomenti di maggiore rilevanza, come un film famoso, è possibile chiedere a una persona di selezionare manualmente le pagine; però, farlo per tutte le possibili ricerche è inimmaginabile. Per questo è necessario un metodo automatico, un algoritmo.

Proviamo a ragionare come un matematico, e cerchiamo di formalizzare il problema: **dato un insieme di pagine web che parlano di un certo argomento, trovare quali sono più importanti.** Dobbiamo innanzitutto dire cos'è una "pagina web". Sicuramente, essa contiene del testo e delle immagini. C'è un altro elemento però: i collegamenti (link) tra una pagina e l'altra. Se indichiamo con delle frecce i collegamenti da una pagina all'altra, otteniamo una struttura come quella della Figura 4, che i matematici chiamano *grafo*.

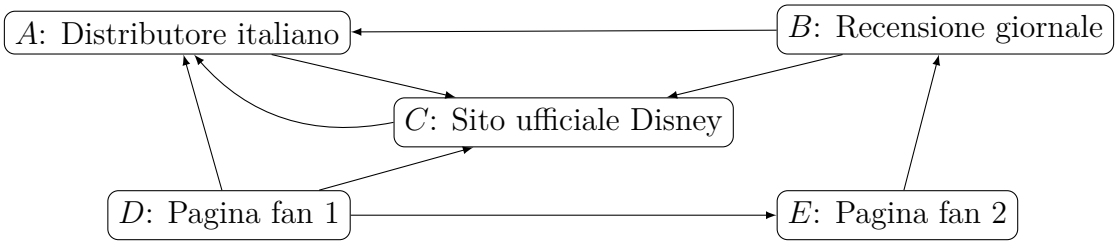


Figura 4: Collegamenti tra una pagina e l'altra, esempio.

Diverse soluzioni Il secondo problema, più difficile, è quello di decidere cosa vuol dire "la più importante". Qui il matematico si ferma, e riconosce che a questa domanda non c'è una risposta univoca, automatica. **Definire l'importanza è una scelta arbitraria.** Una volta fatto, possiamo cercare dei modi di calcolarla, ma prima dobbiamo accordarci su una definizione.

Le prime possibilità che ci possono venire in mente sono le seguenti:

1. Le pagine 'migliori' sono quelle che contengono il nome del film nel titolo, oppure scritto più in grande.

2. Le pagine 'migliori' sono quelle che contengono più spesso il nome del film.

Riuscite ad individuare il grosso problema con criteri di questo tipo? Sono *facilmente manipolabili*: se un fan modifica il testo della sua pagina, riesce facilmente a conquistare il primo posto, scalzando i siti ufficiali. I motori di ricerca della prima generazione, negli anni '90, utilizzavano criteri di questo tipo, e abusi come questo erano all'ordine del giorno. Spesso pagine di 'spam' contenevano solo il nome di un film, ripetuto moltissime volte. Un criterio un po' migliore è il seguente:

3. Le pagine 'migliori' sono quelle verso le quali ci sono molti collegamenti.

Abbiamo introdotto un elemento nuovo: **sfruttare la struttura dei link** per ottenere un rating più affidabile. Manipolare questo criterio è un po' più scomodo, ma ancora fattibile: un fan deve semplicemente creare molte pagine fittizie, che nessuno mai guarderà, ma che contengono un link al suo sito.

Voti più e meno importanti Ci serve invece un criterio che tenga conto che non tutti i collegamenti hanno lo stesso valore: un link proveniente dal sito della Disney 'vale' molto di più che non uno proveniente dal blog di un fan. In uno slogan: *una pagina è importante se viene linkata da pagine importanti.* Questa definizione è ricorsiva: apparentemente, per calcolare l'importanza di una pagina dobbiamo sapere l'importanza di tutte le pagine da cui partono collegamenti a lei, e così via. In realtà, vedremo che per calcolare questo punteggio basta risolvere un sistema di equazioni lineari. Vediamo di formalizzare la cosa: chiamiamo $x_A, x_B, \ldots, x_E$ l''importanza' delle pagine in Figura 4. Il sito B contiene collegamenti ad A e C; possiamo immaginare che esso 'voti' per A e C, o che 'distribuisca' la sua importanza equamente tra di essi:

$$x_C = \tfrac{1}{2}x_B + \ldots \text{altri termini} \ldots, \qquad x_A = \tfrac{1}{2}x_B + \ldots \text{altri termini} \ldots.$$

Analogamente, il D dividerà equamente la sua 'importanza' tra i tre siti verso cui ha link, e A la darà tutta al sito C. Complessivamente, le equazioni sono

$$\begin{cases} x_A = \frac{1}{2}x_B + x_C + \frac{1}{3}x_D, \\ x_B = x_E, \\ x_C = x_A + \frac{1}{2}x_B + \frac{1}{3}x_D, \\ x_D = 0, \\ x_E = \frac{1}{3}x_D. \end{cases} \tag{1}$$

Da sole, esse non sono ancora sufficienti a determinare univocamente l'importanza: dati $x_A, x_B, \ldots, x_E$ che soddisfano il sistema, possiamo moltiplicare tutto per due (o per tre, per quattro...) e ottenere un'altra soluzione altrettanto valida. Aggiungiamo una condizione per fissare il totale:

$$x_A + x_B + x_C + x_D + x_E = 1. \tag{2}$$

Il sistema formato da (1) e (2) ha una sola soluzione: $x_A = x_C = \frac{1}{2}, x_B = x_D = x_E = 0$. Cioè, *le uniche pagine importanti sono A e C*.

Possibili modifiche Vi soddisfa questo risultato? Se la risposta è no, dobbiamo **cambiare la definizione di importanza**. Per esempio, per evitare zeri nella soluzione, possiamo assumere che solo una certa percentuale dell'importanza, diciamo l'85%, venga distribuita in questo modo; il restante 15% viene suddiviso in parti uguali, in modo che nessuno resti a zero. Le equazioni così modificate diventano

$$\begin{cases} x_A = (1-\alpha)\frac{1}{5} + \alpha\left(\frac{1}{2}x_B + x_C + \frac{1}{3}x_D\right), \\ x_B = (1-\alpha)\frac{1}{5} + \alpha x_E, \\ x_C = (1-\alpha)\frac{1}{5} + \alpha\left(x_A + \frac{1}{2}x_B + \frac{1}{3}x_D\right), , \\ x_D = (1-\alpha)\frac{1}{5}, \\ x_E = (1-\alpha)\frac{1}{5} + \alpha\frac{1}{3}x_D. \end{cases} \tag{3}$$

dove abbiamo lasciato scritto esplicitamente il parametro $\alpha = 0.85$. Il sistema formato da (2) e (3) ha soluzione $x_A = 0.43$, $x_B = 0.063$, $x_C = 0.43$, $x_D = 0.030$, $x_E = 0.038$. Ora la soluzione sembra molto più ragionevole! Cattura il fatto che il fan 2 è più importante del fan 1, e che la recensione è più importante di entrambi. Ora siamo pronti per usare questo metodo su problemi più grandi che non uno con cinque sole pagine. Un momento, però...

Metodi di soluzione Come avete risolto i sistemi qui sopra (se avete provato a farlo da soli)? Molto probabilmente avete eliminato una variabile dopo l'altra, oppure avete sommato e sottratto tra loro le equazioni. Questi metodi richiedono un numero di operazioni che cresce come *il cubo del numero di pagine*. Se avessimo 10 pagine invece di 5, calcolarne l'importanza non richiederebbe il doppio del tempo, ma *otto volte tanto*. I numeri diventano difficili da gestire molto presto, anche per un calcolatore. Un computer moderno richiede circa un millisecondo per risolvere un sistema di 100 equazioni lineari in 100 incognite. Anche a questa velocità, però, calcolare l'importanza di *tutte* le pagine che parlano di un certo argomento è difficile: se abbiamo 100.000 pagine, servono 11 giorni. E se invece ne abbiamo 1.000.000?

Un metodo diverso, approssimato ma più efficiente, viene dal dare a queste equazioni un'interpretazione diversa. Supponiamo di guardare un utente annoiato che naviga tra le pagine seguendo ogni volta un link scelto a caso, tutti con la stessa probabilità. Dopo che ha seguito questa procedura per molto tempo, qual è la probabilità di trovarlo su una pagina piuttosto che su un'altra? Per esempio, chiamiamo x_C la probabilità di trovarlo su C. Perchè questo succeda, ci sono tre possibilità: o la pagina precedente era la A, che succede con probabilità x_A; o era nella pagina B (e in questo caso però visto che B ha due link uscenti c'è solo $\frac{1}{2}$ di possibilità che finisca in C), oppure era nella pagina D (e in questo caso ha $\frac{1}{3}$ di possibilità di finire in C). Quindi, abbiamo l'equazione $x_C = x_A + \frac{1}{2}x_B + \frac{1}{3}x_D$. Ma questa equazione l'abbiamo già incontrata in (1), e così quelle delle altre pagine: **le quantità $x_A, x_B, \ldots, x_E$ calcolate più sopra rappresentano le probabilità di trovare un 'utente casuale' in ognuna delle pagine**. L'equazione (2) ci dice per l'appunto che queste probabilità devono sommare a uno. In altre parole, più una pagina è importante, più è facile finirci seguendo collegamenti a caso, il che è ragionevole.

Questo suggerisce un altro metodo di soluzione: simuliamo il comportamento di questo navigatore casuale. Partiamo da una delle pagine, per esempio D. Poi, con un generatore di numeri casuali, scegliamo un link tra quelli presenti sulla pagina, per esempio quello verso E, e seguiamolo. Poi scegliamo un link a caso tra quelli presenti su E (in questo caso uno solo), e continuiamo così. Teniamo traccia della percentuale del tempo che abbiamo passato su ognuna delle pagine, e al crescere di k, queste quantità tenderà con grande probabilità alla soluzione del sistema (1)+(2). Un metodo leggermente più efficiente è questo: partiamo al passo 0 con probabilità uguali di essere su ogni pagina: $[p_A(0), p_B(0), p_C(0), p_D(0), p_E(0)] = [\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]$. Al passo 1, non è complicato calcolare che seguendo i collegamenti le probabilità di essere su ogni pagina sono $[p_A(1), p_B(1), p_C(1), p_D(1), p_E(1)] = [\frac{11}{30}, \frac{1}{5}, \frac{11}{30}, 0, \frac{1}{15}]$. Notate che non è possibile essere nella pagina D, perchè non ci sono collegamenti che puntano ad essa. Al passo 2, le probabilità sono $[\frac{7}{15}, \frac{1}{15}, \frac{7}{15}, 0, 0]$, e al passo successivo arriviamo alla soluzione vera del sistema (1)+(2).

Teletrasporto Riusciamo ad attribuire un significato simile alle equazioni (3)? Sì, e anche questa volta la soluzione è molto elegante: esse corrispondono alle probabilità di incontrare in una certa pagina un navigatore che con probabilità α segue un collegamento, oppure altrimenti (con probabilità $1 - \alpha$) si trasferisce su una pagina scelta a caso uniformemente tra tutte. Provate a scrivere le equazioni corrispondenti a questo comportamento e a verificarlo.

Costo Quante operazioni sono necessarie per calcolare le probabilità al passo $t+1$, date quelle al passo t? Le equazioni sono identiche a quelle di (1) (o (3)), solo

che le quantità a sinistra dell'uguale si riferiscono al passo $t+1$, quelle a destra al passo t: per esempio,

$$p_C(t+1) = p_A(t) + \tfrac{1}{2}p_B(t) + \tfrac{1}{3}p_D(t).$$

C'è un addendo per ogni link *entrante* nella pagina C. Se ci sono k addendi, dobbiamo fare k moltiplicazioni e $k-1$ addizioni. Quindi il numero di operazioni in un 'passo' di questo algoritmo è $(2\ell - 1)n$, dove n è il numero di pagine e ℓ è il numero medio di link entranti in ogni pagina. Per esempio, nel grafo in Figura 4, in A e C arrivano tre link, in B ed E uno, in D nessuno, quindi $\ell = 8/5 = 1.6$ collegamenti.

Potrebbe preoccuparvi questa quantità ignota ℓ: ci sono pagine che ricevono molti collegamenti, per esempio non è irragionevole che la pagina di un film famoso venga raggiunta da centinaia di migliaia di link. Però questa elegante osservazione, che un matematico chiamerebbe 'teorema', vi dimostra che la media non può essere troppo alta:

Teorema 2. *Il numero medio di link entranti è uguale al numero medio di link uscenti da una pagina.*

Dimostrazione. Invece di pensare in termini di medie, pensiamo in termini di totali: ogni freccia ha una 'coda' e una 'punta', quindi il numero totale di frecce che escono da una pagina è uguale al numero totale di frecce che entrano in una pagina. □

Nel nostro esempio, da A, C, E esce un link solo, da B due, e da D tre: in media, 1.6 collegamenti, esattamente come previsto dal teorema. Anche per pagine molto lunghe, il numero di collegamenti presenti su una pagina non supererà qualche decina, quindi ℓ non può essere troppo grande.

Un ultimo dettaglio: quanti passi di questo algoritmo sono necessari per ottenere una buona approssimazione delle 'importanze' vere? è difficile rispondere, anche per un matematico, ma per matrici grandi algoritmi di questo tipo di solito sono più efficienti degli quelli 'diretti' che richiedono n^3 operazioni. Nel caso del sistema (2)+(3), per esempio, già al passo 3 abbiamo quattro cifre significative esatte.

passo	A	B	C	D	E
1	0.3417	0.2000	0.3417	0.0300	0.0867
2	0.4139	0.1037	0.4139	0.0300	0.0385
3	0.4344	0.0627	0.4344	0.0300	0.0385
4	0.4344	0.0627	0.4344	0.0300	0.0385
5	0.4344	0.0627	0.4344	0.0300	0.0385

Hubs and authorities Una variante di questo metodo è un modello chiamato *hubs and authorities*: esso riconosce che ci sono anche pagine che pur non essendo 'buone' contengono link a pagine di buona qualità. Per esempio, una pagina di recensioni conterrà link a buoni film, mentre un buon film magari conterrà solo link a film dello stesso produttore, non necessariamente buoni. Ad ogni pagina quindi associamo due punteggi, che chiamiamo y e z (quindi $y_A, z_A, y_B, z_B, \dots$). I punteggi y indicano quanto una pagina è buona come 'recensione', e i punteggi z indicano quanto sono buoni i suoi contenuti. Una pagina ha un punteggio y più alto quando contiene link a pagine con z alto, e ha un punteggio z alto quando viene puntata da pagine con y alto. Sapreste scrivere delle equazioni per calcolare questi punteggi?

L'importanza dell'importanza Algoritmi come questi sono diventati molto popolari negli ultimi anni; il primo indice a farne uso nell'ambito dei motori di ricerca è stato il *pagerank* di Google, che è stato uno dei fattori che ne hanno determinato il successo. Analisi automatiche di questo tipo compaiono in sempre più siti: i siti di e-commerce vogliono trovare i prodotti più adatti da consigliare ai clienti, i produttori di pubblicità vogliono trovare gli *ad* più rilevanti da mostrarci...

Il ruolo del matematico Trovare il modo migliore di calcolare questi indici è un problema che interessa matematici e informatici a diversi livelli: bisogna trovare buone definizioni di concetti come 'importanza', saperli calcolare in modo efficiente, e dimostrarne le proprietà. Per esempio, un problema che va individuato e risolto è capire come vanno modificati gli algoritmi precedenti se da una pagina non esce nessun link. Oppure: provate a vedere cosa succede se rimpiazzate il grafo di Figura 4 con uno in cui ci sono due insiemi di pagine separati, senza collegamenti tra l'uno e l'altro. Dovreste notare che il sistema (1)+(2) ha infinite soluzioni, mentre il sistema (2)+(3) continua a funzionare. Un matematico è in grado di dimostrare questi risultati e spiegare il comportamento del metodo. Infine, non è poi facile dimostrare che l'algoritmo che abbiamo descritto qui sopra produce sempre una soluzione con tutti gli x positivi o nulli, che è fondamentale per poter assegnare loro il significato di 'importanza'.

Insomma, anche nella scienza del web e dei *big data* (algoritmi per lavorare con grandi quantità di dati) il ruolo del matematico è fondamentale.

dott. Federico Poloni,
Ricercatore presso il Dipartimento di Informatica, Università di Pisa

6 Il Paradosso EPR e la Disuguaglianza di Bell

Alberto Abbondandolo, n.3, Settembre 2016

Conosciamo tutti l'importanza della matematica nella fisica: le teorie fisiche vengono espresse da modelli matematici - per esempio equazioni differenziali - che possono essere analizzati mediante tecniche matematiche più o meno sofisticate, fornendo risultati qualitativi e quantitativi da confrontarsi con quanto emerge dagli esperimenti. Quindi per un fisico la matematica è soprattutto un linguaggio - per formulare le proprie teorie - ed uno strumento - per dedurne delle conseguenze - ma in generale non una fonte di ispirazione.

Vi è però almeno un caso in cui la matematica si è rivelata fondamentale per ideare un esperimento. Si trattava di dare una risposta ad una questione che vedeva contrapposte le idee di Albert Einstein e Max Born. Il problema per la verità sembrava di natura più metafisica che fisica, e come tale impossibile da dirimere sperimentalmente. Invece, una brillante idea puramente matematica di John Bell ha permesso di riportare la questione su un piano fisico ed ha tenuto impegnate diverse generazioni di fisici sperimentali. In questo articolo vogliamo esporre l'idea di Bell, che come vedremo richiede solamente concetti elementari di probabilità e trigonometria. Prima però dobbiamo spiegare quale fosse la materia del contendere.

La meccanica quantistica. La nostra storia ha inizio nel 1900. La fisica dell'Ottocento, che pure aveva avuto un enorme successo nello spiegare numerosissimi fenomeni, comincia a mostrare i suoi limiti. Proprio nel primo anno del nuovo secolo Max Planck formula l'ipotesi, da lui stesso definita "disperata", che la luce sia composta da particelle indivisibili, i *fotoni*. Cinque anni più tardi Albert Einstein usa l'ipotesi di Planck per spiegare l'effetto fotoelettrico. Si è ormai innescata una rivoluzione scientifica che nel giro di un ventennio porterà alla nascita della *meccanica quantistica*, un'opera corale che vede tra i suoi protagonisti i fisici Werner Heisenberg, Paul Dirac, Luis de Broglie, Erwin Schrödinger, Niels Bohr e Max Born.

Seguiamo ora il consiglio di Richard Feynman, secondo il quale niente spiega la meccanica quantistica meglio dell'*esperimento della doppia fenditura*. Si tratta di un esperimento realizzato per la prima volta nel 1909 da sir Geoffrey Ingram Taylor. Una sorgente luminosa a bassissima intensità emette un fotone alla volta. Di fronte alla sorgente c'è uno schermo con due piccole fenditure. Lo schermo blocca gran parte dei fotoni, ma le due fenditure permettono ad alcuni di passare e di andare ad impressionare una lastra fotografica così sensibile da rivelare l'arrivo di un singolo fotone[1].

[1]In realtà per rivelare l'arrivo di singoli fotoni sono necessarie apparecchiature più complicate di una semplice lastra fotografica. Qui però vogliamo evitare di addentrarci in questioni tecniche e

Teniamo lo sguardo fisso sulla lastra: uno alla volta, cominciano ad apparire dei puntini luminosi. Ci aspetteremmo di vedere i puntini formare due macchie più o meno estese in corrispondenza delle due fenditure. Invece, quello che vediamo formarsi sotto i nostri occhi è l'immagine raffigurata in Figura 5: si tratta di tipiche *frange di interferenza*[2].

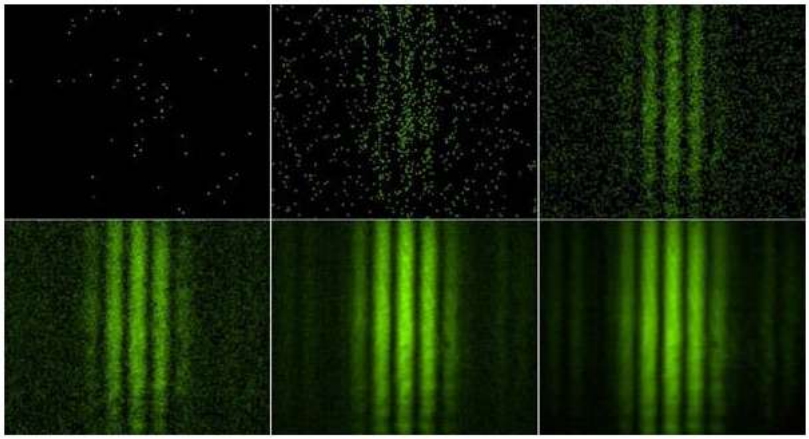


Figura 5: L'esperimento della doppia fenditura. La figura è ricavata dal sito web della Swiss Physical Society, vedi [7].

Un'immagine di questo tipo sarebbe spiegabilissima se l'intensità della luce fosse maggiore. In questo caso, infatti, la luce si comporterebbe come un'onda e le frange sarebbero causate dall'interferenza fra le due onde create dalle due fenditure. Più precisamente, al passaggio dallo schermo l'onda luminosa viene in parte bloccata e passano solamente due onde concentrate in corrispondenza delle due fenditure. Queste due onde si propagano, formando un alternarsi di picchi e valli, ed interferiscono tra loro: dove si incontrano due picchi oppure due valli l'onda risultante è amplificata (interferenza costruttiva), dove si incontrano un picco ed una valle l'onda si cancella (interferenza distruttiva). Le bande luminose sulla lastra corrispondono alle zone di interferenza costruttiva, le bande scure all'interferenza distruttiva.

Ma nel nostro esperimento l'intensità della luce è talmente bassa che i fotoni arrivano uno alla volta, come testimoniano i puntini luminosi che appaiono via via sulla lastra. Le frange di interferenza si manifestano solamente come fatto statistico, dopo che sulla lastra sono arrivati un gran numero di fotoni. La meccanica classica non è in grado di spiegare questo esperimento. Se il singolo fotone è un onda, perchè

ci permetteremo un certo grado di imprecisione nel descrivere gli esperimenti. Il lettore interessato a questi aspetti può consultare [2] e [7].

[2]Un video che mostra questo esperimento è disponibile sulla pagina web che riporta l'articolo [7], cliccando sulla figura a fine testo.

sulla lastra compare un puntino alla volta? E se è una particella, che come tale può passare da una sola delle due fenditure, perchè vediamo frange di interferenza invece che due macchie in corrispondenza delle due fenditure? Al congresso di Solvay del 1926 viene ufficializzata la spiegazione della meccanica quantistica: *il fotone è contemporaneamente un'onda e una particella.* Prima di raggiungere la lastra fotografica è un'onda che passa da entrambe le fenditure, generando due onde che interferiscono tra loro, esattamente come abbiamo spiegato. Quando raggiunge la lastra l'onda *collassa*, ossia appare una particella in un punto a caso, ma con probabilità maggiore nelle zone in cui l'ampiezza dell'onda è grande, minore dove questa ampiezza è piccola. è questa proporzionalità tra l'ampiezza dell'onda e la probabilità di trovare la particella in una data posizione che spiega il formarsi delle frange di interferenza dopo che sono arrivati numerosi fotoni. In altre parole, dobbiamo pensare al fotone come ad una particella, poiché la lastra ce lo rivela come tale, ma dobbiamo ammettere che prima di impressionare la lastra questa particella non abbia una posizione determinata e passi da entrambe le fenditure. Naturalmente a questa spiegazione qualitativa si aggiungono precise equazioni matematiche, che chiariscono da un lato come si propaghi l'onda, dall'altro come avvenga il suo collasso. Il successo della nuova teoria è immediato e subito si riescono a spiegare, qualitativamente e quantitativamente, fenomeni di fronte ai quali la meccanica classica si era arresa.

Le obiezioni di Einstein. La neonata meccanica quantistica non piaceva ad Albert Einstein, che pure ne era stato un precursore con la sua spiegazione dell'effetto fotoelettrico del 1905. Più precisamente, Einstein ne riconosceva i meriti e gli indubbi successi, ma non apprezzava né l'uso della probabilità in una teoria che voleva essere fondamentale né, soprattutto, l'esigenza, necessaria in meccanica quantistica, di dover dividere la realtà fisica in un mondo microscopico, a cui sono associate delle onde, ed un mondo macroscopico, che segue le leggi della meccanica classica e che interagendo con le onde del mondo microscopico ne provoca il collasso. Ecco cosa scriveva in una lettera del 1926 a Max Born, uno dei fondatori della meccanica quantistica ed interlocutore principale di Einstein per vari decenni:

"La meccanica quantistica è degna di ogni rispetto, ma una voce interiore mi dice che non è ancora la soluzione giusta. è una teoria che ci dice molte cose, ma non ci fa penetrare più a fondo il segreto del gran Vecchio. In ogni caso, sono convinto che questi non gioca a dadi con il mondo."

Einstein pensava alla meccanica quantistica, come era stata formulata dai suoi colleghi, come ad una teoria corretta ma deducibile da qualcosa di più fondamentale, un po' come la termodinamica è deducibile, in linea di principio, dalla meccanica di un gran numero di particelle. Per il più grande scienziato del ventesimo secolo la meccanica quantistica è pertanto *incompleta*, nel senso che non fornisce una

descrizione completa di uno stato fisico, un po' come la pressione, il volume e la temperatura non descrivono completamente un gas, una descrizione più accurata del quale richiederebbe la conoscenza di posizione e velocità di ciascuna delle sue molecole.

Nel 1935 Einstein scrive con i colleghi Boris Podolsky e Nathan Rosen un articolo nel quale i tre fisici sostengono di aver *dimostrato* che la meccanica quantistica è incompleta. Per seguire il loro argomento, che presentiamo in una versione leggermente modificata, abbiamo bisogno di ricordare alcuni fatti sulla polarizzazione della luce.

Un raggio luminoso può essere *polarizzato* in una qualsiasi direzione perpendicolare al suo cammino. Lo strumento che mette in evidenza questa proprietà della luce è il *polarizzatore*, ben noto a chi si occupa di fotografia. Si tratta di un dischetto colorato su cui è segnata una direzione. Se un raggio di luce polarizzato di un certo angolo θ incontra un polarizzatore la cui direzione coincide con θ, la luce passa[3]. Se il polarizzatore viene orientato perpendicolarmente a θ, la luce viene fermata. Se l'angolo è intermedio, passa soltanto una frazione della luce. A quel punto la luce passata è polarizzata lungo la direzione del polarizzatore, come si può verificare intercettando il raggio con un secondo polarizzatore.

Secondo la meccanica quantistica anche un singolo fotone possiede una polarizzazione. Questa può avere una direzione precisa oppure può essere indeterminata, esattamente come la posizione. Se il fotone è polarizzato, esso passa con certezza da un polarizzatore con la stessa orientazione, non passa da uno perpendicolare, passa con probabilità[4] $\cos^2\theta$ da un polarizzatore inclinato di un angolo θ rispetto alla propria polarizzazione. Se la polarizzazione del fotone è indeterminata, esso passa da un polarizzatore qualsiasi con probabilità 1/2. Se passa, assume la polarizzazione del polarizzatore.

Veniamo all'esperimento concettuale immaginato da Einstein, Podolsky e Rosen. Una delle conseguenze della meccanica quantistica è che possiamo produrre una sorgente che emette *coppie* di fotoni: i due fotoni di una stessa coppia - chiamiamoli A e B - hanno polarizzazione *indeterminata* ma *identica* e partono in due direzioni opposte. Possiamo verificare il fatto che la polarizzazione sia identica facendo passare i due fotoni per due polarizzatori orientati allo stesso modo e rivelando il loro eventuale passaggio con lastre fotografiche: vedremo un puntino su ciascuna lastra (con probabilità 1/2), oppure nessun puntino (ancora con probabilità 1/2), ma mai un puntino su una lastra e niente sull'altra. Supponiamo ora che i due polarizzatori siano molto lontani tra loro e che il primo (quello che intercetta il fotone A) sia

[3]Di un polarizzatore è rilevante la direzione, non il verso: gli angoli θ e $\theta + \pi$ individuano la stessa direzione. Gli angoli sono espressi in radianti.

[4]Esprimiamo le probabilità come numeri compresi tra 0 e 1: un certo evento ha probabilità p se su N tentativi avviene circa pN volte. Quindi probabilità 0 indica un evento impossibile, probabilità 1 un evento certo, probabilità 1/2 un evento che avviene nel 50% dei casi.

leggermente più vicino alla sorgente del secondo. Secondo la meccanica quantistica, dopo l'emissione i due fotoni hanno polarizzazione indeterminata. Il fotone A è il primo ad incontrare il polarizzatore: in quell'istante la sua indeterminazione collassa, il fotone A assume una polarizzazione ben definita, che coincide con quella del polarizzatore nel caso in cui A passi, con la polarizzazione perpendicolare altrimenti. Il fotone B incontrerà il suo polarizzatore solo un attimo dopo e si comporterà come il fotone A. Ne deduciamo che quando la polarizzazione di A assume un valore definito, il fotone B assume istantaneamente la stessa polarizzazione. Però uno dei postulati della teoria della relatività, elaborata dallo stesso Einstein, ci dice che nessun segnale può propagarsi con velocità maggiore della velocità della luce: se la distanza tra i due fotoni è molto grande, un segnale che parte da A non farà in tempo a raggiungere B prima che questi abbia incontrato il suo polarizzatore. La conclusione di Einstein, Podolsky e Rosen è che i fotoni devono possedere una loro polarizzazione già al momento dell'emissione ed il fatto che a questa la meccanica quantistica non dia alcun valore ne dimostra l'incompletezza.

La disuguaglianza di Bell. All'argomento di Einstein, Podolsky e Rosen, presto noto come "paradosso EPR" dalle iniziali dei suoi ideatori, i fondatori della meccanica quantistica risponderanno che la meccanica quantistica è completa, ma che il paradosso EPR mostra che il postulato della velocità massima dei segnali non ha validità generale. La discussione tra Einstein e Born andrà avanti per anni, senza che il primo riuscisse ad elaborare una teoria che "completasse" la meccanica quantistica, o che il secondo trovasse un argomento convincente per dimostrare che fosse necessario abbandonare il postulato della velocità massima dei segnali.

L'idea per dirimere la questione arrivò solo nel 1964, grazie al fisico irlandese John Bell, quando purtroppo Einstein era già morto da nove anni. L'argomento di Bell, di natura puramente matematica, è sorprendentemente semplice ed elegante e possiamo riassumerlo in queste pagine.

Bell parte dall'esperimento EPR, ma immagina che i due polarizzatori possano essere orientati con angoli diversi tra loro. Fissiamo una volta per tutte un riferimento cartesiano rispetto al quale riferire gli angoli che determinano la direzione dei polarizzatori. Se il primo polarizzatore ha direzione α e il secondo ha direzione β, possiamo determinare sperimentalmente la *correlazione* $\nu(\alpha, \beta)$: si fanno N esperimenti con i polarizzatori orientati come detto, dove N è un numero molto grande; tutte le volte che i fotoni hanno lo stesso comportamento (cioè passano entrambi o non passano entrambi) si somma $+1$; tutte le volte che i fotoni hanno comportamento diverso si somma -1; infine, si divide per N. Il numero ottenuto è $\nu(\alpha, \beta)$. Ad esempio, da quel che sappiamo sul comportamento dei fotoni polarizzati possiamo già prevedere che $\nu(\alpha, \alpha) = 1$ mentre $\nu(\alpha, \alpha + \pi/2) = -1$.

Vediamo quale valore assegni la meccanica quantistica alla correlazione $\nu(\alpha, \beta)$.

I due fotoni hanno polarizzazione indeterminata, ma distribuita in modo equiprobabile in ogni direzione. Se supponiamo che il fotone A raggiunga il suo polarizzatore un istante prima di B, otteniamo che con probabilità $1/2$ la sua polarizzazione assume la direzione di α ed A passa, mentre con uguale probabilità la sua polarizzazione assume direzione ortogonale ad α ed A non passa. Il fotone B assume istantaneamente la stessa polarizzazione di A. Nella prima metà dei casi assume quindi polarizzazione α e passa con probabilità $\cos^2(\beta-\alpha)$. Perciò delle circa $N/2$ emissioni il cui fotone A passa, circa $(N/2)\cos^2(\beta-\alpha)$ delle volte passa anche B e dobbiamo sommare 1, mentre nei rimanenti $(N/2)(1-\cos^2(\beta-\alpha))$ casi B non passa e dobbiamo sottrarre 1. Il contributo a $\nu(\alpha,\beta)$ di queste $N/2$ emissioni è dunque:

$$(N/2)\cos^2(\beta-\alpha) - (N/2)\Big(1-\cos^2(\beta-\alpha)\Big),$$

valore che, usando note formule di trigonometria, può essere riscritto come

$$(N/2)\cos 2(\beta-\alpha).$$

Un calcolo simile (o un argomento di simmetria), mostra che il contributo dell'altra metà delle emissioni (quelle in cui A non passa) è lo stesso, quindi il valore di $\nu(\alpha,\beta)$ previsto dalla meccanica quantistica si ottiene moltiplicando per due il valore sopra e dividendo poi per N, con risultato finale

$$\nu_{\text{MQ}}(\alpha,\beta) = \cos 2(\beta-\alpha).$$

Veniamo all'esperimento immaginato da Bell. Fissiamo tre angoli, $\alpha = 0$, $\beta = \pi/6$, $\gamma = \pi/3$, e consideriamo tre esperimenti diversi, che chiamiamo (α,β), (β,γ), e (α,γ): il primo angolo determina l'orientazione del primo polarizzatore, quello che intercetta il fotone A, il secondo angolo l'orientazione del secondo polarizzatore, raggiunto dal fotone B. Effettuiamo $3N$ esperimenti, scelti a caso ed in modo equiprobabile tra i tre esperimenti (α,β), (β,γ), e (α,γ). Ci aspettiamo che ciascun tipo di esperimento venga eseguito circa N volte. Mettiamo assieme i risultati dei circa N esperimenti di tipo (α,β) ed usiamoli per calcolare la correlazione $\nu(\alpha,\beta)$. Considerando gli altri due sottoinsiemi di esperimenti, troviamo $\nu(\beta,\gamma)$ e $\nu(\alpha,\gamma)$. Riassumiamo i risultati ottenuti in un unica quantità B, definita sommando i primi due numeri e sottraendo il terzo, ossia

$$B = \nu(\alpha,\beta) + \nu(\beta,\gamma) - \nu(\alpha,\gamma).$$

Usando la formula per ν_{MQ} che abbiamo ricavato sopra, concludiamo che il valore che la meccanica quantistica attribuisce a questa quantità è

$$\begin{aligned} B_{\text{MQ}} &= \nu(0,\pi/6) + \nu(\pi/6,\pi/3) - \nu(0,\pi/3) \\ &= \cos\pi/3 + \cos\pi/3 - \cos 2\pi/3 = 3/2. \end{aligned}$$

Vogliamo confrontare questo valore con quello previsto da una qualunque *teoria locale*, ossia una ipotetica teoria come quella che Einstein cercava di elaborare, dove valga il postulato della velocità massima dei segnali. La parola "locale" si riferisce al fatto che quel che avviene in un punto P dell'universo non deve avere alcun effetto istantaneo in un altro punto Q, ma l'effetto su Q può avvenire solamente dopo un intervallo di tempo sufficiente alla luce per viaggiare da P a Q.

In una teoria locale ciascuna coppia di fotoni deve possedere istruzioni su come comportarsi di fronte ad un polarizzatore, comunque orientato. Infatti i due fotoni di una stessa coppia hanno lo stesso comportamento di fronte a due polarizzatori orientati dello stesso angolo, comunque lontani tra loro: non potendo conoscere l'orientazione dei polarizzatori al momento dell'emissione e non potendo comunicare tra loro abbastanza rapidamente quando li incontrano, i due fotoni sono costretti ad aver concordato fra loro il comportamento da tenere di fronte ad una qualunque orientazione del polarizzatore. Etichettiamo ciascuna coppia di fotoni a seconda di come si comporterà di fronte ai tre angoli α, β, γ: questa etichetta è una stringa di tre simboli, ciascuno dei quali è un $+$ oppure un $-$. Il primo simbolo indica il comportamento della coppia di fotoni di fronte ad un polarizzatore orientato di un angolo α: è un $+$ se la coppia passa, un $-$ se non passa. Il secondo simbolo descrive il comportamento di fronte ad un polarizzatore orientato di un angolo β, mentre il terzo riguarda l'angolo γ, con le stesse regole. Ad esempio, i due fotoni di una coppia di tipo $+-+$ passeranno da due polarizzatori orientati di un angolo α oppure γ, mentre non passeranno se l'angolo è β. Questa classificazione suddivide le coppie di fotoni in $2^3 = 8$ tipi.

Nel nostro esperimento avvengono $3N$ emissioni di coppie di fotoni. Indichiamo con N_{+++} il numero di quelle emissioni in cui viene prodotta una coppia di tipo $+++$, ed usiamo una notazione analoga per i numeri delle emissioni degli altri sette tipi di coppie. Otteniamo così otto numeri $N_{+++}, \ldots, N_{---}$, che hanno somma $3N$. Ovviamente, ignorando i dettagli della teoria locale che stiamo considerando, non sapiamo nient'altro di questi otto numeri, alcuni dei quali potrebbero anche essere zero, nel caso in cui la natura vieti l'emissione di coppie di determinati tipi.

Fissiamo l'attenzione sull'esperimento (α, β). Dato che questo esperimento viene compiuto una volta su tre in maniera casuale, è ragionevole aspettarsi che delle N_{+++} coppie di fotoni di tipo $+++$ circa $N_{+++}/3$ siano sottoposte a questo esperimento. Analogamente, circa$N_{++-}/3$ coppie di fotoni di tipo $++-$ si trovano di fronte all'esperimento (α, β) e così via per gli altri sei tipi di coppie. Di tutte le circa N volte in cui viene effettuato l'esperimento (α, β), otterremo un risultato concorde (entrambi i fotoni passano o entrambi non passano) circa

$$\frac{1}{3}(N_{+++} + N_{++-} + N_{--+} + N_{---})$$

volte, mentre otterremo un risultato discorde (un fotone passa, l'altro no) circa

$$\frac{1}{3}(N_{+-+} + N_{+--} + N_{-++} + N_{-+-})$$

volte. Ricordando come è stato definito $\nu(\alpha, \beta)$, troviamo

$$\begin{aligned}\nu(\alpha,\beta) = \frac{1}{3N}(&N_{+++} + N_{++-} + N_{--+} + N_{---} \\ &-N_{+-+} - N_{+--} - N_{-++} - N_{-+-}).\end{aligned}$$

Analogamente, il lettore verificherà facilmente che

$$\begin{aligned}\nu(\beta,\gamma) = \frac{1}{3N}(&N_{+++} + N_{-++} + N_{+--} + N_{---} \\ &-N_{++-} - N_{-+-} - N_{+-+} - N_{--+}),\end{aligned}$$

e

$$\begin{aligned}\nu(\alpha,\gamma) = \frac{1}{3N}(&N_{+++} + N_{+-+} + N_{-+-} + N_{---} \\ &-N_{++-} - N_{+--} - N_{-++} - N_{--+}).\end{aligned}$$

Sommando i primi due numeri e sottraendo il terzo, deduciamo che la nostra ipotetica teoria locale predice il seguente valore di B:

$$\begin{aligned}B_{\text{Loc}} = \frac{1}{3N}(&N_{+++} + N_{++-} + N_{--+} + N_{---} \\ &+N_{+--} + N_{-++} - 3N_{+-+} - 3N_{-+-}).\end{aligned}$$

Questa formula ci permette di limitare in qualche modo il valore di B_{Loc}? Sì: infatti, ricordando che gli otto numeri $N_{+++}, \ldots, N_{---}$ hanno somma $3N$, possiamo pensare a B_{Loc} come alla media aritmetica di $3N$ numeri, dei quali $N_{+-+} + N_{-+-}$ valgono -3 e gli altri valogono 1. Tale media non potrà quindi superare 1: vale cioè la *disuguaglianza di Bell*

$$B_{\text{Loc}} \leq 1.$$

Abbiamo quindi trovato una quantità che puó essere calcolata sperimentalmente, la quantità B, a cui una qualsiasi teoria locale attribuisce un valore non superiore ad 1, mentre la meccanica quantistica le attribuisce il valore $3/2$, che è maggiore di 1. La conclusione è che nell'esperimento immaginato da Bell la previsione della meccanica quantistica non può essere ottenuta da nessuna teoria locale.

Escogitato un modo di dirimere la situazione sul piano teorico, la parola passa ai fisici sperimentali. Tranne che in un caso, l'esperimento di Holt e Pipkin del 1974, tutti gli esperimenti, a partire da quello di Freedman e Clauser del 1972 fino

agli esperimenti effettuati nel 2015 da tre gruppi indipendenti a Delft, Vienna e Boulder, evidenziano una violazione della disuguaglianza di Bell e sono in accordo con la meccanica quantistica. Nei primi esperimenti per la verità questo accordo è opinabile, data la scarsa affidabilità dei rivelatori, ma negli ultimi diventa più evidente. Nell'esperimento di Aspect, Dalibard e Roger del 1982, in particolare, l'orientazione dei polarizzatori viene scelta a caso da un computer dopo che la coppia di fotoni è stata emessa, in modo da escludere la possibilità che la posizione dei polarizzatori possa influenzare quale "tipo" di coppie venga emesso, possibilità che invaliderebbe l'argomento che ci ha portato alla disuguaglianza di Bell: non potremmo più dire che circa 1/3 degli N_{+++} fotoni di tipo $+++$ viene sottoposto all'esperimento (α, β).

La conclusione che dobbiamo trarre, almeno per il momento, è che su questo punto Einstein aveva torto: il mondo è in qualche misura non locale e vi sono segnali che si propagano istantaneamente.

Per saperne di più

[1] A. D. Aczel, *Entanglement*, Raffaello Cortina Editore 2004.

[2] T. L. Dimitrova e A. Weis, The wave particle duality of light: a demonstration experiment, *Am. J. Phys.* 76 (2008), 137–142.

[3] J. Bell, *Speakable and unspeakable in quantum mechanics*, Cambridge University Press 2004.

[4] G. C. Ghirardi, *Un'occhiata alle carte di Dio*, Il Saggiatore 1997.

[5] F. Selleri, *La fisica tra paradossi e realtà*, Progedit 2003.

[6] A. Zeilinger, *Il velo di Einstein*, Einaudi 2006.

[7] A. Weis e T. L. Dimitrova, Wave-particle duality of light for the classroom, *Swiss Physical Society*,

Alberto Abbondandolo,
Professore di Analisi, Fakultät für Mathematik, Ruhr-Universität Bochum

7 La Matematica dei Videogiochi

Marco Franciosi, n.3, Settembre 2016

Introduzione

Nel mondo contemporaneo, una grande opportunità di lavoro per le nuove generazioni viene dal mondo dei videogiochi. Poter lavorare nell'industria dei videogame non è un'eventualità così remota. Partecipare allo sviluppo di un nuovo software interattivo richiede competenze specifiche, ma anche fantasia e capacità comunicativa e può essere ricco di grandi soddisfazioni.

Nella creazione di un videogioco sono importanti le realizzazioni grafiche di ambienti e personaggi e la capacità di modificarle rapidamente. Per affrontare problemi di questo tipo non è sufficiente fare affidamento su computer sempre più potenti. Le nuove capacità di calcolo devono essere sfruttate al massimo. Occorre introdurre concetti matematici avanzati e sviluppare nuovi modi di fare i calcoli per ottenere risultati davvero gratificanti.

Lo schermo è fatto di pixel, ovvero è suddiviso in piccolissimi quadratini. Per una rappresentazione efficace non si può imporre al computer di tener conto di ciascun singolo pixel, separatamente dagli altri. Diventa utile introdurre equazioni capaci di legare i vari pixel e di adattarsi ai cambiamenti. Ad esempio, se si cambia il punto di vista, il paesaggio deve cambiare immediatamente. Analogamente il protagonista del gioco deve essere in grado di rispondere rapidamente agli stimoli del giocatore.

Alla base dei moderni software di grafica ci sono i concetti di interpolazione polinomiale, di parametrizzazione ed il loro utilizzo nell'ambito della geometria proiettiva.

Interpolazione polinomiale

L'interpolazione polinomiale è un metodo per disegnare delle curve o superfici (che corrispondono a equazioni polinomiali) capaci di approssimare una determinata forma, passando per certo numero di punti predeterminati. Il loro utilizzo permette al computer di memorizzare solamente le equazioni date e con velocissimi calcoli realizzare le curve o le superfici necessarie per costruire l'immagine con una precisione che può essere cambiata di volta in volta. Le idee matematiche che stanno alla base di questa teoria sono semplici e possono essere facilmente comprese dagli studenti delle scuole superiori. Il vantaggio nell'utilizzo dei polinomi risiede nella relativa facilità con cui il computer esegue i calcoli necessari nel loro utilizzo. Il punto di partenza per poter sviluppare tali idee si basa sulla nozione fondamentale di "parametrizzazione". Parametrizzare una curva significa descriverla mediante l'utilizzo di un parametro "esterno". Ad esempio se si considera il percorso di una

nave che si muove da un porto all'altro, la traiettoria percorsa può essere descritta mediante l'utilizzo del parametro tempo "t": ad ogni istante possiamo individuare la latitudine e la longitudine della nave stessa e poi possiamo disegnare il tragitto percorso segnando volta per volta le coordinate. Ovvero parametrizzare una curva corrisponde a disegnarne l'evoluzione al variare di un parametro "t".

Per capire come si arriva a sviluppare i primi esempi di interpolazione cominciamo spiegando come si può descrivere un segmento mediante l'uso di un parametro "t". Consideriamo nel piano Cartesiano il punto P di coordinate $(1, 2)$ e il punto Q di coordinate $(3, 6)$ e il segmento che li congiunge.

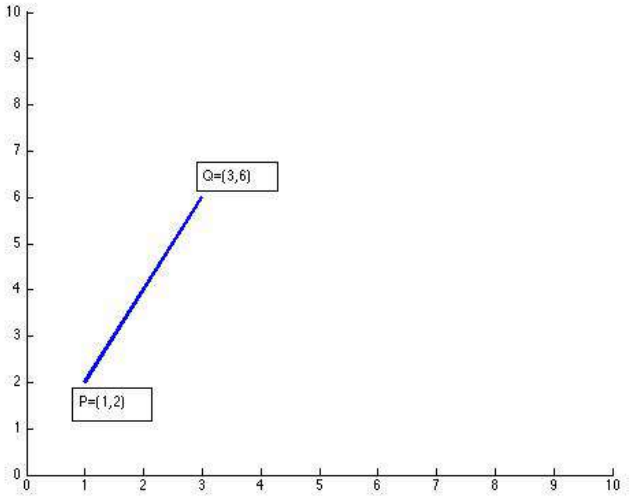


Figura 6: segmento $\overline{PQ}$

Usando la notazione vettoriale il segmento $\overline{PQ}$ può essere descritto come il cammino di un punto materiale che parte da P e si muove con velocità costante data dal vettore differenza $Q - P$.

Si ha pertanto la seguente descrizione parametrica del segmento

$$\begin{pmatrix}1\\2\end{pmatrix} + t \cdot \left[\begin{pmatrix}3\\6\end{pmatrix} - \begin{pmatrix}1\\2\end{pmatrix}\right] \quad con \;\; 0 \leq t \leq 1$$

ovvero

$$\overline{PQ} = (1-t) \cdot \begin{pmatrix}1\\2\end{pmatrix} + t \cdot \begin{pmatrix}3\\6\end{pmatrix} \quad con \;\; 0 \leq t \leq 1$$

Consideriamo adesso un terzo punto R di coordinate $(4, 8)$ e la spezzata generata P, Q ed R, ovvero l'unione dei due segmenti $\overline{PQ}$ e $\overline{QR}$.

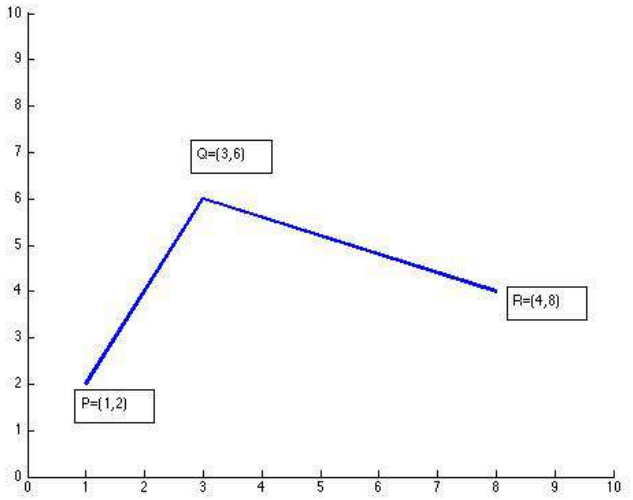


Figura 7: spezzata $\overline{PQR}$

Per costruire una curva liscia capace di approssimare la spezzata PQR iteriamo la costruzione precedente utilizzando un punto $P(t)$ sul primo segmento e un punto $Q(t)$ sul secondo.

Dal punto di vista algebrico otteniamo

$$\begin{cases} x = (1-t)^2 \cdot 1 + 2t(1-t) \cdot 3 + t^2 \cdot 4 \\ y = (1-t)^2 \cdot 2 + 2t(1-t) \cdot 6 + t^2 \cdot 8 \\ 0 \leq t \leq 1 \end{cases}$$

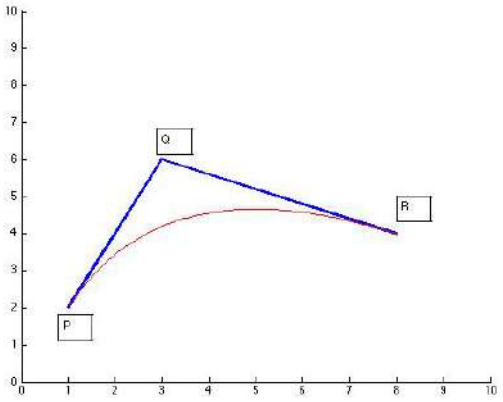


Figura 8: curva che approssima la spezzata PQR

La curva ottenuta segue il profilo della spezzata e negli estremi è tangente ai due segmenti. Se abbiamo invece un quarto punto S, possiamo iterare nuovamente il precedente argomento.

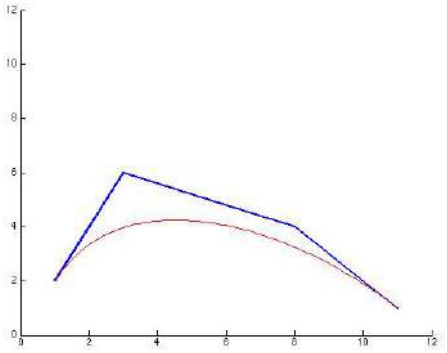


Figura 9: curva che approssima la spezzata $PQRS$

Riepilogando la curva che approssima la spezzata PQR è data dall'equazione (vettoriale)

$$\begin{aligned}&\Big[(1-t)Q+tR\Big]=\\&(1-t)^2\cdot P+2t(1-t)\cdot Q+t^2\cdot R\qquad\text{con}\;\;0\le t\le 1\end{aligned}$$

mentre la curva che approssima la spezzata $PQRS$ è data dall'equazione (vettoriale)

$$(1-t)^3\cdot P+3t(1-t)^2\cdot Q+3t^2(1-t)\cdot R+t^3\cdot S$$

con $0\le t\le 1$.
Le equazioni così ottenute sono determinate da polinomi fissati moltiplicati per i coefficienti dati dalle coordinate dei punti. Se si sposta un punto basta cambiare i corrispondenti coefficienti e si ottiene la nuova curva. Pertanto queste curve si prestano molto bene ad essere manipolate mediante trasformazioni (affini) del piano. Il computer per trasformarle ha bisogno solamente di avere le informazioni relative alle coordinate dei punti: attraverso un piccolo database che raccoglie le equazioni dei polinomi di base può ridisegnare la nuova curva in una frazione di secondo. Per chi è appassionato di algebra vale la pena sottolineare come nascono i polinomi usati nelle equazioni. Si tratta semplicemente di considerare lo sviluppo delle potenze del binomio $((1-t)+t)$:

$$\begin{aligned}((1-t)+t)^2&=(1-t)^2+2t(1-t)+t^2\\((1-t)+t)^3&=(1-t)^3+3t(1-t)^2+3t^2(1-t)+t^3\end{aligned}$$

(si noti che per qualsiasi valore di t il risultato è sempre 1). In generale, i monomi ottenuti dallo sviluppo della potenza ennesima del binomio sono i polinomi utilizzati

per creare una curva interpolante n+1 punti. Questa semplice osservazione permette di considerare tali monomi come i mattoni fondamentali per sviluppare una teoria completa e di facile applicazione.

Geometria Proiettiva

La prospettiva nasce dalla necessità di rappresentare in modo coerente lo spazio usuale (tridimensionale) su un dipinto o in generale su una superficie piana (bidimensionale). I moderni software dedicati all'analisi e allo sviluppo di immagini e soprattutto i videogiochi, sono basati in parte sulle idee esposte dai maestri italiani del rinascimento e trattate dai grandi geometri italiani nei primi anni del '900.

L'idea di fondo di questa teoria nasce dalla necessità di esplicitare gli aspetti matematici che stanno alla base della prospettiva lineare. Per prospettiva lineare si intende la rappresentazione su un piano ottenuta come intersezione dello stesso piano con il cono che ha il vertice nell'occhio e la base nell'oggetto da raffigurare. Per codificare tale rappresentazione occorre tenere conto che l'osservatore non vede i punti nello spazio dove realmente essi sono, ma vede solamente la luce che essi proiettano. Pertanto vengono identificati tutti i punti che stanno su una retta passante per l'occhio dell'osservatore (che consideriamo puntiforme). Questa semplice idea racchiude in se tutti i concetti sufficienti per costruire un modello algebrico del piano proiettivo.

Poniamo coordinate (x, y, z) nello spazio (z è la coordinata che indica la profondità) e l'occhio dell'osservatore nell'origine $O = (0, 0, 0)$. Una retta passante per O e per un punto P di coordinate (a, b, c) può essere descritta parametricamente dalle equazioni:

$$\begin{cases} x = ta \\ y = tb \\ z = tc \end{cases}$$

Nel nostro modello tutti i punti che giacciono sulla retta vengono rappresentati dallo stesso punto sulla superficie di coordinate (x, y) della rappresentazione. Pertanto l'idea naturale è quella di identificare tutti i punti che stanno su una retta passante per l'origine, ovvero di identificare un punto di coordinate (a, b, c) con un punto di coordinate (ta, tb, tc). Ad esempio le terne $(1, 2, 3)$ e $(3, 6, 9)$ rappresentano lo stesso punto del piano proiettivo.

Così come nel piano Cartesiano l'insieme delle coppie di numeri reali (x, y) permette di descrivere tutti i punti del piano "euclideo", in questo modo possiamo descrivere tutti i punti del piano proiettivo mediante terne di numeri reali (x, y, z) con la condizione che non siano tutti e tre nulli e con l'identificazione $(x, y, z) \sim (tx, ty, tz)$. Queste coordinate si chiamano coordinate omogenee.

Il modello che abbiamo costruito rappresenta una sorta di "espansione" del piano euclideo nel senso che tutti i punti di coordinate $(x, y, 1)$ possono venire identificati con i punti del piano Cartesiano di coordinate (x, y). Cosa succede ai punti che sono a "90 gradi" rispetto all'occhio dell'osservatore? Questi punti non possono essere identificati con alcun punto del piano Cartesiano, poiché la retta passante per l'occhio dell'osservatore non interseca il piano dell'immagine. L'insieme di tutti questi punti costituisce quella che si chiama "retta all'infinito". I punti sulla retta all'infinito hanno coordinate omogenee $(x, y, 0)$. In questo modo possiamo pensare al piano proiettivo come l'unione dei punti "al finito", descritti dalle coordinate $(x, y, 1)$, con i punti "all'infinito", descritti dalle coordinate $(x, y, 0)$. Un punto all'infinito corrisponde al punto di intersezione di due rette parallele.

Il vantaggio nell'utilizzo delle coordinate omogenee risiede da un lato nella semplificazione delle procedure di calcolo necessarie per realizzare le trasformazioni del piano, e dall'altro nell'interpretazione dei punti del piano come immagine dei punti dello spazio. Infatti un punto nello spazio di coordinate (x, y, z) lo possiamo proiettare nel piano nel punto di coordinate $(x/z, y/z, 1)$. Ovvero nella tela del pittore, o ancor di più nello schermo del computer, il punto di coordinate $(x/z, y/z)$ rappresenta l'immagine del punto dello spazio 3D di coordinate (x, y, z).

Le trasformazioni dei punti nello spazio, quali ad esempio traslazioni o rotazioni (movimenti tipici di un corpo rigido che si muove) vengono lette come semplici trasformazioni dei punti del piano, facilmente calcolabili con qualsiasi tipo di computer. Analogamente, il cambiamento di punto di vista non è nient'altro che una trasformazione lineare dello spazio tridimensionale che può essere facilmente scritta mediante le coordinate (omogenee) del piano attraverso l'utilizzo di matrici.

Parametrizzazioni nel piano proiettivo

E' possibile creare curve di interpolazione nel piano proiettivo attraverso l'uso delle coordinate omogenee. Si può pensare a una curva di questo tipo come immagine sul piano di una curva nello spazio tridimensionale. Infatti, considerando nello spazio 3D una curva espressa mediante le equazioni

$$\begin{cases} x = p(t) \\ y = q(t) \\ z = r(t) \end{cases}$$

(dove $p(t), q(t), r(t)$ sono polinomi nella variabile t), è facile vedere che la sua immagine nel piano è descritta dai punti di coordinate

$$(x/z, y/z, 1) = (p(t)/r(t), q(t)/r(t), 1).$$

Ovvero, identificando tali punti con i punti del piano otteniamo una parametrizzazione espressa mediante funzioni razionali (cioè frazioni di polinomi) del parametro "t".

Considerando il piano come una porzione del piano proiettivo possiamo utilizzare le coordinate omogenee. Il vantaggio delle coordinate omogenee risiede nel fatto che è possibile moltiplicare per il denominatore comune tutti e tre i valori delle coordinate e quindi esprimere tale punto nella forma più facilmente computabile $(p(t), q(t), r(t))$. Dal punto di vista computazionale è la stessa parametrizzazione della curva nello spazio, però rappresenta una curva nel piano proiettivo. Questa forma è più maneggevole dal punto di vista dei calcoli. In effetti è possibili compiere operazioni mediante l'utilizzo di matrici ottenendo con rapidità una nuova curva dello stesso tipo. In questo modo i calcoli si fanno utilizzando tutte e tre le coordinate. Solamente alla fine si effettua la divisione per la terza coordinata ottenendo una curva che possiamo rappresentare sullo schermo del computer.

Per capire meglio questi concetti vediamo come si parametrizza la circonferenza del piano di centro l'origine e raggio 1. Tale circonferenza ha equazione $x^2 + y^2 = 1$ Per parametrizzarla consideriamo un fascio di rette passanti per il punto $Q = (0, -1)$. Usiamo "t" come parametro per descrivere il coefficiente angolare delle rette del fascio. Ogni retta del fascio interseca la circonferenza nel punto Q e in un altro punto $P(t)$ che dipende dal coefficiente angolare t.

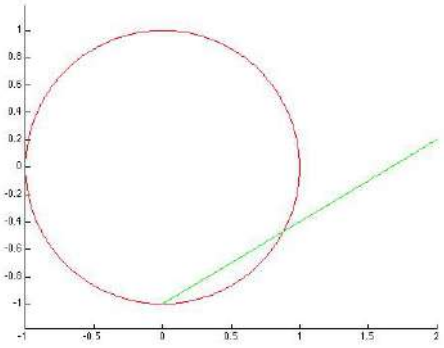


Figura 10: parametrizzazione circonferenza mediante fascio di rette

Con facili calcoli si ottiene che le coordinate x e y del punto $P(t)$ sono espresse dalle seguenti espressioni.

$$\begin{cases} x = \frac{2t}{1+t^2} \\ y = \frac{t^2-1}{1+t^2} \end{cases}$$

In questo modo otteniamo una parametrizzazione razionale della circonferenza (senza il punto U di coordinate $(0,1)$ che corrisponde all'intersezione con una retta parallela all'asse delle y). In coordinate omogenee si legge

$$\begin{cases} x = 2t \\ y = t^2 - 1 \\ z = 1 + t^2 \end{cases}$$

In conclusione mediante l'utilizzo di funzioni razionali (cioè di frazioni di polinomi) è possibile migliorare il livello di approssimazione di una curva interpolante e disegnare curve fondamentali come circonferenze e ellissi.

In coordinate omogenee tali curve si rappresentano mediante polinomi e quindi si prestano con molta facilità ad essere manipolate da programmi di grafica.

Conclusioni

Nei moderni programmi di grafica si utilizzano le curve NURBS (Non-Uniform Rational B-Splines). Le curve NURBS si ottengono mediante raffinamenti delle tecniche illustrate precedentemente. Per realizzare queste curve si aggiungono due ulteriori passi alla realizzazione vista precedentemente. Partendo dalla curva spezzata originale, in primo luogo vengono dati dei pesi, eventualmente diversi, ai vertici della spezzata. Un peso grande significa una maggiore capacità di attrazione della curva verso siffatto vertice. In secondo luogo, per ottimizzare il calcolo del computer si percorre la curva con velocità alta nei tratti che non danno preoccupazioni (ad esempio percorsi quasi rettilinei), mentre si rallenta nei tratti in cui si ha un cambio di direzione o una curvatura alta. Questo risultato si ottiene suddividendo l'intervallo di tempo $[0,1]$ in piccoli intervalli ciascuno di lunghezza diversa, a seconda del grado di accuratezza necessario per realizzare la porzione di curva voluta, e percorrendo tale intervallo con velocità inversamente proporzionale alla sua lunghezza. Le idee che stanno alla base di tali procedimenti sono però le stesse utilizzate da Piero della Francesca nel rinascimento per realizzare i suoi quadri e da Castelnuovo e Enriques per studiare le curve e le superfici algebriche. Questo a mio giudizio è un buon esempio per evidenziare l'importanza di una solida base scientifica e culturale capace di fornire gli strumenti per affrontare le moderne tecnologie.

Marco Franciosi,
Ricercatore di Geometria presso il Dipartimento di Matematica, Università di Pisa

8 Proprietà di Riscalamento e Dimensione

Emanuele Paolini, n.4, Febbraio 2017

Le proprietà di riscalamento delle misure sono un concetto molto semplice che però spesso viene ignorato nei percorsi scolastici. In queste note tenteremo di mettere in evidenza questo strumento, proponendo alcuni problemi a cui può essere applicato. In particolare vedremo un'interessante dimostrazione del teorema di Pitagora e faremo una digressione nelle dimensioni frazionarie.

Problema 1. *Una normale bottiglia di vino ha una capacità di* 3/4 *di litro. Una bottiglia di tipo* Jéroboam *contiene invece* 3 *litri. Quanto è il rapporto tra le altezze delle due bottiglie?*

Questo problema è molto semplice se affrontato con considerazioni dimensionali. La risoluzione non dipende dalla forma della bottiglia. L'unica informazione (che viene sottointesa) è il fatto che i diversi formati di bottiglia differiscono per una *similitudine*.

Similitudine

Una *similitudine* di fattore q è una particolare trasformazione geometrica che ha la proprietà di modificare ogni lunghezza moltiplicandola per lo stesso fattore q. Se σ è una similitudine, e P, Q sono due punti qualunque, si avrà quindi:

$$d(\sigma(P), \sigma(Q)) = q \cdot d(P, Q)$$

dove $d(P, Q)$ denota la distanza tra i punti P e Q e $d(\sigma(P), \sigma(Q))$ è la distanza dei punti immagine di P e Q dopo aver applicato la similitudine σ.

Le similitudini mantengono gli angoli e i rapporti tra le lunghezze dei segmenti. Di conseguenza si dice che le similitudini mantengono la forma.

I diversi formati di bottiglie differiscono per la dimensione, ma hanno tutti la stessa forma. Il Problema 1 ci chiede quindi di determinare il rapporto di similitudine tra due oggetti, conoscendone il volume.

Proviamo ora a considerare un problema apparentemente più semplice del Problema 1.

Problema 2. *Le patatine* tuberine *vengono normalmente confezionate in un cilindro di altezza 18 cm e diametro 9 cm. Una confezione contiene 50g di patatine. Se le patatine venissero confezionate in un cilindro di dimensioni doppie, quanti grammi di patatine ci aspetteremmo di trovare?*

È sensato supporre che il peso delle patatine sia proporzionale al volume occupato. Sarà quindi sufficiente determinare il rapporto tra i volumi dei due cilindri per ottenere il rapporto tra i pesi.

Sappiamo che il volume v di un cilindro di diametro d e altezza h è dato dalla formula:

$$v = \pi \left(\frac{d}{2}\right)^2 h = \frac{\pi}{4} d^2 h.$$

Come varia il volume se moltiplichiamo le misure del cilindro per uno stesso coefficiente q? Poniamo $H = qh$, $D = qd$. Si otterrà allora

$$\frac{V}{v} = \frac{\frac{\pi}{4} D^2 H}{\frac{\pi}{4} d^2 h} = \frac{\frac{\pi}{4}(qd)^2(qh)}{\frac{\pi}{4} d^2 h} = q^3. \tag{4}$$

Abbiamo quindi osservato che una similitudine di rapporto q modifica il volume dei cilindri di un rapporto q^3. Nel Problema 2 il rapporto di riscalamento è $q = 2$ e dunque possiamo affermare che il cilindro di dimensioni doppie avrà un volume pari a $2^3 = 8$ volte il volume del cilindro piccolo. Possiamo quindi aspettarci che anche il peso del suo contenuto venga moltiplicato per 8 e quindi sia pari a $8 \cdot 50\, g = 400\, g$.

Nell'equazione (4) abbiamo osservato come il cofficiente $\pi/4$ nella formula del calcolo del volume si elide quando facciamo il rapporto tra volumi di solidi simili. Questo ci fa intuire che il coefficiente q^3 di riscalamento del volume non dipende dalla forma del solido. In effetti se facciamo lo stesso calcolo per un cubo $v = \ell^3$, un parallelepipedo $v = abc$ o per una sfera $v = \frac{4}{3}\pi r^3$, otterremo sempre lo stesso risultato:

$$\frac{(q\ell)^3}{\ell^3} = \frac{q^3\ell^3}{\ell^3} = q^3, \qquad \frac{(qa)(qb)(qc)}{abc} = q^3, \qquad \frac{\frac{4}{3}(qr)^3}{\frac{4}{3}r^3} = q^3.$$

Possiamo convincerci che qualunque sia la forma di un solido, un riscalamento di fattore q determina un riscalamento del volume di un fattore q^3. Se lo volessimo dimostrare dovremmo però ricordarci come si definisce il volume di un solido qualunque. Quello che si fa è approssimare il solido tramite piccoli cubetti di lato ℓ. Se il solido si ricopre approssimativamente con N cubetti di lato ℓ, il suo volume sarà circa $v = N\ell^3$. Se il solido viene riscalato di un fattore q, lo si potrà approssimare con N cubetti di lato $q\ell$. Dunque il suo volume sarà circa $V = N(q\ell)^3 = q^3 N\ell^3 = q^3 v$, come avevamo intuito.

Queste considerazioni ci permettono di risolvere il Problema 1 che ci chiedeva qual è il rapporto tra le altezze della bottiglia Jéroboam da 3 litri e la normale bottiglia da $0,75$ litri. Il rapporto tra le altezze è pari al fattore q di riscalamento. Visto che il rapporto tra i volumi è $3/(3/4) = 4 = q^3$ si trova $q = \sqrt[3]{4} \approx 1,59$.

Dimensione di una misura

L'esponente 3 nel fattore q^3 di riscalamento del volume identifica la dimensione della misura di volume. Diremo che il volume V è una misura di dimensione 3 in

quanto se un qualunque solido X viene riscalato di un fattore q si ha $V(qX) = q^3V(X)$ (dove abbiamo indicato con $V(X)$ il volume del solido X e abbiamo denotato con qX il riscalamento di X di un fattore q). Similmente l'area è una misura di dimensione 2 in quanto riscalando una superficie di un fattore q la sua area viene moltiplicata per un fattore q^2. La lunghezza è invece una misura di dimensione 1 in quanto se una curva viene riscalata di un fattore q la sua lunghezza viene moltiplicata per $q = q^1$.

Una *misura* è una funzione m che associa ad ogni figura geometrica F un numero $m(F)$. Senza entrare in dettagli tecnici, l'unica proprietà veramente rilevante di una misura è la *additività* ovvero la proprietà $m(F \cup G) = m(F) + m(G)$ quando F e G sono figure che non si sovrappongono. Tutte le misure che stiamo considerando sono inoltre invarianti per isometria, cioè $m(F) = m(F')$ se F' è congruente a F.

In generale una *misura* m si dice avere dimensione d se per ogni figura F si ha

$$m(qF) = q^d m(F).$$

Abbiamo finora osservato che la lunghezza è una misura 1-dimensionale, l'area è 2-dimensionale e il volume è 3-dimensionale. Possiamo sfruttare queste semplici informazioni nei seguenti problemi.

Problema 3. *Daniele ha disegnato il circuito di Monza in scala 1:1000 sul pavimento della terrazza. Sapendo che il circuito reale è lungo* $5793\,m$*, quanto sarà lungo il circuito disegnato da Daniele?*

Contando le piastrelle Daniele ha determinato che l'area racchiusa dal circuito in scala è circa $6,5\,m^2$*. Quanti metri quadri racchiude il vero circuito?*

Le macchinine che Daniele usa per giocare sono invece in scala $1:50$*. Se per dipingere la macchinina Daniele utilizza 1 tubetto di vernice rossa, quanti tubetti gli sarebbero necessari per dipingere la macchina vera?*

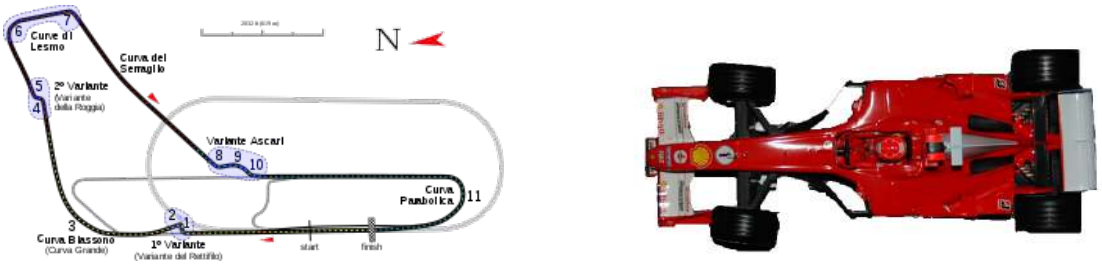


Figura 11: La mappa di un circuito e una macchinina sono esempi di oggetti riscalati. (figura di Will Pitenger e foto di Premnath Kudva, licenze Creative Commons)

Risolviamo i quesiti del Problema 3. Il circuito disegnato da Daniele è ottenuto, da quello reale, per mezzo di una similitudine di fattore $q = 1/1000$. La lunghezza

della pista è una misura 1-dimensionale, e quindi riscala dello stesso fattore q della similitudine. Dunque una curva di lunghezza $5793\,m$, riscalata, risulterà di lunghezza $q \cdot 5793\,m = 5,793\,m$. L'area è invece una misura 2-dimensionale e dunque l'area racchiusa dal circuito in scala è pari a q^2 volte l'area reale. Dunque l'area reale si ottiene moltiplicando per $1/q^2 = 1.000.000$ e risulta quindi pari a $6,5$ milioni di metri quadri (ovvero $6,5$ chilometri quadri). Per quanto riguarda la vernice utilizzata per dipingere la macchinina, possiamo assumere (non avendo maggiori informazioni a disposizione) che questa sia proporzionale all'area della carrozzeria della macchinina. E dunque, come l'area, sarà una misura 2-dimensionale. In questo caso il fattore di scala è $q = 1/50$ e dunque dividere per q^2 significa moltiplicare per $50^2 = 2500$. Stimiamo quindi che sarebbero necessari 2500 tubetti di vernice per dipingere la vettura reale.

Problema 4. *Osserviamo che i fogli di formato A4 (quelli usualmente utilizzati nelle macchine fotocopiatrici o nei quadernoni) hanno come forma un rettangolo che se diviso a metà lungo il lato più lungo dà origine a due fogli di formato A5 (quelli usualmente utilizzati nei quaderni piccoli) che hanno le stesse proporzioni del foglio iniziale. Qual è il rapporto dei due lati di un foglio A4?*

Sapendo che il formato A4 è a sua volta la metà dell'A3, l'A3 metà dell'A2, l'A2 metà dell'A1, l'A1 la metà dell'A0 e sapendo che il foglio di formato A0 ha un'area di 1 metro quadro, calcolare le lunghezze dei due lati di un foglio di formato A4.

Il foglio A4 ha area doppia del foglio A5. Se q è il rapporto di similitudine tra i due formati, essendo l'area una misura 2-dimensionale, si ha dunque $q^2 = 2$ da cui $q = \sqrt{2}$. Significa quindi che il rapporto tra i lati lunghi dei due formati è $\sqrt{2}$, ma il lato lungo del foglio A5 è uguale al lato corto del foglio A4 dunque il rapporto tra i due lati del foglio (sia A4 che, di conseguenza, A5) è $\sqrt{2}$.

Ad ogni suddivisione del foglio A0 l'area si dimezza, quindi il formato A4 ha una area pari a $1/2^4 = 1/16$ dell'area del formato A0 cioè $1/16$ di metro quadro. Chiamata x la lunghezza del lato corto, il lato lungo è $\sqrt{2}x$ (per quanto visto prima) e dunque l'area è $1/16\,m^2 = \sqrt{2}x^2$ da cui si ricava la lunghezza del lato corto

$$x = \frac{1}{\sqrt{16\sqrt{2}}}\,m = \frac{1}{4\sqrt[4]{2}}\,m \approx 21,02\,cm$$

e di conseguenza il lato lungo

$$\sqrt{2} \cdot x \approx 29,73\,cm.$$

Misure 0-dimensionali

Finora ci siamo occupati solamente delle misure di dimensione 1, 2 e 3. Vivendo in uno spazio 3-dimensionale queste sono le dimensioni su cui possiamo avere una diretta esperienza.

Dal punto di visto matematico non c'è però una limitazione fisica. Ha perfettamente senso definire e utilizzare misure di dimensione maggiore alla terza. Non vogliamo qui entrare in questo ambito che sarebbe affascinante, ma ci porterebbe molto lontano dagli altri concetti su cui intendiamo concentrarci. Possiamo però dedicarci per un attimo all'altro estremo dello spettro: quali sono le misure di dimensione 0? La dimensione 0 è quasi banale, ma può essere utile osservare come abbia perfettamente senso e rientri nel contesto generale che stiamo descrivendo.

Problema 5. *La macchinina di formula uno di Daniele è in scala* $1:50$. *Sapendo che la macchinina vera ha quattro ruote, quante ruote ha la macchinina in scala?*

Il concetto, ovvio, che vogliamo mettere in evidenza è il fatto che ci siano delle misure che sono invarianti per riscalamento. Una di queste è il *numero* ovvero la misura che *conta* gli elementi di una figura. Questa risulta essere una misura 0-dimensionale, in quanto il numero di elementi di una figura riscalata di un fattore q viene moltiplicato per $q^0 = 1$ cioè resta invariato.

La dimostrazione del teorema di Pitagora

Al matematico ungherese Paul ErdÅ‘s piaceva immaginare ci fosse un libro "divino" in cui tutti i teoremi matematici venissero dimostrati con procedimenti eleganti e sintetici. La proprietà di riscalamento dell'area ci permette di proporre una dimostrazione, essenziale e sintetica, del teorema di Pitagora.

Dobbiamo prima capire l'essenza dell'enunciato del teorema. Quando si afferma che l'area del quadrato costruito sull'ipotenusa è uguale alla somma delle aree dei quadrati costruiti sui cateti, non è veramente importante che la figura geometrica scelta sia un quadrato. Se ad esempio invece di un quadrato usassimo un pentagono regolare, avremmo comunque che l'area del pentagono costruito sull'ipotenusa è uguale alla somma delle aree dei pentagoni costruiti sui cateti in quanto l'area del pentagono è proporzionale al quadrato del lato, essendo l'area una misura 2-dimensionale (Figura 12).

Dunque il teorema di Pitagora è equivalente ad affermare che una volta scelta una figura qualunque, se questa viene riscalata in proporzione alla lunghezza dell'ipotenusa di un triangolo rettangolo, la sua area risulterà uguale alla somma delle aree delle figure riscalate in proporzione della lunghezza dei cateti. Se il teorema vale per una certa forma fissata, allora varrà per qualunque forma e in particolare per il quadrato.

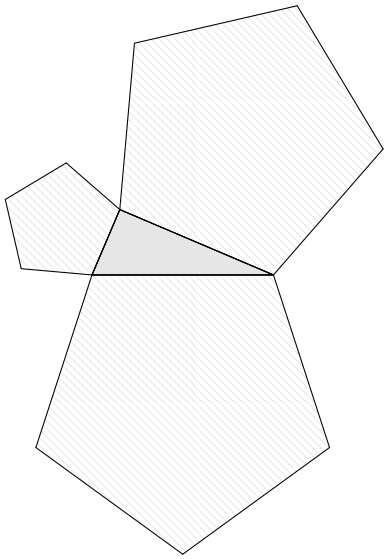


Figura 12: L'area del pentagono costruito sull'ipotenusa è uguale alla somma delle aree dei pentagoni costruiti sui cateti.

Per dimostrare il teorema è quindi sufficiente scegliere una forma opportuna... quella che scegliamo è il triangolo stesso!

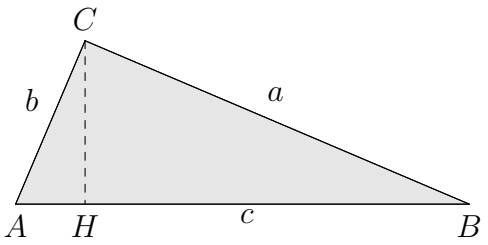


Figura 13: La dimostrazione del teorema di Pitagora può essere fatta semplicemente tracciando l'altezza rispetto all'ipotenusa.

Se ABC è il triangolo con un angolo retto in C, suddividiamo il triangolo in due parti tracciando l'altezza CH rispetto all'ipotenusa. I triangoli ACH e BCH sono simili al triangolo ABC perché sono rettangoli in H e condividono un angolo in A o in B. Tali triangoli sono quindi simili tra loro e ognuno di essi ha come ipotenusa uno dei tre lati del triangolo iniziale ABC. È d'altronde ovvio che l'area del triangolo ABC è uguale alla somma delle aree di ACH e BCH e dunque il teorema è dimostrato.

Possiamo ripetere il ragionamento in maniera più formale. Se chiamiamo a, b le lunghezze dei cateti e c la lunghezza dell'ipotenusa possiamo osservare che il triangolo BCH si ottiene riscalando il triangolo ABC di un fattore a/c mentre il triangolo ACH si ottiene sempre da ABC ma con un fattore di riscalamento pari a b/c. Chiamata $\mathcal{A}$ l'area del triangolo ABC si ha che (visto che l'area è una misura 2-dimensionale) l'area di ACH è pari a $(a/c)^2\mathcal{A}$ e l'area di BCH è pari a $(b/c)^2\mathcal{A}$, dunque si ottiene

$$\left(\frac{a}{c}\right)^2 \mathcal{A} + \left(\frac{b}{c}\right)^2 \mathcal{A} = \mathcal{A}$$

da cui, semplificando $\mathcal{A}$, e moltiplicando ambo i membri per c^2 si ottiene

$$a^2 + b^2 = c^2.$$

Dimensioni frazionarie: i *frattali*

Cerchiamo di determinare ora un metodo per definire la dimensione di una figura geometrica.

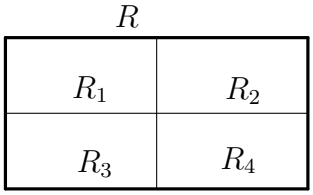


Figura 14: Un rettangolo può essere suddiviso in quattro rettangoli ognuno dei quali è una copia dell'originale riscalato di un fattore 1/2.

Se consideriamo un rettangolo R osserviamo che R può essere suddiviso in quattro rettangoli R_1, R_2, R_3, R_4 simili a R con un fattore di scala $q = 1/2$. Se m è una generica misura d-dimensionale, cioè una misura che soddisfa la relazione $m(qA) = q^d m(A)$, si avrà allora

$$m(R) = m(R_1) + m(R_2) + m(R_3) + m(R_4) = 4m(R/2) = \frac{4}{2^d} m(R)$$

(con $R/2$ si intende il rettangolo che si ottiene riscalando R di un fattore 1/2). Questa relazione risulta banalmente vera se $m(R) = 0$ oppure se $m(R) = \infty$. Ad esempio se $d = 3$ la misura m sarebbe il volume e si avrebbe chiaramente $m(R) = 0$ (un rettangolo è assimilabile ad un parallelepipedo di altezza zero, quindi di volume

zero). Viceversa se scegliessimo $d = 1$ la misura m sarebbe la lunghezza e si avrebbe $m(R) = \infty$ (un rettangolo è assimilabile ad una unione infinita di segmenti, quindi deve avere lunghezza infinita). Se però andiamo a cercare un d per il quale si abbia $m(R) \neq 0$ e $m(R) \neq \infty$, allora possiamo dividere ambo i membri di questa uguaglianza per $m(R)$ ottenendo:

$$1 = \frac{4}{2^d}$$

da cui $2^d = 4$ ovvero $d = \log_2 4 = 2$. Dunque $d = 2$ è l'unica dimensione per la quale il rettangolo può avere una misura finita e non nulla.

Lo stesso risultato si sarebbe ottenuto suddividendo il rettangolo in 9 rettangoli ognuno dei quali riscalato di un fattore 1/3.

Proviamo a ripetere l'esperimento con un cubo C. In questo caso il cubo può essere suddiviso in 8 cubetti riscalati di un fattore 1/2. Si otterrà dunque

$$m(C) = 8m(C/2) = \frac{8}{2^d} m(C)$$

da cui, dividendo per $m(C)$ si ottiene $2^d = 8$ e quindi $d = 3$ come ci saremmo aspettati.

Questo ragionamento non può essere fatto con qualunque figura (almeno non così facilmente). L'importante proprietà che stiamo sfruttando è che queste figure hanno la caratteristica di poter essere suddivise in un certo numero di copie riscalate di sé stesse. Questa proprietà si chiama *autosimilarità* ed è soddisfatta da altre figure molto interessanti: i frattali autosimilari.

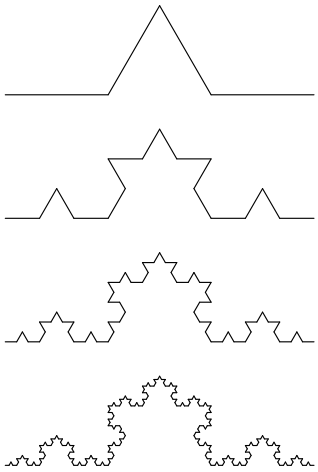


Figura 15: Le prime quattro iterazioni nella costruzione della curva di Koch.

Consideriamo ad esempio la *curva di Koch*. Tale curva si ottiene partendo da un segmento (diciamo di lunghezza unitaria). Il segmento viene suddiviso in tre parti,

si rimuove la parte centrale e la si sostituisce con i due lati del triangolo equilatero la cui base è il segmento rimosso. Quello che si ottiene è una curva spezzata formata da quattro segmenti di lunghezza 1/3. Su ognuno di questi quattro segmenti si può ripetere la stessa operazione: si suddivide in 3 parti, si rimuove la parte centrale e la si sostituisce con due nuovi segmenti di lunghezza 1/9. Questo procedimento può essere ripetuto *all'infinito* fino ad ottenere una figura *frattale* chiamata appunto curva di Koch[5].

Questa figura è *autosimilare* in quanto l'intera figura K è unione di 4 pezzi, ognuno dei quali è una perfetta copia dell'originale, riscalata di un fattore 1/3. Possiamo quindi valutare la dimensione di questo oggetto, come abbiamo fatto con il rettangolo e il cubo. Si ha infatti:

$$m(K) = 4 \cdot m(K/3) = \frac{4}{3^d} \cdot m(K)$$

da cui, supponendo $0 < m(K) < \infty$, si ottiene

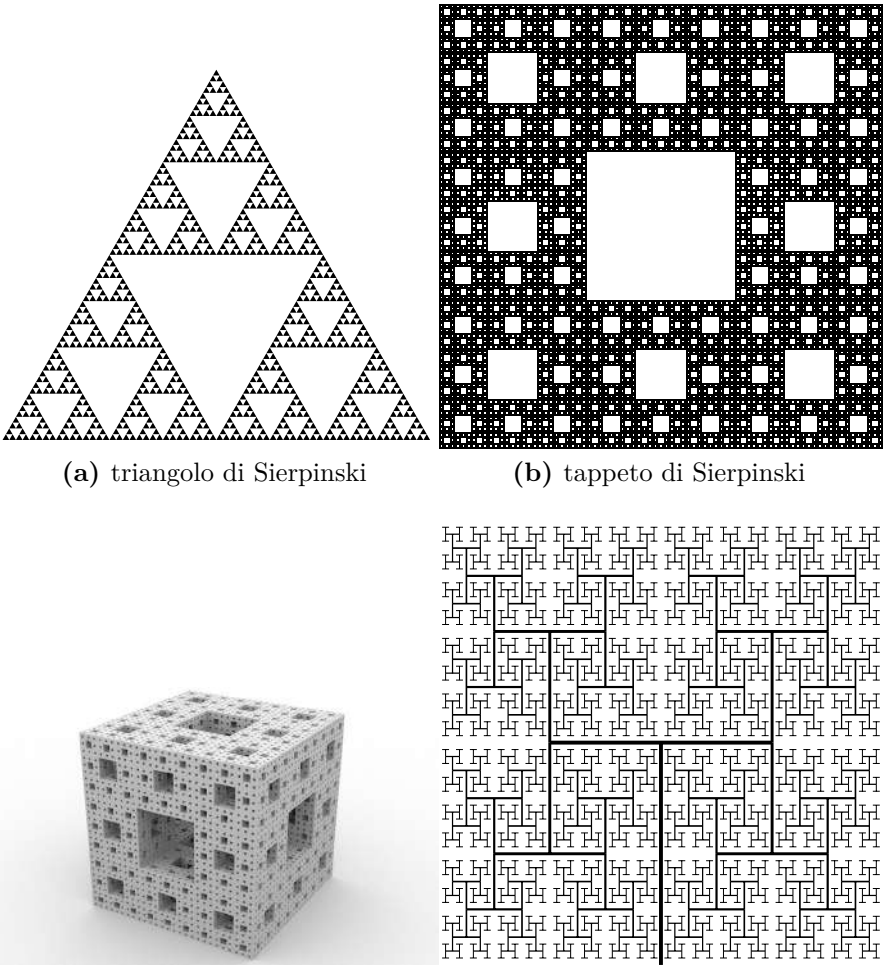
$$1 = \frac{4}{3^d}$$

e quindi $d = \log_3 4 \approx 1.26$. Quello che abbiamo trovato è dunque una figura di dimensione frazionaria, intermedia tra 1 e 2. Si può in effetti verificare che la curva ottenuta ha lunghezza infinita (si provi, per esercizio, ad esprimere la lunghezza della n-esima iterata nella costruzione e si faccia il limite di tale lunghezza per $n \longrightarrow \infty$). D'altra parte ha area nulla. È però possibile definire una misura intermedia tra la lunghezza e l'area che valuta la misura di questa curva dando un risultato finito. In effetti per qualunque $d \in \mathbb{R}$, $d \geq 0$ è possibile definire una misura $\mathcal{H}^d$ (misura di Hausdorff) che abbia dimensione d. Nei casi particolari $d = 1$, $d = 2$, $d = 3$ questa misura coincide effettivamente con la lunghezza, l'area e il volume. Nel caso $d = 0$ questa misura non è altro che la misura che *conta* il numero di punti. Per valori non interi di d le misure di Hausdorff ci permettono di misurare i frattali.

Lasciamo per esercizio il divertimento di determinare la dimensione delle seguenti figure autosimilari (nel caso dell'antenna, il calcolo richiede qualche accortezza in più, il risultato corretto è $d = 2$).

Emanuele Paolini,
Professore Associato, Dipartimento di Matematica di Pisa

[5]Cosa voglia dire esattamente "ripetere all'infinito" e come mai in questo processo "al limite" si ottenga veramente qualcosa, è un fatto assolutamente non banale che può essere affrontato solamente in un corso avanzato di matematica.



(a) triangolo di Sierpinski **(b)** tappeto di Sierpinski

(c) spugna di Menger **(d)** antenna frattale

Figura 16: Frattali autosimilari.

9 Leonhard Euler e il Problema dei Ponti

Elia Saini, n.4, Febbraio 2017

Da qualche giorno una comitiva di turisti tedeschi è in visita a Roma. Durante una delle loro interminabili passeggiate pomeridiane uno di essi domanda ai compagni se sia possibile intraprendere un'escursione che attraversi una sola volta ciascuno dei ponti sul Tevere.

I più sportivi del gruppo accolgono immediatamente la sfida e propongono di procedere per tentativi: in fin dei conti Roma è veramente una splendida città da visitare! Altri, in verità piuttosto pigri, giudicano questa proposta poco pratica. Essi affermano che per verificare tutti i percorsi possibili si debbano effettuare troppe passeggiate.

A questo punto interviene un turista appassionato di storia della matematica che afferma di aver già sentito parlare di un simile rompicapo. Questo enigma, ambientato per la prima volta a KÃ¶nigsberg, venne infatti risolto 300 anni fa dal grande matematico svizzero Leonhard Euler (1707 - 1783), conosciuto in Italia con il nome di Eulero. In Figura 17 è presentata una mappa dell'antica città di

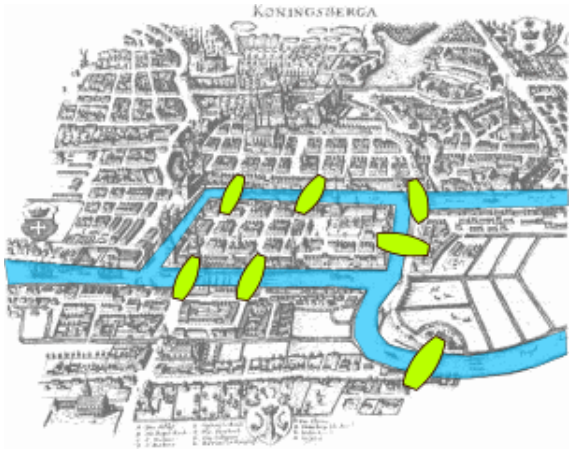


Figura 17: Il percorso del fiume Pregel con i ponti di Königsberg

KÃ¶nigsberg (oggi parte dell'exclave russa di Kaliningrad al confine tra Polonia e Lituania) con evidenziato il percorso del fiume Pregel (oggi Pregolja) e i suoi ponti.

Eulero non solo rispose all'enigma nel caso particolare della città di KÃ¶nigsberg, ma enunciò una regola generale valida per ogni città. Per risolvere questo dilemma Eulero impiegò sei mesi di lavoro e introdusse alcuni strumenti matematici che sono alla base della teoria dei grafi e della topologia moderna. Il quesito enunciato (e risolto) da Eulero può essere formulato nel modo seguente.

Problema. (Dei ponti) *Per qualsiasi città e per qualsiasi disposizione di ponti e rami di fiume, dire se è possibile scegliere un punto di partenza e un punto di arrivo*

e quindi svolgere una passeggiata che attraversi ciascuno dei ponti esattamente una volta.

Passiamo ora ad una trattazione più "matematica" del nostro problema. Armiamoci pertanto di carta, matita, pazienza e buona volontà e iniziamo a ripercorrere alcuni passaggi del ragionamento di Eulero.

Con riferimento alla Figura 18 (il disegno sulla sinistra di questa immagine è

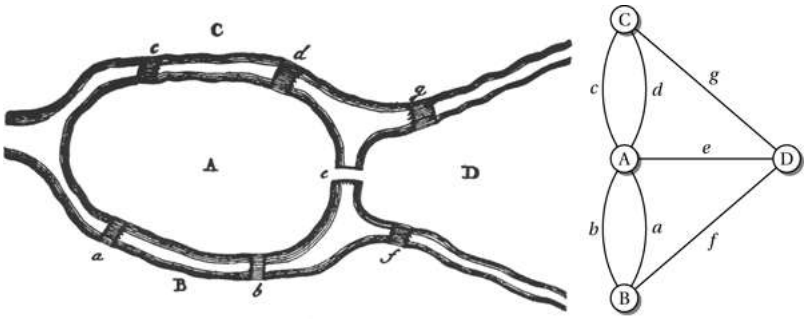


Figura 18: Disegno originale di Eulero e grafo associato

una riproduzione dell'originale di Eulero), consideriamo la città di KÃ¶nigsberg e disegnamo su un foglio uno schema (parte destra) - che chiameremo *grafo* - con dei pallini che corrispondono ai lembi di terra della città e dei segmenti che rappresentano i ponti. Nel linguaggio matematico chiamiamo *vertici* i pallini e *archi* i segmenti.

Un grafo si dice *connesso* se comunque scegliamo due suoi vertici, esiste sempre un cammino lungo gli archi del grafo che congiunga tali vertici. Nel linguaggio comune ciò equivale ad affermare che è possibile disegnare il grafo senza mai staccare la penna dal foglio. Per comprendere questa nozione consideriamo ad esempio Budapest. Fino alla fine del XIX secolo non c'erano ponti che collegassero le due sponde della città. Il grafo era dunque costituito da due vertici e nessun arco. Anche il nome della città è emblematico: Budapest infatti è nata in seguito alla fusione delle due *componenti connesse* Buda e Pest. Se abitassimo in una città con grafo non connesso sarebbe impossibile risolvere il problema dei ponti: potremmo collegare le sponde viaggiando in battello o mongolfiera, ma mai passeggiando.

Possiamo quindi *restringere* il nostro campo d'indagine alle sole città con grafo connesso. Riferendoci alla mappa della città di KÃ¶nigsberg otteniamo il grafo rappresentato nella parte destra della Figura 18. Invitiamo il lettore a disegnare il grafo della propria città. Attenzione: se abitate in una città senza fiumi il vostro grafo sarà costituito da un solo vertice!

Se ci pensiamo un attimo (in matematica questo procedimento si chiama *astrazione*) il problema dei ponti equivale al seguente quesito dal sapore decisamente piú matematico.

Problema. (Del grafo) *Dire se è possibile scegliere un vertice di partenza e un vertice di arrivo e quindi disegnare un grafo connesso senza passare due volte per lo stesso arco.*

Abbiamo dunque *modellizzato* il problema concreto dei ponti e lo abbiamo riformulato in un problema che riguarda grafi, vertici e archi. In Figura 19 sono

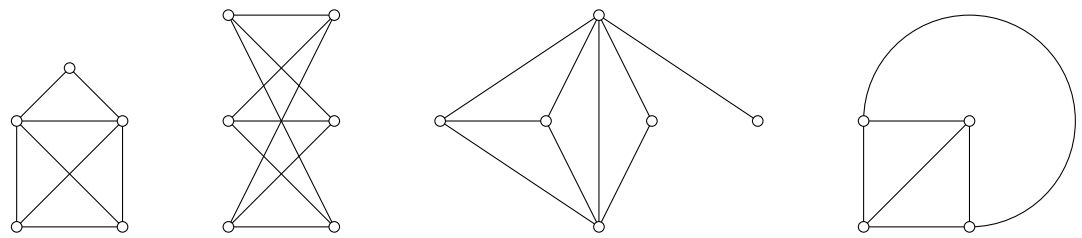


Figura 19: Alcuni esempi di grafi connessi

presentati alcuni esempi di grafi connessi. Per familiarizzare con questo nuovo problema possiamo provare a cercare una soluzione per questi casi particolari. Potrebbe però capitare che per alcuni di essi sia impossibile trovare una soluzione!

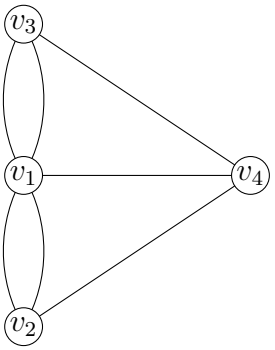
Astraendo il problema della passeggiata, Eulero fu in grado di eliminare gli aspetti incidentali della configurazione di ponti della città di KÃ¶nigsberg e di proporre una regola generale che fosse valida per qualsiasi città. Proprio questo è uno dei punti di forza della sua brillante soluzione.

Per risolvere il problema del grafo seguiamo il seguente *algoritmo*. Questo termine matematico indica una sequenza di operazioni che devono essere seguite passo passo... Esattamente quello che in cucina chiamiamo ricetta!

1. Numeriamo con $v_1, \ldots, v_k$ i vertici del grafo;
2. Per ogni vertice v_j del grafo Γ calcoliamo il grado $\deg_\Gamma(v_j)$ dove il grado è uguale al numero di archi che passano per il vertice v_j;
3. Calcoliamo il numero E_Γ di vertici con grado dispari.

Intuitivamente ogni vertice del grafo corrisponde ad una zona di terraferma, mentre il grado di un vertice è uguale al numero di ponti che partono da tale zona. Per capire questo metodo suggeriamo di ricopiare il disegno presentato in Figura 20 e successivamente di provare a calcolare il numero E_Γ per ciascuno dei grafi riprodotti in Figura 19.

Siamo ora pronti per presentare la soluzione del problema del grafo. Il seguente teorema, enunciato e dimostrato (con alcune piccole imprecisioni) da Eulero in



$$\deg_\Gamma(v_1) = 4$$

$$\deg_\Gamma(v_2) = 3$$

$$\deg_\Gamma(v_3) = 3$$

$$\deg_\Gamma(v_4) = 3$$

$$\boxed{E_\Gamma = 3}$$

Figura 20: Calcolo del numero E_Γ per il grafo della città di KÃ¶nigsberg

una serie di articoli pionieristici pubblicati negli anni quaranta del XVIII secolo, rappresenta uno dei piú straordinari risultati della matematica del periodo illuminista.

Teorema. (Eulero) *è possibile risolvere il problema del grafo se e solo se* $E_\Gamma \leq 2$.

Questo risultato è all'origine di un ramo della matematica contemporanea chiamato *topologia*. La soluzione del problema dei ponti infatti *non* dipende dalle proprietà *metriche* della città, quali ad esempio la distanza tra i ponti, la superficie dei lembi di terraferma, la larghezza delle sponde del fiume. Al contrario, essa dipende unicamente da proprietà *intrinseche* della città, quali ad esempio il numero di ponti che partono da ciascuna zona di terraferma. La topologia è quella branca della matematica che studia le proprietà intrinseche degli oggetti geometrici. Il lavoro di Eulero pose quindi le basi per lo sviluppo di un intero campo di indagine della moderna ricerca matematica.

Riferendoci alla Figura 20 deduciamo che per la città di KÃ¶nigsberg non è possibile svolgere una passeggiata che attraversi ciascuno dei ponti esattamente una volta. Se abitassimo a KÃ¶nigsberg non sarebbe possibile risolvere il problema dei ponti per tentativi: qualunque nostra passeggiata ci riporterebbe almeno una volta sullo stesso ponte. Il metodo di Eulero, sebbene poco sportivo, è estremamente efficace.

Invitiamo il lettore a risolvere il problema dei ponti per la città di Roma e anche per la propria città natale. Buon divertimento!

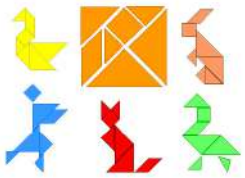
Elia Saini,
Laureato triennale a Pisa, dottorando in Matematica presso l'Università di Friburgo (CH)

10 Il Paradosso di Banach-Tarski

Alessandro Berarducci, n.5, Settembre 2017

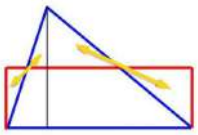
Congruenze per dissezioni

Due figure nel piano sono **congruenti per dissezioni** se una delle due può essere decomposta in un numero finito di pezzi poligonali che possono essere ricomposti (senza alterarne forma e dimensioni) in modo da formare l'altra figura.



Si intende che i vari pezzi non possano avere sovrapposizioni al di fuori dei bordi. La caratteristica principale delle congruenze per dissezioni è che conservano le aree.

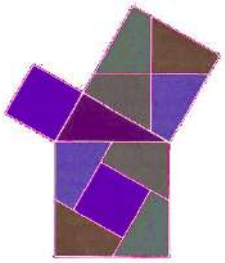
Possiamo ricavare la formula per l'area del triangolo (base per altezza diviso due) usando il fatto che un qualsiasi triangolo è congruente per dissezioni ad un rettangolo con la stessa base e un'altezza pari alla metà dell'altezza del triangolo.



Analogamente si dimostra che ogni parallelogramma è congruente per dissezioni ad un rettangolo (basta tagliare un triangolo rettangolo da uno dei due lati e spostarlo sul lato opposto). Un po' più difficile è dimostrare che ogni rettangolo è congruente per dissezioni ad un quadrato, e componendo queste costruzioni si può in effetti dimostrare che ogni poligono è congruente per dissezioni ad un quadrato. Tutto ciò era in gran parte noto agli antichi greci, mentre una dimostrazione moderna del fatto che due poligoni sono congruenti per dissezioni se e solo se hanno la stessa area si basa sui lavori di Wallace, Bolyai e Gerwien all'inizio del XIX secolo.

Il teorema di Pitagora

Il teorema di Pitagora afferma che, dato un triangolo rettangolo, l'area del quadrato costruito sull'ipotenusa è la somma delle aree dei quadrati costruiti sui cateti. Del teorema esistono molte dimostrazioni, di cui alcune basate sulle congruenze per dissezioni, come nella figura qui sotto.

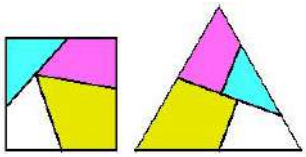


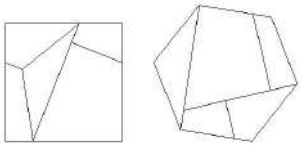
La dimostrazione per dissezioni di Henry Perigal (1801-1898), agente di cambio e matematico dilettante.

Il quadrato grande costruito sull'ipotenusa può essere dissezionato in cinque pezzi, che poi possono essere ricomposti per formare i due quadrati costruiti sui cateti. Sapreste dire cosa occorre verificare per assicurarsi che la dimostrazione sia valida?

Il triangolo e l'esagono regolari

La prova del fatto che poligoni con la stessa area siano congruenti per dissezioni (Teorema di Bolyai-Gerwien) non fornisce stime sul numero minimo di pezzi necessari, tuttavia nei casi concreti ci si può divertire a cercare di minimizzare il numero dei pezzi. Ad esempio per trasformare un triangolo equilatero in un quadrato con la stessa area bastano 4 pezzi, e per un esagono regolare ne bastano 5, ma la suddivisione non è affatto facile da trovare (o anche solo da verificare dopo averla vista)!





Proprietà delle congruenze per dissezioni

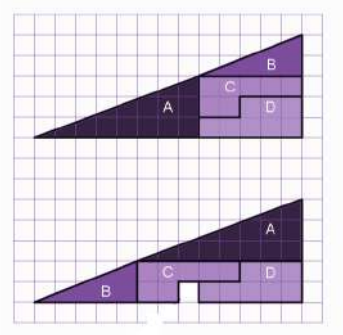
Come semplice esercizio vi propongo di verificare che le congruenze per dissezioni verificano la proprietà transitiva, ovvero se X è congruente per dissezioni a Y e Y è congruente per dissezioni a Z, allora X è congruente per dissezioni a Z. Ad esempio avendo visto che un triangolo equilatero ed un esagono regolari sono entrambi congruenti per dissezioni ad un quadrato, ne segue che un triangolo equilatero è congruente per dissezioni ad un esagono regolare. Sapreste trovare i pezzi necessari?

Molto più difficile è invece dimostrare che se da due figure congruenti dissezioni sottraiamo due figure anch'esse congruenti per dissezioni, le figure risultati sono ancora congruenti per dissezioni. Ad esempio se all'interno di due triangoli uguali facciamo due buchi di forma quadrata e delle stesse dimensioni, ma non

necessariamente posizionati nello stesso modo, le figure risultanti sono ancora congruenti per dissezioni. Questa proprietà sottrattiva può in molti casi semplificare le dimostrazioni che due figure hanno la stessa area.

Trucco o magia?

Abbiamo detto che le congruenze per dissezioni conservano l'area, ma non abbiamo in effetti dato una definizione precisa di area. Sapendo, però, che un poligono è sempre congruente per dissezioni ad un rettangolo (o addirittura ad un quadrato), potremmo essere tentati di definire l'**area di un poligono** come il prodotto delle lunghezze dei lati di un rettangolo congruente per dissezioni al poligono dato. Per assicurarsi che ciò dia una buona definizione, occorrerebbe mostrare che il risultato non dipende da come si scelgono i pezzi della dissezione, e a tal fine sarebbe necessario far vedere che un poligono non può essere congruente per dissezioni ad un poligono più grande che lo contiene. In altre parole ci chiediamo se possa accadere che, attraverso un semplice spostamento di pezzi, si possa far "scomparire" una porzione di una figura come in un gioco di prestigio. Ciò sembra intuitivamente impossibile, ma un famoso puzzle di Dudeney sembra a prima vista mettere in discussione questa certezza:

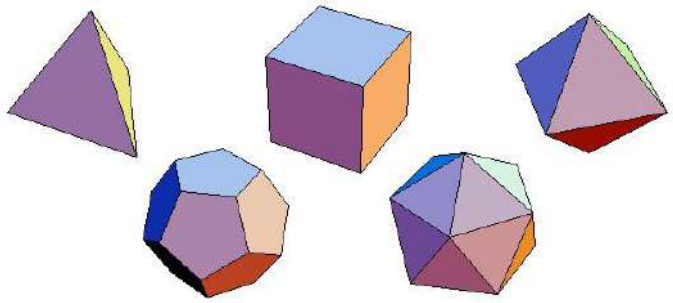


Un puzzle di Henry Ernest Dudeney, 1857-1930

Se non conoscete il trucco, la figura può lasciare sconcertati in quanto sembra mettere in crisi il concetto stesso di area. C'è però un inganno e in effetti si dimostra che simili "sparizioni" non possono capitare. Sapreste trovare l'inganno?

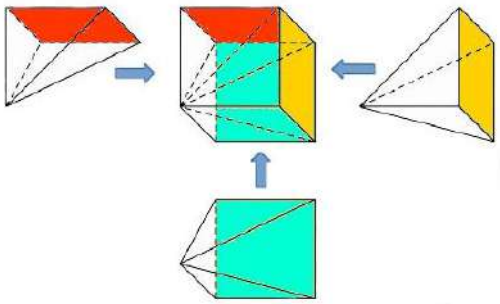
Passiamo alla terza dimensione: volume dei solidi

Abbiamo visto che nel caso dei poligoni l'avere la stessa area equivale all'essere congruenti per dissezioni e sparizioni di pezzi come nel puzzle di Dudeney possono solo essere frutto di inganno. Vediamo ora cosa succede per il volume dei solidi nello spazio tridimensionale.



Volume della piramide

Esattamente come un quadrato può essere diviso in due triangoli rettangoli, un minimo di riflessione mostra come un cubo può essere diviso in tre piramidi rettangole uguali, come nella figura.



Volume della piramide rettangola =
base × altezza / 3

Ne deduciamo che, ammesso che esista una nozione sensata di volume, il volume della piramide rettangola deve essere dato dalla formula " Base × Altezza / 3 " , dove per "Base" intendiamo l'area del quadrato di base. In altre parole, se il lato è lungo 1 metro, il volume della piramide rettangola è 1/3 di metro cubo.

Principio di Cavalieri

Se la piramide ha una forma diversa il ragionamento precedente non si applica. Tuttavia, utilizzando il cosiddetto "principio di Cavalieri" possiamo determinare il volume di una piramide qualsiasi, non necessariamente rettangola o con base quadrata. Per illustrarne il funzionamento, consideriamo due piramidi solide, come nella figura, e poniamole su una stessa base orizzontale. Supponiamo che "affettando" le due piramidi con un piano parallelo alla base orizzontale si ottengano sempre due figure piane della stessa area. Il principio di Cavalieri afferma che in questo caso le due piramidi solide hanno lo stesso volume. Se le riempissimo di acqua, ne conterrebbero la stessa quantità.

Bonaventura Cavalieri, 1598-1647

Applicando il principio, si dimostra che il volume di una qualsiasi piramide è pari al volume di una piramide rettangola con la stessa area di base e la stessa altezza. Quindi il volume della piramide generica si calcola di nuovo con la formula Base $\times$ Altezza /3.

Per giustificare il principio di Cavalieri immaginiamo che le fettine, invece di essere infinitamente sottili, abbiano un certo spessore, molto piccolo rispetto alle dimensioni della piramide. Il volume di una fettina può allora essere approssimato abbastanza bene dall'area della sua base per la sua altezza. Dico "approssimato" anziché calcolato esattamente perché la formula "area di base per altezza" presuppone che i lati delle fettine siano verticali anziché obliqui, dimodoché l'unione delle fettine viene a formare una struttura a gradini tipo ziggurat. Tuttavia facendo fettine sempre più piccole e sommando i loro contributi, l'errore totale che si commette può essere reso arbitrariamente piccolo, come si può vedere approssimando la piramide dall'interno e dall'esterno con due ziggurat, ottenendo al limite il volume della piramide. Su analoghe considerazioni si basa il principio di esaustione di Archimede, così come la moderna teoria dei volumi e degli integrali.

Il terzo problema di Hilbert

Possiamo estendere dal piano allo spazio il concetto di congruenza per dissezioni, semplicemente richiedendo che i pezzi della scomposizione siano poliedrali anziché poligonali (e che non vi siano sovrapposizioni al di fuori dei bordi). Possiamo allora chiederci se due poliedri con lo stesso volume siano sempre congruenti per dissezioni. In contrasto con quanto avveniva per l'area dei poligoni, la risposta è

però negativa: risolvendo uno dei famosi problemi posti da Hilbert nel 1900, Max Dehn ha infatti dimostrato la cosa seguente:

“ Un cubo e un tetraedro non sono mai congruenti per dissezioni, ovvero non è possibile suddividere un cubo in un numero finito di poliedri che possono essere ridisposti in modo da formare il tetraedro. ”

Max Dehn, 1878-1952

Per chi voglia approfondire, diciamo solo che la dimostrazione si basa sul fatto che se due poliedri X ed Y sono congruenti per dissezioni, allora detti $\alpha_1, \ldots, \alpha_s$ e $\beta_1, \ldots, \beta_r$ i rispettivi angoli diedrali (gli angoli tra due facce adiacenti), esistono numeri interi positivi $m_1, \ldots, m_s$ ed $n_1, \ldots, n_r$ rispettivamente, tali che la differenza tra $m_1\alpha_1 + \ldots + m_s\alpha_s$ e $n_1\beta_1 + \ldots + n_r\beta_r$ è un multiplo intero di π (vedi [2]). Nel caso del tetraedro e del cubo, gli angoli diedrali sono rispettivamente $\alpha = \arccos(1/3)$ e $\beta = \pi/2$, e siccome α/β si dimostra essere irrazionale, una tale relazione non può sussistere, onde la non congruenza per dissezioni.

Il principio di continuità

Alla luce del risultato di Dehn, potrebbe sorgere il dubbio se esista un cubo dello stesso volume di un tetraedro. Una risposta positiva è però fornita dal principio di continuità: se teniamo fisso il tetraedro e ingrandiamo progressivamente un cubo inizialmente molto piccolo, dobbiamo necessariamente passare da cubi di volume decisamente inferiore a cubi di volume decisamente superiore a quello del tetraedro, e per continuità il volume dovrà passare per tutte le misure intermedie, assumendo anche esattamente quella del tetraedro.

Equiscomposizioni

Possiamo chiederci se il teorema di Dehn continui a valere rilassando la richiesta che i pezzi della dissezione siano poliedri e ammettendo quindi pezzi più complicati, come ad esempio quelli dei disegni di Escher (o meglio, l’equivalente in 3D).

Una suddivisione complicata del quadrato.

Per precisare le regole del gioco diamo alcune definizioni. Due figure geometriche sono **congruenti** se i punti dell’una corrispondono a quelli dell’altra tramite

un'isometria, ovvero una corrispondenza che preserva le distanze. Nel caso di figure spaziali (ovvero in $\mathbb{R}^3$), questo significa che le due figure sono uguali, salvo il fatto che sono situate in modo diverso nello spazio, possibilmente ruotate, traslate o capovolte l'una rispetto all'altra (incluso il caso in cui una delle figure sia come l'immagine allo specchio dell'altra).

Diciamo che due figure X ed Y sono **equiscomponibili**, se è possibile partizionare X ed Y nello stesso numero finito di pezzi in modo che ciascun pezzo di X sia congruente al corrispondente pezzo di Y.

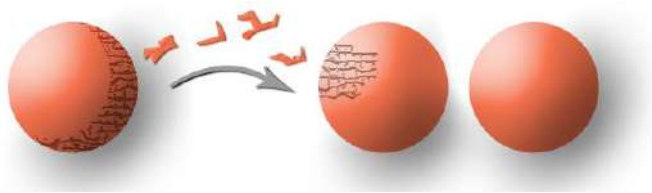
La differenza tra le equiscomposizioni e le congruenze per dissezioni è duplice: da un lato nelle equiscomposizioni si ammettono pezzi di forma arbitraria (non necessariamente poliedrali); dall'altro però si insiste sul fatto che i pezzi formino una partizione, ovvero siano del tutto disgiunti tra loro, non consentendo sovrapposizioni anche solo limitate ai bordi (che invece nelle congruenze per dissezioni erano trascurati).

Si può dimostrare che se X ed Y sono congruenti per dissezioni sono anche equiscomponibili (ovvero si possono eliminare le sovrapposizioni dei bordi, ma al prezzo di usare pezzi più complicati). La domanda è se le equiscomposizioni possano risultare uno strumento più flessibile per dimostrare l'uguaglianza di due volumi. Come vedremo nella prossima sezione, se non mettiamo limiti a quanto possano essere complicati i pezzi, le equiscomposizioni sono però talmente flessibili da risultare applicabili anche a figure con volume diverso!

10.1 Teorema di Banach-Tarski

Il teorema di Banach-Tarski stabilisce che dati due solidi qualsiasi con parte interna non vuota, essi sono sempre equiscomponibili, ovvero è possibile partizionare il primo in un numero finito di pezzi che possono essere ricomposti in modo da formare il secondo.

In particolare, è possibile partizionare un solido sferico in 5 parti che possono essere ricomposte in modo da formare due sfere dello stesso volume della sfera data.



E' inoltre possibile, contro ogni ragionevole aspettativa, dividere un solido sferico della dimensione di una biglia, in un numero finito di pezzi che possono

essere ricomposti per formare una sfera solida del diametro del sole!

In altre parole, mentre con le congruenze per dissezioni non ce la si fa anche quando ce la si dovrebbe fare (teorema di Dehn), con le equiscomposizioni, ce la si fa anche quando non ce la si dovrebbe fare (teorema di Banach-Tarski)!

I due risultati mostrano che il fatto di avere lo stesso volume non equivale né all'esistenza di una congruenza per dissezioni, né all'esistenza di una equiscomposizione. A differenza di altri paradossi, come quello di Zenone o quello del mentitore, destinati a rimanere interrogativi problematici, quello di Banach-Tarski è un vero e proprio teorema matematico, nonostante sia così incredibile da essere chiamato paradosso.

Cosa è il volume?

A questo punto occorrerebbe chiedersi quale sia la corretta definizione matematica di volume. Questo è un argomento importante che però non posso approfondire in queste note, limitandomi ad accennare al fatto che la definizione si basa sul concetto di approssimazione (limiti, integrali), come nella discussione relativa al principio di Cavalieri.

Il teorema di Banach-Tarski mostra in ogni caso che non è possibile assegnare in modo ragionevole un volume a tutte le figure spaziali: ai pezzi della equiscomposizione paradossale della sfera non è possibile assegnare un volume (essi non sono "Lebesgue-misurabili"), altrimenti otterremmo il risultato, questo sì paradossale, che una sfera solida ha lo stesso volume dell'unione di due sfere uguali alla prima.

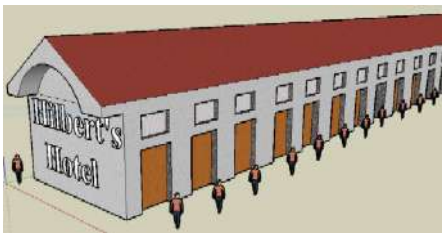
Passiamo alla dimostrazione

La seconda parte di questo intervento è più impegnativa e richiede una certa concentrazione, ma non vi scoraggiate, anche una lettura superficiale può dare i suoi frutti. La dimostrazione (a grandi linee) del teorema di Banach-Tarski sarà infatti l'occasione, o il pretesto, per introdurre informalmente qualche concetto importante della matematica, in particolare il concetto di gruppo. Per facilitare la lettura le sezioni sono monotematiche e solo alla fine sarà chiaro (almeno spero) come amalgamare i vari ingredienti.

Vedremo che i pezzi dell'equiscomposizione paradossale della sfera sono così complicati che non è possibile farsene una semplice immagine visiva, a differenza di quanto avveniva nel caso delle congruenze per dissezioni. Tuttavia, l'esistenza concettuale dei vari pezzi riposa su principi che oggi sono comunemente accettati dai matematici, tra cui il cosiddetto "assioma della scelta" (servirà per scegliere un insieme di rappresentanti delle orbite di certe azioni gruppali). Se si trascurano i bordi, vi sono anche versioni del paradosso che non usano l'assioma della scelta (teorema di Dougherty e Foreman).

L'albergo di Hilbert

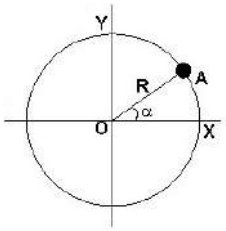
Uno dei più noti paradossi dell'infinito può essere illustrato dal cosiddetto "Albergo di Hilbert". Esso ha infinite stanze, tante quante i numeri interi non negativi $0, 1, 2, 3, \ldots$, e sono tutte occupate.



All'arrivo di un nuovo cliente l'albergatore riesce ad alloggiarlo con dei semplici spostamenti di stanza. Come fa? La risposta non è difficile: basta chiedere a ciascun cliente di spostarsi nella stanza successiva (1 nel 2, 2 nel 3, 3 nel 4, eccetera), in modo che la stanza 0 si liberi e possa accogliere il nuovo cliente.

Un primo paradosso geometrico

Con la stessa idea dell'albergo di Hilbert possiamo dimostrare che un disco è equiscomponibile con il disco stesso privato di uno dei raggi. Più precisamente un disco è scomponibile in due pezzi R, S che possono essere ruotati in modo da formare lo stesso disco privato di un raggio (tenendo però il centro).



Si usa la tecnica dell'"albergo di Hilbert". Scegliamo l'angolo α in modo che $\alpha, 2\alpha, 3\alpha, 4\alpha, \ldots$ siano tutti distinti (basta scegliere α in modo che α/π sia irrazionale). Sia R l'insieme dei raggi che hanno coordinata angolare pari ad uno degli $n\alpha$ (n intero positivo), e sia S la parte restante del disco. Ora risistemiamo i pezzi R, S come segue: S lo lasciamo fermo, mentre R lo ruotiamo di un angolo α. In tal modo il raggio $n\alpha$ si sposta in posizione $(n+1)\alpha$ e il raggio di angolo α sparisce.

Questo paradosso è meno sorprendente di quello di Banach-Tarski perché il disco meno un raggio ha la stessa area di tutto il disco, mentre nel paradosso di Banach-Tarski si riescono ad alterare i volumi. Tuttavia abbiamo fatto un primo passo.

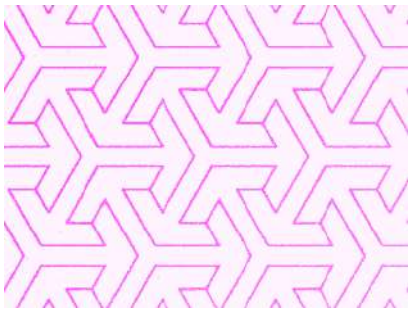
Il concetto di gruppo

Uno dei concetti più importanti dell'algebra moderna è quello di gruppo. Tra gli esempi più importanti vi sono i gruppi di "movimenti", ad esempio le mosse possibili su un cubo di Rubik. In un gruppo è sempre definita una composizione,

che nel caso del cubo di Rubik significa effettuare due mosse una dopo l'altra, e un inverso, che nel nostro esempio corrisponde fare la mossa opposta che fa tornare dove eravamo.

Altri esempi di gruppi sono dati dalle isometrie del piano e dello spazio. Ricordiamo che le isometrie sono le trasformazioni che non alterano le distanze, e nel caso del piano includono le rotazioni intorno ad un punto, le traslazioni, e i ribaltamenti intorno ad un asse.

Se ci restringiamo ai movimenti che preservano certe figure o simmetrie otteniamo un sottogruppo, ovvero un insieme di movimenti incluso in quello di partenza, stabile per composizione e inversi. Per familiarizzarci con la nozione analizziamo il sottogruppo dei movimenti del piano che conservano la tassellazione del piano raffigurata qui accanto.



L'analisi delle simmetrie mostra che il sottogruppo include la rotazione ρ in senso antiorario di $2\pi/3$ radianti (120 gradi) intorno al punto centrale, la rotazione di $4\pi/3$ radianti (240 gradi) ottenuta applicando due volte ρ (denotata $\rho\rho$, o ρ^2), il movimento nullo, denotato "1", i ribaltamenti intorno agli assi di simmetria, ad esempio il ribaltamento β rispetto alla retta di pendenza $\pi/3$ passante per il centro, la traslazione in orizzontale τ, che sposta ogni asse di simmetria non orizzontale in quello accanto a destra, e tutti i movimenti che si ottengono da questi e dai loro inversi componendoli tra loro.

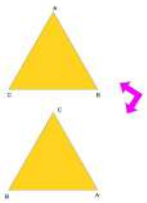
Seguiamo la convenzione (non del tutto standard) che i movimenti si leggano da sinistra a destra, ad esempio $\beta\rho$ significa fare prima β poi ρ. Come esercizio potete verificare che $\rho^3 = 1$ (ruotare di 360 gradi lascia tutti i punti del piano dove sono), $\beta^2 = 1$, e $\beta\rho = \rho^2\beta$ (ribaltare e poi ruotare di 120 gradi equivale a ruotare di 240 gradi e poi ribaltare).

Dato un elemento x del gruppo, il suo inverso x^{-1} è quell'elemento tale che $x^{-1}x = xx^{-1} = 1$. Ad esempio, visto che $\rho^3 = 1$ e che possiamo scrivere $\rho^3 = \rho^2\rho$, otteniamo $\rho^{-1} = \rho^2$, come si verifica anche direttamente osservando che ruotare di 240 gradi in senso antiorario (ρ^2) equivale a ruotare di 120 gradi nel verso opposto (ρ^{-1}).

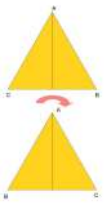
Come esercizio potete verificare che componendo opportunamente ρ, τ, β possiamo ottenere una traslazione lungo una direzione non orizzontale.

Il grafo di Cayley di un gruppo

Dato un gruppo, il suo grafo di Cayley rappresenta in modo visivo quali successioni di composizioni danno lo stesso risultato. Per semplicità consideriamo il gruppo D3 delle simmetrie del triangolo, che può essere generato da una rotazione a e un ribaltamento b, come in figura.



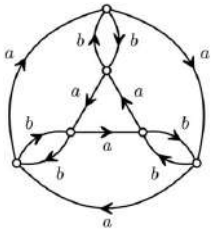
rotazione (a)



ribaltamento (b)

Si tratta di un gruppo simile a quello visto in precedenza salvo che non sono presenti le traslazioni in quanto il centro del triangolo deve rimanere fermo.

Nel gruppo D3 vi sono strade diverse che portano allo stesso risultato. Ad esempio $a^3 = 1$, $b^2 = 1$, $ba = a^2b$. Queste relazioni sono rappresentate nel grafo di Cayley da cammini diversi ma con gli stessi nodi di partenza e arrivo.



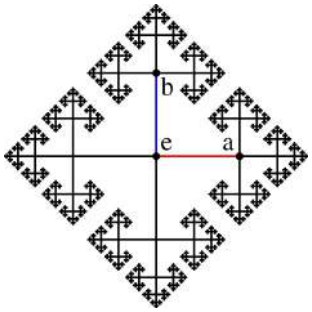
Grafo di Cayley di D3

Possiamo pensare ai sei nodi del grafo come alle sei possibili posizioni ABC, CAB, BCA, CBA, BAC, ACB del triangolo, mentre le frecce rappresentano i movimenti. Come si vede dal diagramma fare tre volte a a partire da qualsiasi posizione riporta alla posizione di partenza.

Gruppi liberi

Un gruppo generato da certi elementi $a, b, \ldots$ si dice **libero** se nel corrispondente grafo di Cayley strade diverse con lo stesso punto di partenza portano a risultati diversi (per evitare banalità si escludono i casi in cui nella successione di mosse sia presente un generatore affiancato al suo inverso). Il gruppo D3 non è quindi libero in quanto ad esempio $ba = a^2b$. In un gruppo libero con due generatori a, b il grafo di Cayley dovrebbe avere la seguente forma.

Grafo di Cayley del gruppo libero F2
con due generatori a, b.



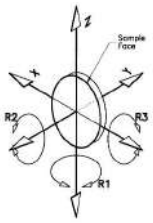
La mossa a fa andare a destra, a^{-1} a sinistra, b in alto, b^{-1} in basso. Strade diverse a partire dal centro e, come ad esempio b e aba^{-1}, portano a nodi diversi.

Si dimostra che due isometrie nel piano non possono mai generare un gruppo libero (questo dipende dal fatto che il gruppo delle isometrie del piano è "risolubile"), ma esistono invece due rotazioni a, b nello spazio tali che il gruppo da loro generato è libero, e questo si rivelerà decisivo per il teorema di Banach-Tarski.

Gruppi di movimenti nello spazio

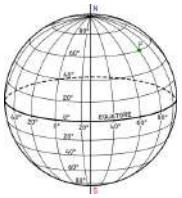
Concentriamoci sulle isometrie dello spazio che fissano l'origine delle coordinate, escludendo quindi le traslazioni. Tali movimenti portano una sfera centrata nell'origine in se stessa. Si dimostra che ogni tale movimento ha lo stesso effetto di una rotazione intorno ad un asse che passa per l'origine. In particolare, comunque muoviamo una sfera, lasciandone fisso il centro, esisteranno due punti sulla superficie sferica che alla fine saranno nella stessa posizione di dove erano all'inizio, ovvero i due punti antipodali dove l'asse di rotazione incontra la superficie.

Il gruppo delle isometrie dello spazio coincide dunque con il gruppo delle rotazioni della sfera in se stessa e si chiama $SO(3, \mathbb{R})$. La sfera viene denotata con il simbolo S^2 e diremo che il gruppo "agisce" sulla sfera.



Orbita di un punto

Dato un punto P in uno spazio e un gruppo G che agisce sullo spazio, l'orbita di P consiste dell'insieme dei punti dove può arrivare P per effetto di uno degli elementi del gruppo.



Consideriamo ad esempio il sottogruppo G di $SO(3,\mathbb{R})$ dato dalle rotazioni di un angolo arbitrario intorno all'asse z. Dato un punto P sulla sfera diverso da uno dei due poli, la sua orbita è allora il parallelo che passa per P, mentre l'orbita di ciascuno dei due poli si riduce ad un solo punto.

Un insieme **trasversale** (rispetto alle orbite) è un insieme che contiene uno ed un solo punto per ogni orbita. Nel nostro esempio un possibile insieme trasversale è dato da un qualsiasi meridiano M. Se scegliamo un sottogruppo più complicato di $SO(3,\mathbb{R})$, trovare un insieme trasversale può richiedere l'assioma della scelta.

Rotazioni indipendenti

Consideriamo due rotazioni a, b di pari ampiezza $\theta = \arccos(1/3)$ intorno agli assi z ed x rispettivamente. Per chi conosca le matrici, si tratta delle rotazioni rappresentate dalle matrici seguenti

$$a = \begin{pmatrix} \frac{1}{3} & -\frac{2\sqrt{2}}{3} & 0 \\ \frac{2\sqrt{2}}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & -\frac{2\sqrt{2}}{3} \\ 0 & \frac{2\sqrt{2}}{3} & \frac{1}{3} \end{pmatrix}$$

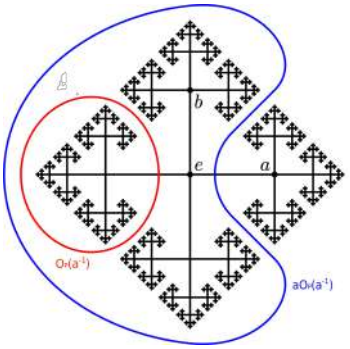
Ai nostri fini, l'unica cosa da sapere è che le rotazioni a, b generano un gruppo libero $F_{a,b}$ isomorfo ad F_2, ovvero il suo grafo di Cayley è identico a quello di F_2 visto nella Sezione 10.1.

Ogni elemento g di $F_{a,b}$ è una rotazione intorno ad un certo asse, vi saranno sulla sfera due punti "eccezionali" dati dall'intersezione dell'asse di rotazione di g con la superficie sferica. Consideriamo l'insieme D di tutti i punti eccezionali al variare di g in $F_{a,b}$. Ora scegliamo un punto P sulla sfera che non appartenga a D (visto che D è numerabile, basta scegliere un punto a caso e con probabilità $1 = 100\%$ non apparterrà a D). L'orbita di P rispetto a $F_{a,b}$ sarà un insieme O_P di punti della sfera in corrispondenza biunivoca con i nodi del grafo di Cayley di F_2, ovvero P verrà mosso in punti diversi da ogni diversa successione delle rotazioni a, b, a^{-1}, b^{-1} (si escludono le successioni contenenti un generatore e un suo inverso affiancati).

Scomposizioni paradossali di un'orbita

Chiamiamo $O_P(a)$ l'insieme dei punti dell'orbita O_P che si ottengono applicando a P una composizione di elementi di $F_{a,b}$ che "inizia" per a, ad esempio la successione di rotazioni abb. Definiamo similmente $O_P(a^{-1}), O_P(b), O_P(b^{-1})$. Questa divisione

si fa al netto delle cancellazioni, ad esempio $a^{-1}abab$ appartiene a $O_P(b)$ perché la a iniziale si cancella con a^{-1} e resta bab.



Nella figura l'orbita più piccola è $O_P(a^{-1})$, mentre la più grande è $aO_P(a^{-1})$.

L'orbita O_P si lascia scrivere come unione disgiunta dei cinque sottoinsiemi $\{P\}, O_P(a), O_P(a^{-1}), O_P(b), O_P(b^{-1})$ corrispondenti rispettivamente al centro e ai quattro "petali" del diagramma di Cayley "centrato in P".

L'osservazione fondamentale è che O_P si lascia anche scrivere come unione di $O_P(a)$ e $aO_P(a^{-1})$ (ovvero $O_P(a^{-1})$ ruotato tramite a), oppure come unione di $O_P(b)$ ed $bO_P(b^{-1})$. In tal modo otteniamo una decomposizione "paradossale" dell'orbita O_P. Ciascuno dei due sottoinsiemi disgiunti $A = O_P(a) \cup O_P(a^{-1})$ e $B = O_P(b) \cup O_P(b^{-1})$ è equiscomponibile con l'intera orbita O_P.

Teorema di Hausdorff

Se scegliamo un insieme M trasversale a tutte le orbite (qui serve l'assioma della scelta) e ripetiamo il ragionamento precedente simultaneamente per tutte le orbite dei vari punti di M, otteniamo una equiscomposizione paradossale di $S^2 - D$, dove S^2 è la superficie sferica e D è l'insieme dei punti dove S^2 incontra gli assi delle rotazioni corrispondenti agli elementi di $F_{a,b}$ (vedi Sezione 10.1).

Felix Hausdorff, 1868-1942

Teorema di Banach-Schröder-Bernstein

Per passare dalla equiscomposizione paradossale di $S^2 - D$ a quella di S^2 e poi a quella di una sfera solida, conviene introdurre qualche notazione. Scriviamo $X \sim Y$ se X ed Y sono equiscomponibili, ovvero se è possibile partizionare X ed Y nello stesso numero finito di pezzi in modo che ciascun pezzo di X sia congruente al corrispondente pezzo di Y. Scriviamo infine $X \preceq Y$ se X è equiscomponibile con un sottoinsieme di Y.


Stephan Banach
1892-1945

Il teorema di Banach-Schröder-Bernstein permette di semplificare notevolmente il compito di verificare che due figure sono equispomponibili. Esso afferma che se $X \preceq Y$ e $Y \preceq X$ allora $X \sim Y$, analogamente al fatto che la congiunzione di due disuguaglianze fornisce una uguaglianza. La dimostrazione non è difficile e si può trovare in [1].

Una figura geometrica X si dice **paradossale** se esistono due sottoinsiemi disgiunti A, B di X tale che $X \sim A \sim B$. Utilizzando queste notazioni, il teorema di Hausdorff dice dunque che $S^2 - D$ è paradossale, ovvero è possibile dividere $S^2 - D$ in due pezzi A e B da ciascuno dei quali possiamo ricostruire $S^2 - D$ tramite delle equiscomposizioni.

Teorema di Banach-Tarski

Per ottenere una decomposizione paradossale di tutto S^2 basta mostrare che $(S^2 - D) \sim S^2$. Si utilizza a tal fine la tecnica dell'albergo di Hilbert per riassorbire D tramite un'opportuna rotazione ρ intorno ad un asse passante per il centro della sfera tale che $D, \rho(D), \rho^2(D), \rho^3(D), ...$ siano tutti disgiunti.

Dalla paradossalità di S^2 si ottiene facilmente la paradossalità della sfera solida meno il centro (basta aggiungere i raggi). La paradossalità dell'intera sfera solida si ottiene infine riassorbendo il centro con una rotazione non periodica per un asse non passante per il centro medesimo.

Alessandro Berarducci,
Professore Ordinario di Logica presso il Dipartimento di Matematica
dell'Università di Pisa

Riferimenti bibliografici

[1] Grzegorz Tomkowicz and Stan Wagon. The Banach-Tarski Paradox. Cambridge University Press. Seconda edizione, 2016.

[2] David Benko. A new approach to Hilbert's third problem. The American Mathematical Monthly. Vol. 114, No. 8 (Oct., 2007), pp. 665-676

11 La Matematica dei Viaggi Spaziali

Daniele Serra, n.5, Settembre 2017

Il 1957 ha segnato un punto di svolta nella storia dell'umanità: quando il 4 Ottobre di quell'anno l'Unione Sovietica ha comunicato di aver lanciato in orbita il primo satellite artificiale, lo Sputnik, si è aperta la grande era dell'esplorazione spaziale. Da allora la nostra conoscenza del Sistema Solare e dell'Universo è cresciuta enormemente. Innumerevoli sono state le sonde spaziali interplanetarie che hanno visitato gli altri pianeti e i corpi minori che ruotano attorno al Sole, non ultima la sonda spaziale della missione New Horizons (NASA), che ha mandato a Terra le prime immagini ravvicinate del pianeta nano Plutone.

In queste pagine cerchiamo di studiare dal punto di vista rigoroso della Meccanica Celeste, in un contesto semplificato ma significativo, il problema di inviare una sonda spaziale a partire dalla stazione di lancio sulla Terra fino a un altro corpo celeste (pianeta, pianeta nano, asteroide, cometa).

11.1 Orbite ellittiche

Il problema del decifrare la forma delle traiettorie dei corpi celesti ha origini molto antiche e nella Storia sono state date innumerevoli soluzioni; alcune di queste erano molto ingegnose, seppure errate. Quando si è passati dalla concezione geocentrica del Sistema Solare a quella eliocentrica, tutto è cambiato. Un enorme contributo è stato dato da Johannes Kepler con la formulazione delle sue tre leggi, formulate a partire dallo studio dell'enomre mole di *osservazioni* raccolte dal suo maestro Tycho Brahe.

> Prima Legge di Kepler, 1608: I pianeti si muovono su orbite ellittiche di cui il Sole occupa uno dei due fuochi.

Questa affermazione è molto accurata, ma non è perfettamente rappresentativa della realtà. I pianeti, infatti, durante il loro moto attorno al Sole, sono perturbati, seppure in misura minore, dall'attrazione degli altri pianeti. Questo implica che le traiettorie non sono ellissi perfette. Anzi, è stato dimostrato dal matematico francese Henri Poincaré che non è possibile trovare una formula per descrivere il moto vero di un pianeta del Sistema Solare! Per questo nei problemi e nelle applicazioni spesso si fa ricorso all'approssimazione ellittica prescritta da Kepler.

Un'altra osservazione che possiamo fare è che le leggi di Kepler sono enunciate per i corpi che ruotano attorno al Sole, ma ovviamente valgono in generale per i corpi che ruotano attorno alla Terra o attorno a Giove, e così via.

Ricordiamo qualche nozione sull'ellisse:

Definizione 11.1. L'ellisse è il luogo geometrico dei punti per i quali è costante la somma delle distanze da due punti fissi detti fuochi.

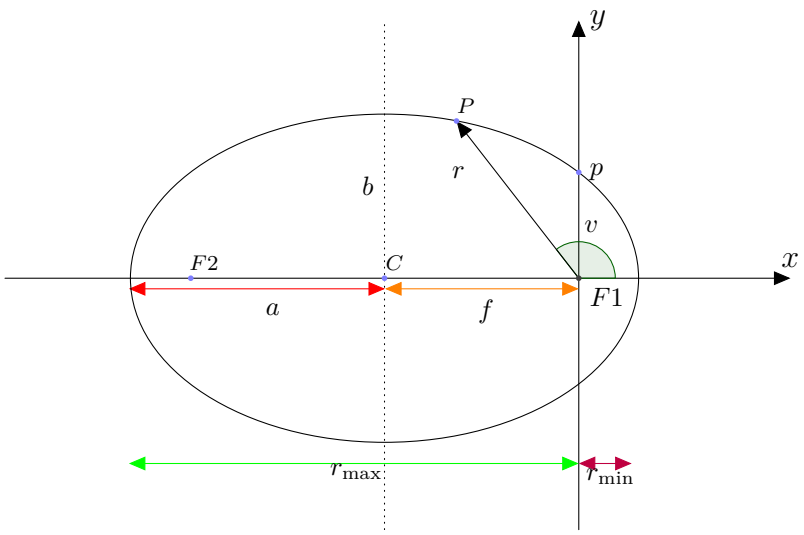


Figura 21: Esempio di ellisse.

È utile studiare l'ellisse in un sistema di riferimento cartesiano in cui l'origine coincide con uno dei due fuochi (Figura 21). Ci accorgiamo facilmente che ogni punto sull'ellisse è identificato dalla sua distanza r dall'origine e dall'angolo v (*anomalia vera*) che tale segmento forma con l'asse orizzontale delle x. Possiamo scrivere l'equazione in *forma polare* dell'ellisse, che dice il valore di r corrispondente a un dato valore di v:

$$r(v) = \frac{a(1-e^2)}{1+e\cos v},$$

dove a si chiama *semiasse maggiore* ed e si chiama *eccentricità* e sono dei parametri che caratterizzano l'orbita. Ad esempio, l'orbita della Terra attorno al Sole ha $a \sim 150$ milioni di km e $e \sim 0.017$.

A partire dall'equazione in forma polare dell'ellisse si possono calcolare interessanti quantità:

Pericentro: è la distanza minima dal fuoco F_1 (il Sole), e si ottiene per un valore dell'anomalia vera uguale a 0, cioè $r_{\text{peri}} = r(0) = a(1-e)$ - per la Terra questo numero vale ~ 147 milioni di chilometri;

Apocentro: è la distanza massima dal fuoco F_1 (il Sole), e si ottiene per un valore dell'anomalia vera uguale a π, cioè $r_{\text{apo}} = r(\pi) = a(1+e)$ - per la Terra questo numero vale ~ 152 milioni di chilometri.

Osserviamo che una circonferenza è il caso limite di un'ellisse con eccentricità $e = 0$.

11.2 Sonde spaziali e missili balistici

Aggiungiamo adesso un po' di dinamica alla descrizione finora puramente geometrica delle traiettorie. La forza con cui un corpo di massa M attrae un altro corpo di massa m rispetta la legge di Newton ed ha intensità

$$F = \frac{GMm}{r^2},$$

dove G è una costante ed è la *costante di Newton*, mentre r è la distanza tra i due corpi. Tale forza ammette un'*energia potenziale* (si dice che è conservativa), per cui possiamo scrivere l'*energia totale del sistema*:

$$E = \frac{1}{2}m\mathrm{v}^2 - \frac{GMm}{r},$$

dove v è la velocità del secondo corpo che ruota intorno al primo, e r è la sua distanza dal primo corpo. L'energia totale ha una proprietà fondamentale. Per spiegare quale, notiamo che in linea di principio, poiché sia la velocità v che la distanza r sono delle grandezze che cambiano durante il moto, allora anche l'energia dovrebbe cambiare durante il moto. In realtà questo non accade: anche se r e v cambiano, E rimane sempre la stessa durante il moto! Si dice che E è un *integrale primo* del sistema. In realtà l'energia non è l'unico integrale primo, ma ne esiste un altro, chiamato *momento angolare*:

$$J = mr\mathrm{v}\sin(\phi),$$

dove ϕ è l'angolo compreso tra il vettore raggio e il vettore velocità.

Anche se apparentemente non c'entra nulla, in realtà la conservazione del momento angolare è strettamente legata alla

> Seconda Legge di Kepler, 1609: Il raggio vettore che unisce il centro del Sole col centro del pianeta spazza aree uguali in tempi uguali[6].

[6]Una conseguenza importante di questa legge è illustrata nella Figura 22. Le aree dei settori sottesi dagli archi AB e CD sono uguali, sebbene CD sia più lungo di AB: il pianeta, dovendole spazzare in tempi uguali, è costretto ad aumentare la velocità quando percorre CD. Concludiamo che i pianeti si muovono più velocemente quando sono più vicini al Sole e più lentamente quando sono più lontani.

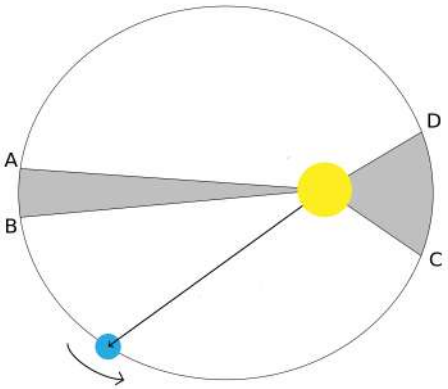


Figura 22: Conseguenze della seconda legge di Kepler. Fonte: Wikipedia.

Una prima applicazione della presenza dei due integrali primi è la seguente: se E e J non cambiano durante il moto del pianeta attorno al Sole, allora sia che le calcoliamo al pericentro, sia che le calcoliamo all'apocentro, sono uguali:

$$\begin{cases} E_{\text{peri}} = E_{\text{apo}} \\ J_{\text{peri}} = J_{\text{apo}}. \end{cases}$$

Calcolando queste quantità e risolvendo il sistema di equazioni che ne viene fuori (provaci!), si ottiene una nuova espressione per l'energia:

$$E = -\frac{GMm}{2a}. \tag{5}$$

Questa nuova formula dice una cosa molto importante: dati due corpi celesti di massa M e m, l'energia dipende esclusivamente dal semiasse maggiore dell'orbita e non dalla sua eccentricità. Inoltre, ci dà una formula per il calcolo della velocità di un corpo in orbita attorno a un altro corpo: uguagliando le due formule per l'energia, si ottiene

$$\frac{1}{2}m\mathrm{v}^2 - \frac{GMm}{r} = -\frac{GMm}{2a} \iff \mathrm{v} = \sqrt{GM\left(\frac{2}{r} - \frac{1}{a}\right)}. \tag{6}$$

Quindi per conoscere la velocità v di un corpo in orbita a un certo tempo, basta conoscere il semiasse maggiore dell'orbita a (che è costante) e la sua posizione r in quell'istante di tempo.

Grazie alla formula (5), possiamo ricavare un fatto molto importante: supponiamo di voler mettere in orbita attorno alla Terra un satellite in orbita circolare a 500 km dalla superficie terrestre; poiché il raggio della Terra è $R_\oplus \sim 6371$ km, il semiasse maggiore dell'orbita del satellite attorno alla Terra sarà pari ad $a \sim 6871$ km. Come si vede anche dalla Figura 23, allo stesso valore del semiasse maggiore corrispondono, oltre all'orbita circolare, anche delle orbite ellittiche (basta cambiare l'eccentricità). Per valori abbastanza alti di e, si ottengono orbite che intersecano la superficie terrestre, cioè orbite per cui se si lancia un corpo, sicuramente andrà ad impattare la Terra in un altro punto: un missile balistico intercontinentale! Questo spiega l'allarme creatosi al lancio dello Sputnik: l'Unione Sovietica aveva sì mostrato di saper mandare in orbita un satellite, ma allo stesso tempo aveva implicitamente mostrato di poter lanciare, con la stessa energia e con lo stesso lanciatore, un missile che potesse colpire un altro Paese.

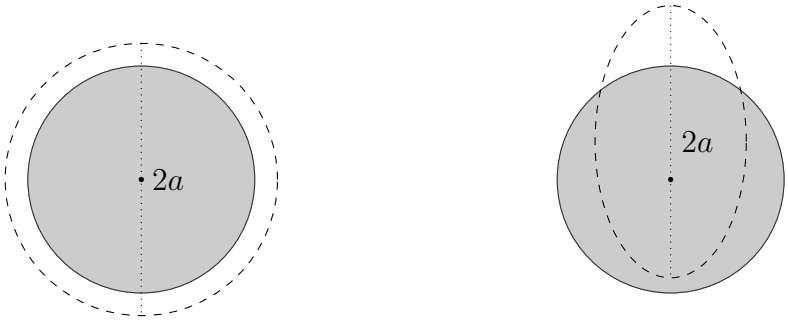


Figura 23: Satellite o missile balistico? Tanto l'energia è la stessa.

11.3 Andiamo su Giove!

Vogliamo progettare una missione spaziale interplanetaria; dobbiamo lanciare un satellite che esca dall'orbita terrestre e raggiunga un altro pianeta, ad esempio Giove. Supponiamo per semplicità che sia la Terra che Giove siano su orbite circolari e siano a_1 il semiasse maggiore dell'orbita terrestre e sia a_2 il semiasse maggiore dell'orbita gioviana. Vogliamo rispondere alle seguenti domande:

1. Come facciamo?
2. Quanto carburante dobbiamo spendere?
3. Quanto tempo ci vuole?

Vediamo subito come arrivare a Giove. In riferimento alla Figura 24, possiamo:

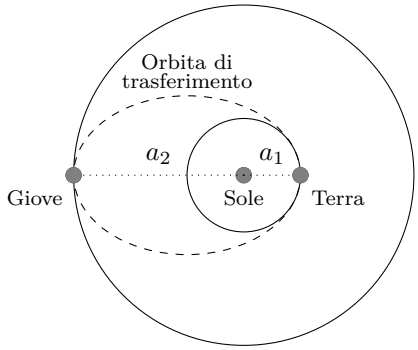


Figura 24: Orbita di trasferimento di Hohmann per una sonda trasferita dall'orbita terrestre all'orbita gioviana.

- accendere i razzi dalla Terra e inserire la sonda in un'orbita ellittica attorno al Sole con semiasse maggiore $\frac{a_1+a_2}{2}$;

- una volta arrivata nei pressi di Giove, accendere di nuovo i razzi per inserire la sonda in'un orbita circolare attorno al Sole di semiasse maggiore a_2.

Questo tipo di trasferimento si chiama *trasferimento alla Hohmann*, dal nome dell'ingegnere spaziale tedesco Walter Hohmann che l'ha inventata.

Per rispondere alla seconda domanda, ricorriamo all'*equazione del razzo di Ciolkovskij*, che mette in relazione il propellente speso durante una manovra orbitale con la variazione di velocità Δv avvenuta con la manovra stessa. Per conoscere il propellente usato, basta quindi calcolare la velocità della sonda prima della manovra e la velocità dopo. Nel caso della nostra orbita di trasferimento verso Giove si tratta di accendere i razzi due volte, corrispondenti a due cambi di velocità: uno, che chiamiamo $\Delta \mathrm{v}_1$, per mettere il satellite dall'orbita terrestre in orbita ellittica e uno, $\Delta \mathrm{v}_2$, per toglierlo dall'orbita ellittica e inserirlo nell'orbita di Giove.

Calcoliamo $\Delta \mathrm{v}_1$. Osserviamo che prima della manovra lo spacecraft si trova in orbita circolare attorno al Sole di semiasse maggiore a_1, quindi la sua velocità per la (6) è

$$\mathrm{v}_T = \sqrt{\frac{GM_{\mathrm{Sole}}}{a_1}}.$$

Dopo la manovra, questo si trova in orbita ellittica attorno al sole di semiasse $(a_1 + a_2)/2$, a distanza a_1 dal Sole, quindi la sua velocità è, sempre per la (6),

$$\mathrm{v}_{\mathrm{peri}} = \sqrt{GM_{\mathrm{Sole}} \left(\frac{2}{a_1} - \frac{2}{a_1 + a_2} \right)}.$$

Concludiamo che

$$\Delta \mathrm{v}_1 = \mathrm{v}_{\mathrm{peri}} - \mathrm{v}_T = \sqrt{\frac{GM_{\mathrm{Sole}}}{a_1}} \left(\sqrt{\frac{2a_2}{a_1 + a_2}} - 1 \right).$$

Sostituendo i valori $GM_{\mathrm{Sole}} \simeq 1.327 \times 10^{20}$ m^3/s^2, $a_1 \simeq 1.5 \times 10^1 1$ m e $a_2 \simeq 7.8 \times 10^{11}$ m, si ottiene $\Delta \mathrm{v}_1 \simeq 8.7$ km/s.

Per esercizio, verifica che $\Delta \mathrm{v}_2$ è dato da

$$\Delta \mathrm{v}_2 = \sqrt{\frac{GM_{\mathrm{Sole}}}{a_2}} \left(1 - \sqrt{\frac{2a_1}{a_1 + a_2}} \right)$$

e che nel caso Terra-Giove questo vale $\Delta \mathrm{v}_2 \simeq 5.6$ km/s, per un consumo totale di $\Delta \mathrm{v} = 14.3$ km/s.

Per rispondere all'ultima domanda, abbiamo bisogno della

> Terza Legge di Kepler (1619): I quadrati dei periodi di rivoluzione dei pianeti sono proporzionali ai cubi dei semiassi maggiori.

Come tradurre in linguaggio matematico questa legge? Se T è il periodo di rivoluzione di un certo corpo celeste attorno al Sole (cioè il tempo che impiega a compiere un giro attorno al Sole) e a è il semiasse maggiore della sua orbita, allora esiste una costante k tale che

$$T^2 = ka^3.$$

Tale costante è uguale per tutti i corpi celesti che orbitano il Sole e vale $4\pi^2/GM_{\mathrm{Sole}}$. È facile quindi calcolare il tempo di percorrenza dell'orbita di Hohmann: corrisponde alla *metà* del periodo di rivoluzione dell'orbita con semiasse maggiore $(a_1 + a_2)/2$, cioè

$$T = \frac{1}{2} \sqrt{\frac{4\pi^2}{GM_{Sole}} \left(\frac{a_1 + a_2}{2} \right)^3} \simeq 1000 \text{ giorni}.$$

Le orbite di Hohmann hanno una proprietà matematica interessante: esse sono dei minimi locali nello spazio di tutte le orbite di trasferimento, rispetto alla quantità di propellente utilizzato. Questo fa sì che siano le orbite più usate in meccanica spaziale. Tuttavia il fatto che non sono dei minimi globali rende possibile l'esistenza di altre orbite più convenienti. Ad esempio, in alcuni casi (se $a_2 \geq 12a_1$, quindi se il pianeta che si vuole raggiungere è molto lontano) è meno costoso in termini di carburante un tipo di trasferimento con tre manovre invece di due: prima si manda la sonda in un punto che si trova oltre il pianeta, poi la si riporta indietro fino al pianeta, quindi la si inserisce in orbita. Sembra strano, ma può essere dimostrato rigorosamente! Naturalmente questi tipo di trasferimenti a tre manovre impiegano molto più tempo rispetto ai trasferimenti di Hohmann.

Un ultimo aspetto importante è costituito dal fatto che Giove si muove, quindi affinché la sonda possa incontrare il pianeta, bisogna scegliere il momento del lancio in modo che il pianeta si trovi nella stessa posizione della sonda all'arrivo di quest'ultima sull'orbita gioviana. Si parla in questo caso di *finestre di lancio*, cioè giorni specifici in cui una sonda può essere lanciata affinché riesca a incontrare il pianeta al termine della traiettoria prescelta.

11.4 Esplorazione planetaria

Una volta che abbiamo messo una sonda attorno a Giove, cosa si fa? È proprio qui che comincia la ricerca planetaria. Sappiamo ancora poco dei giganti gassosi del Sistema Solare (Giove, Saturno, Urano e Nettuno) e in particolare vorremmo studiarne la composizione interna, il campo magnetico, le interazioni con i loro satelliti naturali. Nel caso particolare di Giove, è attualmente in corso la missione spaziale Juno (NASA), che dal Luglio 2016 sta orbitando il pianeta gassoso e sta raccogliendo dati e inviandoli a Terra. Al Dipartimento di Matematica dell'Università di Pisa il Gruppo di Meccanica Celeste si occupa dell'analisi dei dati di *tracking* della sonda Juno con l'obiettivo di determinarne il campo di gravità. Queste informazioni sono cruciali per i geofisici, affinché possano determinare con accuratezza la struttura interna del pianeta. Conoscere com'è fatto il pianeta più grande del Sistema Solare, infatti, aiuta a capire come si sono formati gli altri pianeti, inclusa la Terra. Questo potrebbe in ultima istanza condurre a comprendere sotto quali condizioni si può formare la vita e individuare altri pianeti nell'Universo in cui forme di vita possono svilupparsi o possono essersi già sviluppate.

Daniele Serra,
Assegnista di Ricerca in Fisica Matematica presso il Dipartimento di Matematica dell'Università di Pisa

Per saperne di più:
`https://it.wikipedia.org/wiki/Meccanica_celeste`
`https://en.wikipedia.org/wiki/Transfer_orbit`
`https://www.missionjuno.swri.edu`
`http://poisson.dm.unipi.it/~dserra/downloads/settimana_matematica_2017.pdf`

12 La Teoria dei Grafi Nascosta Intorno a Noi

Ludovico Battista, n.6, Febbraio 2018

Uno dei più grandi interrogativi dell'uomo moderno, insieme allo scopo della vita e all'esistenza di altre forme di vita nell'universo, è la ragione per cui la matematica viene studiata. Sarebbe semplicistico, per non dire arrogante, provare a esaurire questo argomento in poche pagine. Quello che ci proponiamo di fare in quest'articolo è presentare tre giochi la cui soluzione può essere resa in modo chiaro ed affascinante con l'uso di strumenti matematici intuitivi; perché, pur non essendo degli esteti, ai matematici piace l'eleganza.

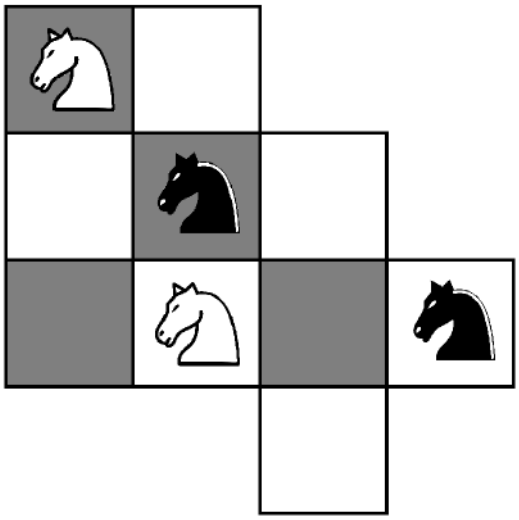


Figura 25: La posizione iniziale.

12.1 Il puzzle dei quattro cavalli, variante

Gli scacchi sono uno sport educativo e divertente; è soprannominato "Il gioco dei re", e non sta a noi decidere se questo dipenda dal fatto che permette di passare molto tempo seduti a non far nulla o dalle sue sfaccettature più strategiche, che si addicono a delle cariche che prendono decisioni importanti.

Oggi vogliamo proporre un rompicapo su questo tema: immaginate di avere quattro cavalli su una scacchiera insolita, due bianchi e due neri, come in figura 25. Scopo del gioco è riuscire a scambiare di posizione i cavalli bianchi e quelli neri. Le regole sono semplici: il movimento del cavallo è come negli scacchi[7] e due cavalli non possono occupare la stessa casella.

[7]Il cosiddetto movimento *a L*: due caselle verticalmente e una orizzontalmente o viceversa.

Prima di andare avanti nella lettura dell'articolo, sarebbe istruttivo provare a pensare a una strategia per risolvere questo problema. Badate bene: andare a tentativi è una strategia! Spesso non è quella più conveniente, ma nei rompicapi è sempre la prima che proviamo ad applicare.

I più tenaci, i più fortunati e i più scaltri dovrebbero essere riusciti nell'impresa, ma anche in caso contrario si può procedere nella lettura. Fino ad ora la matematica potrebbe non aver giocato nessun ruolo, ma si possono sempre complicare le cose: sapreste dire quale è il numero minimo di mosse necessario per risolvere tale problema? Dare una risposta a questa domanda potrebbe risultare piuttosto complicato, ma soprattutto, stavolta, andare a tentativi non può aiutare in alcun modo.

La figura 26 potrebbe darvi una piccola mano, soprattutto se avete già avuto a che fare con l'oggetto che stiamo per presentare. Quando girerete pagina, però, la soluzione vi sarà evidente; vi chiedo perciò di prendervi il vostro tempo: il modo di scansare un ostacolo ci rimane tanto più impresso quanto più esso ci appare insormontabile.

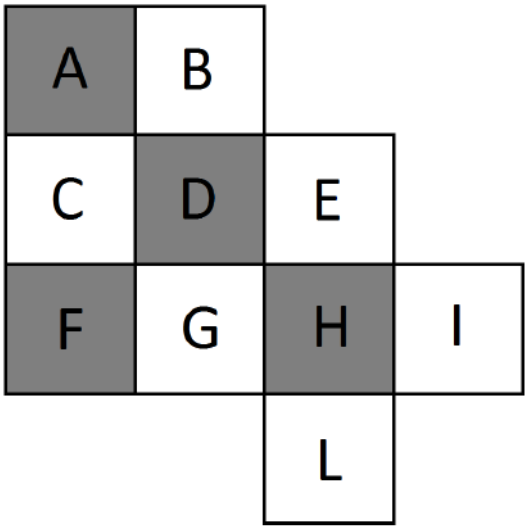
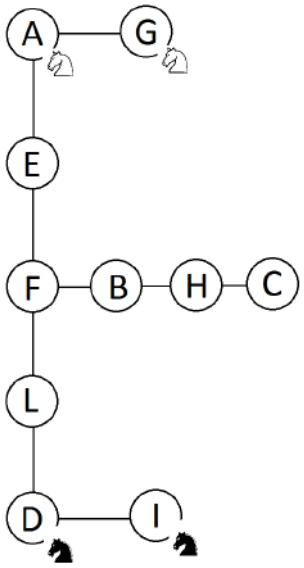


Figura 26: "Delle lettere sulla scacchiera? In che modo dovrebbero aiutarmi?"

Un modo di risolvere il problema è il seguente: disegniamo le lettere su un foglio, e colleghiamo due lettere fra loro se un cavallo può passare, in una mossa, dall'una all'altra. Quello che abbiamo costruito viene chiamato *grafo*: un insieme di *vertici* che sono collegati da alcuni *archi*. Nella fattispecie, il nostro grafo è *non orientato*, perché gli archi non hanno un verso di percorrenza.



Con questo grafo, oltre a risolvere molto facilmente il puzzle, è anche possibile rispondere al nostro problema! Possibile, ma non semplice: il metodo migliore sembra impiegare 26 mosse, ma chi ci assicura che non si possa fare di meglio? In matematica, così come nella vita, ci si trova spesso di fronte ad affermazioni che ci sembrano evidenti. Molte volte abbiamo ragione, ma le più costruttive sono quelle in cui ci sbagliamo; è perciò importante trovare ragioni che definiremmo *inconfutabili*. Ci riuscite?

12.2 Videocassette

Una cosa che viene spesso rinfacciata alle nuove generazioni è l'incapacità di giocare con ciò che si ha in casa. Non siamo più capaci di "fare d'un fil di lana una collana". Per il prossimo gioco abbiamo bisogno di oggetti parallelepipedali, grandi $0.5 \times 1 \times 2$ centimetri, preferibilmente di legno... Ma delle videocassette andranno benissimo.

Prendete delle videocassette. Lo scopo del gioco è il seguente: dovete riuscire a disporle in modo tale che ogni videocassetta tocchi tutte le altre tranne una. Cominciate a provare con quattro.

Una sola precisazione non essenziale: con "toccare" intendo che il contatto avviene su una faccia della videocassetta, non su uno spigolo o su un vertice. Questo dovrebbe aiutare in certe situazioni.

La soluzione con quattro videocassette dovrebbe essere piuttosto immediata. In ogni caso, alcune possibili soluzioni sono proposte in figura 27.

Provate ora con tre videocassette. So che questo è un giornalino, quindi alcuni di voi stanno solo *fingendo* di avere tre videocassette tra le mani. Anche la capacità di astrazione è molto importante, ma non è questo il cuore del discorso: che voi abbiate o meno tre videocassette tra le mani, non dovreste riuscire nell'intento. Sareste capaci di dimostrare che è impossibile con un argomento semplice?

Alziamo la posta. Prendete cinque o sei videocassette e provate. Vi do un indizio: solo in uno di questi due casi una soluzione è possibile, riuscite a trovarla? Per conoscerla proseguite nella lettura!

Ecco, con sei videocassette è possibile. Con cinque no! Il motivo è il seguente:

Proposizione 1. Non è possibile disporre 5 videocassette in modo che ognuna di

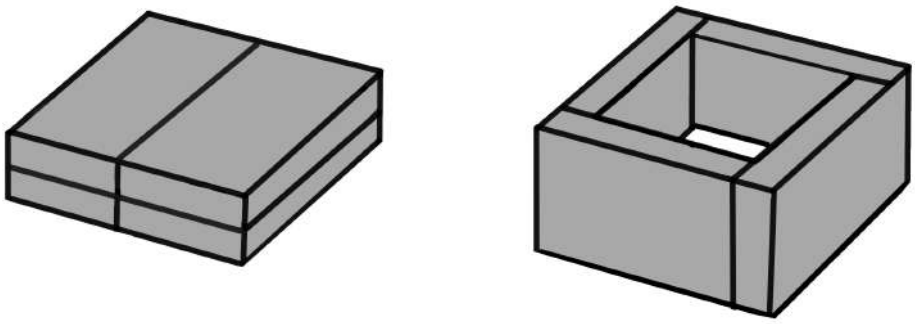


Figura 27: Due possibili soluzioni del problema con 4 videocassette.

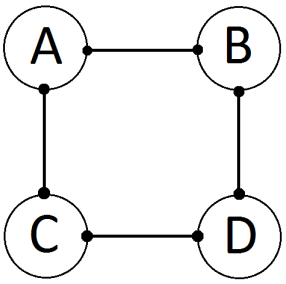


Figura 28: Il grafo associato alle configurazioni di videocassette in figura 27. In entrambi i casi è lo stesso. Ogni vertice ha valenza 2, e la somma delle valenze di tutti i vertici è 8.

esse tocchi tutte le altre tranne una.

Dimostrazione. Supponiamo che tale configurazione esista. Ad essa possiamo associare un grafo nel seguente modo: per ogni videocassetta costruiamo un vertice, e creiamo un arco tra il vertice A e il vertice B se e solo se la videocassetta A tocca la videocassetta B; un esempio di grafo di questo tipo è presente in figura 28. Definiamo ora la *valenza* di un vertice come la somma del numero di archi che hanno un estremo in esso. Contiamo ora in due modi la somma delle valenze di tutti i vertici:

- se immaginiamo di partire dal grafo senza archi e di aggiungerne uno alla volta, ad ogni passo la somma delle valenze dei vertici aumenta di 2. Tale numero, dunque, è pari;

- d'altra parte, da ogni vertice devono uscire esattamente 3 archi, e dunque la somma delle valenze sarà $5 \cdot 3 = 15$.

Questi risultati non sono compatibili! Questo vuol dire che una tale configurazione di videocassette non esiste.

□

Questa dimostrazione è chiaramente applicabile a qualsiasi numero dispari. Quello che abbiamo fatto ci permette di sottolineare alcune caratteristiche costituzionali del ragionamento matematico:

- Lavorare con numeri piccoli aiuta, a volte, a trovare risultati generali. La dimostrazione che abbiamo proposto per cinque videocassette si applica subito a qualsiasi numero dispari. D'altra parte, quella più naturale che si può fornire per tre videocassette non è facilmente generalizzabile. Procedere per tentativi può essere una buona idea o una perdita di tempo, e non è facile capire in che situazione ci si trova.

- La matematica ci permette di sublimare alcune caratteristiche e di ottenere risultati utilizzando soltanto queste ultime. Non prendete sotto gamba questo concetto! Ad esempio, passare da una configurazione di videocassette al grafo associato ci permette di isolare il concetto di *contatto*, ma ci fa perdere delle informazioni. Se le videocassette fossero un miliardo, sarebbe possibile creare un grafo con un miliardo di vertici, ognuno collegato a tutti gli altri tranne uno (sareste capaci di dimostrarlo?), ma non sarebbe comunque possibile risolvere il vero problema, per un motivo più *materiale* (anche qui, sapreste abbozzarne uno?).

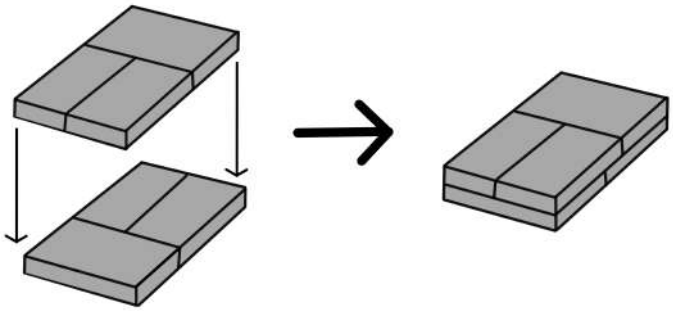


Figura 29: Una possibile soluzione del problema con sei videocassette.

12.3 Sim

Finora abbiamo presentato una *struttura*, i grafi, e abbiamo usato loro caratteristiche elementari. Stavolta presentiamo un concetto più complicato, lo facciamo a partire da un gioco per due giocatori: il Sim.

Il primo giocatore ha una matita e il secondo una penna[8], e ci sono i sei vertici di un esagono su un foglio. A questo punto ogni giocatore, a turno, colora un segmento che unisce due vertici. Lo scopo del gioco è NON disegnare un triangolo del proprio colore con vertici sui vertici iniziali dell'esagono. Chiaramente si è obbligati a disegnare almeno un segmento finché non saranno presenti tutti quelli possibili!

Se farete qualche tentativo a questo gioco, vi accorgerete che non finisce mai in parità. Come mai?

Proposizione 2. Se coloriamo con due colori tutti gli archi del grafo completo su sei vertici[9], comparirà almeno un triangolo monocromatico.

Dimostrazione.
Diamo dei nomi ai vertici, chiamiamoli A, B, C, D, E, F. Consideriamo i 5 archi uscenti da A, ossia $\overline{AB}, \overline{AC}, \overline{AD}, \overline{AE}, \overline{AF}$. Almeno tre di questi hanno lo stesso colore, per il principio dei cassetti[10]. Supponiamo, per semplicità, che siano $\overline{AB}, \overline{AC}, \overline{AD}$, e che siano colorati con la matita. Ora ci sono due possibilità:

- C'è almeno uno tra $\overline{BC}, \overline{CD}, \overline{BD}$ che è colorato con la matita. In questo caso, esiste un triangolo colorato con la matita di vertici A e i due estremi di questo segmento.

[8] Questi saranno i nostri *colori*; se avete il blu e il rosso il disegno sarà più accattivante.

[9] Ossia, un grafo non orientato con sei vertici in cui ogni possibile coppia è collegata da un arco.

[10] Formalmente, usiamo la forma forte di questo principio. Per una referenza, si veda [2].

- Tutti e tre $\overline{BC}, \overline{CD}, \overline{BD}$ sono colorati con la penna. In questo caso sono proprio B, C e D ad essere i vertici di un triangolo, stavolta colorato con la penna.

□

Andiamo più a fondo nello studio di questo problema. Possiamo ad esempio chiederci se il gioco possa funzionare anche con 5 vertici invece che con 6. La risposta è negativa, e una dimostrazione si può trovare in figura 30. Un'altra domanda che apre la strada a una interessante dimostrazione è la seguente: se volessimo giocare in tre, con tre colori, quanti vertici ci servirebbero? Non siamo pronti per rispondere esattamente a questa domanda[11], ma possiamo fare qualche passo.

12.4 Numeri di Ramsey

Nella sezione seguente la trattazione assumerà uno stile molto più rigoroso; il concetto che stiamo per introdurre può apparire, infatti, piuttosto complicato.

Dati t numeri $n_1, \ldots, n_t$, ci chiediamo se esiste un numero naturale k tale che, dato un grafo completo su k vertici i cui archi sono colorati con t colori $c_1, \ldots, c_t$, esiste almeno un sottoinsieme di n_1 vertici tutti collegati tra loro con archi di colore c_1, oppure almeno un sottoinsieme di n_2 vertici tutti collegati tra loro con archi di colore c_2, e così via.

[11] Anche se è nota la risposta, ossia 17, la dimostrazione non è alla nostra portata.

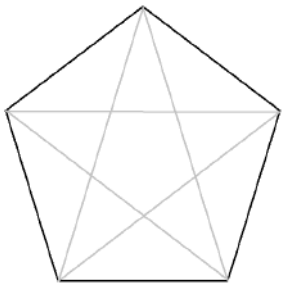


Figura 30: Una 2-colorazione del pentagono che non contiene triangoli monocromatici.

Non è detto che questo numero esista, ma se ne esiste uno esiste *il minimo numero* che ha tale proprietà[12]. In tal caso tale minimo si indica con $R(n_1, \ldots, n_t)$ e si chiama *numero di Ramsey* della t-upla $n_1, \ldots, n_t$.

Teorema 1 (di Ramsey). Per ogni scelta di numeri naturali $n_1, \ldots, n_t$, esiste il numero di Ramsey della t-upla $n_1, \ldots, n_t$.

Dimostrazione. Come abbiamo già fatto notare, ci basta dimostrare che esiste almeno un numero naturale k tale che il grafo completo su k vertici ha la proprietà richiesta; l'esistenza di un minimo sarà una banale conseguenza. Procediamo per induzione su t:

- Se $t = 2$, vogliamo dimostrare che per ogni coppia n, m, esiste un numero con tale proprietà. In particolare procederemo per induzione su $n + m$. Il passo base è chiaro: se $n + m = 2$ allora ho $R(1, 1) = 1$. Ora ci serve notare che, anche se solo uno tra n e m è 1, allora possiamo prendere il grafo completo su un solo vertice e la richiesta è soddisfatta. Dunque, $R(n, 1) = R(1, n) = 1$ per ogni n. Per il passo induttivo, procediamo così: se uno tra n e m è 1, allora abbiamo già notato che $R(n, 1) = 1$ o $R(1, m) = 1$. Se sono entrambi maggiori di 1, allora dimostriamo che ci basta un numero di vertici pari a

  $$R(n, m - 1) + R(n - 1, m).$$

  Visto che entrambi questi numeri di Ramsey hanno somma della 2-upla inferiore a $n + m$, avremo concluso. Consideriamo un grafo completo su $R(n, m - 1) + R(n - 1, m)$ vertici, e isoliamo un vertice v. Dividiamo, come in figura 31, il resto dei vertici del grafo in due insiemi M ed N, dove i vertici in M sono quelli collegati a v con la penna e quelli in N sono quelli collegati a v con la matita. Inoltre, poiché il numero dei vertici del grafo è

  $$R(n, m - 1) + R(n - 1, m) = |N| + |M| + 1,$$

  vale almeno una tra le due disuguaglianze $|M| \geq R(n, m - 1)$ o $|N| \geq R(n - 1, m)$. Nel primo caso, abbiamo due possibilità. O in M c'è un sottoinsieme di n vertici tutti connessi tra loro con una matita (e dunque c'era anche nel grafo iniziale) o ce n'è uno di $m - 1$ vertici tutti collegati tra loro con una penna, a cui si può aggiungere v per trovarne uno di cardinalità m. Nel secondo caso la dimostrazione è analoga.

  Abbiamo dunque dimostrato che

  $$R(n, m) \leq R(n, m - 1) + R(n - 1, m).$$

[12] Questo fatto deriva dal *principio di buon ordinamento* dei numeri naturali: un qualsiasi sottoinsieme non vuoto di $\mathbb{N}$ ammette un minimo. Per approfondire, vedi [3].

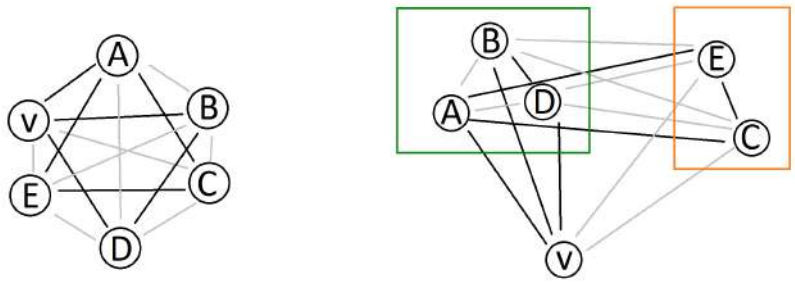


Figura 31: Il processo di isolamento di un vertice.

- Vediamo il passo induttivo (ossia $t-1 \Rightarrow t$). Dimostriamo che, nel caso io abbia una t-upla, basta prendere come numero di vertici il numero di Ramsey della $(t-1)$-upla $(n_1, \ldots, n_{t-2}, R(n_{t-1}, n_t))$, che esiste per ipotesi induttiva. Consideriamo infatti un grafo su $R(n_1, \ldots, R(n_{t-1}, n_t))$ vertici e coloriamo i suoi archi con t colori. Ora *confondiamo* gli ultimi due colori, e troviamo così una colorazione con $t-1$ colori. Per la definizione di $R(n_1, \ldots, R(n_{t-1}, n_t))$, c'è almeno un sottoinsieme di n_1 vertici tutti collegati tra loro con archi di colore c_1, oppure almeno un sottoinsieme di n_2 vertici tutti collegati tra loro con archi di colore c_2, ..., oppure almeno un sottoinsieme di $R(n_{t-1}, n_t)$ vertici tutti collegati tra loro con archi di colore c_{t-1}. Nei primi $t-2$ casi, siamo a posto. Nell'ultimo caso, basta distinguere di nuovo i due colori. Poiché questi ultimi colorano un grafo completo di $R(n_{t-1}, n_t)$ vertici, c'è almeno un sottografo di colore c_{t-1} su n_{t-1} vertici oppure c'è almeno un sottografo di colore c_t su n_t vertici. Il che conclude la dimostrazione.

□

Siamo ora pronti per dare più ufficialmente la seguente definizione:

Definizione 1 (Numero di Ramsey). Dati t numeri $n_1, \ldots, n_t$, definiamo $R(n_1, \ldots, n_t)$ il numero minimo di vertici tale che, dato un grafo completo su $R(n_1, \ldots, n_t)$ vertici i cui archi sono colorati con t colori $c_1, \ldots, c_t$, esiste almeno un sottoinsieme di n_1 vertici tutti collegati tra loro con archi di colore c_1, oppure almeno un sottoinsieme di n_2 vertici tutti collegati tra loro con archi di colore c_2, e così via.

Quello che abbiamo fatto nella proposizione 2 è stato dimostrare che $R(3,3) \leq 6$, infatti abbiamo dimostrato che, se coloriamo un grafo completo su 6 vertici con due colori, c'è almeno un sottoinsieme di 3 vertici con tutti i lati colorati del primo colore, oppure c'è almeno un sottoinsieme di 3 vertici con tutti i lati colorati del secondo colore. La figura 30 ci dimostra che $R(3,3) > 5$, e dunque $R(3,3) = 6$. In realtà, ci sarebbe da fare una osservazione: la figura 30 non ci dice soltanto che

cinque vertici non soddisfano quella proprietà, ci dice anche che non la soddisfano quattro vertici: basta "dimenticare" un vertice e cancellare gli archi di cui tale vertice era un estremo; in tal modo si ottiene un grafo completo su quattro vertici che non può avere alcun triangolo monocromatico, perché non c'era neanche prima! Chiaramente si può andare avanti per dimostrare che non basta nessun numero inferiore di vertici.

A prescindere dal teorema, che può essere di difficile comprensione, il risultato è emblematico: siamo riusciti a dire che $R(n_1, \ldots, n_t)$ esiste sempre, ma non siamo neanche vicini al calcolarlo effettivamente. Anche se gli n_i sono molto piccoli, tali valori sono ancora sconosciuti; per i più curiosi, una tabella è contenuta alla pagina web [1].

Per dare un'idea di quanto sia difficile trovare risultati precisi, Joel H. Spencer racconta nel suo libro *Ten Lectures on the Probabilistic Method* la seguente storia:

"Erdős[13] *ci chiese di immaginare che un esercito alieno, molto più potente di noi, arrivasse sulla Terra chiedendoci di fornire il valore di* $R(5, 5)$*; in caso contrario, avrebbero distrutto il nostro pianeta. Lui affermò che, in quel caso, avremmo dovuto impiegare tutti i nostri computer e tutti i matematici nel tentativo di trovare quel valore. Supponiamo, invece, che ci avessero chiesto il valore di* $R(6, 6)$*. In questo caso, a suo parere, avremmo dovuto prepararci a combattere."*

12.5 Conclusioni

In definitiva, abbiamo preso in esame tre aspetti della matematica che spesso si intrecciano fino a confondersi: la rielaborazione di dati atta a renderli più intuitivi, l'isolamento delle caratteristiche importanti di un problema, la curiosità che porta alla ricerca per generalizzare risultati che sono semplici nel piccolo.

A cosa è servito tutto questo? Quando il problema che abbiamo davanti è chiaro, come nei primi due casi, la matematica ci ha aiutato a risolverlo. Nell'ultimo caso, lo slancio verso la ricerca è qualcosa che non ha sempre un fine. Lo facciamo perché ci piace sapere cosa possiamo dire, vedere dove possiamo arrivare, trovare i nostri limiti e superarli. E, chi lo sa, se gli alieni ci chiederanno il valore di $R(4, 4)$, sapremo già cosa rispondere.

Ludovico Battista,
laureato triennale in matematica

[13] Paul Erdős, 1913-1996; uno dei matematici più prolifici di sempre. A causa del suo atteggiamento molto eccentrico, le storie sul suo conto sono innumerevoli. Per i più curiosi, si veda [5].

Riferimenti bibliografici

[1] https://en.wikipedia.org/wiki/Ramsey_theorem

[2] https://en.wikipedia.org/wiki/Pigeonhole_principle#Strong_form

[3] https://en.wikipedia.org/wiki/Well-ordering_principle

[4] https://en.wikipedia.org/wiki/Graph_theory

[5] https://en.wikipedia.org/wiki/Paul_Erd%C5%91s

13 Sistemi Lineari, Applicazioni e Aspetti Computazionali

Dario A. Bini e Beatrice Meini, n.7, Settembre 2018

Sommario

I sistemi di equazioni lineari hanno una storia molto antica e una teoria ben consolidata. Sono comunque oggetti ancora molto importanti per la ricerca matematica, in quanto modellizzano svariate situazioni del mondo reale, che fanno parte della vita di ogni giorno, ed è grazie alla risoluzione efficiente di sistemi di equazioni che possiamo ad esempio mettere facilmente a fuoco una foto sfocata, mediante un programma di foto ritocco.

In questa nota rivisitiamo i sistemi lineari, ponendo particolare attenzione alle applicazioni e agli aspetti computazionali, tracciando anche un percorso storico.

13.1 Sistemi di equazioni lineari

Un problema abbastanza familiare che si incontra nelle scuole superiori è quello di risolvere un sistema di equazioni del tipo

$$\begin{cases} ax + by = e \\ cx + dy = f \end{cases}$$

dove a, b, c, d, e, f sono numeri assegnati e dove si cercano soluzioni x, y che soddisfino simultaneamente le due equazioni. Sappiamo che, a seconda dei valori dei dati a, b, c, d, e, f, un sistema di questo tipo può non avere soluzioni, ne può avere infinite o, nella situazione generalmente più desiderata, ammette *una sola soluzione*. Ad esempio il sistema

$$\begin{cases} 3x + 4y = 11 \\ 5x + 6y = 17 \end{cases}$$

ha unica soluzione soluzione $x = 1$, $y = 2$.

Uno dei motori più significativi che alimentano la ricerca in matematica è la curiosità intellettuale, e tra i desideri più frequenti per un matematico c'è la spinta a generalizzare ed estendere il più possibile i concetti e gli strumenti disponibili. Per questo viene naturale considerare sistemi di tre equazioni e tre incognite o più in generale di n equazioni e n incognite, dove $n > 1$ è un numero intero qualsiasi. In questo caso, non avendo un numero sufficiente di lettere per rappresentare i coefficienti di questi sistemi conveniamo di usare degli indici per descrivere coefficienti, termini noti e incognite. Denotiamo allora x_j, per $j = 1, \ldots, n$, le n incognite, con $a_{i,j}$ il coefficiente di x_j nella equazione i-esima e con b_i il termine

noto nell'equazione i-esima. In questo modo il sistema si lascia scrivere nella forma seguente

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1 \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2 \\ \dots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n = b_n. \end{cases} \tag{7}$$

In sintesi il sistema viene univocamente individuato da una tabella di $n \times n$ coefficienti $A = (a_{i,j})_{i,j=1,\dots,n}$, che è chiamata *matrice*, e da una n-upla di termini noti $b = (b_i)_{i=1,\dots,n}$. In forma compatta il sistema viene scritto semplicemente come

$$Ax = b$$

dove $x = (x_i)_{i=1,\dots,n}$ è la n-upla delle incognite e la giustapposizione di A e x sintetizza la parte sinistra dell'espressione (25). Se vogliamo ricordare la struttura delle equazioni possiamo scrivere il sistema anche con la notazione

$$\sum_{j=1}^{n} a_{i,j}x_j = b_i, \quad i = 1, 2, \dots, n,$$

dove la notazione $\sum_{j=1}^{n}$ significa sommare tutti i termini per j che scorre da 1 fino a n. Nessuno ci vieta di considerare sistemi lineari n equazioni in m incognite, con $m \neq n$, cioè del tipo

$$\sum_{j=1}^{m} a_{i,j}x_j = b_i, \quad i = 1, 2, \dots, n,$$

oppure di infinite equazioni e infinite incognite, basta per questo considerare l'espressione

$$\sum_{j=1}^{\infty} a_{i,j}x_j = b_i, \quad i = 1, 2, \dots$$

purché la somma di un numero infinito di addendi dia un risultato finito.

13.2 Applicazioni

Sono innumerevoli i problemi del mondo reale che si riducono alla risoluzione di sistemi lineari. In questa sezione vogliamo presentare alcuni esempi, in forma semplificata, che sono molto significativi come rappresentanti di problemi reali.

13.3 Problemi di equilibrio

Problemi di equilibrio intervengono in numerose applicazioni del mondo reale. Vediamo una applicazione della chimica e una della fisica.

Bilanciamento di una equazione chimica

Consideriamo la reazione chimica

$$C_2H_6 + O_2 \to CO_2 + H_2O.$$

Bilanciare la reazione significa trovare i valori x, y, z e t tali che il numero di atomi di ciascun elemento sia lo stesso in entrambi i membri dell'equazione:

$$xC_2H_6 + yO_2 \to zCO_2 + tH_2O.$$

Questo porta al sistema di equazioni lineari

$$\begin{cases} 2x - z = 0 \\ 6x - 2t = 0 \\ 2y - 2z - t = 0 \end{cases}$$

la cui soluzione generale è

$$\begin{cases} y = \frac{7}{2}x \\ z = 2x \\ t = 3x. \end{cases}$$

Poiché siamo interessati a soluzioni intere, scegliamo $x = 2$, dunque otteniamo $y = 7$, $z = 4$, $t = 6$ e l'equazione bilanciata è

$$2C_2H_6 + 7O_2 \to 4CO_2 + 6H_2O.$$

Problemi più complessi richiedono la soluzione di sistemi di equazioni lineari con un numero molto più alto di incognite.

Collana appesa a due estremità

Una collana è costituita da n perle collegate da un filo elastico, dove la forza esercitata tra due perle è proporzionale alla distanza delle perle. Le perle sono soggette anche alla forza peso. Qual è la configurazione di equilibrio se appendiamo la collana a due estremi fissati?

Fissiamo un sistema di assi cartesiani con l'origine nell'estremo sinistro in cui è fissata la collana, indichiamo con (x_i, y_i) le coordinate della i-esima perla, per $i = 1, \ldots, n$, e con (b_1, b_2) le coordinate dell'estremità destra a cui è fissata la collana. Sulla perla i-esima, se $i = 2, \ldots, n-1$, agisce la forza di gravità e le forze elastiche esercitate dalle due perle contigue. Più precisamente, se k è la costante elastica del filo, nella direzione dell'asse x agiscono le forze

$$-k(x_i - x_{i+1}) - k(x_i - x_{i-1}),$$

mentre nella direzione dell'asse y agiscono le forze

$$p_i - k(y_i - y_{i-1}) - k(y_i - y_{i+1}),$$

dove p_i è la forza peso che agisce sulla perla i-esima. Sulla prima e l'ultima perla agiscono, rispettivamente, le forze nella direzione dell'asse x

$$\begin{aligned} &- kx_1 - k(x_1 - x_2), \\ &- k(x_n - x_{n-1}) - k(x_n - b_1), \end{aligned}$$

mentre nella direzione dell'asse y

$$\begin{aligned} &p_1 - ky_1 - k(y_1 - y_2), \\ &p_n - k(y_n - y_{n-1}) - k(y_n - b_2), \end{aligned}$$

assumendo per semplicità che la forza elastica esercitata dalle due estremità fisse abbia la stessa costante elastica k.

La figura 32 sintetizza le 3 forze che agiscono su una generica perla della collana.

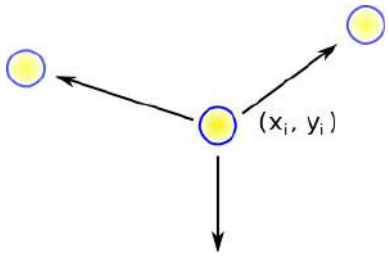


Figura 32: Le tre forze che agiscono sulla perla i-esima

Imponendo la condizione di equilibrio, cioè che la somma delle forze sia zero, lungo l'asse x otteniamo il sistema di equazioni lineari

$$\begin{cases} 2x_1 - x_2 = 0, \\ -x_{i-1} + 2x_i - x_{i+1} = 0, \quad i = 2, \ldots, n-1, \\ -x_{n-1} + 2x_n = b_1, \end{cases}$$

mentre lungo l'asse y otteniamo il sistema

$$\begin{cases} 2y_1 - y_2 = p_1/k, \\ -y_{i-1} + 2y_i - y_{i+1} = p_i/k, \quad i = 2, \ldots, n-1, \\ -y_{n-1} + 2y_n = b_2 + p_n/k. \end{cases}$$

Questi sono due sistemi di n equazioni in n incognite che hanno un'unica soluzione, che ci fornisce le coordinate (x_i, y_i) delle perle nella condizione di equilibrio. La

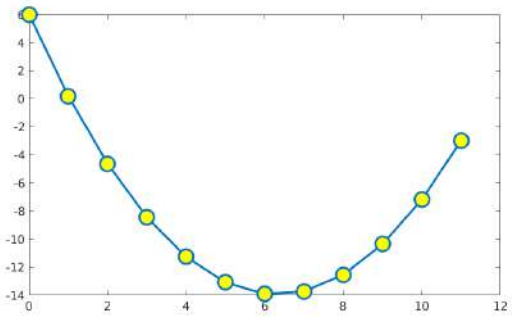


Figura 33: Posizione di equilibrio di una collana formata da 12 perle di uguale massa appesa alle due estremità.

figura 33 mostra la posizione di equilibrio di una collana formata da 12 perle di uguale massa dove l'estremo sinistro ha quota 6 e quello destro ha quota -3.

Possiamo estendere questa analisi per studiare la deformazione di una rete a molle di un letto. Stavolta ogni punto della rete è individuato da tre coordinate (x_i, y_i, z_i). Se la rete ha n maglie, in totale abbiamo n^2 punti da individuare, che si ottengono risolvendo tre sistemi di equazioni lineari, ciascuno di n^2 equazioni e n^2 incognite. In particolare, se $n = 10^3$ abbiamo 10^6 incognite!

13.4 Analisi di reti e problemi del Web

Una rete (meglio conosciuta come *network*) è costituita da un insieme di nodi e di archi che connettono i nodi. Ad esempio:

- Reti del web: l'insieme dei nodi è l'insieme della pagine web e gli archi rappresentano i link da una pagina all'altra.

- *Social networks*: gli archi rappresentano le connessioni tra varie persone (i nodi) in termini di amicizia, collaborazione, abitudini,....

- Modelli di traffico: l'insieme di nodi è costituito da punti di smistamento di un traffico (ad esempio gli incroci in un traffico stradale, oppure gli aeroporti) e gli archi sono le connessioni tra i vari nodi.

La dinamica nel tempo di queste reti può essere studiata risolvendo sistemi di equazioni lineari. Nel caso del problema del Page-Rank di Google, trattato in [3], il sistema di equazioni lineari ha dimensione circa 10 miliardi.

13.5 Modelli economici

Wassily Leontief ha ricevuto il premio Nobel nel 1973 per il suo modello di sviluppo dell'economia, che può essere interpretato mediante sistemi di equazioni lineari. Nel modello più semplice si studia un sistema economico formato da n industrie, in cui ciascuna industria consuma dei beni prodotti dalle altre industrie o da essa stessa (ad esemio un impianto che genera energia elettrica utilizza per la produzione un po' dell'energia che genera). Il sistema economico è detto chiuso se soddisfa le proprie necessità, cioè nessun prodotto entra o esce dal sistema economico. Indichiamo con $m_{i,j}$ il numero di unità che l'industria s_i produce affinché l'industria s_j produca una unità. Se p_k rappresenta la produzione totale dell'industria s_k, allora $m_{i,j}p_j$ rappresenta il numero di unità prodotte dall'industria s_i e consumate dall'industria s_j. Dunque il numero totale di unità prodotte dall'industria s_i è data da

$$m_{i,1}p_1 + m_{i,2}p_2 + \cdots + m_{i,n}p_n.$$

Per avere un'economia bilanciata, la produzione totale di ciascuna industria deve essere uguale al suo consumo totale. Questo porta al sistema di equazioni lineari

$$\begin{cases} m_{1,1}p_1 + m_{1,2}p_2 + \cdots + m_{1,n}p_n = p_1 \\ m_{2,1}p_1 + m_{2,2}p_2 + \cdots + m_{2,n}p_n = p_2 \\ \cdots \\ m_{n,1}p_1 + m_{n,2}p_2 + \cdots + m_{n,n}p_n = p_n. \end{cases}$$

Ad esempio, supponiamo che l'economia di una certa regione dipenda da tre industrie: un'industria che produce servizi, una che produce energia elettrica e un'industria petrolifera. Dopo aver monitorato le attività di queste tre industrie, si osserva che:

- per produrre una unità di servizi, l'industria che produce servizi consuma 0.3 unità della propria produzione, 0.3 unità di elettricità e 0.3 unità di petrolio;
- per produrre una unità di elettricità, l'industria che produce elettricità ha bisogno di 0.4 unità di servizi, 0.5 unità di petrolio e 0.1 unità della propria produzione;
- l'industria petrolifera, per produrre una unità di petrolio, ha bisogno di 0.3 unità di servizi, 0.6 unità di elettricità e di 0.2 unità della propria produzione.

Vogliamo calcolare quanto ciascuna industria debba produrre per soddisfare le richieste proprie e delle altre industrie, assumendo che nessun prodotto entri e esca dal sistema. Quindi il consumo totale di ciascuna industria deve essere uguale alla sua produzione totale. Indicando con p_1, p_2 e p_3 rispettivamente la produzione

dell'industria di servizi, dell'industria che produce energia elettrica e dell'industria petrolifera, si ottiene il sistema di equazioni lineari

$$\begin{cases} 0.3p_1 + 0.3p_2 + 0.3p_3 = p_1 \\ 0.4p_1 + 0.1p_2 + 0.5p_3 = p_2 \\ 0.3p_1 + 0.6p_2 + 0.2p_3 = p_3, \end{cases}$$

la cui soluzione generale è $p_1 = 0.82s$, $p_2 = 0.92$, $p_3 = s$, con s qualsiasi. Poiché cerchiamo soluzioni positive, possiamo ad esempio scegliere $s = 100$, quindi $p_1 = 82$ unità, $p_2 = 92$ unità e $p_3 = 100$ unità.

Questo modello economico assume che non ci siano beni che entrino o escano dal sistema, ma nella realtà questo succede raramente. Generalmente un'industria produce anche dei beni per richieste esterne al sistema, ad esempio l'industria petrolifera produce del petrolio per l'esportazione. In tal caso, indicando con d_i la richiesta di produzione all'industria s_i per l'esterno, l'equazione che deve soddisfare l'industria i-esima è

$$m_{i,1}p_1 + m_{i,2}p_2 + \cdots + m_{i,n}p_n + d_i = p_i.$$

13.6 Restauro di immagini

Un'immagine digitale, ad esempio una fotografia, viene codificata da una o più tabelle $m \times n$ di numeri (matrici), che rappresentano i pixel che formano l'immagine. Ad esempio, una foto in bianco/nero viene rappresentata con una tabella di $m \times n$ numeri, il cui numero in posizione (i, j) rappresenta l'intensità luminosa del puntolino (pixel) di coordinate (i, j) nella foto. Di solito il nero si rappresenta con 0 e il bianco con 255. Una foto a colori è rappresentata da tre tabelle di numeri: una per il rosso, una per il verde e una per il blu (codifica RGB).

Un problema che si presenta in tantissime applicazioni, ad esempio nello studio di foto scattate da satelliti, è la rimessa a fuoco di immagini sfocate. Se interpretiamo un'immagine come la sovrapposizione di tante immagini fatte di un solo puntolino, allora l'immagine sfocata è la somma delle sfocature di ogni singolo puntolino dell'immagine originale. Inoltre, la sfocatura di un singolo puntolino, è essa stessa un'immagine, che rappresenterà una piccola macchiolina. La figura 34 mostra l'immagine di un puntolino luminoso e la sua sfocata.

Nel caso di un'immagine bianco/nero, la sfocatura di un puntolino bianco su sfondo nero sarà una matrice che ha pochi elementi diversi da 0, che si concentrano intorno al puntolino. Chiamiamo con $A = (a_{i,j})_{i,j}$ la matrice che descrive il punto sfocato e, per convenienza, facciamo scorrere gli indici i e j da $-k$ a k, dove $2k+1$ è l'ampiezza della "patacca" e dove il centro della "patacca" ha indici $i = 0$, $j = 0$. Se chiamiamo $X = (x_{i,j})_{i=1,\ldots,m,j=1,\ldots,n}$ la matrice che rappresenta l'immagine originale,

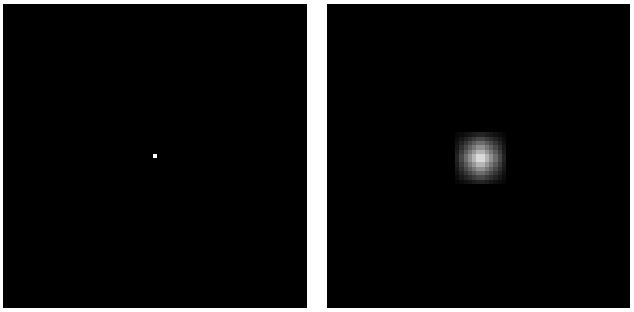


Figura 34: Immagine di un puntolino luminoso e la sua sfocata

e $S = (s_{i,j})_{i=1,\ldots,m,j=1,\ldots,n}$ la matrice che rappresenta quella sfocata, vale la relazione

$$s_{i,j} = \sum_{p=-k}^{k} \sum_{q=-k}^{k} a_{p,q} x_{i-p,j-q}, \quad i = 1, \ldots, m, j = 1, \ldots, n,$$

dove si assume $x_{i,j} = 0$ se $i \notin \{1, \ldots, m\}$, $j \notin \{1, \ldots, n\}$. Dunque, sfocare un'immagine significa calcolare gli $s_{i,j}$, mentre rimettere a fuoco significa risolvere un sistema lineare. Il numero di equazioni e di incognite è dato dal numero $m \cdot n$ di pixel dell'immagine. Per rimettere a fuoco una fotografia di 20 Megapixel occorre risolvere un sistema con 20 milioni di incognite!

Nella figura 35 è riportata una immagine originale, una versione sfocata e il risultato della rimessa a fuoco ottenuta risolvendo in modo approssimato il sistema lineare che definisce la sfocatura.

Figura 35: Immagine originale, immagine sfocata e immagine rimessa a fuoco risolvendo un sistema lineare

13.7 Come si risolve un sistema lineare

Per le innumerevoli applicazioni, il problema del calcolo della soluzione di un sistema di equazioni lineari è stato affrontato sotto i seguenti aspetti:

- dal punto di vista algoritmico, con lo sviluppo e l'analisi di metodi di risoluzione che in particolare permettano di risolvere sistemi di grandi dimensioni in tempi di calcolo ragionevoli;

- dal punto di vista numerico, con lo studio della propagazione degli errori dovuti all'uso dell'aritmetica *floating point* dei calcolatori.

Nelle sezioni seguenti andiamo a analizzare questi aspetti.

13.8 Complessità computazionale

Per risolvere un sistema di due equazioni e due incognite conosciamo il *metodo di sostituzione* che consiste nel ricavare una incognita, ad esempio la y, dalla prima equazione, sostituirla nella seconda equazione in modo da ottenere una equazione di primo grado nella sola x. Risolta questa equazione possiamo così ricavare la y. Questo metodo è noto come metodo di Gauss. Un altro metodo, noto come metodo di Cramer, che viene spesso presentato per la sua eleganza, permette di scrivere le soluzioni x e y come rapporto di due *determinanti*.

Entrambi questi metodi possono essere estesi al caso di un sistema di n equazioni in n incognite. Ma quante operazioni aritmetiche dobbiamo svolgere per applicare questi metodi? Questa domanda non è banale poiché ci aiuta a capire quanto tempo dobbiamo aspettare affinché un computer fornisca la soluzione di un sistema. Nonostante i computer disponibili attualmente siano molto veloci e possano eseguire miliardi di operazioni al secondo, può accadere di dover aspettare anni o addirittura millenni prima di avere la soluzione calcolata da un computer. Allora è opportuno sapere il costo di un metodo di risoluzione come funzione di n.

Il costo del metodo di Cramer e del metodo di Gauss è valutabile quantomeno nella parte più costosa. Infatti il metodo di Gauss, risolve un sistema eseguendo circa $\frac{2}{3}n^3$ operazioni aritmetiche, mentre il metodo di Cramer, in cui i determinanti sono calcolati con la regola di Laplace dello sviluppo per righe, ha un costo di circa $2 \cdot 3 \cdot 4 \cdots n(n+1) = (n+1)!$ operazioni aritmetiche. Le due funzioni hanno un andamento molto diverso. Nella tabella 1 si riportano i tempi di esecuzione stimati dei due algoritmi eseguiti su un PC che esegue un miliardo di operazioni al secondo. È sorprendente osservare che per risolvere un sistema di 50 equazioni col metodo di Cramer occorrono circa 5×10^{42} miliardi di anni mentre per il metodo di Gauss bastano pochi millesimi di secondo! A confronto il tempo trascorso dall'universo dal Big Bang ad oggi, che è di circa 13.7 miliardi di anni, diventa trascurabile rispetto a quello impiegato dal metodo di Cramer.

Il metodo di Gauss richiede ancora tempi di risoluzione ragionevoli se la dimensione n del sistema è moderatamente grande, ad esempio, alcune migliaia, ma se ad esempio n è dell'ordine di 20 milioni come nel problema della rimessa a fuoco di una

n	Gauss	Cramer
5	0.08milionesimi di secondo	0.7 milionesimi di secondo
10	0.6 milionesimi di secondo	4 secondi
20	530 milionesimi di secondo	160 migliaia di anni
50	8.3 millesimi di secondo	5×10^{42} miliardi di anni
10^3	66 secondi	1.3×10^{2546} miliardi di anni
10^6	2114 anni	2.6×10^{5565699} miliardi di anni
10^9	2114 miliardi di anni	*

Tabella 1: Tempi stimati per risolvere un sistema di n equazioni e n incognite con un PC usando i metodi di Gauss e di Cramer.

fotografia di 20 Megapixel che è venuta sfocata, o dell'ordine di 10 miliardi come nel caso del problema del PageRank, il tempo di calcolo richiederebbe migliaia di anni anche usando un super computer.

In questi casi la fantasia del matematico deve escogitare metodi specifici che sfruttino le peculiarità di questi sistemi come accade appunto per i due problemi citati.

Se n è molto grande può essere conveniente utilizzare metodi iterativi [4], che non calcolano esattamente la soluzione $x_1, \ldots, x_n$ del sistema, ma calcolano n successioni $x_1^{(\ell)}, \ldots, x_n^{(\ell)}$, $\ell \geq 0$, tali che $\{x_i^{(\ell)}\}_{\ell \geq 0}$ converge a x_i per $\ell \longrightarrow \infty$, per $i = 1, \ldots, n$. Sotto opportune ipotesi, è sufficiente un valore piccolo di ℓ affinché $x_i^{(\ell)}$ sia una buona approssimazione di x_i. L'idea è dunque calcolare $x_i^{(j)}$, $j = 0, \ldots, \ell$, con ℓ moderatamante piccolo, e prendere $x_i^{(\ell)}$ come approssimazione di x_i. In generale, ciascuna iterazione, cioè il calcolo di $x_1^{(j)}, \ldots, x_n^{(j)}$ dati $x_1^{(j-1)}, \ldots, x_n^{(j-1)}$, richiede circa $2n^2$ operazioni aritmetiche. Dunque se ℓ è il numero di iterazioni effettuate, il tempo di calcolo cresce come ℓn^2, che è decisamente più basso rispetto a quello del metodo di Gauss se ℓ è molto più piccolo della dimensione n.

Il costo per iterazione può essere abbassato se la matrice A, che definisce il sistema, ha della particolari caratteristiche. Ad esempio, per il problema del PageRank, A è una matrice così detta *sparsa*, cioè ha pochi elementi $a_{i,j}$ diversi da zero rispetto alla dimensione n, dunque una iterazione richiede circa kn operazioni, dove k è il numero di elementi diversi da zero [1]. La sparsità della matrice A è dovuta al fatto che una generica pagina del Web non contiene link a *tutte* le pagine esistenti ma solo ad *alcune*. Per cui in ogni riga di A ci sono pochi elementi uguali a 1 e tutti gli altri uguali a zero.

Per il problema della ristrutturazione di immagini, la matrice ha una struttura così detta *Toeplitz*, per cui una iterazione può essere calcolata mediante algoritmi veloci basata sulla Fast Fourier Transform (FFT), con un numero di operazioni

proporzionale a $n \log n$ [2].

13.9 Propagazione degli errori

Un altro problema computazionale non banale riguarda gli errori. Infatti, facendo eseguire un metodo di risoluzione a un computer e utilizzando la velocissima aritmetica *floating point*, accade che ad ogni operazione vengono generalmente introdotti errori molto piccoli, ma che possono sommarsi e amplificarsi durante il calcolo e rovinare completamente il risultato. Questi piccoli errori sono dovuti al fatto che i numeri *floating point* si rappresentano con un numero finito di cifre, per cui occorre spesso troncare a questo valore il numero di cifre dei risultati calcolati ad ogni passo. Al termine del calcolo, il computer restituisce dei numeri $\widetilde{x}_1, \ldots, \widetilde{x}_n$, che sono una approssimazione della soluzione esatta $x_1, \ldots, x_n$.

Ad esempio, il sistema

$$\begin{cases} \varepsilon x_1 + (1+\varepsilon) x_2 = 1 + 2\varepsilon \\ x_1 + \varepsilon^{-1} x_2 = 1 + \varepsilon^{-1} \end{cases} \tag{8}$$

dove $\varepsilon \neq 0$ è un numero assegnato, ha una unica soluzione data da $x_1 = 1$, $x_2 = 1$. La Tabella 2 mostra, per diversi valori di ε, i valori di $\widetilde{x}_1$ e $\widetilde{x}_2$ calcolati da un PC utilizzando l'eliminazione di Gauss e l'errore di approssimazione, definito come

$$\max \left\{ |x_1 - \widetilde{x}_1|, |x_2 - \widetilde{x}_2| \right\}.$$

Si osserva come i risultati deteriorino quando ε diventa piccolo. In particolare, se $\varepsilon = 10^{-8}$, il valore di $\widetilde{x}_1$ è addirittura negativo. Dalla tabella si osserva che, dividendo ε per 10, l'errore viene circa moltiplicato per 100 o, in altri termini, il numero di cifre corrette diminuisce di 2.

ε	$\widetilde{x}_1$	$\widetilde{x}_2$	errore
10^{-4}	1	1	0
10^{-5}	1.000001110223025	0.999999999988897	$1.11 \cdot 10^{-6}$
10^{-6}	0.999777955395075	1.000000000222045	$2.22 \cdot 10^{-4}$
10^{-7}	1.022204460492503	0.9999999977795540	$2.22 \cdot 10^{-2}$
10^{-8}	-1.220446071454774	1.000000022204461	2.22

Tabella 2: Soluzione effettivamente calcolata $\widetilde{x}_1$, $\widetilde{x}_2$ del sistema (8) e errore $\max\{|1-\widetilde{x}_1|, |1-\widetilde{x}_2|\}$, per diversi valori di ε

H. Hotelling negli anni 40 del secolo scorso dimostrò che, in generale, quando si applica il metodo di Gauss gli errori generati ad ogni operazione possono crescere

esponenzialmente in funzione di n. In questo modo il metodo sarebbe completamente inutilizzabile poichè produrrebbe dei risultati completamente rovinati dall'errore. In realtà la situazione non è così catastrofica se si utilizzano particolari tecniche per controllare la crescita degli errori. Nel 1948, A. Turing introdusse delle strategie, successivamente affinate da J. H. Wilkinson, dette *strategie di pivoting*, per scegliere quale incognita ricavare da quale equazione da sostituire nelle altre in modo tale che gli errori generati dall'aritmetica floating point siano tenuti sotto controllo. Succesivamente, A. Householder introdusse un metodo di risoluzione che ha un costo leggermente più alto ($\frac{4}{3}n^3$ operazioni aritmetiche) che non produce alcuna amplificazione degli errori.

Ad esempio, vogliamo risolvere il sistema di n equazioni e n incognite

$$\begin{cases} 0.1x_1 + x_n = 1.1 \\ x_1 + 0.1x_2 = 1.1 \\ x_2 + 0.1x_3 = 1.1 \\ \dots \\ x_{n-1} + 0.1x_n = 1.1 \end{cases} \tag{9}$$

che ha come soluzione $x_1 = 1, x_2 = 1, \dots, x_n = 1$. Calcoliamo la soluzione al PC utilizzando l'eliminazione di Gauss con e senza strategie di pivoting, per diversi valori di n. La Tabella 3 mostra, al variare di n, l'errore di approssimazione, definito come $\max_{i=1,\dots,n} |x_i - \widetilde{x}_i|$, dove $\widetilde{x}_i$ è la soluzione calcolata dal PC, con e senza strategie di pivoting. Si osserva come i risultati deteriorino quando n aumenta se non si usano strategie di pivoting, invece i risultati hanno errori molto bassi e quasi indipendenti da n utilizzando pivoting.

n	Con Pivoting	Senza Pivoting
5	0	$9.09 \cdot 10^{-13}$
10	$2.22 \cdot 10^{-16}$	$1.32 \cdot 10^{-7}$
15	$2.22 \cdot 10^{-16}$	$1.82 \cdot 10^{-2}$
20	$1.11 \cdot 10^{-16}$	$9.61 \cdot 10^{2}$
25	$1.11 \cdot 10^{-16}$	$1.13 \cdot 10^{8}$

Tabella 3: Errore $\max_{i=1,\dots,n} |1 - \widetilde{x}_i|$ della soluzione del sistema (9), calcolata con e senza strategia di pivoting, per diversi valori di n

13.10 Cenni storici

La comparsa dei sistemi di equazioni lineari, in base alle conoscenze storiche attuali, risale alla cultura cinese di oltre 2500 anni fa. I primi tentativi documentati

di risolvere sistemi si trovano nel libro cinese Chiu-chang Suan-shu (nove capitoli sull'aritmetica) che si stima fosse stato scritto intorno al 200 AC. All'inizio dell'ottavo capitolo è descritto il problema:
tre fasci di grano di buona qualità, due di media qualità e uno di cattiva qualità sono venduti a 39 dou; due fasci di buona qualità, tre di media e uno di cattiva qualità sono venduti a 34 dou; un fascio di buona qualità, due di media e tre di cattiva qualità sono venduti a 26 dou. Quali sono i costi di ciascun fascio di buona, media e cattiva qualità?

In chiave moderna il problema si formula con il sistema

$$\begin{cases} 3x + 2y + z = 39 \\ 2x + 3y + z = 34 \\ x + 2y + 3z = 26. \end{cases}$$

Nel libro il problema viene affrontato mettendo dei bastoncini colorati di bambù, che rappresentano i coefficienti, in un tavoliere e le righe del tavoliere vengono manipolate secondo delle regole opportune.

Documentazioni successive si hanno dopo quasi due millenni, quando il matematico giapponese S. Kowa (1642–1708) migliora la tecnica cinese introducendo il concetto attualmente noto come determinante. Circa nello stesso periodo il matematico tedesco G. W. Leibniz (1646–1716) sviluppa in modo indipendente il suo concetto di determinante. Sembra che sia nel lavoro di Kowa che in quello di Leibniz sia contenuta quella che poi verrà chiamata la regola di Cramer per risolvere sistemi, scoperta poi da G. Cramer (1704–1752).

Tra il 1750 e il 1900 viene scritto molto sul concetto di determinante, esso diventa lo strumento più importante per risolvere sistemi lineari. Nel 1809 il matematico tedesco C. F. Gauss usa il metodo di sostituzione per risolvere sistemi, chiamato solo recentemente metodo di eliminazione gaussiana. Nel 1930, W.H. Richardson sviluppa metodi iterativi.

Gli aspetti computazionali dell'algebra lineare cominciano ad essere studiati con J. von Neumann (1903–1957), matematico ungherese emigrato in America negli anni '40. L'avvento dei calcolatori dà un grande impulso alla ricerca di metodi di risoluzione efficienti. Nel 1932 M. Picone crea il corso di *Calcoli Numerici e Grafici: sistemi lineari, non lineari, integrazione, analisi di Fourier* presso la Scuola di Scienze Statistiche e Attuariali di Roma. A proposito dei sistemi Picone scrive: "il problema può essere considerato come la divisione tra numeri in piu dimensioni... Mentre per la divisione esistono macchine automatiche, per i sistemi lineari non è così ... Recentemente il prof. Malloc dell'Università di Cambridge ha costruito una macchina elettrica molto originale che può risolvere sistemi fino a 10 equazioni e incognite." Nel 1946, L. Fox, C. Goodwin, A. Turing e J. H. Wilkinson risolvono un sistema 18×18 con una calcolatrice da tavolo in due settimane.

Nel 1943 H. Hotelling (1895–1973) dimostra che il metodo di Gauss è inaffidabile se usato con una aritmetica approssimata come è quella dei computer, a causa della grande amplificazione degli errori. Nel 1947 von Neumann e Goldstein dimostrano che il metodo di Gauss è affidabile se applicato a matrici definite positive. È J. H. Wilkinson (1919–1986) a fare per primo un'analisi sistematica e rigorosa degli errori e a dimostrare che il metodo di Gauss, accompagnato da opportune strategie di pivoting, è comunque affidabile ed efficiente. Gli aspetti computazionali dell'algebra lineare hanno un grandissimo sviluppo a partire dal secolo scorso, con notevoli contributi del matematico americano G. H. Golub (1932–2007).

Un enorme impulso alla ricerca è stato dato dalle applicazioni. I modelli matematici nelle scienze applicate, nel calcolo scientifico e nell'ingegneria si riconducono alla risoluzione di grossi sistemi di equazioni lineari. Maggiore è l'accuratezza del modello e maggiore è il numero di equazioni e incognite del sistema. Allo studioso è richiesta la continua individuazione e analisi di metodi eficienti che permettano di trattare problemi di dimensioni sempre più grandi in tempi di calcolo contenuti.

Lo sviluppo di internet, in particolare l'analisi delle reti complesse, ha moltiplicato i problemi allo studio e aumentato significativamente l'importanza e l'interesse dei metodi dell'algebra lineare.

Dario A. Bini e Beatrice Meini,
Professori Ordinari presso il Dipartimento di Matematica di Pisa

Riferimenti bibliografici

[1] A.N. Langville and C.D. Meyer, Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton, NJ, 2012.

[2] M.K. Ng, Iterative methods for Toeplitz systems. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004.

[3] F. Poloni, La matematica dell'importanza. Il giornalino degli Open Days, n° 2. 2016

[4] R.S. Varga, Matrix iterative analysis. Springer Series in Computational Mathematics, 27. Springer-Verlag, Berlin, 2000.

14 Matematica del Conteggio

Filippo Disanto, n.7, Settembre 2018

L'idea in queste pagine è di illustrare alcuni degli scopi e dei metodi della combinatoria enumerativa, quella parte della matematica che—detto molto in generale—si occupa di contare il numero di possibili configurazioni che un sistema discreto può assumere, al variare di uno o più parametri scelti. Partiamo con un esempio standard che ci permette di accennare ad alcuni strumenti di base.

14.1 Enumerazione di alberi binari orientati

Un albero binario orientato si sviluppa a partire da una radice, con due figli per ogni nodo non terminale (Fig. 36A). Il termine "orientato" indica che stiamo considerando anche l'orientamento destra/sinistra dei vari rami. Come si vede nella Figura 36A, per $n = 1, 2, 3, 4$ ci sono rispettivamente $1, 1, 2, 5$ alberi binari orientati con n foglie. Quanti alberi ci sono per un valore di n arbitrario? Vediamo come dare una formula esatta introducendo alcune tecniche di conteggio.

Denotiamo con a_n il numero di alberi con n foglie—quindi $a_1 = 1, a_2 = 1, a_3 = 2, a_4 = 5$, e così via. La soluzione al nostro problema può essere trovata attraverso la *funzione generatrice*

$$A(z) = a_1\, z + a_2\, z^2 + a_3\, z^3 + a_4\, z^4 + \dots$$

associata alla sequenza (a_n). Senza preoccuparci troppo di problemi di convergenza, interpretiamo la nostra funzione $A(z)$ come una somma infinita puramente simbolica, in cui le potenze z^n sono dei segnaposti per i numeri a_n.

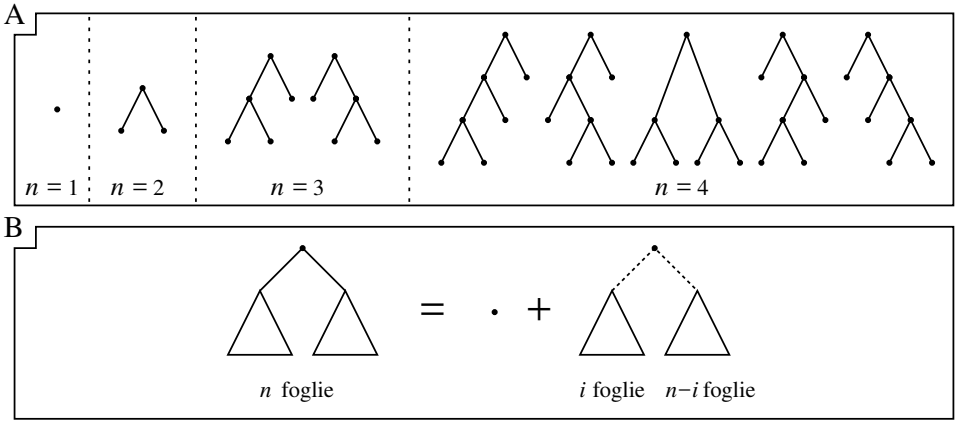


Figura 36: Alberi binari orientati. **(A)** Alberi con $1 \leq n \leq 5$ foglie. **(B)** Decomposizione di un albero con $n \geq 1$ foglie.

Il primo passo verso la soluzione consiste nel trovare un' equazione per $A(z)$. L'idea è di tradurre in equazione la naturale decomposizione degli alberi che stiamo considerando. In Figura 36B, si vede che per un albero binario t con n foglie ci sono due opzioni possibili: o t consiste di una sola foglia, cioè $n = 1$, oppure, se $n > 1$, t si decompone nei due sottoalberi attaccati alla radice di t. Questi avranno rispettivamente i ed $n - i$ foglie, per un certo valore $1 \leq i \leq n - 1$. Per $n > 1$, tale decomposizione si traduce direttamente nella ricorrenza $a_n = a_1 \cdot a_{n-1} + a_2 \cdot a_{n-2} + \cdots + a_{n-1} \cdot a_1$, che ci fornisce

$$\begin{aligned}
A(z) &= a_1\, z + a_2\, z^2 + a_3\, z^3 + a_4\, z^4 + \dots \\
&= a_1\, z + (a_1\, z \cdot a_1\, z) + (a_1\, z \cdot a_2\, z^2 + a_2\, z^2 \cdot a_1\, z) \\
&\quad + (a_1\, z \cdot a_3\, z^3 + a_2\, z^2 \cdot a_2\, z^2 + a_3\, z^3 \cdot a_1\, z) + \dots \\
&= a_1\, z + a_1\, z\, (a_1\, z + a_2\, z^2 + a_3\, z^3 + \dots) \\
&\quad + a_2\, z^2\, (a_1\, z + a_2\, z^2 + \dots) + a_3\, z^3\, (a_1\, z + \dots) + \dots \\
&= a_1\, z + a_1\, z\, A(z) + a_2\, z^2\, A(z) + a_3\, z^3\, A(z) + \dots \\
&= a_1\, z + A(z) \cdot (a_1\, z + a_2\, z^2 + a_3\, z^3 + \dots) = a_1\, z + A(z) \cdot A(z) \\
&= z + A(z)^2.
\end{aligned}$$

Trovata l'equazione che cercavamo, possiamo in questo caso anche risolverla ottenendo

$$A(z) = \frac{1 - \sqrt{1 - 4z}}{2}.$$

L'espansione in serie di potenze di $A(z)$ per $z = 0$ ci mostra che $A(z) = A(0) + \frac{A'(0)}{1!}\, z + \frac{A''(0)}{2!}\, z^2 + \frac{A'''(0)}{3!}\, z^3 + \frac{A''''(0)}{4!}\, z^4 + \dots$, dove $A(0) = 0$ e per $n \geq 1$

$$\frac{A^{(n)}(0)}{n!} = a_n.$$

Verifichiamo quest'ultima uguaglianza per i primi valori di n. Poichè $A'(z) = (1 - 4z)^{-1/2}$ si ha che $\frac{A'(0)}{1!} = 1 = a_1$, poi $A''(z) = 2\,(1 - 4z)^{-3/2}$ ci dice che $\frac{A''(0)}{2!} = 1 = a_2$, siccome $A'''(z) = 12\,(1 - 4z)^{-5/2}$ abbiamo che $\frac{A'''(0)}{3!} = 2 = a_3$, ed infine $A''''(z) = 120\,(1 - 4z)^{-7/2}$ implica che $\frac{A''''(0)}{4!} = 5 = a_4$.

Attraverso semplici operazioni di derivazione, l'espressione trovata per la funzione generatrice $A(z)$ ci permette quindi di calcolare il valore dei numeri a_n per un n arbitrario. In particolare, si verifica direttamente che si ha

$$a_n = \frac{1}{n}\binom{2n-2}{n-1}$$

Infatti, da come si comportano le prime derivate di $A(z)$, si vede che per $n \geq 1$ il valore di $A^{(n+1)}(0)$ può essere calcolato ricorsivamente come $A^{(n+1)}(0) = 2\,(2n -$

1) $A^{(n)}(0)$. Da quest'ultima relazione, dividendo entrambi i membri per $(n+1)!$, otteniamo la ricorrenza

$$a_{n+1} = \frac{2\,(2n-1)}{n+1}\,a_n,$$

che è soddisfatta dalla formula per a_n indicata sopra.

Dall'ultima ricorrenza trovata si nota che, per n sufficientemente grande, il rapporto a_{n+1}/a_n tende al valore 4. Questo ci dice che la quantità a_n che stiamo studiando ha una crescita esponenziale del tipo 4^n. La cosa interessante è che avremmo potuto dedurre questa proprietà asintotica direttamente dalla espressione trovata per $A(z)$. Detto in termini semplici, è infatti un fenomeno generale che una sequenza numerica cresca come $(1/\rho)^n$ quando ρ è il più grande numero reale tale che in tutti i punti complessi $z = a + i \cdot b$ di modulo $|z| = \sqrt{a^2+b^2} < \rho$ la funzione generatrice della sequenza risulta avere un andamento "regolare", privo di spigolature o lacerazioni (Fig. 37). Nel caso della funzione $A(z)$, a causa della

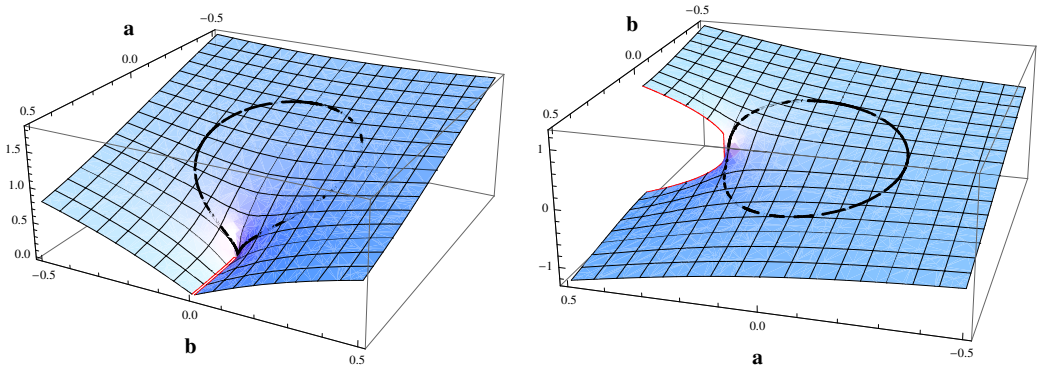


Figura 37: Parte reale (a sinistra) e parte immaginaria (a destra) della funzione $\sqrt{1-4z}$, con $z = a + i \cdot b$. Nella regione $|z| = \sqrt{a^2+b^2} < 1/4$ (dentro il laccio), la funzione $\sqrt{1-4z}$ ha un andamento "regolare".

presenza della radice $\sqrt{1-4z}$ abbiamo $\rho = 1/4$, da cui è possibile dedurre la crescita esponenziale del tipo $(1/\rho)^n = 4^n$ per la sequenza (a_n).

La corrispondenza tra la crescita asintotica di una sequenza e le proprietà analitiche della funzione generatrice associata è ciò che rende la *combinatoria analitica* [1] uno strumento particolarmente utile per risolvere problemi enumerativi.

14.2 Configurazioni in alberi di specie

Dopo aver accennato alle funzioni generatrici ed al legame tra le loro proprietà analitiche e la cresicita asintotica della sequenze numeriche associate, mostriamo brevemente un' applicazione di queste tecniche in ambito biologico-computazionale.

Un *albero di specie* è un albero binario che descrive le relazioni ancestrali tra specie o popolazioni (Fig. 38). I rami dell'albero rappresentano popolazioni che evolvono nel tempo. Andando dal basso verso l'alto (indietro nel tempo), due diverse popolazioni confluiscono in uno stesso ramo che rappresenta la popolazione ancestrale ad entrambe. Gli intervalli temporali di questo processo corrispondono alle lunghezze dei vari rami dell'albero di specie. Ad esempio, in Figura 38

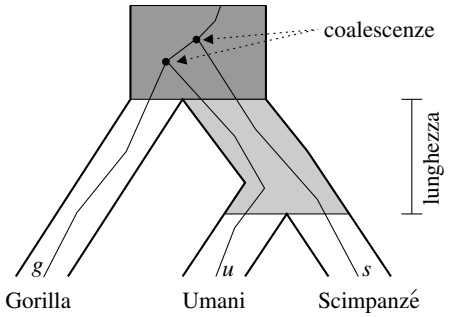


Figura 38: Un albero di specie con all'interno un albero filogenetico realizzato prendendo tre copie g, u, s dello stesso gene appartenenti ad individui delle tre specie considerate. Il gene u trova un progenitore comune con il gene g prima di trovarlo con il gene s.

si considerano tre specie: Gorilla, Umani e Scimpanzé. Umani e Scimpanzé confluiscono in una stessa popolazione che, dopo un certo numero di generazioni misurato dalla lunghezza del rispettivo ramo, si congiunge con la popolazione dei Gorilla. Secondo tale rappresentazione, Umani e Scimpanzé sono in più stretta relazione tra loro di quanto lo siano con i Gorilla.

Può succedere—ed in realtà succede spesso—che seguendo la storia evolutiva di singole parti di genoma prese da individui delle popolazioni considerate, l'*albero filogenetico* che se ne ricava non sia in accordo con l'albero di specie. In Figura 38, prendendo un gene che regola la stessa funzione in tre individui g, u, s delle tre specie in esame, l'albero filogenetico che si sviluppa all'interno dell'albero di specie ci dice che il gene preso dall'umano u trova un antenato comune con il gene preso dal gorilla g prima che le rispettive linee genetiche possano coalescere con la linea proveniente dal gene preso dallo scimpanzé s. Questo tipo di fenomeno di discordanza, detto di "incomplete lineage sorting" [2], è correlato alle quantità temporali in gioco nell'evoluzione delle specie considerate: più sono corti i rami dell'albero di specie e più è probabile che a livello di singole porzioni di genoma l'albero filogenetico sia non concorde.

Per calcolare la probabilità di tali eventi e, più in generale, per determinare con quale probabilità un certo albero filogenetico può prodursi all'interno di un dato albero di specie, si usano modelli stocastici ed algoritmi la cui complessità può essere descritta attraverso lo studio della numerosità di certe strutture combinatorie che codificano i modi diversi con cui un albero filogenetico può disporsi interna-

mente ad un albero di specie. Di queste strutture se ne trovano di diversi tipi, a seconda dell'algoritmo che si considera. Accenniamo brevemente alle *configurazioni ancestrali*, restringendoci al caso in cui l'albero filogenetico e l'albero di specie siano concordi.

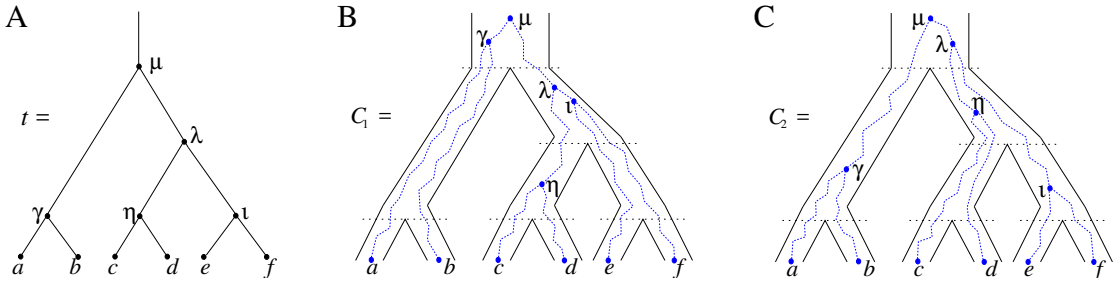


Figura 39: Configurazioni diverse dello stesso albero filogenetico in un albero di specie.

Supponiamo di avere un albero di specie ed un albero filogenetico che hanno la forma dell'albero t in Figura 39A. L'albero filogenetico può configurarsi dentro l'albero di specie in vari modi, ad esempio come in Figura 39B o Figura 39C. La configurazione C_1 si differenzia dalla configurazione C_2 in base a quali sono i rami dell'albero di specie in cui avvengono le diverse coalescenze dell'albero filogenetico interno (denotate in lettere greche). Nella configurazione C_1, le coalescenze dell'albero interno che avvengono nella radice dell' albero di specie sono γ e μ. Nella configurazione C_2, le coalescenze che avvengono nella radice dell'albero di specie sono invece μ e λ. I due insiemi $\{\gamma, \mu\}$ e $\{\lambda, \mu\}$ sono due delle possibili configurazioni ancestrali alla radice dell'albero di specie t. Un'altra possibile configurazione ancestrale alla radice è ad esempio $\{\gamma, \eta, \iota, \lambda, \mu\}$, quando tutto l'albero interno si realizza nella radice dell'albero di specie.

In generale, la quantità da studiare è il numero $c(t)$ di possibili configurazioni ancestrali alla radice di un albero di specie t. Per alberi di specie t che sono massimamente asimmetrici, la quantità $c(t)$ cresce in maniera polinomiale rispetto al numero di foglie in t. In altri casi, $c(t)$ cresce in modo esponenziale. Cosa succede se prendiamo un albero di specie t scelto a caso con probabilità uniforme tra tutti quelli che hanno n foglie? Usando una tecnica simile a quella adottata per l'enumerazione degli alberi binari orientati, si può far vedere che in media $c(t)$ cresce asintoticamente come

$$\mathbb{E}[c(t)] \sim \sqrt{\frac{3}{2}} \left(\frac{4}{3}\right)^n.$$

14.3 Conclusioni

Abbiamo brevemente accennato ad alcuni scopi e metodi della combinatoria enumerativa: la matematica del conteggio. Oltre alle possibili applicazioni, un

aspetto interessante di questa disciplina è dato dalla semplicità con cui i problemi possono essere formulati. Attenzione però, non sempre le soluzioni sono altrettanto semplici da trovare!

Filippo Disanto,
Ricercatore presso il Dipartimento di Matematica di Pisa

Riferimenti bibliografici

[1] P. Flajolet, R. Sedgewick. 2009. *Analytic Combinatorics.* Cambridge: Cambridge University Press.

[2] D. Venema. 2013. *Evolution Basics: Species Trees, Gene Trees and Incomplete Lineage Sorting.* https://biologos.org/blogs/dennis-venema-letters-to-the-duchess
/evolution-basics-species-trees-gene-trees-and-incomplete-lineage-sorting/

15 Quanto tempo ci vuole per mescolare un mazzo di carte?

Alessandra Caraceni, n.8, Gennaio 2019

...e quanto per completare una collezione di figurine? Quanto per mandare in rovina un giocatore d'azzardo?

Lo studio di questi problemi trova le proprie radici nelle origini settecentesche della teoria della probabilità, ma porta ancora oggi sempre nuovi frutti, nella forma di ulteriori risultati, tecniche e applicazioni.

Alla ricerca di risposte, ci imbatteremo in alcuni degli strumenti più efficaci che i matematici abbiano elaborato per catturare con i loro modelli un fenomeno del reale che a lungo li aveva elusi: il caso.

15.1 Le catene di Markov

Consideriamo un problema che all'apparenza ha poco a che fare col mescolare mazzi di carte: sfidiamoci a testa o croce. A turno, lanciamo una moneta; diciamo che, con grande generosità, lascerò che cominciate voi. Se esce testa, sarò costretta a darvi un euro; se esce croce, sarete voi a darmi un euro. Poi tocca a me, che lancerò una moneta per determinare se sarete voi a dovermi dare un euro o viceversa, e avanti così. Quando uno di noi rimane a corto di euro, il gioco termina.

Come "modellizzare" questo gioco in modo da poterlo analizzare? La quantità di cui ci interessa tenere traccia nel tempo è, chiaramente, quanti euro ciascuno di noi due possieda; in verità, assumeremo che il numero totale di euro rimanga invariato (diciamo che è n all'inizio, di cui voi ne possedete a e io $n-a$); è così sufficiente conoscere il totale e tenere traccia dei vostri euro per poter ricostruire quanti ne abbia io. Ad ogni "mossa" del gioco, supponendo che voi abbiate $0 < k < n$ euro, il lancio della moneta (che sia il mio turno o il vostro) determina se al turno successivo avrete $k+1$ o $k-1$ euro; se la moneta non è truccata, questi due eventi hanno uguale probabilità. Se però avete 0 o n euro, il gioco è finito: rimarrete in questo stato per sempre.

Detto X_t il numero di euro che possedete dopo t turni (dove, diciamo, $X_0 = a$), si tratta per $t > 0$ di una quantità aleatoria che dipende dalla sequenza dei risultati dei lanci che abbiamo effettuato, con la proprietà che, se conosciamo il valore di X_t, sappiamo precisamente quali siano le probabilità per X_{t+1} di assumere ciascuno dei valori possibili; ovvero, il valore di X_{t+1} dipende da quello di X_t, ma non da come il gioco si sia svolto in precedenza: non importa che io sia stata fortunata per tutta la partita; il $t+1$-esimo lancio della moneta determinerà in maniera del tutto imparziale se il vostro capitale aumenta o diminuisce.

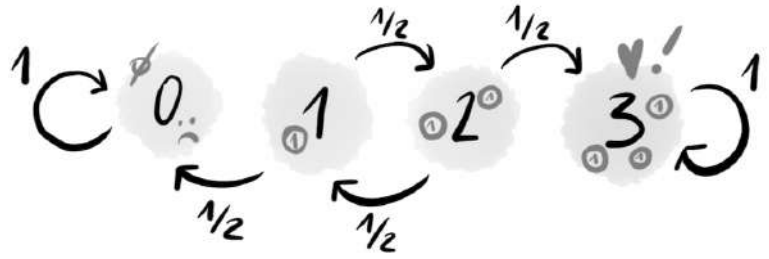


Figura 40: I quattro stati della catena X_t nel caso $n = 3$ e le relative probabilità di transizione.

Un processo X_t con questa proprietà prende il nome di *catena di Markov*; l'insieme dei valori possibili per X_t, che nel nostro caso è $\{0, 1, \ldots, n\}$, è detto *insieme degli stati*. L'intero processo può essere descritto semplicemente dichiarando, per ogni coppia di stati, la probabilità che, se la catena si trova nel primo stato al tempo t (o al "passo" t, o dopo la t-esima mossa), si trovi nel secondo stato al tempo $t + 1$[14]. Un modo molto naturale di rappresentare una catena di Markov è quello in Figura 40.

Una branca molto vivace della probabilità si occupa di studiare il comportamento asintotico, cioè dopo tempi molto grandi, delle catene di Markov. È chiaro che, per la nostra catena X_t, quello che immaginiamo debba succedere dopo tanto tempo è trovarci nello stato n o nello stato 0. Questo ha a che fare con il fatto che questi due stati sono raggiungibili dal nostro stato iniziale a e che, se la catena parte da 0 o da n, vi rimarrà indefinitamente (si chiamano, incidentalmente, stati *di assorbimento*).

Una domanda naturale da porsi è quale sia la probabilità di andare prima o poi a finire in 0 a seconda del vostro capitale a di partenza e di quello totale n. La risposta ha una forma sorprendentemente semplice e vi invito a tentare di scoprirla, ma non è questo il genere di questione del quale ci occuperemo in questo articolo. Quello che vogliamo invece studiare è il tempo necessario perché la catena raggiunga uno fra lo stato 0 e lo stato n (si dice *diventi stazionaria* o *raggiunga l'equilibrio*): voglio almeno una stima di quanto tempo dovrò dedicarvi per questo gioco; basta il tempo di un caffè? Devo cancellare i miei piani per la prossima settimana?

Torneremo presto ad analizzare il gioco proposto, ma prima vediamo perché quel che abbiamo detto finora si presti effettivamente ad insegnarci qualcosa sui mazzi di carte e sui modi di mescolarli: perché è possibile parlare di mescolate in termini di catene di Markov?

[14]per la verità questa probabilità potrebbe anche dipendere da t, ma noi considereremo solo catene in cui questo non avviene, cosiddette *omogenee* nel tempo.

Beh, immaginate un mazzo di n carte; questo sarà inizialmente ordinato in un certo modo, uno di $n! = n \cdot (n-1) \cdot \ldots \cdot 2 \cdot 1$ modi diversi[15]. Per mescolarlo sarà necessario compiere una sequenza di mosse in qualche modo "casuali" che modifichino questo ordinamento. Immaginare questa sequenza come una catena di Markov è molto naturale: il nuovo ordinamento del mazzo dopo una mossa di mescolamento dipenderà sicuramente solo da quello immediatamente precedente e non dalle mosse fatte in passato. In più, una mossa di mescolamento dovrebbe essere "abbastanza casuale" perché, se il mazzo è già mescolato (cioè non conosciamo il suo ordinamento, che può essere uniformemente uno degli $n!$ possibili), questo rimanga ben mescolato dopo un'ulteriore mossa; e "abbastanza efficace" perché, dato un qualunque ordinamento iniziale, sia possibile con una sequenza di mosse raggiungerne qualunque altro (altrimenti il mazzo potrebbe non mescolarsi mai!).

Queste richieste sulle mosse sono essenzialmente sufficienti perché, se consideriamo la catena Y_t il cui insieme degli stati è quello degli $n!$ ordinamenti delle carte e i cui passaggi da uno stato all'altro sono dati dalle mosse di mescolamento, alla lunga il mazzo risulti ben mescolato qualunque sia Y_0; ovvero ci aspettiamo che, per t abbastanza grande, la probabilità che Y_t sia un qualunque ordinamento fissato sia, almeno all'incirca, $1/n!$.

Ma quale sarebbe un esempio di "catena di mescolamento" per un mazzo di carte? Una singola mossa potrebbe consistere, per esempio, nel prendere la carta in cima al mazzo e reinserirla "a caso" al suo interno. Questo metodo, spesso chiamato per ovvi motivi *top-to-random*, è il primo che analizzeremo; non sarà particolarmente efficiente, ma è semplice da modellizzare e dovrebbe essere – almeno intuitivamente – chiaro che alla lunga sarà efficace nel mescolare il mazzo. In ogni caso dimostreremo quest'ultimo fatto e analizzeremo in dettaglio il tempo necessario per raggiungere la stazionarietà nella Sezione 15.4. Nella Sezione 15.5 concluderemo con l'analisi di un metodo di mescolamento molto più simile a uno reale, che Bayer e Diaconis sono riusciti a trattare in un articolo nel 1992 che fece notizia perfino nel mondo non matematico, guadagnando loro un posto in prima pagina sul New York Times!

Ma per adesso torniamo al nostro gioco di testa o croce; per semplicità, analizzeremo solo il caso $n = 3$, che è il primo caso interessante. La catena X_t che stiamo considerando è, c'è da dire, un po' particolare per il fatto di avere *due* possibili esiti "stazionari": potete vincere tutti i 3 euro o perderli tutti. Un po' perché è in generale preferibile studiare catene che abbiano un'unica distribuzione stazionaria e un po' per mostrare come sia possibile, scegliendo un insieme degli stati diverso,

[15]Quello che voglio dire è che ci sono n possibilità per la carta che sta in cima al mazzo; scelta quella, $n-1$ possibilità per la successiva, e così via fino all'ultima. Quindi il numero di ordinamenti possibili è il prodotto dei numeri da 1 a n, detto n *fattoriale* o $n!$.

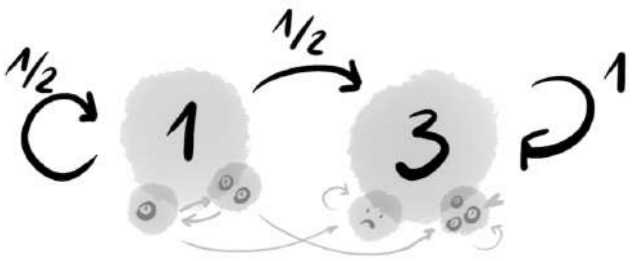


Figura 41: I due stati della catena $\widetilde{X}_t$; ciascuno "ingloba" due diversi stati della catena originale X_t.

attaccare lo stesso problema analizzando catene diverse, modifichiamo leggermente il nostro modello.

Se quella che ci interessa è la durata del gioco e non la probabilità che questo termini con una vostra vittoria o sconfitta, potremmo decidere di prendere come stato il valore assoluto della differenza fra il mio e il vostro numero di euro. Ma questo può valere solamente 1 o 3, e quando diviene 3 il gioco termina! La nuova catena $\widetilde{X}_t$, che stiamo definendo come $|X_t - (3 - X_t)| = |2X_t - 3|$, è quindi semplicissima: la probabilità di passare da 3 a 3 è 1 e quella di passare da 3 a 1 è zero (una volta raggiunta differenza 3, non possiamo più giocare e questa si conserva per sempre). D'altra parte, se uno di noi ha un euro e l'altro due, a ogni mossa c'è una probabilità di $1/2$ che chi ha meno euro ne guadagni uno (lo stato rimane lo stesso) e $1/2$ che lo perda (si passa allo stato 3), vedi Figura 41.

Supponiamo di partire dallo stato $\widetilde{X}_0 = 1$ e sia T il primo tempo tale che $\widetilde{X}_T = 3$; allora la probabilità che T valga k, che chiamerò $\mathbb{P}(T = k)$, è la probabilità di rimanere nello stato 1 per $k - 1$ volte e di uscirne alla k-esima, cioè $\frac{1}{2^k}$.

Come calcolare dunque la durata media del gioco, cioè il valore medio di T? Si tratta, appunto, di calcolare una media dei valori possibili per T che dia a ciascuno un peso proporzionale alla sua probabilità di verificarsi. Se T prendesse un certo numero finito di possibili valori, ciascuno con la medesima probabilità, il suo valor medio, anche detto "valore atteso" di T e spesso denotato con $\mathbb{E}(T)$, sarebbe semplicemente la media aritmetica dei valori possibili. In questo caso vogliamo dare a ciascun valore k un peso dato dalla probabilità che si abbia $T = k$, ovvero moltiplicarlo per 2^{-k}.

In pratica, vorremmo conoscere o approssimare la quantità

$$1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + \dots,$$

ma disgraziatamente non è ovvio come riuscirci!

Quello che conviene fare è usare seguente trucco: si può calcolare $\mathbb{E}(T)$, anziché come somma dei valori $k\,\mathbb{P}(T = k)$ per $k \geq 1$, come somma dei valori $\mathbb{P}(T > s)$ per

$s \geq 0$, ovvero come somma delle probabilità che T assuma un valore strettamente maggiore di ciascun numero naturale fissato. Le due somme danno lo stesso risultato perché, dato che $\mathbb{P}(T > s)$ è la somma di $\mathbb{P}(T = k)$ per $k > s$, in entrambe le versioni l'addendo $\mathbb{P}(T = k)$ compare esattamente k volte (nella prima è scritto esplicitamente e nella seconda è "nascosto" dentro ai termini $\mathbb{P}(T > 0)$, $\mathbb{P}(T > 1)$, ..., $\mathbb{P}(T > k - 1)$, che sono esattamente k).

Nel nostro caso, visto che $\mathbb{P}(T > k)$ è la probabilità che per (almeno) k volte chi ha meno euro vinca il lancio di moneta e vale quindi $\frac{1}{2^k}$, abbiamo così

$$\mathbb{E}(T) = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 1 + 1 = 2.$$

Il caso generale in cui abbiamo n euro in totale è più difficile da trattare e più adatto a essere risolto con metodi un po' diversi; tuttavia, è ancora possibile ottenere un valore preciso per $\mathbb{E}(T)$. Nel caso siate curiosi, si ha che $\mathbb{E}(T) = a(n-a)$, dove a è il vostro capitale iniziale e $n - a$ il mio (non a caso abbiamo trovato $\mathbb{E}(T) = 2$ e siamo partiti con uno e due euro). Questo significa che, se partiamo con capitali analoghi, ci aspettiamo che il gioco duri in media un numero di mosse dell'ordine del quadrato degli euro coinvolti.

Tutto ciò che abbiamo imparato in questa sezione ci servirà per stimare il tempo necessario per mescolare un mazzo di carte con la mossa top-to-random. Ma prima ancora ci aspetta un nuovo problema, simile sotto vari aspetti a quello degli euro giocati a testa o croce; dimenticatevi la partita a carte per adesso: è ora di andare a fare la spesa!

15.2 Il collezionista

> *Manuel è un avido collezionista. Ogni volta che spende 25 euro al supermercato OGrande, ottiene in regalo una statuina raffigurante uno di venti personaggi della saga di Star Wars. Ovviamente si tratta ogni volta di un personaggio casuale, e ancor più ovviamente Manuel è determinato a completare la sua collezione. Assumendo che non possa procurarsi i personaggi in altro modo, quanto dovrà spendere in media per riuscire a collezionarli tutti?*

A prima vista, questo problema non sembra aver molto a che fare con gli algoritmi per mescolare mazzi di carte. Ma abbiate pazienza, miei giovani Padawan! Risulterà che ne abbia eccome.

Anzitutto, non vi sorprenderà il fatto che per modellizzare l'impresa di Manuel si possano tirare in ballo le catene di Markov; e nemmeno, immagino, che il "tempo" della nostra catena sia naturalmente scandito a colpi di 25 euro di spesa.

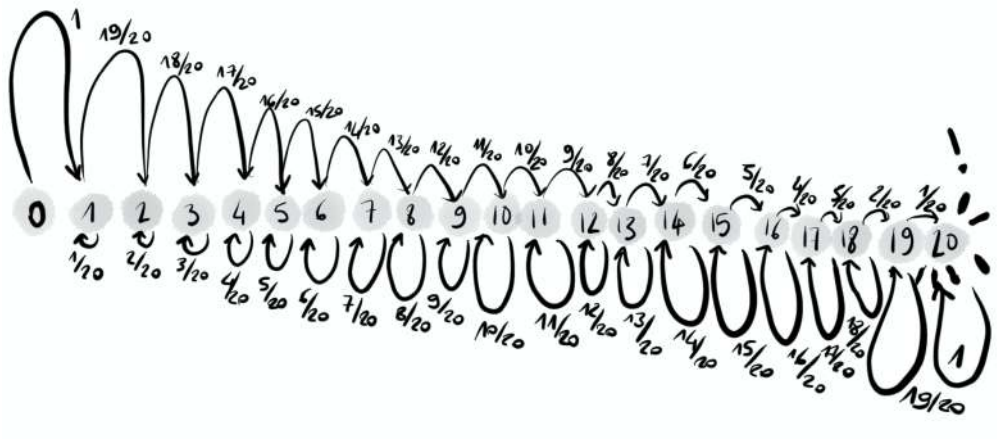


Figura 42: La catena associata all'evoluzione della collezione di Manuel.

Potremmo stabilire che gli stati della catena siano tutte le possibili collezioni di Manuel e che il valore X_t della catena al tempo t sia proprio la collezione di Manuel al momento in cui raggiunge i $25t$ euro di spesa, ovvero le t statuine; ad esempio si potrebbe avere:

$$X_0 = \emptyset, X_1 = \text{  }, X_2 = \text{  }, X_3 = \text{  }, X_4 = \text{  }, \ldots$$

Questo ci consentirebbe senz'altro di analizzare il problema della stima del tempo T tale che X_T sia la prima collezione a contenere 20 personaggi distinti, ma è un modello inutilmente ricco (e l'insieme degli stati non è nemmeno finito!).

Possiamo invece ignorare completamente ogni aspetto delle collezioni "parziali" di Manuel eccetto il numero di personaggi distinti che possiede, in modo da considerare una catena i cui stati sono gli elementi dell'insieme $\{0, 1, \ldots, 20\}$. In effetti, dallo stato i (Manuel ha i personaggi distinti) è possibile solamente rimanere allo stato i (Manuel spende 25 euro e ottiene una statuina che possiede già) o allo stato $i + 1$ (i prossimi 25 euro di spesa procurano a Manuel un nuovo personaggio).

Supponiamo che si abbia $X_t = i$; a prescindere da quante statuine abbia Manuel in totale (cioè da quanto valga t), dato che i personaggi sono equiprobabili, la probabilità che al prossimo passaggio ne ottenga uno degli i che ha già è $\frac{i}{20}$; la probabilità che ne ottenga uno nuovo è $\frac{20-i}{20}$. Possiamo quindi riassumere la situazione, come nella sezione precedente, con lo schema in Figura 42.

A prescindere dallo stato iniziale X_0 (nel nostro caso $X_0 = 0$), dopo un tempo eventualmente molto lungo che chiameremo T_{20} la catena raggiungerà lo stato 20 e in esso rimarrà per sempre. La domanda che ci poniamo è: quanto vale in media T_{20} (ovvero, se volete, quanto vale $\mathbb{E}(T_{20})$)?

Andiamo per gradi: se chiamiamo T_1 il tempo necessario per raggiungere lo stato 1, abbiamo che $T_1 = 1$. Ma che dire di T_2, cioè il tempo necessario per avere

due personaggi distinti? Per ogni k, non è difficile calcolare la probabilità che si abbia $T_2 - T_1 > k$: è la probabilità che si abbia

$$X_{T_1} = X_{T_1+1} = X_{T_1+2} = \ldots = X_{T_1+k} = 1,$$

ovvero che Manuel ottenga sempre lo stesso personaggio per k volte consecutive, cioè 20^{-k}.

Quello che stiamo facendo non è molto diverso dal calcolo della durata media del gioco della sezione precedente; applicando lo stesso trucco possiamo calcolare il valore medio $\mathbb{E}(T_2 - T_1)$ come la somma di $\mathbb{P}(T_2 - T_1 > k)$ per $k \geq 0$, cioè

$$1 + \frac{1}{20} + \frac{1}{20^2} + \frac{1}{20^3} + \ldots.$$

La somma che si otteneva alla fine della sezione precedente (la somma degli inversi delle potenze di 2) è talmente famosa che ne abbiamo scritto direttamente il valore numerico (cioè 2) senza fare commenti. In questo caso, come possiamo sommare fra di loro le potenze di 1/20? I termini della somma infinita che vorremmo calcolare sono tutti positivi e diventano rapidamente molto molto piccoli. Se credete nella forza e nel fatto che esista per la somma un valore ben definito S, sarete forse disposti a credere che, siccome possiamo riscriverla come

$$1 + \frac{1}{20}\left(1 + \frac{1}{20} + \frac{1}{20^2} + \frac{1}{20^3} + \ldots\right),$$

debba necessariamente aversi $S = 1 + \frac{1}{20}S$, e che il valore che cerchiamo sia $S = \frac{20}{19}$.

Se siete scettici... congratulazioni! Avete la stoffa del vero matematico: non vi resta che calcolare le somme parziali (cioè le approssimazioni finite ottenute sommando fino a ogni valore fissato di k) e convincervi che si tratti di approssimazioni sempre migliori di $\frac{20}{19}$ (cosa che, trattandosi di somme di successioni geometriche, non dovrebbe risultarvi difficile).

Ma a questo punto ci siamo quasi! Calcolare quanto valga in media il tempo $T_{r+1} - T_r$, cioè il tempo trascorso dalla catena nello stato r, non è davvero più difficile. Abbiamo

$$\mathbb{P}(T_{r+1} - T_r > k) = \left(\frac{r}{20}\right)^k$$

(probabilità per Manuel di ottenere uno degli r personaggi che ha già per k volte consecutive) e quindi, sostituendo $\frac{r}{20}$ a $\frac{1}{20}$ nel ragionamento di prima (provate!),

$$\mathbb{E}(T_{r+1} - T_r) = \left(\frac{r}{20}\right)^0 + \left(\frac{r}{20}\right)^1 + \left(\frac{r}{20}\right)^2 + \ldots = \frac{20}{20 - r}.$$

Possiamo ora scrivere

$$T_{20} = T_1 + (T_2 - T_1) + (T_3 - T_2) + \ldots + (T_{20} - T_{19}),$$

dove ogni addendo è un numero aleatorio che rappresenta il tempo necessario, una volta ottenuti r personaggi diversi, per procurarsene uno nuovo. E adesso... beh, il valore di T_{20} in media non sarà che la somma delle medie che abbiamo trovato! Ovvero

$$\frac{20}{19} + \frac{20}{18} + \ldots + \frac{20}{3} + \frac{20}{2} + \frac{20}{1},$$

che fa poco meno di 71. In altre parole, in media Manuel dovrà sborsare la bellezza di quasi 1775 euro al supermercato OGrande per completare la sua collezione!

15.3 Qualche approfondimento sul problema del collezionista

Il problema che abbiamo appena risolto è tipico nell'ambito dello studio della convergenza all'equilibrio delle catene di Markov e aiuta a formalizzare un'intera classe di altri problemi apparentemente diversi. È perciò utile porsi il problema del collezionista più in generale per collezioni con n tipologie distinte di oggetti e chiedersi come vari il tempo necessario per completare la collezione in funzione di n, specialmente quando n è molto grande; a studiare questo problema fra i primi furono figure classiche della matematica quali De Moivre, Eulero e Laplace, il che fa di questa domanda uno dei pilastri della teoria della probabilità.

Il medesimo ragionamento della sezione precedente comporta che, nel caso di n oggetti, il tempo necessario per completare la collezione sia in media uguale a

$$n\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n-1} + \frac{1}{n}\right).$$

La quantità fra parentesi è un numero che, dato un valore specifico di n, potremmo calcolare (si chiama l'n-esimo numero armonico); se però ci accontentiamo di un suo valore approssimato che ci consenta di valutare il tipo di crescita che ha in funzione di n, potremmo dimostrare (e chi di voi conosce i primi rudimenti di analisi matematica ne sarebbe probabilmente capace) che il suo valore non è troppo diverso da $\log n$.

In particolare possiamo dire che il tempo stimato per completare una collezione di n elementi cresce all'incirca come $n \log n$ (più velocemente di n, ma ben più lentamente di n^2); con qualche strumento matematico in più, è possibile rendere questa affermazione ancora più precisa: si può mostrare che il tempo necessario per completare la collezione, sebbene dipenda da quanto siamo fortunati, tenderà con probabilità enorme ad aggirarsi in un certo intervallo intorno a $n \log n$ quando n è molto grande: difficilmente possiamo sperare di fare di meglio, né abbiamo molto da temere.

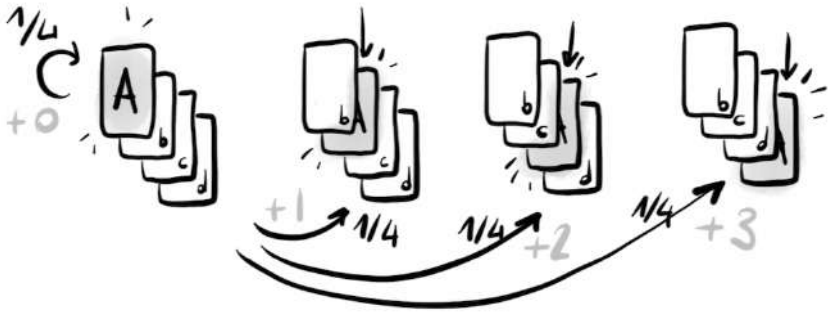


Figura 43

15.4 L'algoritmo top-to-random

È giunto finalmente il momento di occuparci di metodi per mescolare mazzi di carte e in particolare dell'algoritmo top-to-random descritto nella sezione 15.1, che consiste, dato un mazzo di n carte, nel prendere ripetutamente la prima carta e reinserirla in una posizione casuale.

In altre parole, ad ogni "mossa" scegliamo uniformemente un numero k compreso fra 1 e n (quindi ogni scelta ha la stessa probabilità $1/n$) e spostiamo la carta in cima al mazzo in avanti inserendola dopo le $k-1$ carte immediatamente successive (se scegliamo 1, lasciamo l'ordinamento del mazzo così com'è). Per vedere questo algoritmo come una catena di Markov, è sufficiente considerare come insieme degli stati l'insieme di tutti i possibili $n!$ ordinamenti del mazzo; la nostra catena partirà da un certo stato X_0 e, ad ogni iterazione dell'algoritmo, passerà da X_t a un ordinamento X_{t+1} che è identico al precedente se dal mazzo viene rimossa la carta che si trova in cima secondo l'ordinamento X_t. In Figura 43 vedete rappresentate le transizioni di questa catena da uno stato fissato nel caso $n = 4$.

Come detto in precedenza, al crescere del tempo t ci aspettiamo che il mazzo sia "sempre più mescolato", ovvero che la distribuzione data dalla catena si avvicini sempre di più a quella uniforme sull'insieme degli stati. Ma quando possiamo dirci soddisfatti e dichiarare che il mazzo "è quasi sicuramente ben mescolato"?

Una prima idea potrebbe essere quella di ripetere l'operazione top-to-random n volte, quante le carte del mazzo; ma è molto probabile che, visto che non tutte le carte vengono reinserite verso il fondo, ci troveremo a spostare alcune carte più di una volta, e quasi certamente non arriveremo a modificare l'ordine relativo delle ultime carte del mazzo entro n iterazioni. No, n proprio non va. Ma $2n$? n^2? 2^n? In quanto tempo siamo ragionevolmente sicuri di farcela?

A partire dal ragionamento fallito sulle n mescolate possiamo farci venire un'idea su come essere sicuri di aver mescolato per bene.

Supponiamo di segnarci il momento in cui per la prima volta scegliamo il numero

n, cioè mettiamo la prima carta in fondo al mazzo. Diciamo che questo avvenga al tempo t_1; continuiamo imperterriti a mescolare. Adesso segniamoci il tempo $t_2 > t_1$ in cui per la prima volta (dopo t_1) scegliamo n o $n-1$. Siete d'accordo sul fatto che, al tempo t_2, l'ordine delle ultime due carte sia casuale? Dovreste! Infatti, posto che avremo appena fatto la mossa corrispondente a n o a $n-1$, queste due mosse hanno sempre la stessa probabilità di essere effettuate; dunque la carta che al tempo $t_2 - 1$ era in cima al mazzo si trova ora subito dopo o subito prima della carta che al tempo $t_2 - 1$ era in fondo al mazzo, con uguale probabilità.

Oramai avrete capito l'antifona: definiamo t_3 come il primo momento dopo t_2 in cui effettuiamo una fra la mossa n, la mossa $n-1$ e la mossa $n-2$; al tempo $t_3 - 1$ le ultime due carte del mazzo erano "mescolate", i tre possibili punti di inserzione della carta della cima sono equiprobabili, e dunque le ultime tre carte sono in un ordine casuale al tempo t_3. Una volta definiti $t_4 < t_5 < \ldots < t_n$ allo stesso modo... voilà! Possiamo affermare con certezza che al tempo t_n il mazzo "sia mescolato" (e lo è per sempre dopo quel momento).

Ma i tempi che abbiamo definito vi dicono niente?

Visto che abbiamo definito t_n come il primo istante dopo t_{n-1} in cui venga effettuata la mossa corrispondente a n o a $n-1$ o a $n-2$... o a 2 o a 1 (cioè una mossa qualunque), abbiamo semplicemente $t_n - t_{n-1} = 1$. Per quanto riguarda $t_{n-1} - t_{n-2}$, è il tempo che è necessario attendere perché, scegliendo a caso ad ogni iterazione un numero in $\{1, \ldots, n\}$, ne otteniamo uno compreso fra 2 e n, o più semplicemente diverso da 1; allo stesso modo, la quantità $t_{n-2} - t_{n-3}$ è il tempo da attendere per pescare un numero diverso da 1 e da 2... Infine, per far concludere l'intervallo di tempo t_1 va bene soltanto "pescare" il numero n. Se consideriamo questi intervalli di tempo in ordine inverso, quindi, non stiamo facendo altro che tentare di completare una collezione di n statuine!

Abbiamo perciò, esattamente come nel caso generale del problema del collezionista,

$$\mathbb{E}(t_n) = n\left(1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}\right) \approx n \log n.$$

Si noti che, nel caso in cui $n = 52$, si ha che $\mathbb{E}(t_n)$ vale all'incirca 236, ossia siamo riusciti a mostrare che un mazzo di 52 carte è sicuramente completamente mescolato prima di un evento che richiede in media 236 mosse top-to-random per verificarsi.

Cosa? 236 mosse? Che barba!, dirà qualcuno.

Alt!, risponderanno i più attenti fra di voi, *non è mica detto che il mazzo non fosse mescolato prima.*

È in effetti possibile che il mazzo fosse già perfettamente mescolato e che non fosse indispensabile attendere il verificarsi degli eventi con i quali abbiamo definito $t_1, t_2, \ldots, t_n$. Tuttavia, per valori grandi di n si dimostra che non possiamo aspettarci di mescolare il mazzo in meno di $n \log n$ mosse all'incirca: quello che si può scoprire è che, se effettuiamo un numero sensibilmente minore di mosse

top-to-random, con grande probabilità la carta che all'inizio era l'ultima del mazzo si troverà ancora fra le ultime $n/\log n$ carte, cosa che invece in un mazzo ben mescolato accade con probabilità $\frac{1}{n} \cdot \frac{n}{\log n} = \frac{1}{\log n}$, cioè molto molto piccola.

D'altra parte, forse c'è un motivo se nessuno, di fronte a un vero mazzo di carte, si metterebbe ad applicare la mossa top-to-random...

15.5 Il riffle shuffle

A conclusione di questo articolo ci getteremo un'impresa più difficile, quella di analizzare un modello di algoritmo di mescolamento molto più realistico e quindi un po' più complicato. La dimostrazione che presentiamo non è quella piuttosto sofisticata di Bayer e Diaconis, ma una versione un po' più debole ed elementare elaborata da Diaconis e Aldous, che tuttavia ne cattura gli aspetti fondamentali.

Per mescolare un vero mazzo di carte, una procedura molto comune consiste nel dividerlo in due mazzetti e, tenendo uno dei due mazzetti nella mano sinistra e uno nella destra, ricomporre il mazzo con un gesto di scorrimento dei pollici sulle carte il cui effetto è quello di "compenetrare" un mazzetto nell'altro; le carte di un mazzetto si frappongono a quelle dell'altro in posizioni casuali, ma l'ordine di ogni coppia di carte all'interno di un singolo mazzetto si mantiene.

Modellizzeremo questo algoritmo di mescolamento, spesso chiamato "riffle shuffle", nel seguente modo. Supponiamo di voler mescolare un mazzo di n carte e di partire da un certo ordinamento. Una singola mossa del riffle shuffle consisterà nello scegliere uniformemente a caso una stringa di zeri e uni di lunghezza n; poiché le stringhe possibili sono 2^n, ciascuna avrà probabilità $\frac{1}{2^n}$. Supponiamo che la stringa abbia k zeri e $n-k$ uni; allora separiamo le prime k carte del mazzo dalle ultime $n-k$ e le riordiniamo inserendo le k carte del mazzetto "superiore" in corrispondenza degli zeri della stringa e quelle del mazzetto "inferiore" in corrispondenza degli uni della stringa, mantenendo l'ordine interno di ciascun mazzetto.

Ad esempio, supponiamo di avere $n = 8$, etichettiamo le carte con numeri interi positivi e partiamo dall'ordinamento $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8$; la mossa corrispondente alla stringa 01101000 consiste nel separare i mazzetti $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5$ (corrispondente ai 5 zeri) e $6 \cdot 7 \cdot 8$ (corrispondente ai 3 uni) e ricomporli come

$$\begin{matrix} 0 \cdot 1 \cdot 1 \cdot 0 \cdot 1 \cdot 0 \cdot 0 \cdot 0 \\ 1 \cdot {}_6 \cdot {}_7 \cdot 2 \cdot {}_8 \cdot 3 \cdot 4 \cdot 5 \end{matrix} .$$

Intuitivamente ci aspettiamo che il riffle shuffle sia un algoritmo più efficiente di quello dato dalla mossa top-to-random e che un numero abbastanza piccolo di mosse (se non altro rispetto a 236...) sia sufficiente nella realtà per essere ragionevolmente sicuri di aver raggiunto un buon livello di mescolamento. Ma *quanto* piccolo?

Per rispondere useremo alcune tecniche che sono centrali nell'area della matematica che si occupa di stimare i "tempi di mescolamento", ovvero i tempi di convergenza delle catene di Markov verso l'equilibrio.

Anzitutto, troveremo più comodo, anziché stimare in quanto tempo il mazzo risulti mescolato via mosse del riffle shuffle, stimare in quanto tempo lo si riesca a mescolare con mosse "inverse": dati un ordinamento e una stringa di zeri e uni, costruiremo l'ordinamento successivo come quello che si ottiene prendendo le carte in corrispondenza degli zeri (nell'ordine in cui si trovano) e mettendole in cima al mazzo. Dovrebbe essere chiaro che questa mossa sia l'inversa di quella descritta in precedenza, ma anche che il problema di stimare il tempo di mescolamento del riffle shuffle sia esattamente equivalente a quello di stimare il tempo di mescolamento di questo riffle shuffle "inverso", che chiameremo, ehm... *elffir shuffle* da ora in poi.

A questo punto, introduciamo l'idea chiave della nostra stima – o meglio, della stima di Diaconis e Aldous! – che è quella di costruire un cosiddetto "accoppiamento". La strategia è questa: supponete di avere il vostro mazzo da mescolare con un certo ordinamento fissato delle n carte, ma supponete anche di avere a disposizione un mazzo premescolato, il cui ordinamento è uniformemente casuale fra gli $n!$ possibili. Adesso immaginate che io vi dia una ricetta per tradurre ogni mossa dell'elffir shuffle da effettuare sul primo mazzo in una singola mossa di elffir shuffle corrispondente da effettuare sul mazzo premescolato. Purché questa ricetta sia fatta in modo che le mosse da effettuare sul mazzo premescolato siano uniformi, l'ordinamento del mazzo premescolato cambierà, ma rimarrà uniformemente casuale (dopotutto, effettuare una mossa del riffle shuffle o dell'elffir shuffle su un mazzo mescolato lo mantiene mescolato).

Ma adesso, se io sono in grado di fornirvi una ricetta tale che, dopo un certo tempo, si possa dire che il primo mazzo ha precisamente lo stesso ordinamento del secondo mazzo, a quel punto anche il primo mazzo sarà per forza mescolato!

Ed è proprio questo che intendo fare. Avete il vostro mazzo di n carte da mescolare e accanto un secondo mazzo di carte immaginario (o più immaginario del primo!) premescolato. Supponete di voler effettuare la mossa corrispondente a una certa stringa di zeri e uni sul primo mazzo; la stringa da far corrispondere per il secondo mazzo è ottenuta così: sbirciate quale sia la prima carta del secondo mazzo e andate a cercarla nel primo; se questa corrisponde a uno zero nella stringa scelta (si sposta nella parte superiore del mazzo) le facciamo corrispondere uno zero anche nella stringa che costruiamo per il secondo mazzo (la stringa comincerà per zero); viceversa, se le corrisponde un uno, scriviamo un uno anche nella stringa della mossa "immaginaria". Facciamo lo stesso per tutte le carte del secondo mazzo: associamo loro la stessa cifra – zero o uno – che è loro associata nella stringa applicata al primo mazzo, indipendentemente dal fatto che avranno posizioni diverse nei due mazzi.

Questa ricetta fa corrispondere a ciascuna stringa di zeri e uni una stringa diversa e quindi assegna probabilità $\frac{1}{2^n}$ a ogni stringa da applicare al secondo mazzo, che rimane perciò correttamente mescolato (ogni ordinamento mantiene la stessa probabilità).

Supponiamo di effettuare una serie di mosse di elffir shuffle sul primo mazzo (e le corrispondenti sul secondo). A ogni singola carta (ad esempio alla donna di picche) toccherà per ogni mossa uno zero o un uno, a seconda che questa vada a far parte della sezione superiore o inferiore nel mazzo; notate che, per come abbiamo costruito la ricetta per la corrispondenza delle stringhe, ad ogni iterazione le verrà assegnata la stessa cifra nel primo e nel secondo mazzo (e in effetti potreste sospettare che questa ricetta fosse elaborata apposta per ottenere questa proprietà...).

Per ogni carta, segnatevi la successione di zeri e uni che le viene assegnata mano a mano che iterate mosse di elffir shuffle. Naturalmente è possibile che, dopo per esempio cinque mosse, sia la donna di picche che il re di fiori abbiano avuto l'assegnazione delle cifre $0, 0, 1, 0, 1$, in sequenza; in quel caso non è possibile dire a priori se venga prima la donna o il re dopo le cinque mosse di elffir shuffle: dipende dal loro ordine nel mazzo iniziale. Ma se le sequenze di zeri e uni associate alle due carte sono diverse (ad esempio: la donna ha avuto $1, 0, 1, 0, 0$ e il re ha avuto $0, 0, 0, 1, 0$) sappiamo dire quale si trovi sopra e quale sotto dopo le cinque smazzate, a prescindere dal loro ordine iniziale; nel nostro esempio, sappiamo che dopo la penultima smazzata la donna si è trovata sopra al re (stavano in due sezioni diverse del mazzo); poiché la quinta ha posto le due carte all'interno dello stesso gruppo (quello da mettere sopra, ma questo non è importante), il loro ordine reciproco non è cambiato. Più in generale, se due carte hanno una sequenza di zeri e uni diversa, la carta che si trova più in alto è quella che ha uno zero come ultima cifra "diversa"; ma soprattutto, le due carte risulteranno necessariamente nello stesso ordine nel mazzo "reale" e nel perpetuamente mescolato mazzo "immaginario"!

Ma un momento: questo significa, per quanto detto prima, che non appena le sequenze di zeri e uni associate a ogni singola carta del mazzo sono tutte diverse, siamo sicuri che l'ordinamento del mazzo sia lo stesso del mazzo immaginario, e quindi uniformemente casuale!

Non ci resta che stimare dopo quante smazzate questo evento si verifichi, ma questo non è difficile. Ogni mossa di elffir shuffle, dato che produce una stringa casuale di zeri e uni, associa indipendentemente ad ogni carta una cifra uniformemente casuale; date due carte fissate, qual è la probabilità che, dopo t mosse, le sequenze associate siano uguali? Si tratta della probabilità che le t cifre assegnate alla seconda carta replichino perfettamente quelle della prima, cioè 2^{-t}.

D'altra parte, vi è un modo molto semplice di stimare la probabilità che al tempo t almeno una coppia di carte abbia la stessa sequenza di zeri e uni; tale probabilità non può essere maggiore della somma delle probabilità che questo

avvenga per ciascuna coppia. Dato che le coppie sono $n(n-1)/2$ (perché?[16]), stiamo dicendo che

$$\mathbb{P}\left(\begin{array}{c}\text{dopo } t \text{ iterazioni ci sono}\\ \text{due carte con la stessa sequenza}\end{array}\right) \leq n(n-1)2^{-t-1}.$$

Detto T il numero di iterazioni necessarie perché per la prima volta tutte le sequenze siano distinte si ha quindi

$$\mathbb{P}(T > t) \leq n(n-1)2^{-t-1} \leq n^2 2^{-t-1}.$$

Ma abbiamo mostrato che al tempo T il mazzo "reale" ha precisamente lo stesso ordine del premescolato mazzo "immaginario"; ovvero, qualunque mazzo sarà certamente ben mescolato dopo T iterazioni dell'elffir shuffle, o, equivalentemente, del riffle shuffle.

Possiamo ora scegliere (in funzione di n) un valore di t che ci dia una probabilità che consideriamo soddisfacente di aver mescolato il mazzo dopo t iterazioni. Ad esempio, dopo $2\log_2 n$ smazzate, abbiamo che il mazzo è ben mescolato con probabilità almeno $1 - n^2 2^{-2\log_2 n - 1}$, cioè $1/2$; dopo $2\log_2 n + 2$ smazzate, la probabilità sarà almeno $7/8$ e dopo $2\log_2 n + 6$ più del 99%. Nel caso di un mazzo di 52 carte, per esempio, abbiamo dimostrato che dopo 14 iterazioni del riffle shuffle il mazzo sarà ben mescolato con probabilità maggiore del 90%. Una bella differenza rispetto alle centinaia di iterazioni richieste dall'algoritmo top-to-random!

Riferimenti bibliografici

[1] D. Aldous, J. A. Fill. Reversible Markov Chains and Random Walks on Graphs, 2002, disponibile all'indirizzo `http://www.stat.berkeley.edu/~aldous/RWG/book.html`

[2] D. Bayer, P. Diaconis. Trailing the Dovetail Shuffle to its Lair, 1992, The Annals of Applied Probability, n. 2, p. 294–313

[16] Scegliamo la prima carta fra le n possibili, la seconda fra le $n-1$ diverse dalla prima; dividiamo per due perché la coppia (donna di picche, re di fiori) è in effetti la stessa di (re di fiori, donna di picche).

16 Problemi classici e moderni in Teoria dei Numeri

Roberto Dvornicich, n.8, Gennaio 2019

16.1 Introduzione

Lo scopo di questo lavoro è di presentare lo stato dell'arte relativamente ad alcuni problemi classici della teoria dei numeri.

È molto difficile descrivere esattamente cos'è la teoria dei numeri, perché la ricerca in questo settore si è allargata in un gran numero di filoni diversi, seppure spesso con interazioni reciproche. Credo perciò che sia più utile evidenziare alcuni di questi filoni:

1. lo studio dei *numeri primi*, delle loro proprietà e della loro *distribuzione*;
2. lo studio delle *equazioni diofantee* (cioè le equazioni per cui non si ricercano tutte le soluzioni reali, ma solo quelle con numeri interi);
3. lo studio dell'*approssimazione diofantea* (cioè della possibilità di approssimare un numero non razionale, per esempio π, mediante frazioni);
4. lo studio delle *proprietà aritmetiche di insiemi di numeri più complessi dei numeri interi*, ma che hanno caratteristiche simili, sia finiti che infiniti, e l'estensione dei problemi diofantei a questi insiemi;
5. le *applicazioni* dell'aritmetica ai *sistemi di trasmissione digitale di dati* (codici e crittografia).

Nel seguito ci occuperemo di toccare alcuni di questi argomenti nella maniera più elementare possibile. Dovendo necessariamente fare delle scelte, abbiamo scelto di trattare:

- l'ipotesi di Riemann sulla distribuzione dei numeri primi
- le congetture "famose" (primi gemelli e congettura di Goldbach)
- i numeri "famosi" (e e π)
- equazioni diofantee: l'ultimo teorema di Fermat ed altro
- test di primalità e algoritmi di fattorizzazione
- applicazioni

16.2 Numeri primi

L'ipotesi di Riemann (ma sarebbe meglio dire *congettura* di Riemann) è probabilmente il più grande problema aperto della teoria dei numeri. Questa congettura, dovuta appunto a Riemann, risale al 1859.

Per capire di cosa si tratta, facciamo un passo indietro.

Quanti sono i numeri primi? Tutti sanno che sono infiniti. Ma la domanda ha ancora un senso se viene posta in maniera più precisa. Per ogni numero reale positivo x, definiamo $\pi(x)$ il numero dei primi che sono compresi fra 1 e x.

Come varia $\pi(x)$ in funzione di x?

Per esempio, abbiamo la seguente tabella:

x	$\pi(x)$	$x/\pi(x)$
1000	168	6.0
1000000	78498	12.7
1000000000	50847534	19.7

Si può notare che:

(a) la *percentuale* dei numeri primi fra 1 e x decresce con x;

(b) approssimativamente, il rapporto $x/\pi(x)$ è una funzione lineare del numero di cifre di x.

In effetti, esiste un certo argomento euristico, basato su un modello naturale di probabilità (essenzialmente, si assume che, per un generico intero k, gli eventi *"k è divisibile per m"* e *"k è divisibile per n"*, dove m ed n sono primi fra loro, siano eventi indipendenti), che induce alla seguente

Congettura 1. *La probabilità che un numero n sia primo è $\frac{1}{\log n}$, dove il logaritmo è il logaritmo naturale, ossia fatto rispetto alla base e (costante di Nepero).*

Questo fatto è stato dimostrato, ed è il succo del cosiddetto **Teorema dei numeri primi**.

Teorema 3.

$$\pi(x) \sim \int_2^x \frac{1}{\log t} dt \sim \frac{x}{\log x}$$

Il simbolo $\sim$ (asintotico a) sta a denotare un'approssimazione, ma un'approssimazione molto precisa.

Il teorema dei numeri primi si può enunciare in una forma che si dimostra essere del tutto equivalente alla prededente "pesando" ogni numero primo p con un peso uguale a $\log p$:

Teorema 4. $\theta(x) := \sum_{p \leq x} \log p \sim x$.

L'ipotesi di Riemann riguarda la bontà di questa approssimazione.

Essa dice: se il modello probabilistico che abbiamo inventato funziona, esso dovrebbe seguire le leggi della probabilità. La probabilità (legge dei grandi numeri) dice per esempio che se facciamo una serie successiva di n lanci di monete (testa o croce) non solo ci aspettiamo che circa la metà dei lanci diano testa e metà croce, ma anche che lo scostamento rispetto a questo valore atteso sia piccolo. Lo scostamento previsto è circa $\sqrt{n}$.

Sarà vero anche nel nostro caso? È vero, cioè, che nella nostra formula approssimata l'errore che facciamo è al massimo quello che dovrebbe essere, ossia all'incirca $\sqrt{\frac{x}{\log x}}$? O, nella formulazione equivalente, $\sqrt{x}$?

La validità di questa tesi sembra molto plausibile, ed avrebbe conseguenze rilevanti per la conoscenza di un gran numero di problemi collegati ai numeri primi.

Purtroppo, a tutt'oggi, non conosciamo la verità. Non sappiamo se la congettura di Riemann sia vera o falsa, anche se ci sono vari indizi a favore. Il primo indizio è di carattere "filosofico": un modello probabilistico che si discosta da quello "naturale" dovrebbe essere motivato da fenomeni distorsivi speciali, che nessuno ha mai riscontrato.

Ma ci sono indizi più convincenti dal punto di vista numerico.

Ci sono vari modi equivalenti di formulare la congettura di Riemann: il più famoso di essi è che un insieme infinito di punti del piano (gli zeri non banali della funzione zeta di Riemann) sia in realtà costituito da punti che giacciono tutti *su una medesima retta verticale*.

Allora si può verificare se almeno qualcuno di questi punti giace effettivamente su questa retta. Si sono fatti dei calcoli:

I primi 24.000.000.000.000 punti controllati giacciono effettivamente sulla retta!

Per molti osservatori esterni questa evidenza numerica costituisce una prova oggettiva, ma purtroppo non è così per i matematici. Un'eccezione è sempre possibile, e non sarebbe il primo caso nella storia quello in cui si è trovata un'eccezione non prevista anche quando tutti credevano che ormai non fosse più ragionevolmente possibile pensare che ci fosse.

I numerosissimi tentativi di dimostrazione della congettura hanno dato solo risultati parziali, come per esempio quello di garantire che almeno il 40% dei punti giace sulla retta (specificando cosa vuol dire il 40% di *infiniti* punti).

Se vi volete cimentare a dimostrarla voi, per dir così, in casa, ecco una formulazione alternativa semplice ma assolutamente esatta dell'ipotesi di Riemann, che tutti possono capire. Consideriamo il minimo comune multiplo di tutti di numeri

fra 1 ed n e chiamiamolo $M(n)$. Si sa, dal teorema dei numeri primi, che

$$M(n) \sim e^n$$

ossia che

$$\log M(n) \sim n.$$

È vero che l'approssimazione scritta sopra rispetta le leggi normali della probabilità, ossia che la differenza fra $\log M(n)$ *ed* n *è al massimo (all'incirca)* $\sqrt{n}$*?*

16.3 I primi gemelli

Il problema dei primi gemelli è quello di stabilire se esistano infinite coppie di numeri primi della forma $(n, n+2)$. Per esempio, le coppie

$$(3,5), (5,7), (11,13), (17,19), (29,31), (41,43)$$

sono coppie di primi gemelli.

Purtoppo, anche questo è un problema aperto. È interessante comunque notare che, se il modello probabilistico dei numeri naturali delineato prima fosse adeguato, allora si potrebbe non solo dimostrare che esistono infinite coppie di primi gemelli, ma anche dire "quante" sono.

Denotiamo con $\pi_2(x)$ il numero di coppie di primi gemelli $(n, n+2)$ fatte con numeri minori o uguali a x. Allora si ha la formula euristica (congetturale)

$$\pi_2(x) = 1,320326... \times \frac{x}{\log^2 x}.$$

Anche qui si sono fatti alcuni calcoli; l'ultimo risultato disponibile riguarda un'analisi di tutti i numeri che hanno fino a 15 cifre decimali. L'errore percentuale che dà la formula congetturale è inferiore a un milionesimo!!!

16.4 La congettura di Goldbach

La congettura di Goldbach dice che

Congettura 2. *Ogni numero pari maggiore o uguale a 4 si può scrivere come la somma di due numeri primi.*

Esiste una congettura analoga per i numeri dispari:

Congettura 3. *Ogni numero dispari maggiore o uguale a 7 si può scrivere come somma di tre numeri primi.*

L'attenzione attuale, specialmente da parte dei dilettanti della matematica, è rivolta verso il problema che riguarda i numeri *pari*. Infatti il problema relativo ai numeri dispari è stato recentemente risolto in maniera definitiva (2013).

IL CASO DISPARI: $2n + 1 = p_1 + p_2 + p_3$.

Analizziamo però come mai la soluzione definitiva sia solo così recente. Nel 1937 il matematico russo Vinogradov ha dimostrato il seguente

Teorema 5. *Tutti i numeri dispari "abbastanza grandi" si possono scrivere come somma di tre numeri primi.*

Cosa si intende con "abbastanza grandi"? Analizzando la dimostrazione di Vinogradov, si vede che essa funziona per tutti i numeri maggiori o uguali di una costante incredibilmente grande, $3^{3^{15}}$ (un numero con quasi 7 milioni di cifre decimali). Si potrebbe pensare che, utilizzando il computer, si possano trattare tutti i numeri minori di $3^{3^{15}}$ e quindi arrivare ad un teorema valido per tutti i dispari maggiori o guali a 7.

Sembra facile, ma non lo è. Vediamo perché.

Supponiamo di voler verificare tramite un computer che tutti i numeri dispari che la dimostrazione lascia in sospeso si possono esprimere come somma di tre numeri primi. Se $x = p_1 + p_2 + p_3$ allora è chiaro che almeno uno degli addendi deve essere maggiore o uguale di $x/3$. Questo significa che dobbiamo avere a disposizione una tavola di numeri primi che arrivi almeno fino a $1/3 \times 3^{3^{15}}$. Ammesso che abbiamo a disposizione i mezzi teorici per farlo, c'è un problema: *gli atomi dell'universo sono "solo"* 10^{80}!

Fortunatamente, dopo Vinogradov altri matematici hanno via via migliorato la costante di riferimento, fino ad abbassarla (Helfgott, 2013) ad un livello accettabile per i nostri computer.

IL CASO PARI: $2n = p_1 + p_2$?

Risolto il caso dispari, è su questo caso che si concentra l'attenzione di molti appassionati. Infatti per questo caso non esiste un teorema analogo a quello del caso dispari. I fatti definitivamente dimostrati hanno una validità minore. Ecco due esempi. Il primo riguarda una variazione del problema:

Teorema 6. *Ogni numero pari "abbastanza grande" si può scrivere come somma di due numeri dei quali uno è sicuramente primo e l'altro o è primo oppure è il prodotto di due numeri primi.*

Anche qui "abbastanza grande" è *troppo* grande per poter verificare tutti i casi esclusi dal teorema.

Il secondo esempio riguarda il numero di possibili eccezioni alla validità della congettura. Definiamo

$$E(x) := \#\{n \leq x \,|\, 2n \neq p_1 + p_2\}.$$

Teorema 7. *Per ogni $\varepsilon > 0$ esiste una costante $C = C(\varepsilon)$ tale che*

$$E(x) \leq C x^{\frac{1}{2}+\varepsilon}.$$

Da questo teorema di deduce che "quasi tutti" i numeri pari si possono scrivere come somma di due numeri primi.

16.5 I numeri famosi

I numeri famosi e e π (ma il discorso vale per molti altri numeri di uso quotidiano in matematica) non si possono scrivere con esattezza usando il sistema decimale, perché si avrebbe bisogno di infinite cifre. Non si può nemmeno specificare una regola che permetta di calcolare tutte le cifre, come per esempio per

$$\frac{1}{11} = 0,090909090909...$$

perché tale regola non esiste.

Il motivo risiede nel fatto che essi sono definiti con processi di *limite* e non semplicemente tramite le usuali quattro operazioni.

La domanda è: si possono definire esattamente questi numeri usando strumenti puramente *algebrici* (come le quattro operazioni, i radicali, eccetera) ma senza strumenti analitici?

Per specificare il problema abbiamo bisogno della seguente definizione:

Definizione. *Un numero (reale o complesso) si dice* algebrico *se è radice di un polinomio non nullo a coefficienti interi.*

Per esempio, $\sqrt{3}$ è radice del polinomio $x^2 - 3$ e un numero α che soddisfi la relazione $\alpha^5 - \alpha - 1 = 0$ è algebrico (anche se non si riesce a scrivere tramite radicali).

Il problema quindi diventa:

Problema. *I numeri e e π sono algebrici?*

La risposta è NO per entrambi i casi.

Il risultato non è inatteso, nel senso che si può verificare che, preso un numero "a caso", è quasi certo (la probabilità è uguale a 1) che la risposta sia no per questo numero.

Tuttavia, come è facilmente intuibile, è assai complicato escludere che un certo numero possa essere radice di uno qualsiasi fra gli infiniti polinomi a coefficienti interi.

La soluzione del problema relativo ad e è datata 1873 (Hermite), quella relativa a π è datata 1882 (Lindemann). Quest'ultima ha conseguenze su un problema posto già dagli antichi greci, il problema della *quadratura del cerchio*.

La quadratura del cerchio

Dato un cerchio di raggio 1, è possibile costruire con riga e compasso un quadrato la cui area sia uguale a quella del cerchio dato (e cioè π) ?

Si può dimostrare abbastanza facilmente che, in un sistema di riferimento cartesiano, le coordinate di tutti i punti che si riescono a costruire con riga e compasso *sono* soluzioni di un'equazione $f(x) = 0$, dove $f(x)$ è un polinomio a coefficienti interi.

Se si potesse quadrare il cerchio, si potrebbe costruire un quadrato di lato $\sqrt{\pi}$. Ma né π né la sua radice quadrata (questa è una conseguenza relativamente semplice) sono soluzioni di alcuna equazione di questo tipo.

I problemi aperti, tuttavia, sono sempre più di quelli risolti. Come detto, il problema relativo ad e e π è stato risolto, ma, in pratica, solo per loro e per *pochissimi* altri numeri. Per esempio, non si sa risolvere il problema nemmeno per le combinazioni più semplici che si possono fare con questi due numeri, quali $e + \pi$, $e \cdot \pi$, etc.

16.6 Le equazioni diofantee

16.6.1 L'equazione di Fermat

L'equazione diofantea più conosciuta è quella di Fermat:

$$x^n + y^n = z^n.$$

Fermat affermava che, se $n \geq 3$, questa equazione non ha alcuna soluzione con numeri interi ad eccezione di quelle "banali", ossia quelle in cui una delle variabili è uguale a zero (per esempio, $0^5 + 3^5 = 3^5$). La storia di questa equazione è molto lunga, e molto si è speculato sul fatto che Fermat avesse in mente una soluzione del problema da lui stesso posto.

È probabile (ma certamente non è sicuro) che Fermat NON avesse una soluzione. Sta di fatto che il problema è stato risolto solo 350 anni dopo la sua proposizione (Wiles, 1995).

Innanzitutto: qual è l'interesse di sapere se un'equazione come quella di Fermat ha soluzioni, ed eventualmente di conoscere quali?

A questa domanda si potrebbe tranquillamente rispondere: nessuno. Come la stessa cosa si può dire di moltissimi, per non dire quasi tutti, i problemi di matematica. Storicamente, i problemi di matematica sono stati studiati in quanto interessanti *per se stessi*, indipendentemente dalle loro applicazioni pratiche. È altresì vero che molti risultati della matematica *hanno poi avuto* applicazioni pratiche, ma molto spesso applicazioni che non rientravano nell'obiettivo di coloro che vi hanno contribuito, e che non erano nemmeno nella loro immaginazione.

Nel caso del cosiddetto Ultimo Teorema di Fermat (la ricerca delle soluzioni dell'equazione diofantea $x^n + y^n = z^n$) l'interesse puramente speculativo del problema è quello che ha mosso migliaia e migliaia di matematici, professionisti o dilettanti, a dedicarcisi. Col senno di poi si può dire che questo ha contribuito a enormi sviluppi del pensiero matematico, alcuni dei quali hanno avuto *anche* ricadute dal punto di vista delle applicazioni.

Come noto, la dimostrazione di Wiles dell'Ultimo Teorema di Fermat è estremamente lunga e tecnica, e non si può raccontare se non ad un pubblico molto esperto. Perciò ci limitiamo a pochissimi cenni.

Innanzitutto, si tratta di una dimostrazione *per assurdo*.

In secondo luogo, essa usa dei risultati profondi di *geometria*. Che cosa ha a che fare la geometria con un problema puramente aritmetico come questo?

Già negli anni '50 il matematico Frey aveva avuto l'idea di legare l'equazione di Fermat all'equazione di una *c*urva. Una curva, nel piano, si può descrivere tramite un'equazione in due variabili: per esempio, l'equazione $x^2 + y^2 = 1$ descrive i punti di una *circonferenza* (di centro l'origine e di raggio 1).

Frey argomentava così: supponiamo, per assurdo, che l'equazione di Fermat abbia una soluzione (non banale), e che a, b, c siano tre numeri positivi tali che $a^p + b^p = c^p$ (qui l'esponente p è un numero primo diverso da 2, ma si può facilmente vedere che questo è il caso cruciale).

Consideriamo l'equazione

$$y^2 = (x - a^p)(x - b^p)(x + c^p).$$

Le soluzioni di questa equazione formano appunto una curva *a*lgebrica, di un genere speciale: una *curva ellittica*.

I geometri classificano le curve algebriche secondo il loro "grado di complessità", un invariante chiamato "genere". Le coniche, che sono le curve più semplici, hanno

genere 0; le curve ellittiche, che rappresentano il livello successivo di difficoltà, hanno genere 1.

Sulle curve ellittiche si sa moltissimo: in particolare, si sa quando due diverse equazioni definiscono curve ellittiche dello stesso *tipo* (cioè sono isomorfe), e si sanno *classificare* tutti i tipi possibili di curve ellittiche.

Sapendo i coefficienti dell'equazione che descrive la curva ellittica, se ne può dedurre il "tipo" (cioè la classe di isomorfismo).

La dimostrazione consiste, essenzialmente, nel far vedere che, se effettivamente si potessero trovare a, b, c come sopra e quindi si costruisse la curva ellittica relativa, questa curva *non potrebbe rientrare in nessuno dei tipi possibili*.

L'interazione fra geometria ed aritmetica, sviluppata enormemente a partire dalla seconda metà del secolo scorso, è uno dei grossi risultati che si sono avuti anche per merito dello studio dell'Ultimo Teorema di Fermat. In particolare, oggi le equazioni diofantee non si studiano più *una alla volta*, ma si raggruppano in famiglie che descrivono insiemi geometrici dello stesso tipo. Per esempio, Faltings ha dimostrato nel 1983 il seguente teorema:

Teorema 8. *Sia $f(x, y) = 0$ l'equazione di una curva di genere > 1. Allora esistono solo un numero finito di soluzioni dell'equazione $f(x, y) = 0$ con x, y numeri razionali.*

16.6.2 L'equazione di Catalan

Un altro spettacolare risultato recente consiste nella soluzione dell'equazione di Catalan. Catalan (1844) considerava i quadrati

$$1, 4, 9, 16, 25, 36, 49, \ldots,$$

i cubi

$$1, 8, 27, 64, 125, \ldots,$$

le quarte potenze

$$1, 16, 81, 256, \ldots$$

e così via, per riunirli in un'unica successione:

$$1, 4, 8, 9, 16, 25, 27, 36, 49, 64, \ldots$$

Catalan notava che in questa successione ci sono due numeri consecutivi, e cioè 8 e 9.

Ci sono altre coppie di numeri consecutivi in questa successione? Catalan pensava di no. Ed effettivamente Mihăilescu, nel 2001, ha dimostrato che Catalan aveva ragione:

Teorema 9. *Se consideriamo tutti i numeri della forma a^b, dove b è un esponente maggiore o uguale a 2, l'unica coppia di numeri consecutivi è costituita da 8 e 9.*

La spettacolarità della dimostrazione consiste nel fatto che invece, questa volta, si tratta di una dimostrazione *puramente aritmetica*, ed in fondo basata su idee dovute a Kummer intorno alla metà del secolo diciannovesimo (lo stesso Kummer aveva fatto i primi importanti progressi nello studio dell'ultimo teorema di Fermat).

Tuttavia, l'aritmetica dei numeri interi *non basta*: bisogna costruire un'aritmetica su strutture più complesse (i cosiddetti *campi ciclotomici*) ed è lì che si può risolvere il problema.

16.6.3 Risolveremo tutte le equazioni?

Le soluzioni di problemi così antichi in tempi recenti possono far pensare che siamo vicini a risolvere il problema di tutte le equazioni diofantee.

Non è così. Un problema che Hilbert, nel congresso mondiale dei matematici del 1900, aveva posto in una lista di problemi per il ventesimo secolo era il seguente:

Problema n.10 di Hilbert. È possibile trovare un algoritmo che determini se una data equazione diofantea in n incognite abbia soluzione?

Matijaševic, nel 1970, ha risposto di NO. Non esiste, né potrà mai esistere, un modo per risolvere *tutte* le equazioni diofantee.

La dimostrazione di Matijaševic si inquadra nell'ambito della *logica matematica*.

Nel 1936 K. Gödel aveva dimostrato che, nell'usuale sistema di assiomi della matematica, ma anche con qualsiasi altro sistema che si potesse inventare, la matematica ha dei limiti: ci sono degli enunciati di cui non potremo mai dimostrare né che sono veri, né che sono falsi. Si tratta degli enunciati che Gödel ha chiamato *indecidibili*.

Matijaševic ha fatto vedere che esistono dei particolari tipi di equazioni diofantee per cui la questione se abbiano o meno soluzioni è indecidibile.

16.7 Primalità e fattorizzazione

Uno dei problemi basilari della teoria dei numeri consiste nel determinare se un certo numero n è un numero primo; nel caso in cui non lo sia, di determinare la sua scomposizione in fattori primi. Come accenneremo alla fine, questo problema assolutamente teorico e astratto ha incredibili conseguenze pratiche.

È chiaro che, se il numero dato n è relativamente piccolo, chiunque, o a mano o con l'aiuto di un calcolatore, può rispondere alla domanda. Il problema quindi si pone in termini di *complessità*: dato un numero n di k cifre, quante sono le operazioni necessarie per dare una risposta?

In pratica, *quanto tempo* ci vuole?

Gli algoritmi che riguardano i numeri interi vengono classificati in classi che corrispondono a diversi gradi di complessità (tempo necessario per la loro esecuzione).

La classe **P** è la classe dei problemi per i quali il numero di passi necessario per eseguire l'algoritmo è *polinomiale* rispetto al numero di cifre dei numeri interi che si esaminano.

Nel nostro caso, considerando un numero n con k cifre, un algoritmo che decida se n è primo oppure no si dice polinomiale se si può effettuare con un numero di passi non superiore a una *potenza* di k, per esempio k^2 oppure anche k^{100}.

Esaminiamo l'algoritmo più naturale per decidere se un numero n è primo oppure no:

dividiamo n per 2, per 3, per 4, per 5, e così via: se ad un certo punto troveremo una divisione esatta (con resto zero), allora il numero non sarà primo (ed avremo trovato un fattore di n); se invece tutte le divisioni per numeri minori di n (ma in realtà basta fermarsi alla radice quadrata di n) danno resto diverso da zero, allora il numero è primo.

Quindi l'esecuzione dell'algoritmo, almeno nel caso in cui n sia un numero primo, richiede di fare circa $\sqrt{n}$ divisioni. Se n ha k cifre, diciamo che n è dell'ordine di grandezza di 10^k, ci vorranno quindi circa $10^{k/2}$ divisioni. Per k grande, questo numero è molto superiore a una potenza (qualsiasi) di k.

Se ne deduce che l'algoritmo naturale non è polinomiale, ma *esponenziale*.

Sono stati studiati vari altri algoritmi che "accorciano" il tempo di esecuzione: alcuni *deterministici*, ossia che danno la risposta con assoluta certezza, altri *probabilistici*, ossia che hanno una altissima probabilità di dare la risposta esatta. Questi ultimi algoritmi sono ovviamente più veloci dei primi, ma bisogna accontentarsi di un grado, se pur minimo, di incertezza.

Tutti i tipi di algoritmi deterministici conosciuti fino a pochissimo tempo fa erano di tipo esponenziale (in realtà, appena migliore); una combinazione ingegnosa dei tipi deterministico e probabilistico porta a degli algoritmi che, nella grande maggioranza dei casi, si possono eseguire in tempo polinomiale, ma che lasciano un numero di eccezioni per le quali è necessario un tempo esponenziale.

Tra la sorpresa generale dei matematici, tre indiani, Agrawal, Kayal e Saxena (in seguito AKS), hanno trovato nel 2002 un algoritmo deterministico per stabilire se un numero è primo oppure no che funziona in tempo polinomiale.

Di questi matematici solo il primo aveva una certa notorietà internazionale, ma forse più per i suoi studi informatici che per quelli matematici; gli altri due sono suoi giovanissimi allievi.

Ma il vero motivo di sorpresa è un altro: l'idea che sta alla base della formulazione dell'algoritmo è così semplice che sarebbe potuta venire in mente a un qualsiasi studente del primo biennio di matematica.

Invece, nel corso di secoli, non era venuta in mente a nessuno!

L'IDEA DELL'ALGORITMO AKS

Si parte da due fatti legati fra loro e noti da secoli. Primo fatto:

Teorema 10. *Se p è un numero primo, allora vale la congruenza*

$$(x+y)^p \equiv x^p + y^p \bmod p.$$

Inoltre, la questa congruenza non vale se al posto di p si prende un numero non primo.

Ricordiamo che due numeri si dicono *congrui* modulo p se divisi per p danno lo stesso resto. La congruenza enunciata sopra dice che, *se p è un numero primo*, il polinomio $(x+y)^p$ ha un termine uguale a x^p, un termine uguale a y^p e tutti gli altri suoi termini hanno coefficienti divisibili per p.

Secondo fatto (piccolo teorema di Fermat):

Teorema 11. *Se p è un numero primo, allora per ogni intero m vale la congruenza*

$$m^p \equiv m \bmod p.$$

Da questi due fatti elementari AKS deducono il loro teorema, che è un semplice esercizio per un normale studente:

Teorema 12. ***(AKS)*** *Siano n ed a due numeri interi senza fattori comuni. Allora vale la congruenza fra polinomi*

$$(x+a)^n \equiv x^n + a \bmod n$$

SE E SOLO SE n è un numero primo.

La difficoltà di ottenere questo teorema, come detto, non consiste affatto nella sua dimostrazione, ma nella sua *invenzione*: bisogna infatti *immaginare* l'enunciato e le sue possibili applicazioni.

Dal teorema AKS è abbastanza chiaro quello che bisogna fare: dato n, provare a vedere se la cosa è vera, per esempio, per $a = 1$. Detto così, questo richiede ancora un tempo troppo elevato, perché bisognerebbe calcolare tutti i coefficienti del polinomio $(x+a)^n$.

Ma, contando su idee presenti in algoritmi precedentemente sviluppati, si vede che in realtà non occorre considerare i coefficienti uno per uno, ma solo un numero assai più limitato di combinazioni fra di loro, e provare a vedere che cosa succede di queste combinazioni se le si divide per numeri piccoli.

Questo porta ad un algoritmo polinomiale.

Nonostante il risultato teorico sia straordinario, l'algoritmo AKS non viene ancora usato nella pratica. Come mai?

Il fatto è che, per testare se un numero con k cifre è primo oppure no, ci vuole un numero di passi che è circa $C \cdot k^{7,5}$, dove C è una costante molto grande. Anche se il numero di passi necessario per gli altri algoritmi è dato da una formula che è sicuramente peggiore per k molto grande, questa dà un risultato migliore, per via della costante C, quando k è relativamente piccolo (un punto di riferimento attuale è $k = 200$).

16.7.1 Gli algoritmi di fattorizzazione

Quando si usa un test di primalità del tipo di AKS, e si ottiene la risposta "n non è un numero primo", non si individua necessariamente la fattorizzazione di n. Si sa solo che n non soddisfa le proprietà che sono proprie dei numeri primi.

Per avere un algoritmo di fattorizzazione bisogna fare un passo in più. Quello che in realtà serve, se si scopre che un numero n non è primo, è di individuare un suo divisore proprio (cioè diverso da 1 e da n). Infatti, se a è un divisore proprio di n, e dunque $n = ab$ per qualche intero b, si può ripetere l'algoritmo per i numeri a e b al fine di scoprire se essi sono primi o hanno dei divisori propri. Ripetendo questo ragionamento, con numeri via via più piccoli, si riesce a determinare la scomposizione di n in fattori primi.

Gli algoritmi di fattorizzazione oggi disponibili sono sicuramente molto più efficienti dell'algoritmo "naturale" descritto precedentemente. È forse interessante notare che, tra gli algoritmi più efficienti conosciuti, uno fa un uso sistematico delle curve ellittiche, che abbiamo già incontrato nella discussione a proposito dell'ultimo teorema di Fermat.

Tuttavia, se si eccettuano gli algoritmi *ad hoc* che funzionano solo per numeri di una forma molto speciale, la complessità di tutti gli algoritmi noti è sempre *subesponenziale*, del tipo C^{k^α}, dov C e α sono costanti, con $\alpha < 1$.

16.7.2 Le applicazioni

Anche se non si può avere una prova sicura che qualcuno non scopra, prima o poi, un algoritmo di fattorizzazione di complessità polinomiale, al giorno d'oggi la fattorizzazione di un numero rimane uno dei problemi più complessi (nel senso di "*time-consuming*"), ed è su questa convinzione che si basa una delle applicazioni della teoria dei numeri più diffusa, la crittografia.

La crittografia si occupa di trovare dei metodi efficienti per trasmettere dei messaggi, o comunque delle informazioni, in modo codificato, in modo tale che una persona che sia in possesso di uno strumento (chiave di lettura) per decodificare

le informazioni le possa decifrare, ma una persona che non conosca la chiave di lettura no.

L'uso della crittografia è storicamente provato fin dal tempo degli antichi romani, per scopi militari.

Oggi se ne fanno diversi usi: insieme a quello militare e di spionaggio, quelli preponderanti sono per le transazioni di carattere economico, per garantire la privacy, per un controllo di sicurezza dell'identità degli individui ammessi a certi servizi.

Una carta bancomat, un acquisto on-line con una carta di credito, l'uso della password nell'aprire un computer, per scaricare files, per leggere la posta elettronica o per entrare in alcuni siti internet sono esempi quotidiani dell'uso della crittografia.

Un sistema crittografico *efficiente* deve rispondere ai seguenti requisiti:

1. rendere *facile* l'uso del sistema da parte dei suoi utenti autorizzati; in particolare, per il mittente di un messaggio deve essere facile codificarlo, per il ricevente deve essere facile decodificarlo;

2. rendere *estremamente difficile*, per non dire impossibile, decodificare dei messaggi se non si conosce la chiave di interpretazione.

Il primo obiettivo si risolve facilmente trasformando le parole in numeri di formato limitato, usando le congruenze, ed usando le normali operazioni su di esse.

Per raggiungere il secondo obiettivo, pure con molte varianti, la scelta è quella di usare la difficoltà della fattorizzazione dei numeri interi. In pratica, seppure con molte varianti, sia chi codifica che chi decodifica (conoscendo la chiave) deve fare delle semplici operazioni di moltiplicazione. Ma per decodificare è necessaria una chiave, che può essere scoperta solo se si riescono a fattorizzare numeri molto grandi.

A questo criterio è ispirato il primo sistema crittografico a chiave pubblica (detto RSA dagli inventori Rivest, Shamir e Adleman, 1978), che ha ispirato un gran numero di varianti che sono state usate in maniera massiccia negli ultimi quarant'anni.

Ultimamente spuntano all'orizzonte degli algoritmi quantistici, ma questo è un campo che esula da questa trattazione elementare.

Riferimenti bibliografici

[1] L. Childs, *Algebra, un'introduzione concreta*, ETS Editrice.

[2] H. Davenport, *Aritmetica superiore: un'introduzione alla teoria dei numeri*, Zanichelli.

[3] R.L. Graham, D.E. Knuth, O. Patashnik, *Matematica discreta (Principi matematici per l'informatica)*, Hoepli.

[4] G.H. Hardy and E.M. Wright, *An introduction to the theory of numbers*, Oxford University Press.

[5] N. Koblitz, *A course in number theory and cryptography*, Springer Verlag.

[6] P. Ribemboim, *The new book of prime number records*, Springer Verlag.

17 Una chiacchierata con Alessio

Giornalino n.9, Settembre 2019

Il matematico Alessio Figalli, uno dei vincitori della prestigiosa Medaglia Fields nel 2018, risponde alle domande dei ragazzi della Settimana Matematica, svoltasi nel nostro Dipartimento dal 16 al 18 gennaio 2019.

Penso che questa sia una bellissima iniziativa: cerco sempre di partecipare agli incontri con ragazzi come voi e di contribuire a occasioni formative come questa, che vi sta dando la possibilità di scoprire cosa sia la Matematica... il che è un privilegio! Molti non sanno cosa sia: c'è un po' quest'idea del matematico che serve solo a dividere il conto quando andate a mangiare la pizza. Ma in questi giorni scoprirete che la Matematica è molto di più: può essere profonda, creativa, può essere una sfida... e anche un lavoro, cosa di cui non avevo idea quando ero uno studente come voi.

Quando a 16-17 anni facevo tranquillamente il mio liceo classico l'obiet-tivo era quello di "portare a casa il voto", finché non ho scoperto le Olimpiadi della Matematica. E' stato allora che ho cominciato a vedere la Matematica con occhi diversi: non si trattava più semplicemente di imparare regole e applicarle, ma di un processo più complicato in cui sei chiamato a mettere qualcosa di tuo, a volte combinando teoremi "noti" in maniera non ovvia, non standard, a volte proprio inventandoti una soluzione dal nulla.

Ecco, alla fine fare Matematica è una versione più evoluta proprio di questo: è questo che facciamo io e i miei colleghi... con la differenza che purtroppo, mentre davanti a un esercizio che vi viene dato sapete se non altro che una soluzione c'è, nella ricerca questo non è detto: a volte non sai proprio dove andrai a parare!

Ma torniamo ai tempi del liceo: partecipando alle Olimpiadi ho incontrato tanti ragazzi appassionati ed è grazie a loro che ho considerato l'idea di provare a entrare in Normale. Non ci avevo mai pensato prima: io sono romano, a Roma ci sono ben tre università... non è che ti

venga in mente di cambiare città! In ogni caso non credevo di avere molte speranze: mi sono messo a studiare cercando di risolvere i problemi delle ammissioni passate, quelli di Matematica e specialmente quelli di Fisica; e all'inizio la realtà è che li guardi e sul momento ti dici "beh, non li farò mai!". Ma davanti a un problema difficile, che non riesci a fare, ci sono due modi possibili di reagire: o ti dici "non lo so fare, vuol dire che non sono in grado" e abbandoni la sfida, oppure pensi "adesso non lo so fare; vediamo se lo saprò fare fra un mese" e ti dai una possibilità.

Ed è così che è stato per me: leggi la soluzione, piano piano arrivi a capirla, e anche se non riesci a fare il secondo, il terzo, il quarto esercizio... magari al quinto azzecchi la prima parte del ragionamento, e a forza di lavorare e imparare prima o poi risolvi un esercizio per intero. Già dalle Olimpiadi avevo imparato che servono impegno e un po' di sacrificio; non basta il talento, la preparazione è fondamentale. Questo è vero anche nello studio universitario, nella ricerca... è importante accettare che ci sono esercizi difficili e prenderli come una sfida. Racconto spesso che per ogni articolo con cui ho risolto un problema aperto ci sono dieci, quindici problemi che non riesco a risolvere e articoli non conclusi che rimangono nel cassetto.

Ai miei tempi gli scritti di ammissione in Normale si svolgevano proprio in queste aule; dopo lo scritto di Matematica ero sconfortato: ero certissimo che fosse stato un disastro, ma dopo una notte parzialmente insonne mi ripresentai per lo scritto di Fisica, e fatto quello me ne tornai a Roma sconsolato. Tornai a Pisa dieci giorni dopo per tentare l'ammissione a Ingegneria al Sant'Anna; fu in quell'occasione che scoprii il mio nome fra quelli degli ammessi all'orale, affissi nell'ingresso della Scuola Normale. Fu una rivelazione: a me non era così chiaro cosa volessi fare – Matematica, Ingegneria... – ma il momento di gioia che provai nello scoprire di essere passato agli orali fu così grande che mi sorpresi a pensare che avrei fatto Matematica in ogni caso, anche se non fossi stato ammesso.

Ma fui ammesso. Seguirono quattro anni della mia vita passati fra questi banchi, giorno dopo giorno; posso dire in tutta sincerità di essere stato molto fortunato: ho avuto ottimi docenti qui, l'Università di Pisa dà una formazione fantastica. Io sono normalista, ma non va dimenticato che i normalisti seguono tutti i corsi all'Università di Pisa, con l'eccezione di alcuni corsi aggiuntivi tenuti all'interno della Scuola.

Durante i miei studi scoprii l'esistenza di programmi di scambio. Alcuni di voi conosceranno l'Erasmus, che consente di trascorrere un periodo all'estero durante gli studi universitari; la Normale ha accordi particolari con le Ecoles Normales francesi, di cui mi avvalsi per trascorrere un periodo di sei mesi a Lione. Fu un'esperienza importante, l'occasione di conoscere un secondo mondo: consiglio a chiunque di passare un periodo all'estero!

Finiti i sei mesi della mia borsa di scambio, tornai a Pisa a laurearmi: era il momento di decidere cosa fare della mia vita. Oggi una laurea in Matematica consente a chi lo voglia di entrare direttamente nel mondo del lavoro con una vasta gamma di scelte, dall'informatica alla finanza... Io mi sentivo a mio agio in ambito accademico e decisi di tentare il concorso di ammissione al dottorato.

Con il dottorato si inizia a fare ricerca, si entra in un mondo diverso dallo studio. Non è solo questione di potenzialità: nel mondo della ricerca bisogna saper affrontare la frustrazione. La ricerca di dottorato comunque può essere molto soddisfacente, è un'esperienza che consiglio a chiunque sia appassionato. Nel mio caso il dottorato andò molto bene, dopo un iniziale periodo duro.

Poi ho lavorato in Francia (grazie all'esperienza all'Ecole Normale), in America e infine in Svizzera, a Zurigo, dove vivo da quasi tre anni.

Ecco, questa è un po' la mia storia; ma al di là di questa piccola introduzione vorrei lasciare spazio alle vostre domande: non siate timidi!

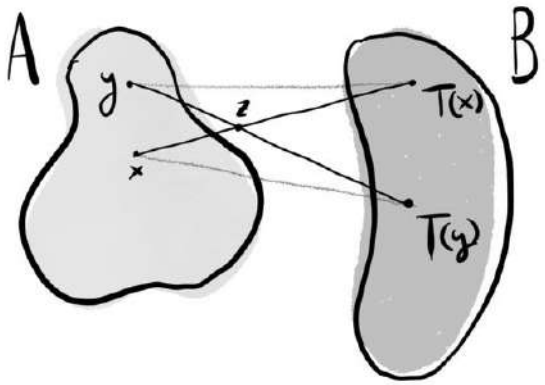


Figura 44: La disuguaglianza triangolare mostra come i raggi di trasporto non possano intersecarsi senza contraddire l'ottimalità.

Quali sono i tuoi interessi di ricerca?

La Matematica è molto varia: si può provare a dividerla in macroaree, dalla geometria all'algebra all'analisi alla probabilità, con tante sfaccettature all'interno. Dopo vari dubbi da studente ho deciso di orientarmi verso l'analisi; un problema che mi affascinò molto durante la mia laurea magistrale e che poi usai come argomento per la mia tesi di laurea fu quello del trasporto ottimale.

Si tratta di un problema che ha un'origine molto antica. Alla fine del 1700 Gaspard Monge, un matematico francese, lavorò su questioni di questo tipo: suppongo di avere delle miniere; queste miniere producono del materiale edile con il quale voglio costruire delle fortificazioni, creare degli avamposti di guerra (ricordate che la Francia a fine '700 viveva un momento espansionistico sotto Napoleone); come posso trasportare "in materia ottimale" il materiale dalle miniere alle fortificazioni? Naturalmente dobbiamo anzitutto specificare cosa significhi per noi ottimale: per Monge, l'intenzione era quella di minimizzare un costo totale, dove il costo di trasportare un'unità di materiale da un punto a un altro era semplicemente proporzionale alla distanza fra i due punti. Notate poi che la questione è complicata dal fatto che ciascun avamposto avrà bisogno per essere costruito di una specifica quantità di materiale, e a sua volta ciascuna miniera ne produce una quantità fissata. Come distribuisco le unità di materiale prodotte dalle miniere sui vari avamposti?

Monge elaborò uno studio teorico di questo problema che a posteriori giudicheremmo abbastanza elementare, ma colse un aspetto molto importante, che vi posso mostrare facilmente. Supponete per semplicità di voler trasportare il materiale dai punti della zona A ai punti della zona B del piano. Immaginate di decidere di portare ciò che si trova nel punto x sul punto $T(x)$ (T sta per "trasporto") e ciò che si trova nel punto y su $T(y)$, come nella Figura 44. Questa scelta può essere ottimale? La risposta è no, per il semplice fatto che il segmento da x a $T(x)$ e il segmento da y a $T(y)$ si intersecano! Per la disuguaglianza triangolare, infatti, risulta conveniente mandare x su $T(y)$ e y su $T(x)$: abbiamo $|x-T(x)|+|y-T(y)| = |x-z|+|z-T(x)|+|y-z|+|z-T(y)| > |x-T(y)|+|y-T(x)|$, dove $|a-b|$ è la distanza fra il punto a e il punto b e z è il punto d'intersezione fra i segmenti sopra citati.

Quello che abbiamo appena mostrato si può esprimere in questi termini: i raggi di trasporto non si intersecano mai. A partire da questa idea, il problema di Monge si riduce a identificare questi raggi unidimensionali lungo i quali i punti possono viaggiare e poi, passando da un problema

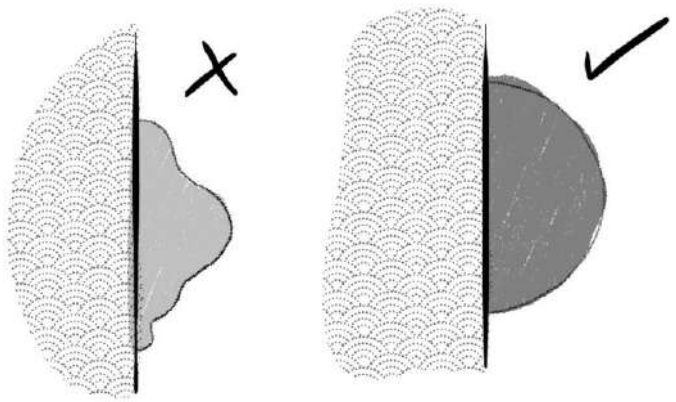


Figura 45: Ottimalità del semicerchio.

nel piano a un problema sulla retta, più semplice da risolvere, determinare la specifica destinazione di ciascun punto lungo il suo raggio.

La storia del trasporto ottimale ovviamente non finisce qui; ma l'episodio specifico che affascinò me è molto più moderno ed è legato alla tradizione matematica di Pisa, in particolare alla figura di Ennio de Giorgi.

De Giorgi è un matematico che si è occupato molto di problemi geometrici, in particolare fra le altre cose di quelle che si chiamano "superfici minime". Un suo risultato fondamentale è la disuguaglianza isoperimetrica; in altre parole, la ragione per cui le bolle sono rotonde!

Ma anche la ragione per cui tante città hanno una forma circolare. Conoscete la storia di Didone? Didone arriva sulle coste dell'Africa, dove vuole fondare una città; le viene offerta tanta terra quanta ne può ricoprire con una pelle di bue. Lei taglia la pelle di bue in tante strisce sottili e le cuce insieme a formare una corda molto lunga; con questa corda (di una certa lunghezza fissata) vuole racchiudere l'area maggiore possibile. Ora, Didone si trova sulla costa dell'Africa, per cui possiamo immaginare che il suo problema sia quello di racchiudere una porzione del semipiano con la sua corda, come in Figura 45. La forma migliore che Didone possa dare alla sua regione di semipiano, quella di area massima data la lunghezza della corda, è quella di un semicerchio. È dentro a questo semicerchio ottimale che sorgerà Cartagine.

Se il problema fosse stato quello di racchiudere una regione del piano di area massima possibile con una curva di lunghezza fissata, o quello equivalente di tracciare una curva di lunghezza minima possible che racchiuda un'area fissata nel piano, la soluzione sarebbe stata il cerchio.

Lo stesso problema si può porre nello spazio: a volume fissato, qual è la forma che minimizza la superficie? E questo è esattamente il problema della forma delle bolle di sapone: quando una bolla che avete prodotto si chiude, l'aria che contiene è quella che vi avete soffiato all'interno; la forma sulla quale si assesterà la bolla è quella che minimizza l'energia di tensione, che è proprio proporzionale all'area superficiale. Insomma: la soluzione al problema nel caso delle tre dimensioni è, come si vede dalle bolle di sapone, proprio la sfera. Da qui un matematico può porsi la stessa domanda nel caso di quattro, cinque... quante dimensioni vuole!

La dimostrazione rigorosa della disuguaglianza isoperimetrica (in tre dimensioni, e anche più in generale) è più difficile di quanto non lo sia quella del caso due-dimensionale: la prima valida in dimensione qualunque è degli anni '50 ed è dovuta appunto a De Giorgi.

Da studente scoprii l'esistenza di questo teorema matematico che mi piacque molto; e una cosa che mi affascinò fu la connessione inaspettata data dal fatto che una dimostrazione di questo

teorema sfrutta proprio il trasporto ottimale. Ora, qui andiamo un po' sul complicato; l'idea di base è quella di trasportare le particelle d'aria da — diciamo — una bolla "immaginaria" di forma qualunque e di un certo volume fissato a una bolla sferica dello stesso volume; e la quantità da minimizzare, il "costo" associato alle particelle d'aria spostate, non è la distanza come nel problema di Monge, ma il suo quadrato.

Insomma: da studente sono stato veramente affascinato da queste idee e connessioni, e da allora mi sono dedicato proprio all'applicazione di risultati di calcolo delle variazioni, e più in generale tecniche analitiche, a problemi geometrici.

Hai avuto difficoltà venendo dal liceo classico?

Sì! Ma per fortuna le difficoltà non durano troppo.

Per le Olimpiadi della Matematica, che sono state per me l'inizio di questa avventura, non c'è una vera differenza fra formazione classica e scientifica: le nozioni utili di geometria, combinatoria, un po' di algebra — cose che mi sono messo a studiare da solo — non fanno parte in ogni caso dei programmi scolastici. Per entrare in Normale per me il problema più grosso è stato l'esame di Fisica: con tanta buona forza di volontà presi un libro di testo di Fisica del liceo scientifico e provai a colmare le lacune che potevo: il mio regalo di postmaturità a me stesso!

Una volta entrato in Università, poi, il mio problema era soprattutto psicologico: in realtà tante cose che vengono fatte agli ultimi anni di liceo scientifico e non al liceo classico vengono riviste in maniera più approfondita al primo anno di università, in particolare durante il corso di Analisi; la didattica è organizzata in modo che anche uno studente del classico possa frequentare Matematica! Ma all'inizio può essere scoraggiante vedere che gli altri sanno fare tutto e tu ancora non sai fare nulla; per esempio quando venivano assegnati esercizi che sono standard per un liceo scientifico io vedevo tanti ragazzi che non erano minimamente in difficoltà davanti al "solito studio di funzione", mentre io non sapevo proprio da dove cominciare! Ma si tratta di affrontare il problema, o si è perso in partenza; io per fortuna ho incontrato gente molto disponibile, ma soprattutto l'insegnamento qui era strutturato molto bene: accanto alle lezioni c'erano sempre esercitazioni; gli esercitatori ci davano esercizi da fare e avevano delle ore di ricevimento. Nel mio caso andai nello studio di Ariela Briani (era esercitatrice di Analisi ai tempi, adesso è in Francia) e dissi che proprio non sapevo da dove cominciare: non sapevo come si facesse uno studio di funzione; lei si mise lì e spiegò a me come agli altri ragazzi venuti con me a ricevimento quello di cui avevamo bisogno.

La realtà è che dopo già un semestre il problema non si pone più. E a posteriori ho anche notato un vantaggio. Durante un periodo trascorso a Pisa del mio dottorato di ricerca sono stato esercitatore a Ingegneria Edile e Architettura, sia per Analisi che per Geometria; avevo degli studenti veramente in gamba e fin dall'inizio ho cercato di far passare il messaggio che è molto importante fare gli esercizi anche per chi sa già le cose; se non ci si esercita, poi lo si paga all'esame! Purtroppo non tutti mi hanno ascoltato e ricordo benissimo che proprio i migliori, nonostante i tanti voti altissimi che abbiamo dato a fine esame, non sono riusciti a superare il 10. Quindi ecco, voglio dire che in questo aver fatto il classico aiuta: non ti dà la spocchia di già sapere le cose!

Com'è stata la tua esperienza alle Olimpiadi della Matematica?

Iniziai a partecipare al quarto anno di liceo; incominciai a studiare per conto mio: esisteva un sito gestito dall'Unione Matematica Italiana in cui tanti ragazzi scrivevano problemi e soluzioni, e all'inizio soprattutto seguivo i problemi postati e risolti dagli altri. Come dicevo prima, uno all'inizio i problemi magari non li sa fare; ma leggendo le soluzioni scopre nuove tecniche, vede che gli altri nominano teoremi che non conosce, scopre cosa siano e poi alla lunga impara a usarli.

Al penultimo anno di liceo andai dal mio preside, dato che la mia scuola non era neanche iscritta al Progetto Olimpiadi, e dissi che a me sarebbe piaciuto partecipare; lui accettò di iscrivere la scuola e io partecipai alle gare individuali, arrivando al livello nazionale. Fu una bellissima esperienza, anche perché i miei compagni di scuola erano al classico proprio perché odiavano la Matematica: parlare di Matematica con loro sarebbe stato impensabile! Alla gara nazionale a Cesenatico incontrai invece tanti ragazzi appassionati, alcuni dei quali erano già all'ultimo anno e si stavano preparando per provare a entrare in Normale, da cui in un certo senso parte tutta la storia che vi ho raccontato! Partecipai alle gare individuali durante gli ultimi due anni di liceo e ad una gara a squadre solo una volta, a Roma, all'ultimo anno. Io credo che sia molto bello trovare qualcuno con cui si ha sintonia per studiare e prepararsi insieme, ma nel mio caso il mio giro di amici non era fatto di appassionati di Matematica: ero da solo per necessità; e anche per i ragazzi della squadra del liceo che avevamo formato la Matematica non era proprio la priorità. Ma è utile e bello incoraggiarsi a vicenda: in due s'impara molto più in fretta!

Ti sei mai pentito della scelta fatta?

Uhm... no! Certo a volte penso che sarebbe bello saperne di più di tante altre cose. Per esempio sono rimasto molto colpito quando ci fu la scoperta legata alle onde gravitazionali: ci sono stati tantissimi seminari sul tema fatti in giro e andando ad ascoltarli li ho trovati di un affascinante mostruoso. Poi però bisogna anche rendersi conto che risultati come questi sono i grandi traguardi di una disciplina; in realtà ci sono migliaia e migliaia di fisici che lavorano e non è che uno si trovi facilmente a fare scoperte sensazionali; e anche chi lavora per esempio sulle onde gravitazionali magari sta lì non so quanti anni ad ascoltare nella speranza che arrivi questo segnale.

In realtà io mi sono accorto che la Matematica è giusta per me; a me piace la sua stabilità: la Matematica è quella, le cose sono vere o sono false; non è che qualcuno possa tirare fuori dal nulla un'altra teoria, magari non compatibile... la Matematica è molto onesta in questo senso, e questo l'ho molto apprezzato negli anni anche a livello della sua comunità; se dimostro un teorema, può non piacere, però l'ho dimostrato. Questo mi ha sempre dato un senso di tranquillità.

Ma con il progredire della tecnologia la Matematica diventa obsoleta?

Beh, la tecnologia va avanti, ma la Matematica va avanti anche grazie alla tecnologia... e la tecnologia grazie alla Matematica!

Voi vivete nell'era post-Google: inserite parole chiave in questo motore di ricerca che piace tanto, lui riesce a trovarvi il sito giusto. Vi assicuro che negli anni '90 non era così: inserendo le parole chiave non si trovava mai davvero quello che si stava cercando. E la Matematica che c'è dietro Google, l'algoritmo che "decide" come selezionare i siti, è molto interessante. A partire dalle

parole chiave voglio trovare proprio il sito che sto cercando... non è una questione banale. Non basta semplicemente trovare tutti i siti che contengono quelle parole chiave! Anche se per esempio stessi cercando una frase completa, non voglio che Google mi proponga un sito scritto da una persona completamente inaffidabile, che nessuno ha mai visitato; e in effetti questo non avverrà mai: l'algoritmo produce una gerarchia; un sito condiviso, a cui tanti altri siti puntano, avrà un peso maggiore rispetto a uno scritto da una persona a caso. In tutto questo dobbiamo gestire un margine di incertezza, considerare il problema sotto un aspetto probabilistico e accettare la presenza di fluttuazioni; infatti magari il risultato che cerco non sarà il primo o il secondo, ma il quarto.

Poi adesso si parla tanto di intelligenza artificiale, di big data... Questi sono problemi di Matematica! Si tratta di capire, per esempio, come insegnare a un computer a riconoscere immagini. A questo proposito, una ragazza che si è dottorata a Zurigo recentemente occupandosi proprio di questo problema ha raccontato questa storia interessante all'inizio della sua tesi. Come si insegna a un computer a riconoscere immagini? Un po' come a un bambino: gli si fornisce una quantità di dati, migliaia di immagini, e gli si dice "questa è una mucca", "questo è un cane", "questo è un gatto"... Poi in immagini nuove il computer deve riconoscere degli oggetti. Ma un esempio classico che ancora ci mette in difficoltà, in cui dopo aver identificato migliaia e migliaia di immagini il computer cade in errore, è questo: prendo una mucca e la metto su una spiaggia. Il computer resta completamente scioccato perché non ha mai visto una mucca sulla spiaggia, e quindi non riconosce la mucca! Qui si tratta di capire per esempio come riconoscere una parte dell'immagine isolandola dal contesto; e recentemente, cosa che mi ha abbastanza affascinato, sono stati proposti dei metodi per farlo che hanno a che fare proprio col trasporto ottimale! L'idea è quella di pensare alle immagini come composte da pixel e confrontarle cercando di spostare i pixel di un'immagine su quelli di un'altra in modo "ottimale", valutando così quanto siano vicine.

La Matematica è anche questo: la tecnologia va avanti, ma la Matematica certo non resta indietro!

Hai mai pensato ad applicazioni pratiche della tua ricerca?

Allora, c'è da dire che il trasporto ottimale è una disciplina della Matematica all'interno della quale ci sono tantissime sottoaree. C'è gente che lo studia da un punto di vista molto applicato; per esempio nell'urbanistica: voglio trasportare le persone da casa al luogo di lavoro; conosco la distribuzione delle persone perché so dove sono costruite le abitazioni e conosco le dimensioni degli uffici; voglio capire quale sia la maniera migliore di costruire infrastrutture. Ma il trasporto ottimale interviene anche in biologia: se l'obiettivo è pompare il sangue dal cuore perché arrivi dove deve arrivare nel nostro corpo — e il corpo vuole farlo in un certo senso in maniera ottimale, senza "sprecare" chilometri di vene! — si sviluppa un sistema circolatorio strutturato come il nostro, con arterie più grandi e poi ramificazioni in vasi sempre più piccoli, fino ai capillari... esattamente le stesse strutture si vedono negli alberi, se ci fate caso: il nutrimento si raccoglie nel tronco, viene tirato su e poi ridiviso.

Insomma, esistono applicazioni concrete; ce ne sono anche di più avanzate: machine learning, intelligenza artificiale e così via.

La verità è però che io ho studiato problemi abbastanza puri, nel senso che ho usato il trasporto ottimale per studiare problemi tipo "bolle di sapone" come quelli di cui vi parlavo prima. Dopotutto, se fossi stato più motivato dalle applicazioni magari sarei andato a lavorare nel privato.

Ma bisogna capire che, anche dal punto di vista delle applicazioni, lasciare "libero" il cervello può essere molto interessante. Un altro esempio di applicazione legato sempre al trasporto

ottimale è quello della meteorologia. Negli anni '90 un meteorologo ha scoperto che, sotto certe condizioni, le nuvole si muovono in maniera "ottimale"; ovvero, diciamo che io guardo una nuvola (composta da tante particelle), e la riguardo un secondo dopo (quindi vedo un'altra configurazione di particelle), e mi chiedo come ciascuna particella si sia spostata tra il primo e il secondo istante di tempo; ecco, il cammino di ciascuna di queste particelle segue una legge del trasporto ottimale. Scoperto questo, si può cercare di lavorare dal punto di vista del trasporto ottimale — cosa che ho fatto — per risolvere queste equazioni della meteorologia. Dalla scoperta di questo concetto alla soluzione del problema ci sono voluti circa vent'anni. Ora, se tutti fossimo sempre alla ricerca di applicazioni immediate, magari nessuno si sarebbe mai reso conto di questa connessione; questo stesso meteorologo in realtà lavora in un centro di ricerca. Lasciare la gente libera di pensare un po' in grande può essere fonte di tante scoperte.

Questo comunque l'abbiamo già visto nella realtà di tutti i giorni: si potrebbero fare molti esempi sui quali potete trovare facilmente informazioni in rete: quello della crittografia, che non esisterebbe senza teoremi astratti di algebra che i matematici hanno dimostrato nel '700, quello della codifica digitale di audio e immagini, basata su scoperte di Fourier del 1800...

Io mi diverto con la Matematica pura; ma la Matematica pura ha dimostrato di poter portare a invenzioni rivoluzionarie! Con questo non voglio dire che con i miei risultati si cambierà il mondo; io ho ricevuto un premio per la Matematica assegnato da matematici: da una comunità pura che apprezza teoremi puri. Per noi la priorità è quella di far progredire la Matematica stessa — come entità, come teoria... — con l'idea che possa fare del bene, anche portare a lungo termine dello sviluppo. Ma più di tutto siamo mossi dal senso della conoscenza!

Hai altre passioni oltre alla Matematica? Quali?

Diciamo che ho avuto altri passatempi: al liceo ero molto amante dello sport, ma una volta entrato in Normale per i primi anni era dura trovare il tempo; un po' di sacrifici anche in questo senso li ho dovuti fare! Mi è sempre piaciuto molto giocare a calcio (anche se purtroppo — ne parlavo proprio oggi a pranzo con i colleghi — ultimamente giocare mi fa sentire gli anni!). E ho molti hobby: mi piace sciare, andare in montagna...

Col tempo ho apprezzato sempre di più il fatto di riuscire a "staccare": con grandi passioni come quella per la Matematica c'è il grosso rischio di non riuscire a staccare il cervello, cosa che alla lunga può nuocere, specialmente se sei concentrato su un problema per il quale magari hai preso una strada di soluzione sbagliata. Ogni tanto bisogna svuotare e ripartire, come un reboot per il computer: prendersi un weekend in cui il sabato si risponde alle mail e la domenica è completamente libera, per esempio.

Una cosa che cerco di fare è seguire il calcio: a Zurigo ci sono riuscito di meno, ma a Austin avevamo organizzato l'A. S. Roma Austin Fanclub, ero riuscito a coinvolgere un sacco di umanisti!

Insomma: è buono avere hobby e dedicarsi ad altre cose. Io personalmente non ho avuto altre passioni che fossero al livello di quella per la Matematica.

Qual è la motivazione che ti spinge nella ricerca?

Un po' questo spirito della conoscenza che ho nominato prima. Mettiamola così: potrei dire, "secondo me fra tutte le curve del piano il cerchio è quella che racchiude la massima area a perimetro fissato". Sembra ragionevole, no? Beh, sarebbe bello dimostrarlo. Fin da quando ero studente — lo spirito critico del matematico! — mi sarebbe venuto da dire: bello, sembra ragionevole, ma siamo sicuri che sia proprio così? Non potremmo esserci persi qualcosa? Una cosa è sembrare, una cosa è avere una certezza.

Io sono sempre stato attratto da problemi che a me sembravano affascinanti (e questo è un giudizio personale! Non per niente nella Matematica abbiamo così tante sottoaree); d'altra parte, fra i problemi che mi affascinano ho cercato di essere anche pragmatico nella scelta di quelli a cui dedicarmi, perché alcuni possono essere più curiosità personali, altri più strutturali e potenzialmente interessanti per tutta la comunità.

La scelta dei propri problemi di ricerca in ogni caso è molto difficile e imparare a farla fa parte della maturazione di un matematico: i miei studenti di dottorato, già laureati, vengono da me a chiedermi un problema su cui lavorare, non lo scelgono da soli; io do loro delle proposte e della letteratura da leggere, loro tornano magari non del tutto convinti — mi dicono che vorrebbero qualcosa di più geometrico, o meno geometrico, o con un po' di probabilità... — e a quel punto si aggiusta il tiro. Io da studente ho capito abbastanza presto di essere particolarmente interessato all'Analisi, ma per capire in che direzione andare al suo interno ho dovuto esplorare molto!

Una lettura che per te è stata importante?

Una che mi è piaciuta durante il liceo è stata il libro "Che cos'è la Matematica?", di Courant e Robbins, penso sia un classico. Non sono mai stato un grande lettore ma lo lessi con molto piacere!

Trova più facilmente lavoro un laureato in Matematica o uno in Fisica?

Ehm, non sono un'agenzia di collocamento!

Io conosco bene il mondo matematico e molto meno quello fisico, ma una cosa che vi posso dire è questa: oggi i matematici vengono assunti spesso al di là delle loro conoscenze in Matematica. E' la forma mentis che sviluppano che è molto apprezzata nel mondo del lavoro: ho amici che, dopo una laurea, un dottorato, addirittura un postdoc in Matematica, sono andati a lavorare a Google o nella finanza; e sono stati assunti nonostante non avessero alcuna conoscenza pregressa del mestiere che andavano a fare! Questo perché in genere un matematico, una volta assimilati i concetti di base, ha un approccio ai problemi che piace: tira fuori l'idea giusta, magari è proprio intrigato dall'ottimizzare certi processi... Insomma, quella che è molto apprezzata, vedo, è la versatilità della mente matematica; e naturalmente non è detto che un fisico non possa averla!

Per quanto riguarda il mondo accademico, credo che la quantità di opportunità sia comparabile. Quello che posso dire è che, dovendo insegnare la Matematica a tutti — matematici, fisici, ingegneri! — i matematici sono richiesti in tutte le università.

Secondo te la Matematica si scopre o si inventa?

Quando ero giovane qualcuno mi disse che "come matematico ero molto ingegnere": ho una visione molto pragmatica. Per me la Matematica è uno strumento che nel tempo è stato sviluppato dall'uomo per descrivere la natura. Ad esempio, nel caso della geometria euclidea, uno mette degli assiomi perché sembrano naturali, e da quelli dimostra il Teorema di Pitagora. Però la verità è che in questo processo noi ci lasciamo ispirare molto dalla natura: se io disegno un triangolo rettangolo su questo vale il teorema di Pitagora; e fra le altre cose sappiamo che la somma dei suoi angoli interni è 180°. Ragionevole: magari pure in un universo parallelo vale la stessa cosa. Però se considero la cara vecchia Terra — una sfera, diciamo — e dal polo Nord

scendo lungo un meridiano fino all'equatore, poi mi faccio 2000 km lungo l'equatore e ritorno al polo Nord con un altro meridiano, ho formato un triangolo con due angoli di 90° più un terzo angolo... la somma è ben più di 180° e il Teorema di Pitagora ha dei problemi a valere! Ecco, stiamo parlando di una geometria non euclidea; la Matematica sviluppa teorie sempre più generali proprio per modellizzare situazioni di questo genere, in cui magari teoremi che ci sembravano ovvi non sono più veri. Ma le geometrie non euclidee le abbiamo create proprio perché le abbiamo "viste"! Insomma, io penso che noi sviluppiamo la Matematica a partire dalla nostra esperienza e da quello che ci circonda: magari in un mondo diverso svilupperemmo una Matematica diversa a partire da assiomi diversi. Questa è la mia visione: chiaramente si tratta di una questione filosofica e molto personale; altri potrebbero rispondere in modo completamente diverso!

18 Frazioni e quaderni a quadretti

Carlo Carminati e Giulio Tiozzo, n.9, Settembre 2019

18.1 Introduzione

Cominciamo proponendo alcune domande e problemi legati ad oggetti matematici che tutti conoscono sin dalle scuole elementari. Vi suggeriamo di tenere a mano un foglio a quadretti: potrebbe tornarvi utile!

18.2 Mission impossible

Ecco un celebre rompicapo, attribuito al matematico e scrittore inglese Lewis Carroll. La figura qui sotto induce a pensare che sia possibile ricoprire un rettangolo 5x13 ricomponendo i pezzi di una piastrella quadrata 8x8.

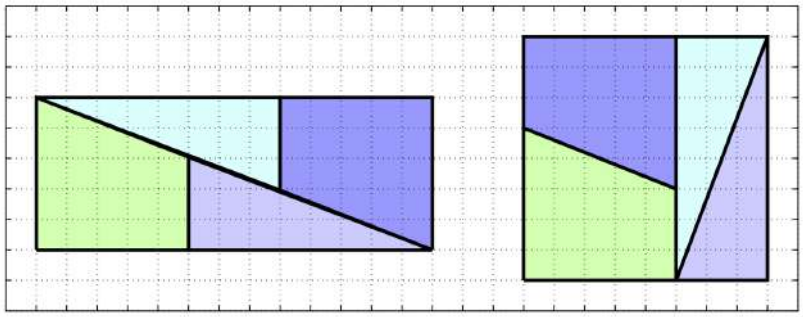


Figura 46: Dov'è l'imbroglio?

La cosa è paradossale, dato che le aree non coincidono: $5 \times 13 = 65 = 8^2 + 1$... Tra qualche pagina spiegheremo dove sta l'inghippo.

18.3 Le frazioni (non) crescono sugli alberi

Le frazioni, o più precisamente i numeri che si scrivono come rapporto di due interi, si chiamano *numeri razionali*; l'insieme di tutti i numeri razionali si indica col simbolo $\mathbb{Q}$. Lavorare con le frazioni è facile, a patto di ricordare due cose fondamentali:

(i) la somma di due frazioni p/q e p'/q' si calcola mediante la formula:

$$\frac{p}{q} + \frac{p'}{q'} = \frac{pq' + p'q}{pq};$$

(ii) mai dividere per 0.

Nel seguito di questa sezione vedremo che "trasgredendo" a queste regole possiamo scoprire cose interessanti e formulare domande non banali. Cominciamo defininendo una nuova operazione che indichiamo col simbolo $\oplus$:

$$\frac{p}{q} \oplus \frac{p'}{q'} = \frac{p + p'}{q + q'}.$$

Per esempio $\frac{1}{3} \oplus \frac{1}{2} = \frac{2}{5}$ ... esattamente quello che **non** si deve assolutamente fare se si vuole calcolare la somma usuale[17]. Osserviamo infatti che, in generale, se p, q, p', q' sono interi positivi la quantità $\frac{p}{q} \oplus \frac{p'}{q'}$ rappresenta un valore intermedio tra gli 'addendi' $\frac{p}{q}$ e $\frac{p'}{q'}$ (quindi non coincide **mai** con la vera somma); chiameremo *frazione mediana*[18] il risultato dell'operazione $\oplus$.

Se abbiamo una lista L di frazioni positive possiamo definire una lista derivata aggiungendo alla lista L, tra ogni coppia di elementi adiacenti, la loro frazione mediana; in questo modo partendo da una lista di $n+1$ elementi ne otteniamo una di $2n+1$ elementi.

Partiamo ora dalla lista che contiene le "frazioni"[19] $L_0 := [\frac{0}{1},\ \frac{1}{0}]$ e consideriamo la lista derivata L_1, la derivata della derivata L_2, e così via:

$$\begin{array}{ll} L_0 & [\frac{0}{1},\ \frac{1}{0}] \\ L_1 & [\frac{0}{1},\ \frac{1}{1},\ \frac{1}{0}] \\ L_2 & [\frac{0}{1},\ \frac{1}{2},\ \frac{1}{1},\ \frac{2}{1},\ \frac{1}{0}] \\ L_3 & [\frac{0}{1},\ \frac{1}{3},\ \frac{1}{2},\ \frac{2}{3},\ \frac{1}{1},\ \frac{3}{2},\ \frac{2}{1}, \frac{3}{1},\ \frac{1}{0}] \\ L_4 & [\frac{0}{1},\ \frac{1}{4},\ \frac{1}{3},\ \frac{2}{5},\ \frac{1}{2},\ \frac{3}{5},\ \frac{2}{3},\ \frac{3}{4},\ \frac{1}{1},\ \frac{4}{3},\ \frac{3}{2},\ \frac{5}{3},\ \frac{2}{1},\ \frac{5}{2}, \frac{3}{1},\ \frac{4}{1},\ \frac{1}{0}] \\ & \ldots \end{array} \tag{10}$$

In questo modo otteniamo una sequenza infinita di liste finite, dove ciascuna è contenuta nella successiva: $L_0 \subset L_1 \subset L_2 \subset L_3 \subset ....$

Possiamo anche rappresentare gli elementi così generati come nodi di un albero binario infinito, chiamato albero di Stern-Brocot, dove all'enne-simo livello compaiono i 2^{n-1} elementi generati

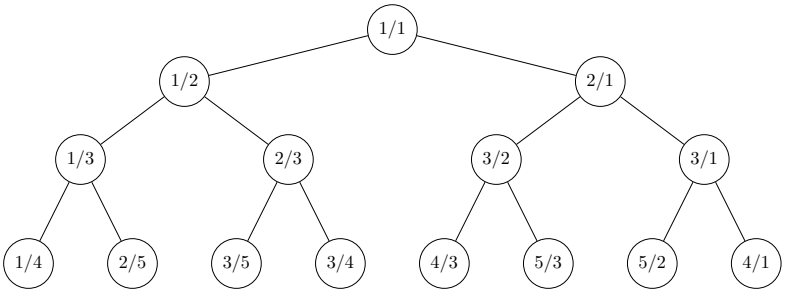


Figura 47: Ramificazioni iniziali dell'albero di Stern-Brocot.

dall'ennesima derivazione. Si noti che tutti gli elementi che stanno nel sottoalbero sinistro sono frazioni comprese tra 0 ed 1, inoltre l'applicazione $p/q \mapsto q/p$ corrisponde alla simmetria di asse verticale che scambia il sottoalbero sinistro con quello destro.

Ci possiamo chiedere se ogni frazione positiva p/q compaia nell'albero, se le frazioni che compaiono siano sempre in forma ridotta ai minimi termini (ovvero se p e q sono primi tra loro), e se possano esserci ripetizioni.

La definizione ed il lemma che seguono saranno la chiave per dare risposta a tutte queste domande:

[17] Nella letteratura anglosassone l'operazione che abbiamo indicato con $\oplus$ viene talvolta indicata scherzosamente col termine *freshman sum* (la somma della matricola).

[18] Attenzione: la frazione mediana non coincide, in generale, con la media aritmetica.

[19] Ovviamente la prima rappresenta il numero 0, mentre la seconda è solo una scrittura formale.

Definizione 1. Due frazioni $\frac{p}{q} < \frac{m}{n}$ sono una *coppia irriducibile* se si ha che

$$mq - np = 1. \tag{11}$$

Lemma 1. Se le frazioni $\frac{p}{q} < \frac{m}{n}$ sono adiacenti in una delle liste $L_0, L_1, L_2, L_3, \ldots$ allora sono una coppia irriducibile. Inoltre la frazione mediana $\frac{p+m}{q+n}$ è il razionale con denominatore minimo contenuto nell'intervallo aperto $(p/q, m/n)$.

Verificare che tutti gli elementi adiacenti nelle prime liste sono coppie irriducibili è un'operazione meccanica (p.es. 3/5 e 2/3 sono adiacenti nella quarta lista generata, e infatti $2 \times 5 - 3 \times 3 = 1$), ma per dimostrare che questa proprietà vale sempre dovremo utilizzare un procedimento induttivo. Per il momento ci limitiamo ad osservare che (11) implica che le frazioni che compaiono nelle liste sono tutte ridotte, e sono ordinate in maniera strettamente crescente (di conseguenza ogni frazione compare nell'albero al più una volta). In effetti dal lemma segue che ogni frazione positiva compare in corrispondenza di qualche nodo dell'albero, ma questo lo vedremo dopo nella sezione 18.5, dove dimostriamo anche il lemma 1.

Un'altra domanda interessante è la seguente: data una frazione p/q, a quale livello dell'albero la troviamo? Qual è il percorso per raggiungerla? Nella sezione 18.5 proveremo a rispondere anche a queste domande.

18.4 Linee rette nell'era digitale

Lo schermo del computer è un'area rettangolare divisa in minuscole celle (i *pixel*); per rappresentare una linea sullo schermo il computer colora i pixel che vengono attraversati. Questo fatto è evidente quando una linea retta viene rappresentata su uno schermo a bassa risoluzione: in figura 48, a scopo esplicativo, raffiguriamo una retta su una quadrettatura abbastanza grossolana.

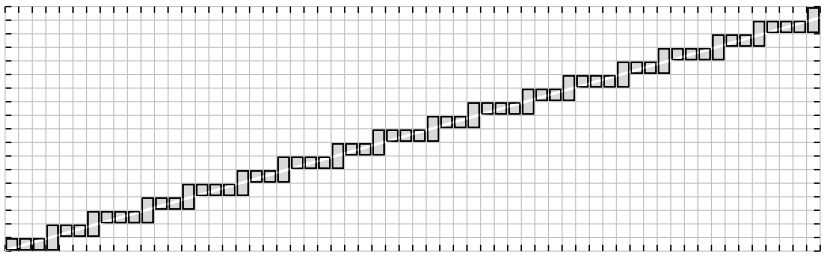


Figura 48: Una retta nel piano e la sua 'digitalizzazione' su una griglia. Possiamo codificare questa retta mediante la sequenza □ □ □ ▯ □ □ ▯ □ □ □ ▯ □ □ ▯ □ □ □ ▯ □ □ ▯ □ □ □ ▯ ...

Una retta di coefficiente angolare compreso tra 0 ed 1 può essere composta assemblando opportunamente le tessere □ e ▯, incollandole in corrispondenza di un lato verticale (vedi figura 48). In maniera analoga, una retta con coefficiente angolare maggiore di 1 potrà essere assemblata incollando tessere di forma □ o □□ incollate lungo i lati orizzontali.

Alcune domande naturali sono le seguenti:

1. Quali sequenze di tessere possiamo trovare lungo una retta? Per esempio non è difficile convincersi che nessuna retta conterrà mai la sequenza □ □ ▯ ▯.

2. Quali rette danno luogo ad una sequenza periodica?

3. Data la sequenza di tessere associata ad una certa retta, posso ricavare informazioni sul suo coefficiente angolare?

Il secondo di questi quesiti non è di difficile soluzione (invitiamo chi legge a provare a rispondere autonomamente), degli altri due tratteremo più avanti.

18.5 Prospettiva proiettiva

Torniamo alle domande che abbiamo formulato nella sezione 18.3, ed in particolare al lemma 1, che sarà la chiave di tutto.

Partiamo osservando che la condizione (11) di essere una coppia irriducibile (ovvero $mq-np=1$) ha un'interessante interpretazione geometrica[20]. Infatti, come si può verificare direttamente calcolando la differenza di aree di figure elementari, la quantità $mq-np$ rappresenta l'area del parallelogramma generato nel piano dai vettori[21] (q,p) e (n,m); questi vettori individuano due rette di coefficiente angolare rispettivamente $\frac{p}{q}$ e $\frac{m}{n}$, e la frazione mediana $\frac{p+m}{q+n}$ risulta essere il coefficiente angolare del vettore risultante dalla somma tra (q,p) e (n,m). Osserviamo anche che tra i punti a coordinate intere interni al rettangolo $[0,q+n]\times[0,p+m]$, i punti (q,p) e (n,m) sono quelli più vicini alla diagonale.

Queste considerazioni mostrano che, anche se siamo partiti parlando di frazioni, tutte le questioni si possono riformulare in maniera molto più trasparente usando i vettori. Più precisamente: siamo interessati allo *spazio proiettivo*, ovvero l'insieme delle *direzioni* individuate dai vettori. L'opera-zione $\oplus$ è quindi una somma sotto mentite spoglie, e se non riconosciamo le consuete proprietà di una somma è solo perché stiamo interpretando il risultato non come un *vettore* ma come una *direzione*. È una questione di prospettiva: i vettori $(1,1)$ e $(101,100)$ sono molto lontani se considerati come vettori, ma corrispondono a direzioni quasi indistinguibili (per lo meno ad occhio nudo). Anche la frazione 'degenere' $1/0$ trova adeguata spiegazione in questo contesto, visto che corrisponde all'inoffensivo vettore $(0,1)$, che punta in direzione verticale.

La figura 49 dovrebbe suggerire la soluzione del paradosso esposto in § 18.2: la suddivisione del rettangolo nasconde, sotto il tratto spesso dei bordi neri dei tasselli, un parallelogramma di area 1, che corrisponde all'area che sembra svanire quando si ricompongono i pezzi nella piastrella quadrata. L'occhio umano è un formidabile supporto per la nostra intuizione, ma risulta poco affidabile per fare stime quantitative precise (è difficile accorgersi, ad occhio, di una discrepanza inferiore al 2%). Per questo motivo è sempre bene affidarsi a dimostrazioni rigorose[22].

Dimostrazione del Lemma 1. È immediato verificare che tutti gli elementi adiacenti nelle prime liste sono coppie irriducibili. Supponiamo ora di aver dimostrato la proprietà (11) per tutti gli elementi adiacenti di una certa lista L_k; osserviamo quindi che due elementi adiacenti nella

[20]Chi ha dimestichezza con le matrici riconoscerà facilmente che $mq-np$ è il determinante della matrice $\begin{pmatrix} q & n \\ p & m \end{pmatrix}$.

[21]Nel seguito identificheremo sistematicamente il punto di coordinate (q,p) col vettore che va dall'origine a (q,p).

[22]I programmi di calcolo numerico che girano sui computer moderni utilizzano la doppia precisione, che garantisce circa 15 cifre decimali significative: molto meglio dell'occhio umano ma pur sempre una precisione finita, che a volte risulta comunque insufficiente, dato che in matematica non di rado capita di aver a che fare con quantità mostruosamente piccole (o grandi).

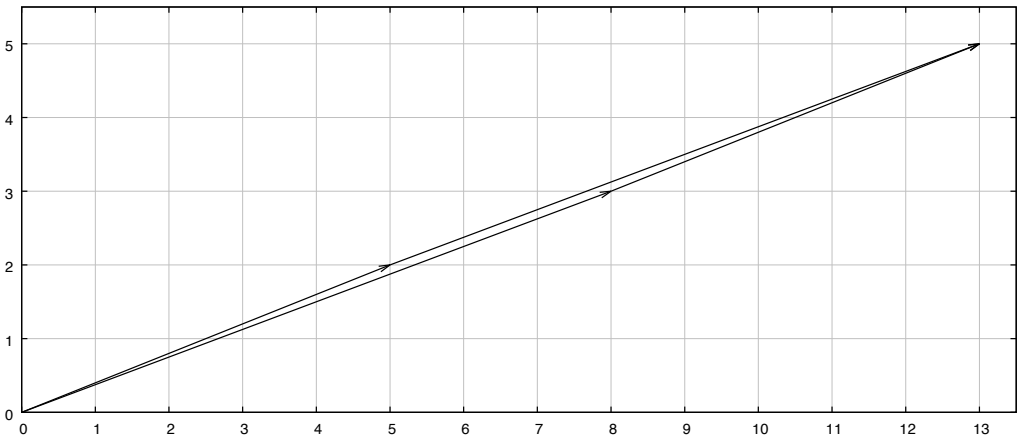


Figura 49: Il parallelogramma generato dai vettori $(8,3)$ e $(5,2)$ ha area $8 \times 2 - 5 \times 3 = 1$, ma è molto schiacciato sulla diagonale, individuata dalla direzione del vettore $(13,5)$.

lista L_{k+1} formano una coppia della forma

$$\left(\frac{p}{q}, \frac{p+m}{q+n}\right) \quad \text{oppure} \quad \left(\frac{p+m}{q+n}, \frac{m}{n}\right)$$

con p/q, e m/n elementi adiacenti di L_{k+1}. Ma in entrambi i casi queste sono coppie irriducibili. Infatti nel caso della prima coppia abbiamo che $(p+m)q-(q+n)p = mq-np = 1$; analogamente nel caso della seconda si verifica $(p+m)n-(q+n)m = mq-np = 1$.

Pertanto la proprietà (11) si propaga per contagio da una lista a quella successiva, e dunque deve valere per tutte le liste[23].

La minimalità del denominatore della frazione mediana si comprende meglio impostando il problema in termini di vettori: dobbiamo infatti verificare che tra tutti i punti a coordinate intere che stanno all'interno dello spicchio del primo quadrante compreso tra le semirette individuate dai vettori (q,p) e (n,m), il vettore $(q+n, p+m)$ è quello di ascissa minima. In altre parole dobbiamo verificare che, se

$$\frac{p}{q} < \frac{a}{b} < \frac{m}{n} \tag{12}$$

e vale

$$\lambda \begin{pmatrix} q \\ p \end{pmatrix} + \mu \begin{pmatrix} n \\ m \end{pmatrix} = \begin{pmatrix} b \\ a \end{pmatrix} \tag{13}$$

con λ e μ numeri reali, allora $b \geq q+n$.

Ma possiamo invertire (13) ottenendo che

$$\begin{cases} \lambda = nb - ma \\ \mu = -pb + qa. \end{cases}$$

Quindi sia λ che μ sono numeri interi, e di conseguenza la condizione (12) implica che $aq - pb \geq 1$ e $nb - ma \geq 1$, ovvero $\lambda \geq 1$ e $\mu \geq 1$. Deduciamo quindi che

$$b = \lambda q + \mu n \geq q + n,$$

[23]Qui, volendo essere formali, dovremmo far riferimento al principio di induzione matematica.

e vale l'uguaglianza se e solo se $\lambda = \mu = 1$: ciò conclude la dimostrazione del lemma. Notiamo che da questa dimostrazione segue pure un altro fatto interessante: il parallelogramma generato dai vettori (q, p) e (n, m) non contiene alcun nodo a coordinate intere al suo interno.

A questo punto non è difficile dimostrare che ogni frazione a/b appartiene a tutte le liste L_k per tutti gli indici più grandi di un certo k_0. A questo scopo osserviamo che se $a/b \notin L_k$, posso definire $I_k = (\frac{p_k}{q_k}, \frac{m_k}{n_k})$ come l'intervallo aperto che ha come estremi i due elementi adiacenti di L_k tali che $\frac{p_k}{q_k} < \frac{a}{b} < \frac{m_k}{n_k}$. È facile constatare che gli intervalli I_k formano una successione di intervalli incapsulati $I_{k+1} \subset I_k$ e, per la minimalità della frazione mediana, deve sempre essere $b \leq q_k + n_k$. D'altra parte la quantità $q_k + n_k$ forma una successione strettamente crescente in k, pertanto può rimanere sotto la soglia b solo per un numero finito di valori dell'indice k. Questo vuol dire che ci sarà un indice k_0 per cui si avrà che $a/b \in I_{k_0} = (\frac{p}{q}, \frac{m}{n})$ e $a/b = \frac{p+m}{q+n}$, così che $a/b \in L_{k_0+1}$, e ciò termina la dimostrazione.

Nel seguito chiameremo *genitori* della frazione a/b la coppia irriducibile di frazioni p/q e m/n che compare nella dimostrazione appena terminata, ovvero quella che soddisfa le condizioni $mp - qn = 1$ e $\frac{a}{b} = \frac{p+m}{q+n}$. Questa definizione è ben posta ed ha anche un significato geometrico intrinseco: le frazioni p/q e m/n sono i coefficienti angolari dei vettori (q, p) e (n, m), che corrispondono ai due punti a coordinate intere più vicini alla diagonale del rettangolo $[0, b] \times [0, a]$. Pertanto un problema interessante è il seguente: data una frazione a/b, come si determinano i suoi genitori?

18.6 L'albero dei periodi

Torniamo ora alle questioni che abbiamo accennato nella sezione 18.4: vogliamo codificare una retta in base al modo con cui questa taglia i tasselli di una griglia quadrata. Per semplicità nel seguito ci limiteremo a considerare rette del tipo $y = \alpha x + \beta$ con coefficiente angolare $\alpha \in [0, 1]$.

Per formalizzare questo problema conviene definire una sequenza bi-infinita composta da caratteri di un alfabeto binario: chiamiamo *sequenza di taglio* della retta r la successione $(..., \xi_{-2}, \xi_{-1}, \xi_0, \xi_1, \xi_2, ...)$ dove $\xi_n = 1$ se la retta r taglia una retta orizzontale del reticolo nella striscia $n - 1 \leq x \leq n$, e $\xi_n = 0$ altrimenti.

Per il momento assumeremo anche che la retta sia *non degenere*, ovvero che non passi per alcun nodo a coordinate intere; è facile vedere che in tal caso questa codifica coincide con quella accennata in nella sezione 18.4 mediante la corrispondenza $\square \mapsto 0$ e $\boxminus \mapsto 1$. Al di là di considerazioni di carattere tipografico, preferiamo utilizzare l'alfabeto $\{0, 1\}$ invece che $\{ \square , \boxminus \}$ perché tale scelta rende facile esprimere la codifica a partire dall'equazione della retta. Infatti è immediato verificare che

$$\xi_k = \lfloor \alpha k + \beta \rfloor - \lfloor \alpha(k-1) + \beta \rfloor \tag{14}$$

dove $\lfloor x \rfloor$ indica il più grande intero che non supera x (p.es. $\lfloor \sqrt{5} \rfloor = 2$, $\lfloor 3 \rfloor = 3$). Per tutte le rette non degeneri tale definizione corrisponde a quella che abbiamo dato nella sezione 18.3, ma in realtà la formula (14) è definita anche per una retta degenere (anche se in questo caso non è evidente l'interpretazione geometrica). Avremmo anche potuto scegliere un'espressione leggermente diversa:

$$\xi_k = \lceil \alpha k + \beta \rceil - \lceil \alpha(k-1) + \beta \rceil \tag{15}$$

dove $\lceil x \rceil$ denota il più piccolo intero non inferiore ad x (p.es. $\lceil \sqrt{5} \rceil = 3$, $\lceil 3 \rceil = 3$). Queste due espressioni coincidono su tutte le rette non degeneri, ma non su quelle degeneri che possiamo

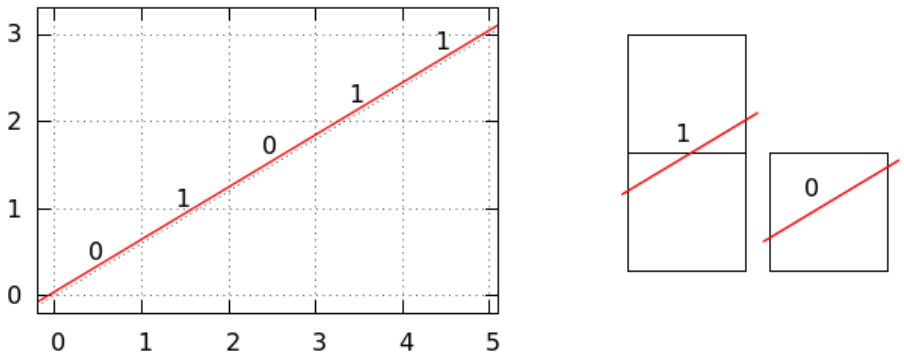


Figura 50: La sequenza di taglio di una retta.

pensare come rette dotate di una doppia codifica; indicheremo rispettivamente con (ξ_k^+) e (ξ_k^-) le due codifiche di una retta ottenute con le formule (14) e (15), rispettivamente; chiameremo queste codifiche sequenza di taglio *superiore* ed *inferiore*.

Per dare un'interpretazione geometrica di questa doppia codifica può essere utile, in via provvisoria, codificare una retta degenere utilizzando l'alfabeto ternario $\{0, 1, X\}$, dove si procede come per le rette non degeneri, tranne per il fatto che si usa il simbolo X in adiacenza di un passaggio per un nodo della quadrettatura. In questo modo le cifre X capitano sempre in coppia. La codifica di una retta parallela e molto vicina a quella data si ottiene trasformando tutte le coppie XX in 10 o in 01 a seconda che la nuova retta passi sopra o sotto.

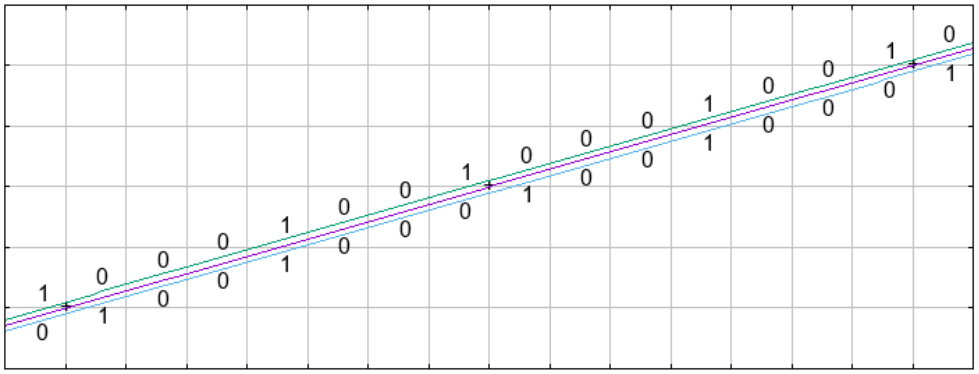


Figura 51: Sequenze di taglio per una retta appena sopra ed una appena sotto una retta degenere; le due sequenze differiscono solo in corrispondenza di un nodo sulla retta degenere.

Una retta degenere con coefficiente angolare razionale passa per infiniti nodi della quadrettatura e, di conseguenza, la sua codifica conterrà infiniti blocchi del tipo XX. P.es. la retta $y = \frac{p}{q}x$ interseca i nodi della griglia a coordinate intere nei punti del tipo (kq, kp) con k intero (e solo in questi punti, se p e q sono primi fra loro). Osserviamo inoltre che la stringa compresa tra due

blocchi XX consecutivi è palindroma, per evidenti ragioni di simmetria; ciò ha come conseguenza il fatto che le due stringhe $(\xi_1^{\pm}, ..., \xi_q^{\pm})$ contenenti i primi q elementi delle due sequenze di taglio di una retta degenere si ottengono una dall'altra semplicemente invertendo l'usuale verso di lettura. Questo fatto può essere interpretato in relazione alla seguente proprietà geometrica: invertendo il verso di lettura della sequenza di taglio di una retta non degenere r si ottiene la sequenza di taglio della retta in posizione simmetrica ad r rispetto all'origine.

Per brevità nel seguito chiameremo *retta razionale* una retta con coefficiente angolare razionale; la sequenza di taglio di una retta razionale è particolarmente semplice, infatti è periodica. Si noti che è facile risalire al coefficiente angolare p/q di una retta razionale a partire dal periodo della sua sequenza di taglio: infatti q è la lunghezza del periodo, mentre p è il numero di volte che la cifra 1 appare nel periodo.

Al variare del parametro reale c, le rette razionali $y = \frac{p}{q}x + c$ hanno sequenze di taglio che differiscono solo per una traslazione, di conseguenza hanno tutte lo stesso periodo (a meno di permutazioni cicliche). Anzi, nella famiglia $y = \frac{p}{q}x + c$ le rette degeneri sono tutte e sole quelle per cui $c = k/q$ con k intero, e tutte le rette che sono comprese tra due rette degeneri consecutive hanno sequenza di taglio identica.

Per una retta razionale di coefficiente p/q si può quindi provare un'altra curiosa relazione tra le sequenze di taglio (ξ_k^+) e (ξ_k^-) definite dalle equazioni (14) e (15): esse si ottengono una dall'altra mediante una traslazione, ovvero esiste un intero $q_1 \in (0, q)$ tale che $\xi_k^- = \xi_{k+q_1}^+$. Non è difficile rendersi conto che q_1 deve essere il denominatore di un genitore p_1/q_1 di p/q: infatti la sequenza di taglio superiore (ξ_k^+) è la sequenza di taglio di ogni retta del tipo $y = \frac{p}{q}x + c$ per $0 < c < 1/q$, e il nodo di coordinate intere (q_1, p_1) sta proprio sulla retta degenere $y = \frac{p}{q}x + \frac{1}{q}$.

Se (ξ_k^+) è la sequenza di taglio superiore della retta degenere r di coefficiente angolare $\alpha = p/q < 1$ allora $(\xi_1^+, ..., \xi_q^+)$ è minimo, nell'ordine lessicografico, rispetto ad ogni sua permutazione ciclica. Infatti basta confrontare $(\xi_1^+, ..., \xi_q^+)$ con i primi elementi $(\xi_1', ..., \xi_q')$ della sequenza di taglio di una retta non degenere $y = \frac{p}{q}x + c'$ con $c' \in (0, 1)$: se i due periodi non coincidono, chiamiamo k_0 il primo indice per cui $\xi_{k_0}' \neq \xi_{k_0}^+$, avremo che in corrispondenza delle linee verticale $\{x = k\}$ con $k < k_0$ le due rette tagliano il lato dello stesso tassello della quadrettatura mentre per $k = k_0$ tagliano il lato di tasselli differenti, e dato che r' sta sopra r avremo $\xi_{k_0}' = 1$ e $\xi_{k_0}^+ = 0$, il che dimostra la tesi. Un argomento del tutto analogo mostra che $(\xi_1^-, ..., \xi_q^-)$ è più grande (nell'ordine lessicografico) di qualunque sua permutazione ciclica.

Chiameremo *periodo standard* della retta di coefficiente angolare p/q la stringa $W_{p/q} := (\xi_1^+, ..., \xi_q^+)$. Osserviamo che se $mq - np = 1$ allora

$$W_{\frac{p}{q} \oplus \frac{m}{n}} = W_{\frac{p}{q}} W_{\frac{m}{n}} \tag{16}$$

dove al membro destro abbiamo semplicemente la concatenazione di due stringhe. Il motivo per cui vale questa proprietà è il seguente: nel triangolo di vertici $(0, 0)$, (q, p) e $(q + n, p + m)$ non ci sono nodi a coordinate intere, di conseguenza la sequenza di taglio superiore del segmento che va dall'origine a $(q + n, p + m)$ è la concatenazione della sequenza di taglio superiore del segmento che va dall'origine a (q, p) con quella del segmento che va da (q, p) a $(q + n, p + m)$.

Possiamo quindi partire da una lista contenente i periodi $W_{0/1} = 0$ e $W_{1/1} = 1$ per generare liste sempre più complete, aggiungendo i periodi corrispondenti alle frazioni medianti (in modo

Figura 52: In questo caso le due sequenze di taglio coincidono prima di $k_0 = 3$, mentre $\xi'_3 = 1$ e $\xi^+_3 = 0$.

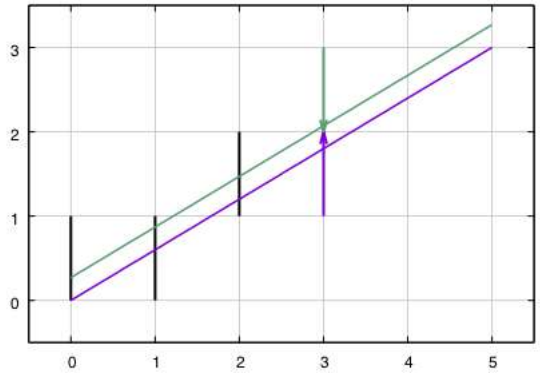
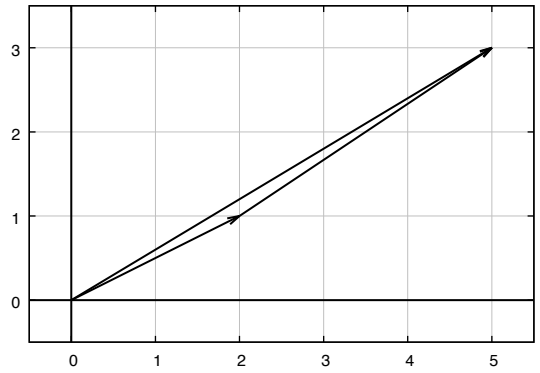


Figura 53: Abbiamo $p/q = 1/2$ e $m/n = 2/3$; si verifica quindi che $W_{3/5} = 01011$ è concatenazione di $W_{1/2} = 01$ con $W_{2/3} = 011$.



del tutto analogo a quanto fatto nella sezione 18.3):

$$\begin{array}{ll} F_1 & [0,\ 1] \\ F_2 & [0,\ 01,\ 1] \\ F_3 & [0,\ 001,\ 01,\ 011,\ 1] \\ F_4 & [0,\ 0001,\ 001,\ 00101,\ 01,\ 01011,\ 011,\ 0111,\ 1] \\ & \ldots \end{array}$$

L'unica differenza rispetto alla costruzione della sezione 18.3 è che prendiamo in considerazione solo periodi relativi a frazioni $p/q \leq 1$ (p.es. le frazioni corrispondenti ai periodi di F_4 sono

$$\left[\frac{0}{1},\ \frac{1}{4}, \frac{1}{3},\ \frac{2}{5},\ \frac{1}{2},\ \frac{3}{5},\ \frac{2}{3},\ \frac{3}{4},\ \frac{1}{1}\right],$$

che è la prima metà della lista L_4). Possiamo anche ordinare i periodi corrispondenti alle frazioni tra 0 ed 1 mediante una struttura ad albero che riflette quella della metà sinistra dell'albero di Stern-Brocot.

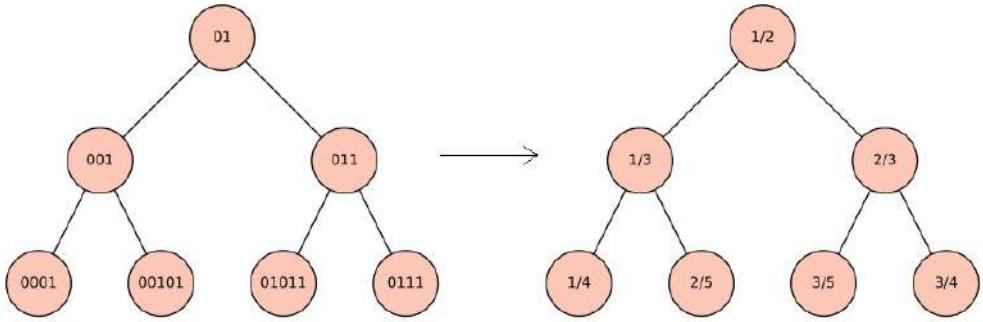


Figura 54: Prime ramificazioni dell'albero delle frazioni e dell'albero dei periodi.

L'equazione (16) mostra anche che ogni periodo standard ammette una fattorizzazione "canonica" (p.es. il periodo standard 0010101 è concatenazione dei due periodi standard 00101 e 01, e questo è l'unico modo di scrivere 0010101 come concatenazione dei due periodi standard).

C'è anche un altro modo per descrivere l'albero dei periodi, ovvero mediante gli operatori di sostituzione U_0 ed U_1 definiti da

$$U_0 : \begin{cases} 0 \mapsto 0 \\ 1 \mapsto 01 \end{cases} \qquad U_1 : \begin{cases} 0 \mapsto 01 \\ 1 \mapsto 1 \end{cases}$$

Applicando U_0 alle stringhe che compaiono nell'albero dei periodi tutti i blocchi di zeri compresi tra due 1 consecutivi si allungano di una unità (in particolare non ci sono più blocchi 11), e l'intero albero viene mandato nel sottoalbero di sinistra (p.es. $U_0(01) = 001$, $U_0(01011) = 00100101$) mentre applicando U_1 si allungano i blocchi di 1 tra due zeri consecutivi (in particolare spariscono i blocchi del tipo 00) e l'intero albero viene mandato nel sottoalbero di destra. Questo fatto si verifica facilmente sui primi elementi dell'albero, ed utilizzando l'induzione possiamo mostrare che questo vale a tutti i livelli dell'albero dei periodi.

Osserviamo che sul sottoalbero di sinistra, che contiene sequenze dove la cifra 1 è isolata, possiamo applicare l'inverso dell'operatore U_0, accorciando di un carattere i blocchi di 0 consecutivi; simmetricamente, sul sottoalbero di destra possiamo applicare l'inverso dell'operatore U_1,

accorciando di un carattere i blocchi di 1 consecutivi. Pertanto possiamo facilmente individuare la posizione di qualunque periodo all'interno dell'albero procedendo a ritroso, fino ad arrivare alla radice 01:

$$00010010001001001 \overset{U_0^{-1}}{\mapsto} 001010010101 \overset{U_0^{-1}}{\mapsto} 0110111 \overset{U_1^{-1}}{\mapsto} 01011 \overset{U_1^{-1}}{\mapsto} 001 \overset{U_0^{-1}}{\mapsto} 01$$

Da questo deduciamo che $00010010001001001 \in F_6$ e per raggiungerlo partendo da 01 devo scendere due volte verso sinistra, poi due volte verso destra, ed infine una volta verso sinistra.

Conoscere la posizione di un periodo nell'albero è utile anche per dare un algoritmo per determinare la sua fattorizzazione canonica. Per esempio, utilizzando le informazioni che abbiamo ottenuto sopra possiamo dedurre la fattorizzazione canonica di 00010010001001001 da quella di 01:

$$00010010001001001 = U_0U_0U_1U_1U_0(01) = U_0U_0U_1U_1U_0(0)\ U_0U_0U_1U_1U_0(1)$$

da cui si ricavano i fattori

$$U_0U_0U_1U_1U_0(0) = 0001001 \qquad \text{e} \qquad U_0U_0U_1U_1U_0(1) = 0001001001.$$

Fin qui abbiamo parlato prevalentemente di rette con coefficiente angolare razionale, ma quanto detto sopra ha conseguenze interessanti anche per rette con coefficiente angolare α qualunque. Non è difficile dimostrare che se $qm - pn = 1$ allora $\alpha \in (p/q, m/n)$ se e solo se la sequenza di taglio di una retta di coefficiente angolare α si può esprimere come una concatenazione infinita (non necessariamente periodica, se α è irrazionale) delle stringhe $W(p/q)$ e $W(m/n)$. Questo fatto può essere utilizzato per dare una stima dall'alto e dal basso di α qualora sia noto un segmento della sequenza di taglio della retta.

Le sequenze di taglio di rette di coefficente $\alpha \notin \mathbb{Q}$ si chiamano *sequenze sturmiane*; esse non sono periodiche, ma rappresentano le sequenze binarie infinite di complessità minima (dopo le periodiche). Infatti, data una sequenza binaria infinita (ξ_k), per ogni intero naturale N possiamo considerare l'insieme $\mathcal{L}_N$ di tutti i blocchi di N caratteri consecutivi che compaiono all'interno della sequenza data. Per una generica sequenza binaria l'insieme $\mathcal{L}_N$ ha certamente un numero finito di elementi (può avere al massimo 2^N elementi), tuttavia per la sequenza di taglio di una retta non razionale siamo molto lontani da questo limite teorico. Infatti una sequenza è sturmiana se e solo se $\mathcal{L}_N$ ha esattamente $N + 1$ elementi per ogni N intero positivo.

18.7 Ancora domande!

Spesso lo sforzo per sviscerare un argomento produce più domande che risposte. Non è necessariamente un male, pertanto ne elenchiamo alcune.

1. Partendo da 1/2 e scendendo a zig-zag lungo l'albero di Stern-Brocot troviamo:

$$\frac{1}{2} \searrow \frac{1}{3} \nearrow \frac{2}{5} \searrow \frac{3}{8} \nearrow \frac{5}{13} \searrow \frac{8}{21} \nearrow \frac{13}{34} \searrow \dots$$

   Ogni coppia di elementi consecutivi di questa successione è irriducibile; tali coppie determinano una successione di intervalli incapsulati che si stringono attorno ad un valore non razionale. Determinarlo.

2. Mostrare che qualunque cammino con infiniti zig-zag determina una successione di intervalli incapsulati che individua univocamente un numero reale.

3. Dire se la stringa 000101 può apparire nella sequenza di taglio di una retta.

4. Il paradosso della sezione 18.2 era legato alla presenza del parallelogramma fantasma generato dai vettori $(8, 3)$ e $(5, 2)$. Questi vettori corrispondono a due frazioni nel cammino a zig-zag dell'esercizio 1. Utilizzare questo fatto per riprodurre lo stesso paradosso in tassellazioni di dimensione maggiore (p.es. ricomponendo opportuni tasselli di un quadrato 13×13 nel rettangolo 21×8 – in questo caso è l'area del quadrato ad essere leggermente maggiore...).

5. Per ogni intero positivo chiamiamo F_n la lista (ordinata in maniera crescente) di tutte le frazioni p/q che, scritte in forma ridotta, hanno denominatore $q \leq n$. Per esempio

$$F_5 = \left[\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{2}{5}, \frac{1}{3}, \frac{1}{2}, \frac{3}{5}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}\right].$$

   Mostrare che ogni coppia di elementi adiacenti in F_n è irriducibile.

6. Proponiamo un algoritmo per trovare i genitori di una frazione $p/q \in [0, 1]$. Una possibile strategia si basa sugli sviluppi in frazione continua; ne diamo un esempio nel caso di $7/10$.

   Quozienti parziali. Determiniamo lo sviluppo in frazione continua di $7/10$ applicando ripetutamente l'algoritmo di divisione con resto:

$$\begin{cases} 10 = 7 \times \mathbf{1} + 3, \\ 7 = 3 \times \mathbf{2} + 1, \\ 3 = 1 \times \mathbf{3} + 0. \end{cases} \qquad \text{Di conseguenza:} \quad \frac{7}{10} = \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{3}}}.$$

   Genitore 1. Il primo genitore è ottenuto dalla frazione continua di $7/10$ privata dell'ultimo termine:

$$\cfrac{1}{1 + \cfrac{1}{2}} = \frac{2}{3}.$$

   Genitore 2. Si ottiene "per differenza": $\frac{7-2}{10-3} = \frac{5}{7}$. Si noti che $\dfrac{2}{3}$ e $\dfrac{5}{7}$ sono una coppia irriducibile: $5 \times 3 - 2 \times 7 = 1$.

   Mostrare che questo algoritmo permette di trovare i genitori di qualsiasi numero razionale.

7. Si consideri l'albero dei razionali diadici, associato alle liste

$$D_1 = [0, 1], \quad D_2 = [0, 1/2, 1], \quad D_3 = [0, 1/4, 1/2, 3/4, 1],$$
$$D_4 = [0, 1/8, 1/4, 3/8, 1/2, 5/8, 3/4, 7/8, 1], ...$$

   Per ogni elemento dell'albero disegniamo nel quadrato $[0, 1] \times [0, 1]$ il punto che ha come ordinata l'elemento considerato e come ascissa il corrispondente elemento nel sottoalbero sinistro dell'albero di Stern-Brocot (p.es. alcuni punti saranno dati da $(1/2, 1/2)$, $(1/3, 1/4)$, $(3/5, 5/8)$,...). Mostrare che i punti così ottenuti stanno sul grafico di una funzione continua e strettamente crescente (grafico che può essere visto come "completamento" dei punti così generati).

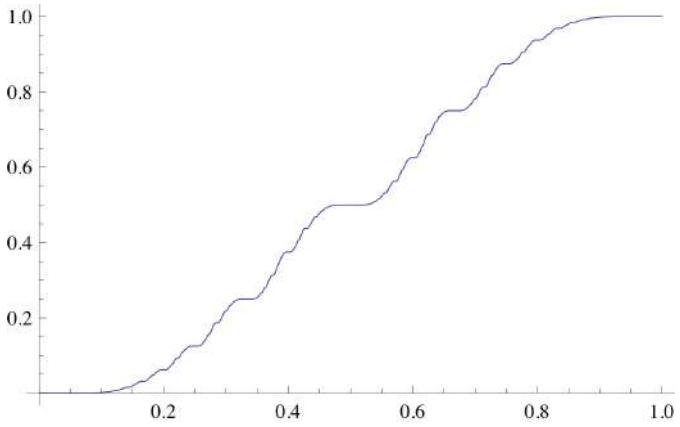


Figura 55: Unendo i puntini si ottiene il grafico della funzione "punto di domanda" di Minkowski, anche conosciuta come "slippery devil's staircase".

8. Giovanni gioca a tirare tiri liberi a basket: fallisce il primo tiro, ma dopo un certo numero di tiri la percentuale di successi supera l'80%. (a) Esiste un momento intermedio in cui la percentuale di successi è esattamente pari a 80% ? (b) Stesso problema, sostituendo 80% con 60%.

9. Un noto teorema del matematico Georg Alexander Pick esprime l'area A di un poligono con vertici a coordinate intere con la formula

$$A = N_i + \frac{N_b}{2} - 1$$

dove N_i è̀del numero di nodi interni e N_b quello dei nodi che stanno sul bordo del poligono. Per esempio se $np - mq = 1$ il triangolo generato di vertici $(0,0)$, (q,p) e (n,m) contiene punti nodali solo in corrispondenza dei vertici, infatti ha area $A = 0 + \frac{3}{2} - 1 = \frac{1}{2}$. Provare quest'ultima formula e dedurne il teorema di Pick per un generale poligono con vertici interi.

19 Il Lemma di Sperner e il Teorema di Monsky

Luca Bruni, n.10, Febbraio 2020

19.1 Introduzione

Cominciamo con l'enunciare un problema geometrico:

Problema 1. Suddividere un quadrato in un numero pari di triangoli di uguale area.

Una semplice soluzione del problema potrebbe essere la seguente: supponiamo di voler dividere il quadrato in $2n$ triangoli della stessa area; allora dividiamo i lati orizzontali del quadrato in n segmenti di lunghezza uguale, tracciamo gli n rettangoli individuati dai punti e suddividiamo ogni rettangolo in due triangoli mediante la diagonale.

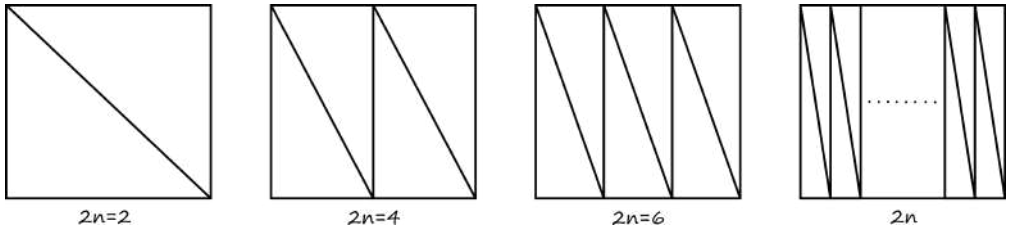


Figura 56: Suddivisione di un quadrato in un numero pari di triangoli con la stessa area.

Proviamo adesso a risolvere il seguente problema il cui enunciato è molto simile:

Problema 2. Suddividere un quadrato in un numero dispari di triangoli di uguale area.

Dopo alcuni tentativi ci rendiamo conto che il problema è notevolmente più complicato del precedente e sembra molto difficile trovarne una soluzione.

Il primo a pensare a questo problema fu, nel 1965, il matematico Fred Richman[24], il quale avrebbe voluto includere questo quesito nel testo di un esame, ma, non riuscendo a risolverlo e non trovando alcuna referenza al riguardo, decise invece di proporlo pubblicamente nella rivista *American Mathematical Monthly*[25]. Pur nella sua semplicissima formulazione, il problema rimase irrisolto per ben cinque anni: il primo a fornire una risposta al riguardo fu il matematico Paul Monsky[26] nel 1970. Il problema non ha soluzione: Monsky dimostrò l'impossibilità di suddividere un quadrato in un numero dispari di triangoli di ugual area combinando alcune tecniche combinatorie e algebriche. Quello che cercheremo di fare in questo articolo è ripercorrere la brillante dimostrazione di Monsky illustrandone i seguenti passi:

- Triangolazioni di poligoni, colorazioni e lemma di Sperner.

[24]Fred Richman, matematico americano, ha ricoperto il ruolo di docente presso la "New Mexico State University" e in seguito presso la "Florida Atlantic University".

[25]Rivista di matematica fondata da Benjamin Finkel nel 1894. Attualmente viene pubblicata 10 volte all'anno dalla Mathematical Association of America. Contiene numerosi articoli di ampio interesse rivolti a tutta la comunità matematica.

[26]Paul Monsky, nato il 17 giugno 1936. Matematico americano, ha ricoperto il ruolo di docente alla "Brandeis University" presso Boston, Massachusetts.

- Colorazione del piano cartesiano e in particolare del quadrato I di coordinate $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$ con l'ausilio della *norma* 2-*adica*.
- L'"area" di un opportuno triangolo all'interno del quadrato è "troppo grande".

19.2 Il Lemma di Sperner

Lo scopo di questa sezione è enunciare e dimostrare il Lemma di Sperner e dare qualche informazione sulle sue applicazioni; anche se a prima vista sembra solamente un fatto curioso, esistono risultati matematici (e non solo) di notevole interesse basati su questo simpatico lemma.

19.2.1 Il Lemma di Sperner

Cominciamo con qualche definizione introduttiva; d'ora in avanti assumeremo sempre che i poligoni siano *semplici*, ovvero che i lati del poligono non si intersechino tra di loro.

Dato un poligono P una *triangolazione di* P è una suddivisione di P in un numero finito di triangoli in modo che ogni due triangoli della suddivisione si intersechino o esattamente in un lato o esattamente in un vertice.

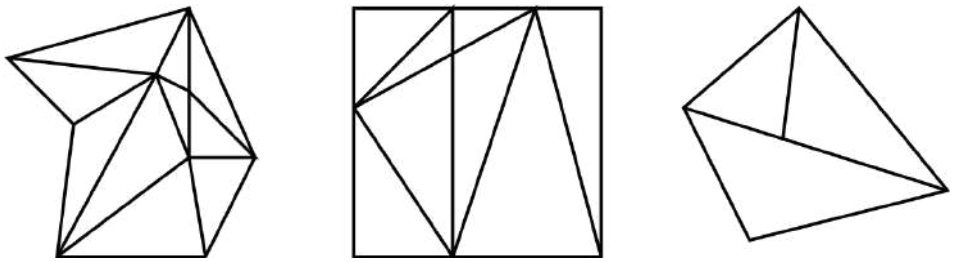


Figura 57: Le prime due figure da sinistra sono esempi di triangolazioni. La terza figura **non** è una triangolazione.

Una triangolazione è detta *colorata* se ogni vertice della triangolazione è colorato con il numero 1, 2 o 3; in questo caso chiameremo un lato della triangolazione 12-*lato*, 23-*lato* o 31-*lato* se i vertici del lato che stiamo considerando sono colorati rispettivamente di (1-2), (2-3), (3-1). Infine, data una triangolazione, diremo che un triangolo è *completo* se ha tutti i vertici di colore diverso (si veda la Figura 58).

Dopo queste piccole nozioni preliminari siamo pronti per enunciare il lemma di Sperner:

Lemma 1 (Sperner, 1928)**.** *Sia P un poligono e sia data una sua triangolazione colorata. Allora il numero di triangoli completi ha la stessa parità del numero di* 12-*lati sul bordo del poligono.*

Prima di inoltrarci nella dimostrazione vera e propria osserviamo questa curiosa conseguenza: se nella triangolazione colorata c'è un numero dispari di 12-lati nel bordo, allora **c'è almeno un triangolo completo**. Pertanto nelle tre figure qui sotto, anche se non conosciamo come sono colorati i vertici centrali, siamo comunque certi di trovare da qualche parte un triangolo completo!

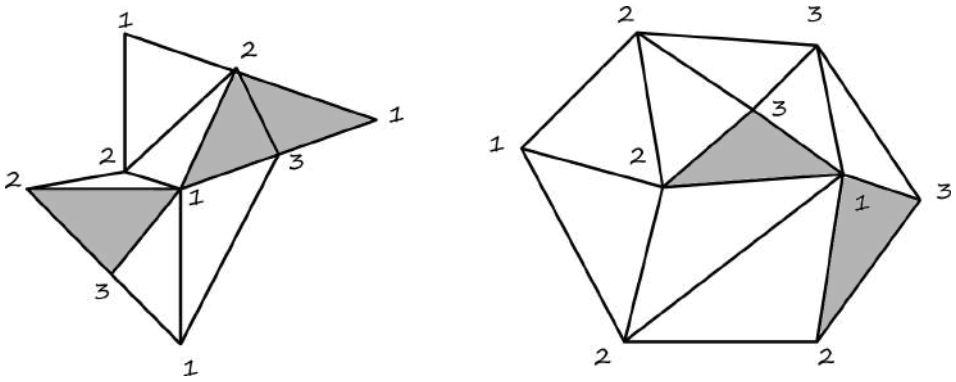


Figura 58: Esempi di colorazioni e di triangoli completi (in grigio).

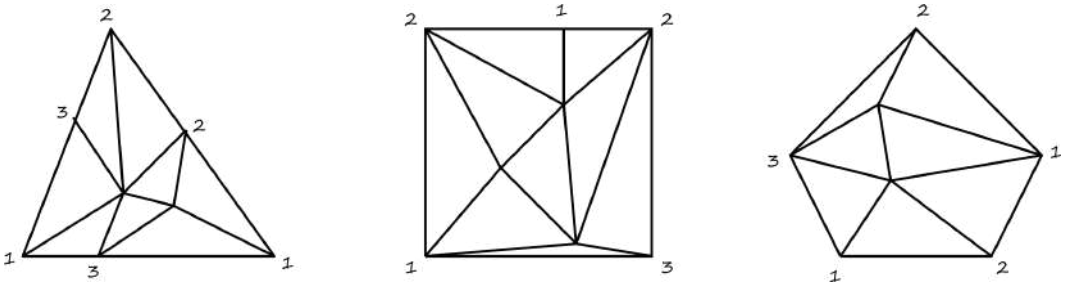


Figura 59: In qualsiasi modo completiamo la colorazione dei poligoni in figura, siamo sicuri che ci sarà almeno un triangolo completo.

Dimostrazione. Si tratta di un doppio conteggio: mettiamo un punto da ogni parte di un 12-segmento (si veda la Figura 60). Quello che vogliamo fare è contare il numero di punti nella parte interna del triangolo e mostrare che la sua parità è uguale sia a quella del numero dei triangoli completi, sia a quella del numero di 12-lati sul bordo. Notiamo che ogni segmento che **non** si trova sul bordo del poligono contribuisce o per 0 punti (se non è un 12-lato) o per 2 punti (se è un 12-lato); invece un segmento che si trova sul bordo contribuisce per 0 punti o per 1 punto a seconda che non sia un 12-lato o che lo sia. Abbiamo in pratica dimostrato che *il numero di punti nell'interno del poligono ha la stessa parità del numero di* 12*-lati che si trova sul bordo.* Adesso contiamo il numero di punti che si trova all'interno di ogni triangolo della triangolazione. I triangoli completi hanno per costruzione un solo 12-lato e di conseguenza al loro interno avranno soltanto un punto. Un qualsiasi altro triangolo della triangolazione, invece, avrà un numero pari di punti al proprio interno come si può facilmente osservare considerando i pochi casi possibili (si veda anche la Figura 60).

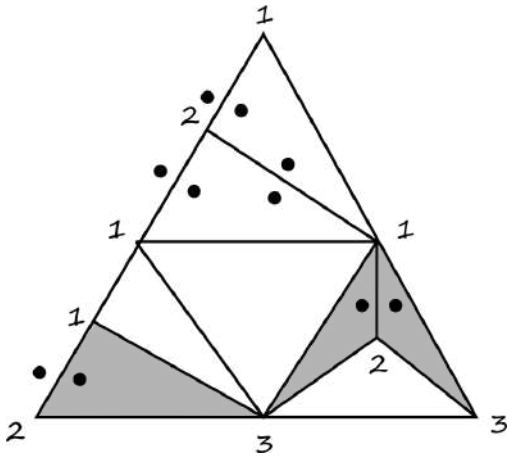


Figura 60: Il procedimento dimostrativo di double-counting nel caso di un triangolo.

Ma allora abbiamo anche mostrato che *il numero di triangoli completi ha la stessa parità nel numero di punti interni del poligono.* Combinando l'informazione appena trovata con quella precedente si ottiene la tesi. □

Basandoci sulla dimostrazione del Lemma di Sperner, è possibile fornire il seguente metodo per "scovare" dove sono i triangoli completi nel caso in cui questi siano in numero dispari: prendiamo un poligono qualunque e facciamone una triangolazione. Coloriamo i vertici in modo che ci sia un numero dispari di 12-lati sul bordo. Grazie all'osservazione fatta sopra siamo sicuri che da qualche parte ci sia un triangolo completo. Immaginiamoci che i triangoli della triangolazione siano isole e costruiamo un *ponte* tra due isole se e solo se il lato che le separa è un 12-lato. Allo stesso modo costruiamo un ponte tra un'isola e l'esterno del poligono se e solo se il lato che li separa è un 12-lato. Se adesso proviamo a percorrere i sentieri formati dai ponti che partono dall'esterno del poligono siamo sicuri di trovare almeno un triangolo completo (lasciamo al lettore il piacere di spiegare perché questo accada).

Prima di andare oltre, vogliamo soffermarci su un altro fatto che sarà cruciale nella dimostrazione del teorema di Monsky: grazie al seguente risultato (e a una opportuna colorazione che vedremo nella prossima sezione) saremo infatti in grado di utilizzare il Lemma di Sperner su una generica triangolazione del quadrato I del piano cartesiano trovando al suo interno un triangolo

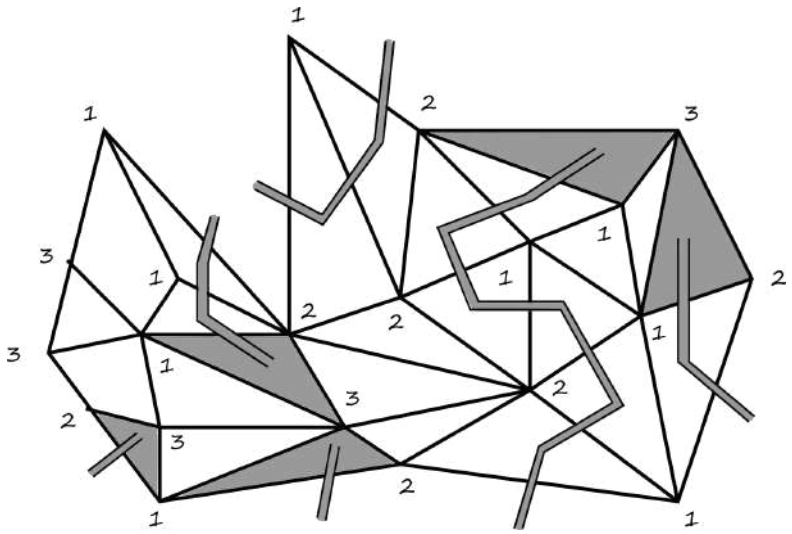


Figura 61: Il metodo dei *ponti* per la ricerca dei triangoli completi.

completo. Anche in questo caso, la dimostrazione è dovuta a Sperner ed è interpretabile come una versione uno-dimensionale del lemma che abbiamo già analizzato.

Lemma 2 (Lemma di Sperner unidimensionale)**.** *Sia S un* 12*-segmento. Suddividiamo S in lati più piccoli colorando i vertici della suddivisione con i colori* 1 *e* 2*; allora il numero di* 12*-lati della suddivisione è dispari.*

Dimostrazione. Si può mostrare questo risultato per induzione sul numero di punti della suddivisione: se il segmento è suddiviso da 0 punti, allora ho esattamente un 12-lato (coincide con S stesso) e dunque la tesi è vera. Supponiamo adesso che la tesi sia vera per ogni suddivisione di n punti e mostriamo che è vera per una di $n+1$. Qualsiasi suddivisione con $n+1$ punti possiamo ottenerla da una suddivisione con n punti aggiungendone semplicemente uno da qualche parte. Analizziamo cosa succede alla parità nei vari casi (si veda la Figura 62):

- *Aggiungiamo un punto di colore* 1 *tra due di colore* 1*:* si formano due 11-segmenti e si cancella un 11-lato e dunque il numero di 12-lati non cambia.

- *Aggiungiamo un punto di colore* 2 *tra due punti di colore* 2*:* analogo al precedente.

- *Aggiungiamo un punto di colore* 1 *tra due di colore* 2 *(o viceversa):* si formano due 12-lati e si cancella un 11-lato. Dunque non si altera la parità del numero degli 12-lati.

- *Aggiungiamo un punto di colore* 1 *tra uno di colore* 1 *e uno di colore* 2 *(o viceversa):* si forma un 12-lato e un 11-lato e si cancella un 12-lato. Dunque il numero di 12-lati non cambia.

Dunque aggiungendo un punto colorato la parità non cambia e possiamo concludere grazie all'ipotesi induttiva. □

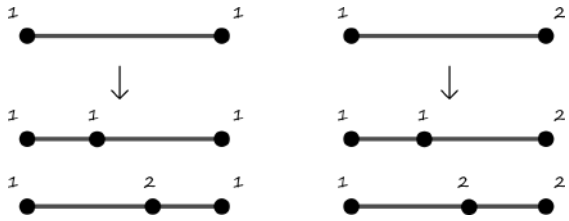


Figura 62: I possibili scenari sopo aver aggiunto un punto alla suddivisione.

19.2.2 Possibili applicazioni

Il risultato enunciato è soltanto la versione bidimensionale del lemma, ma esistono delle generalizzazioni nel caso multidimensionale[27].

Anche se a prima vista sembra solamente un bizzarro risultato, si può dimostrare che il lemma di Sperner è equivalente a un importante risultato di topologia algebrica noto come *Teorema del punto fisso di Brouwer*[28].

Oltre che nell'applicazione che vedremo di seguito, il lemma di Sperner è stato utilizzato anche per implementare alcuni algoritmi per il calcolo di *zeri di funzioni*[29]. Ma non solo: lo si usa anche per la risoluzione di alcuni problemi di *"divisione equa delle risorse"*. Supponiamo che le nostre risorse siano modellizzate da una torta; quello che vogliamo fare è suddividere la torta in tanti pezzi quante sono le persone tra cui vogliamo distribuire le risorse e che ogni persona ritenga che il proprio pezzo sia "migliore" (in base alle proprie valutazioni) di quello di tutti gli altri. Grazie al lemma di Sperner sono stati implementati dei metodi per risolvere questo problema[30].

Un'ultima applicazione è in ambito economico: si può utilizzare tale risultato per trovare situazioni di *"equilibrio"* in transazioni finanziarie.

19.3 Coloriamo il piano

Scopo di questa sezione è riuscire a colorare tutto il piano cartesiano in modo accorto per applicare il lemma di Sperner al quadrato $I = [0,1] \times [0,1]$. Per farlo abbiamo bisogno questa volta di qualche risultato algebrico:

19.3.1 La norma 2-adica

La prima cosa che facciamo è dire che cosa sia la norma 2-adica per l'insieme $\mathbb{Q}$ dei numeri razionali.

Sia $x \in \mathbb{Q} \setminus \{0\}$. Osserviamo che è possibile scrivere x in modo unico come $x = 2^n \frac{a}{b}$ con $a \in \mathbb{N}$ e $b, n \in \mathbb{Z}$ tali che a, b siano primi tra loro e non divisibili per 2. Definiamo la *valutazione 2-adica di x* come

$$v_2(x) = v_2\left(2^n \frac{a}{b}\right) = n.$$

Se $x = 0$ poniamo $v_2(0) = \infty$. Definiamo inoltre la *norma 2-adica di x* come:

$$|x|_2 = 2^{-v_2(x)}$$

[27] Wikipedia, Sperner's Lemma.
[28] Wikipedia, Brouwer fixed-point theorem.
[29] Wikipedia, Simmons-Su protocols.
[30] Wikipedia, Competitive equilibrium.

Nel caso in cui $x = 0$ poniamo $|x|_2 = 0$.

La definizione sembra molto astratta, ma facciamo alcuni esempi per capire meglio cosa succede:

$$\left|\frac{37}{12}\right|_2 = 2^{-v_2(2^{-2}\frac{37}{3})} = 2^2 = 4$$
$$\left|\frac{3}{5}\right|_2 = 2^{-v_2(2^0\frac{3}{39})} = 2^{-0} = 1$$
$$|1024|_2 = 2^{-v_2(2^{10}\cdot 1)} = 2^{-10} = \frac{1}{1024}$$

Facciamo inoltre un'osservazione cruciale che ci servirà più avanti: *la norma 2-adica di un qualsiasi numero intero dispari d è uguale a 1*. Infatti un numero intero dispari è della forma $2^0\frac{d}{1}$ e dunque la sua norma 2-adica vale 1.

La norma 2-adica gode di alcune proprietà algebriche interessanti; ne vediamo alcune che saranno utili più avanti:

1. $|x|_2 > 0$ per ogni $x \in \mathbb{Q} \setminus \{0\}$ e $|0|_2 = 0$.
   Infatti la norma 2-adica è una potenza di 2 e dunque sempre maggiore di zero. Il fatto che $|0|_2 = 0$ segue direttamente dalle definizioni.
2. $|x|_2|y|_2 = |xy|_2$.
   Infatti siano $x = 2^n\frac{a}{b}$ e $y = 2^m\frac{a'}{b'}$ con a, b, a', b' come sopra, allora, poiché $xy = 2^{n+m}\frac{aa'}{bb'}$,
   $$|xy|_2 = 2^{-v_2(xy)} = 2^{-(n+m)} = 2^{-n}\cdot 2^{-m} = 2^{-v_2(x)}2^{-v_2(y)} = |x|_2|y|_2.$$
3. $|x+y|_2 \leq \max\{|x|_2, |y|_2\}$. Inoltre se $|x|_2 < |y|_2$ allora $|x+y|_2 = |y|_2$. Infatti siano $x = 2^n\frac{a}{b}$ e $y = 2^m\frac{a'}{b'}$, allora $x + y = 2^n\big(\frac{a}{b} + 2^{m-n}\frac{a'}{b'}\big)$. Dato che stiamo supponendo $n \leq m$, allora la quantità tra parentesi può contenere fattori 2 solo al numeratore. Di conseguenza,
   $$|x + y|_2 = 2^{-v_2(x+y)} \leq 2^{-n} = \max\{|x|_2, |y|_2\}.$$
   In particolare se vale la disuguaglianza stretta tra le norme, allora il minore uguale diventa un uguale e si ha la tesi.

Concludiamo facendo notare che le proprietà dimostrate possono essere enunciate in un ambito molto più generale: il numero primo 2 non ha nessuna caratteristica speciale rispetto agli altri primi p in questo contesto; è infatti possibile definire una valutazione e una norma p-adica nel medesimo modo e le proprietà sarebbero comunque verificate.

19.3.2 Il teorema di Chevalley

Come già annunciato più volte, il nostro scopo è colorare il piano: lo vogliamo fare mediante la norma 2-adica delle coordinate dei punti del piano. Sorge però un problema: abbiamo definito la norma solamente sui punti del piano che hanno entrambe coordinate le razionali; quello che vorremmo fare è estendere tale valutazione a tutti i punti con coordinate reali e vorremmo che

tale estensione continuasse a verificare le proprietà $(1), (2), (3)$. In alcuni casi questa estensione è naturale: se vogliamo definire $|\sqrt{2}|_2$ osserviamo che deve valere che $|\sqrt{2}|_2|\sqrt{2}|_2 = |2|_2 = \frac{1}{2}$; ne ricaviamo che $|\sqrt{2}|_2 = \frac{1}{\sqrt{2}}$. Con questo semplice sistema possiamo estendere la norma 2-adica a tutte le radici, ma come ben sappiamo i numeri reali non sono solamente radici; come facciamo dunque ad estendere la norma 2-adica a numeri come π? Per fare questo ci viene in aiuto un importante teorema dovuto al matematico Claude Chevalley[31]. Enunciare e dimostrare il teorema nella sua forma generale necessita di strumenti matematici piuttosto avanzanti e pertanto enunciamo, senza dimostrarlo, soltanto il corollario di cui abbiamo bisogno:

Theorem 1 (Chevalley). E' possibile estendere la norma 2-adica definita sui numeri razionali a tutti i numeri reali $\mathbb{R}$ in modo che le proprietà $(1), (2), (3)$ siano ancora verificate.

19.3.3 La colorazione

Grazie alla norma 2-adica siamo finalmente pronti a colorare tutto il piano. Prendiamo un'estensione della norma 2-adica a tutti i numeri reali e partizioniamo il piano in 3 parti che coloriamo rispettivamente di 1, 2 e 3:

$$\begin{aligned} S_1 &:= \{(x,y) \ : |x|_2 < 1, |y|_2 < 1\} \\ S_2 &:= \{(x,y) \ : |x|_2 \geq 1, |x|_2 \geq |y|_2\} \\ S_3 &:= \{(x,y) \ : |y|_2 \geq 1, |y|_2 > |x|_2\} \end{aligned}$$

Il quadrato che vogliamo analizzare è $I = [0,1] \times [0,1]$ che ha vertici nei punti $A = (0,0), B = (1,0), C = (1,1), D = (0,1)$; cominciamo col chiederci di che colore sono questi punti; come già osservato $|0|_2 = 0$ e $|1|_2 = 1$ e dunque otteniamo:

$$\begin{aligned} A = (0,0) \in S_1 \qquad B = (1,0) \in S_2 \\ C = (1,1) \in S_2 \qquad D = (0,1) \in S_3 \end{aligned}$$

Osserviamo che è impossibile realizzare un immagine raffigurante la colorazione in quanto la norma 2-adica si comporta in maniera strana rispetto al numero di cui stiamo facendo la norma: un numero molto grande come 2^{100} ha infatti una norma 2-adica molto piccola e, viceversa, ci sono numeri molto piccoli con norma 2-adica molto grande. Quello che vogliamo fare è però capire di che colore potrebbero essere i punti che stanno nei vari lati del quadrato. I punti del lato $\overline{AB}$ sono caratterizzati dal fatto che la coordinata y è uguale a 0; di conseguenza $|y|_2 = 0$ e dunque sicuramente tali punti **non** possono stare in S_3. Dunque il lato $\overline{AB}$ è formato solo da punti di colore 1 o 2. Con ragionamenti del tutto analoghi si osserva facilmente che:

Segmento	Colori possibili
AB	1 o 2
BC	2 o 3
CD	2 o 3
DA	3 o 1

Prima di inoltrarci nella dimostrazione vera e propria del teorema di Monsky, abbiamo bisogno di qualche altra proprietà della colorazione che abbiamo creato. In particolare vogliamo

[31] Claude Chevalley, 11 febbraio 1909 – 28 giugno 1984. Matematico francese; diede importanti contributi alla teoria dei numeri, alla geometria algebrica, alla teoria dei campi di classi, alla teoria dei gruppi finiti e alla teoria dei gruppi algebrici.

trovare una stima della norma 2-adica dell'area di un triangolo completo nel piano colorato. Per cominciare osserviamo che la colorazione 2-adica è invariante per traslazione rispetto a punti di colore 1 (cioè il colore di un punto non cambia se lo traslo per un vettore che identifica un punto di colore 1). La dimostrazione di questo fatto segue dalle proprietà della norma 2-adica che abbiamo già mostrato: sia $(a, b) \in S_1$, allora $|a|_2 = |-a|_2 < 1$ e $|b|_2 = |-b|_2 < 1$. Verifichiamo ora che se $(x, y) \in S_i$, allora $(x - a, y - b) \in S_i$.

$i = 1$) Per ipotesi $|x|_2 < 1$ e $|y|_2 < 1$. Allora $|x - a|_2 = |x + (-a)|_2$ è uguale alla norma maggiore fra $|x|_2$ e $|a|_2$ per le proprietà della norma. In ogni caso $|x - a|_2 < 1$. In maniera analoga si mostra $|y - b|_2 < 1$.

$i = 2$) Per ipotesi $|x|_2 \geq 1$ e $|x|_2 \geq |y|_2$. Allora $|x|_2 > |a|_2$ e dunque $|x + (-a)|_2 = |x|_2 \geq 1$. Inoltre $|y + (-b)|_2$ è uguale o a $|y|_2$ o a $|b|_2$; dunque in ogni caso $|x - a|_2 \geq |y - b|_2$.

$i = 3$) Per ipotesi $|y|_2 \geq 1$ e $|y|_2 > |x|_2$. Allora $|y|_2 > |b|_2$ e dunque $|y + (-b)|_2 = |y|_2 \geq 1$. Inoltre $|x + (-a)|_2$ è uguale o a $|x|_2$ o a $|a|_2$; dunque in ogni caso $|y - b|_2 > |x - a|_2$.

Un altro utile fatto di cui abbiamo bisogno è che l'area di un triangolo nel piano cartesiano con un vertice nell'origine si calcola con una comoda formula a partire dalle coordinate dei 3 punti: sia T un triangolo nel piano cartesiano con vertici $O = (0, 0)$, $P = (x_P, y_P)$, $Q = (x_Q, y_Q)$, allora:

$$Area(T) = \frac{1}{2}|x_P y_Q - x_Q y_P|.$$

Per convincercene osserviamo il disegno in Figura 63 (a meno di rinominare i punti possiamo supporre di avere proprio questa rappresentazione).

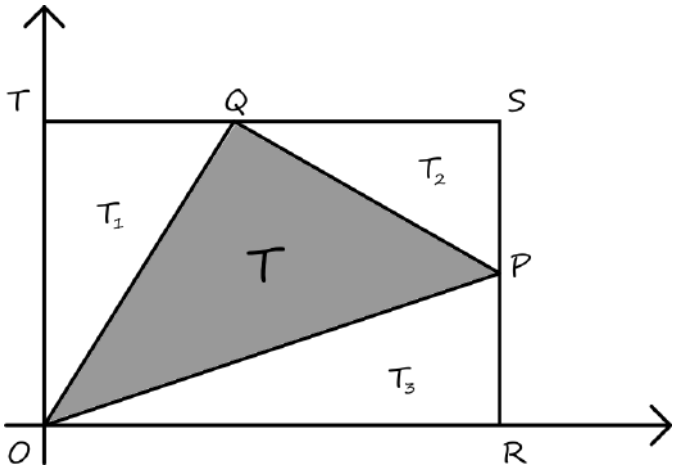


Figura 63: Area del triangolo nel piano cartesiano

Possiamo calcolare l'area del triangolo OPQ come differenza tra l'area del rettangolo $ORST$ e l'area dei triangoli rettangoli T_1, T_2, T_3:

$$Area(T) = Area(ORST) - Area(T_1) - Area(T_2) - Area(T_3) =$$
$$= x_P y_Q - \frac{1}{2}(x_Q y_Q + (y_Q - y_P)(x_P - x_Q) + x_P y_P) =$$
$$= \frac{1}{2}(x_P y_Q - x_Q y_P).$$

19.4 Il teorema di Monsky

19.4.1 Il teorema

Prima di arrivare al teorema vero e proprio osserviamo quest'ultimo interessante fatto preliminare:

Proposizione 1. *Sia T un triangolo completo in $\mathbb{R}^2$ rispetto alla colorazione 2-adica. Allora $|Area(T)|_2 > 1$.*

Dimostrazione. Grazie al fatto che la colorazione è invariante per traslazione rispetto ai punti di colore 1, possiamo supporre che il triangolo T abbia il vertice di colore 1 nell'origine, P di colore 2 e Q di colore 3 (o viceversa: dato che prendiamo il valore assoluto il risultato non cambierà). Per la formula dell'area che abbiamo ricavato si ha che $Area(T) = \frac{1}{2}|x_P y_Q - x_Q y_P|$. Passando dunque alla norma 2-adica si ha:

$$|Area(T)|_2 = \left|\frac{1}{2}\right|_2 |x_P y_Q - x_Q y_P|_2 = 2|x_P y_Q + (-x_Q y_P)|_2 =$$
$$= 2|x_P y_Q|_2 = 2|x_P|_2|y_Q|_2 \geq 2 > 1$$

Nel passaggio dalla prima alla seconda riga abbiamo usato il fatto che, per la colorazione, $|x_P|_2 \geq |y_P|_2$ e $|y_Q|_2 > |x_Q|_2$ e dunque $|x_P y_Q|_2 > |x_Q y_P|_2$. □

A questo punto abbiamo tutti gli ingredienti per poter dimostrare che il problema posto all'inizio dell'articolo non ha effettivamente soluzione:

Theorem 2 (Monsky, 1970). Sia S un quadrato; è possibile triangolare S in m triangoli di area uguale se e solo se m è pari.

Dimostrazione. Abbiamo già visto che se m è pari, allora è possibile suddividere il quadrato.

Viceversa senza perdita di generalità possiamo assumere che il nostro quadrato sia il effettivamente il quadrato I già considerato nel piano cartesiano. A questo punto supponiamo di avere una suddivisione del quadrato in m triangoli di uguale area con m dispari: vogliamo trovare un assurdo. Coloriamo i vertici della triangolazione mediante la colorazione 2-adica del piano. Come già osservato nel paragrafo 3.3, gli unici 12-lati nel bordo della triangolazione si possono trovare nel lato $\overline{AB}$: dato che $\overline{AB}$ ha i vertici colorati in modo diverso, per il lemma di Sperner unidimensionale sappiamo che c'è un numero dispari di 12-lati su $\overline{AB}$ e dunque su tutto il bordo. Dunque per il lemma di Sperner applicato al quadrato sappiamo che deve esistere almeno un triangolo completo nella triangolazione scelta! Chiamiamo T tale triangolo.

Adesso l'area totale del quadrato sarà pari a $\overline{AB} \cdot \overline{BC} = 1 \cdot 1 = 1 = m \cdot Area(T)$ in quanto ogni triangolo ha area uguale a quella di T per ipotesi. Ma passando alla norma 2-adica ($|m|_2 = 1$ poiché m è dispari) si ottiene:

$$1 = |1|_2 = |m \cdot Area(T)|_2 = |m|_2|Area(T)|_2 > 1$$

che è assurdo. □

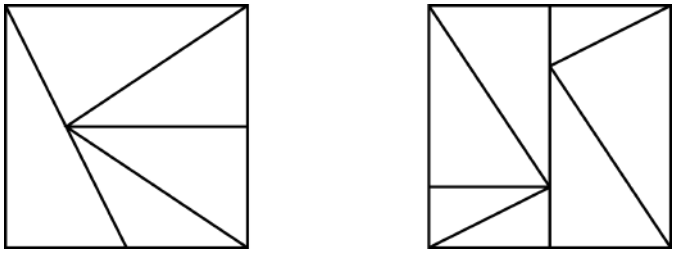


Figura 64: Suddivisioni di un quadrato.

19.4.2 Un piccolo problema

Il lettore più attento avrà notato che, in quanto sopra esposto, abbiamo considerato solamente le triangolazioni e non le *suddivisioni*; in particolare abbiamo considerato non valide le suddivisioni come in Figura 64.

L'argomento che abbiamo utilizzato si può generalizzare anche a queste situazioni, ma la dimostrazione risulta essere un po' più tecnica. Lasciamo ai più temerari una traccia di quello che dobbiamo dimostrare:

1. Ogni retta nel piano colorato non interseca una tra le 3 regioni S_1, S_2, S_3;
2. In una qualsiasi suddivisione in triangoli del quadrato I, esiste sempre un triangolo completo;
3. Conclusione come nel teorema di Monsky.

Per quanto riguarda il punto 1 si tratta di giocare con le proprietà della norma 2-adica; il punto 2 è quello un po' più delicato: data una suddivisione, si deve spezzare il quadrato in più poligoni semplici (in maniera opportuna grazie al punto 1) in cui è possibile applicare il lemma di Sperner; a quel punto, con un argomento di parità simile a quelli già visti, si può concludere che esiste un triangolo colorato.

Riferimenti bibliografici

[1] F. Su, *Rental Harmony: Sperner's Lemma in Fair Division*, The American Mathematical Monthly, Vol. 106, (1999), 930–942.

[2] P. Monsky, *On Dividing a Square into Triangles*, The American Mathematical Monthly, Vol. 77, No. 2, (1970), 161–164.

[3] M. Aigner and G. M. Ziegler, *Proofs from THE BOOK*, Fifth Edition, Springer, 2014, 151–158.

20 Approssimazioni razionali e frazioni continue

Francesco Ballini, n.11, Settembre 2020

In questo articolo cerchiamo di dare una risposta alle seguenti due domande:

1. Qual è il numero razionale più "semplice" compreso tra $\frac{109}{49}$ e $\frac{96}{43}$?
2. Quali sono i numeri razionali che approssimano "meglio" $\sqrt{2}$?

Diciamo che un numero è *razionale* se è esprimibile come frazione, cioè come rapporto fra due interi. Naturalmente bisognerebbe dare definizioni formali anche di "semplice" e di "meglio", cosa che però non faremo in questo articolo: ci limitiamo ad esplorare questi concetti tramite esempi e algoritmi di calcolo, sfruttando (e, speriamo, espandendo) l'idea intuitiva che ne ha il lettore.

20.1 Trovare numeri razionali "piccoli" in un intervallo

Cerchiamo il numero razionale più "semplice" fra $\frac{109}{49}$ e $\frac{96}{43}$. Dato che non sappiamo ancora chi è, chiamiamolo x. Abbiamo:

$$\frac{109}{49} < x < \frac{96}{43}$$

Osserviamo che $2 < \frac{109}{49} < \frac{96}{43} < 3$, pertanto ci sembra ragionevole sottrarre un 2 dall'equazione:

$$\frac{109}{49} - 2 < x - 2 < \frac{96}{43} - 2$$
$$\frac{11}{49} < x - 2 < \frac{10}{43}$$

Notiamo quindi che $0 < \frac{11}{49} < \frac{10}{43} < 1$; non sembra conveniente aggiungere o togliere qualche intero, ma possiamo considerare i reciproci di tutte le quantità in gioco, ovviamente invertendo i versi delle disuguaglianze:

$$\frac{49}{11} > \frac{1}{x-2} > \frac{43}{10}$$

A questo punto possiamo applicare di nuovo la nostra strategia precedente, osservando che $5 > \frac{49}{11} > \frac{43}{10} > 4$. Ma allora togliamo 4 dalla disuguaglianza di sopra, ottenendo:

$$\frac{49}{11} - 4 > \frac{1}{x-2} - 4 > \frac{43}{10} - 4$$
$$\frac{5}{11} > \frac{1}{x-2} - 4 > \frac{3}{10}$$

Abbiamo dunque $1 > \frac{5}{11} > \frac{3}{10} > 0$, quindi possiamo di nuovo considerare gli inversi, anche stavolta invertendo i versi delle disuguaglianze:

$$\frac{11}{5} < \frac{1}{\frac{1}{x-2} - 4} < \frac{10}{3}$$

A questo punto non abbiamo più $\frac{11}{5}$ e $\frac{10}{3}$ compresi tra due interi consecutivi; abbiamo invece $\frac{11}{5} < 3 < \frac{10}{3}$. Perciò, possiamo legittimamente sospettare che, se x è il numero razionale più "semplice" compreso tra $\frac{109}{49}$ e $\frac{96}{43}$, allora anche $\dfrac{1}{\dfrac{1}{x-2}-4}$ sia il numero razionale più "semplice" tra $\frac{11}{5}$ e $\frac{10}{3}$; questo numero è dunque 3 (difatti $\frac{11}{5} = 2.2$ e $\frac{10}{3} = 3.33...$), perciò poniamo:

$$\frac{1}{\dfrac{1}{x-2}-4} = 3$$

da cui

$$\frac{1}{x-2} - 4 = \frac{1}{3}$$
$$\frac{1}{x-2} = \frac{13}{3}$$
$$x - 2 = \frac{3}{13}$$
$$x = \frac{29}{13}$$

Si può effettivamente verificare che $\frac{29}{13}$ sia il numero razionale più "semplice" tra $\frac{109}{49}$ e $\frac{96}{43}$: infatti

$$\frac{109}{49} = 2.22448979...$$
$$\frac{96}{43} = 2.23255813...$$

E, ad esempio, nessuna frazione il cui denominatore sia 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 o 12 può essere compresa tra 2.224 e 2.233.

Cerchiamo ora di vedere la nostra strategia sotto una diversa luce: per quanto riguarda le nostre quantità, abbiamo sottratto interi affinché risultassero comprese fra 0 e 1, per poi trovarne l'inverso e quindi ripetere la nostra strategia. Vediamo cosa succede applicando questo algoritmo singolarmente a $\frac{109}{49}$, senza x. Scriviamo:

$$\alpha = \frac{109}{49}$$

Poiché $2 < \frac{109}{49} < 3$, sottraiamo **2**:

$$\alpha - 2 = \frac{11}{49}$$

Invertiamo:

$$\frac{1}{\alpha - 2} = \frac{49}{11}$$

Poiché $4 < \frac{49}{11} < 5$, sottraiamo **4**:

$$\frac{1}{\alpha-2}-4=\frac{5}{11}$$

Invertiamo:

$$\frac{1}{\frac{1}{\alpha-2}-4}=\frac{11}{5}$$

Poiché $2<\frac{11}{5}<3$, sottraiamo **2**:

$$\frac{1}{\frac{1}{\alpha-2}-4}-2=\frac{1}{5}$$

Invertiamo:

$$\frac{1}{\frac{1}{\frac{1}{\alpha-2}-4}-2}=5$$

Siamo dunque giunti ad un intero, **5**. Proviamo a ricalcolare α (che è $\frac{109}{49}$) preservando la struttura annidata della nostra frazione:

$$\frac{1}{\frac{1}{\frac{1}{\alpha-2}-4}-2}=5$$

$$\frac{1}{\frac{1}{\alpha-2}-4}-2=\frac{1}{5}$$

$$\frac{1}{\frac{1}{\alpha-2}-4}=2+\frac{1}{5}$$

$$\frac{1}{\alpha-2}-4=\frac{1}{2+\frac{1}{5}}$$

$$\frac{1}{\alpha-2}=4+\frac{1}{2+\frac{1}{5}}$$

$$\alpha-2=\frac{1}{4+\frac{1}{2+\frac{1}{5}}}$$

$$\alpha = 2 + \cfrac{1}{4 + \cfrac{1}{2 + \cfrac{1}{5}}}$$

$$\frac{109}{49} = 2 + \cfrac{1}{4 + \cfrac{1}{2 + \cfrac{1}{5}}}$$

Questa scrittura di $\frac{109}{49}$ prende il nome di *frazione continua*. Notiamo che la sequenza 2, 4, 2, 5 è esattamente la sequenza degli interi precedentemente scritti in grassetto, cioè gli interi che abbiamo sottratto per rendere la nostra quantità compresa tra 0 e 1 (e il 5 finale). Tale sequenza ha anche un'altra origine: proviamo a calcolare l'MCD tra 109 e 49 utilizzando l'algoritmo di Euclide. Eseguiamo la prima divisione con resto:

$$109 = \mathbf{2} \cdot 49 + 11$$

Da cui otteniamo 49 e 11. Abbiamo quindi:

$$49 = \mathbf{4} \cdot 11 + 5$$

Da cui 11 e 5, perciò:

$$11 = \mathbf{2} \cdot 5 + 1$$

E proseguiamo con un ulteriore passo con 5 e 1:

$$5 = \mathbf{5} \cdot 1 + 0$$

Otteniamo quindi la stessa sequenza $2, 4, 2, 5$.

Con analoghi conti possiamo calcolare la frazione continua di $\frac{96}{43}$:

$$\frac{96}{43} = 2 + \cfrac{1}{4 + \cfrac{1}{3 + \cfrac{1}{3}}}$$

E quella di $\frac{29}{13}$:

$$\frac{29}{13} = 2 + \cfrac{1}{4 + \cfrac{1}{3}}$$

Anche in questi casi (e in generale per ogni numero razionale), i numeri che appaiono nella frazione continua sono gli stessi numeri ottenuti dall'algoritmo di Euclide.

La scrittura in frazione continua dà una spiegazione elegante del perché sia proprio $\frac{29}{13}$ il numero razionale più "semplice" compreso tra $\frac{109}{49}$ e $\frac{96}{43}$: a questi ultimi corrispondono le sequenze $2, 4, 2, 5$ e $2, 4, 3, 3$ rispettivamente. Affinché un numero razionale x sia compreso tra $\frac{109}{49}$ e $\frac{96}{43}$ la

sua frazione continua deve necessariamente cominciare con $2, 4, 2$ o $2, 4, 3$ (Esercizio: perché?) e la frazione continua più semplice possibile è dunque data dalla sequenza $2, 4, 3$ (infatti la frazione continua che corrisponde a $2, 4, 2$ è minore di $\frac{109}{49}$).

20.2 Approssimare numeri reali con numeri razionali

Vogliamo ora capire in che modo sia possibile approssimare $\sqrt{2}$, che non è razionale (cioè esprimibile come rapporto fra numeri interi), con numeri razionali adeguatamente "semplici". Osserviamo subito che esistono approssimazioni buone a piacere: avendo

$$\sqrt{2} = 1.4142135...$$

possiamo pensare al numero razionale:

$$\frac{1414}{1000} = 1.414$$

Questo approssima $\sqrt{2}$ a meno di un errore di circa 0.0002. Tuttavia, $\frac{1414}{1000}$ non è una delle approssimazioni "migliori": esistono approssimazioni più fini con numeri razionali più "semplici".

Il nostro obiettivo è riciclare le osservazioni sulle frazioni continue fatte precedentemente. Prendiamo ad esempio $\frac{109}{49}$:

$$\frac{109}{49} = 2 + \cfrac{1}{4 + \cfrac{1}{2 + \cfrac{1}{5}}}$$

Osserviamo che, troncando la sua frazione continua, otteniamo successive approssimazioni razionali di $\frac{109}{49}$:

$$2 = 2.000000...$$

$$2 + \frac{1}{4} = \frac{9}{4} = 2.250000...$$

$$2 + \cfrac{1}{4 + \cfrac{1}{2}} = \frac{20}{9} = 2.222222...$$

$$2 + \cfrac{1}{4 + \cfrac{1}{2 + \cfrac{1}{5}}} = \frac{109}{49} = 2.224489...$$

Ovviamente $\frac{109}{49}$ è la "miglior" approssimazione razionale di se stesso, tuttavia, troncare la frazione continua ci fornisce approssimazioni meno fini ma più "semplici".

Vogliamo avere una frazione continua per $\sqrt{2}$. In che modo possiamo costruirla? Certamente l'algoritmo di Euclide non può andare bene; di cosa dovremmo calcolare l'MCD? Prima abbiamo

visto un metodo non basato sul fatto che $\frac{109}{49}$ fosse razionale: togliere l'intero adeguato affinché la nostra quantità risulti compresa tra 0 e 1, calcolare l'inverso e ripetere l'algoritmo. I numeri della frazione continua sono quindi, in ordine, gli interi tolti.

Proviamo ad applicare questa strategia a $\sqrt{2}$. Abbiamo:

$$\sqrt{2}$$

Che è compreso tra 1 e 2. Togliamo quindi **1**, da cui:

$$\sqrt{2} - 1$$

Invertiamo:

$$\frac{1}{\sqrt{2}-1} = \sqrt{2} + 1 \left(\text{poiché } (\sqrt{2}+1)(\sqrt{2}-1) = \sqrt{2}^2 + \sqrt{2} - \sqrt{2} - 1 = 1\right)$$

Pertanto abbiamo $\sqrt{2} + 1$, che è compreso tra 2 e 3, da cui togliamo quindi **2**:

$$\sqrt{2} - 1$$

Invertiamo:

$$\frac{1}{\sqrt{2}-1} = \sqrt{2} + 1$$

Che è compreso tra 2 e 3, da cui togliamo quindi **2**:

$$\sqrt{2} - 1$$

Invertiamo:

$$\frac{1}{\sqrt{2}-1} = \sqrt{2} + 1$$

Che è compreso tra 2 e 3, da cui togliamo quindi **2** e così via...

Pertanto possiamo scrivere la nostra frazione continua di $\sqrt{2}$:

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + ...}}}}$$

Questa frazione continua non è nessun numero "vero": è solo un simbolo. Tuttavia, possiamo convincerci che essa sia davvero legata a $\sqrt{2}$. Supponiamo che la scrittura di frazione continua corrisponda a un vero numero x:

$$x = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + ...}}}}$$

Osserviamo che:

$$x = \boxed{1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + ...}}}}} = 1 + \cfrac{1}{1 + \boxed{1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + ...}}}}}} = 1 + \frac{1}{1 + x}$$

Poiché le due quantità incasellate in rettangoli (entrambe uguali a x) corrispondono alla stessa frazione continua. Abbiamo dunque:

$$x = 1 + \frac{1}{1 + x}$$

$$x(1 + x) = (1 + x) + 1$$

$$x + x^2 = 2 + x$$

$$x^2 = 2$$

Da cui $x = \sqrt{2}$ (Esercizio metamatematico: e se fosse $-\sqrt{2}$?).

Ora che abbiamo la frazione continua di $\sqrt{2}$, possiamo troncarla per ottenere le "migliori" approssimazioni:

$$1$$

$$1 + \frac{1}{2} = \frac{3}{2}$$

$$1 + \cfrac{1}{2 + \cfrac{1}{2}} = \frac{7}{5}$$

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2}}} = \frac{17}{12}$$

$$1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2}}}} = \frac{41}{29}$$

E andando avanti così abbiamo $\sqrt{2} = 1.4142135...$ approssimato da:

$$1 = 1.0000000...$$

$$\frac{3}{2} = 1.5000000...$$
$$\frac{7}{5} = 1.4000000...$$
$$\frac{17}{12} = 1.4166666...$$
$$\frac{41}{29} = 1.4137931...$$
$$\frac{99}{70} = 1.4142857...$$
$$\frac{239}{169} = 1.4142011...$$
$$\frac{577}{408} = 1.4142156...$$

Vediamo ad esempio che $\frac{99}{70}$ approssima $\sqrt{2}$ con un errore di circa 0.00007, quindi meglio di $\frac{1414}{1000}$. In generale, per ogni numero reale al posto di $\sqrt{2}$, troncare la frazione continua produce le approssimazioni "migliori".

Nel nostro caso speciale di $\sqrt{2}$, queste approssimazioni hanno innumerevoli proprietà aritmetiche e possono essere calcolate rapidamente. Supponiamo di avere una "buona" approssimazione razionale di $\sqrt{2}$:

$$\frac{a}{b} \sim \sqrt{2}$$

Questa dà immediatamente un'altra buona approssimazione, prendendone l'inverso:

$$\frac{b}{a} \sim \frac{1}{\sqrt{2}}$$

Ma:

$$\frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{\sqrt{2}} \cdot \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$$

Quindi:

$$\frac{b}{a} \sim \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$$

Perciò abbiamo un'altra approssimazione di $\sqrt{2}$:

$$\frac{2b}{a} \sim \sqrt{2}$$

In particolare, abbiamo che il prodotto delle due approssimazioni è proprio 2:

$$\frac{a}{b} \cdot \frac{2b}{a} = 2$$

Pertanto, abbiamo due approssimazioni di $\sqrt{2}$, una lievemente maggiore di esso e una lievemente minore. Possiamo quindi essere tentati di fare una media fra le due approssimazioni.

Prendiamo ad esempio l'approssimazione $\frac{1}{1}$ di $\sqrt{2}$. Possiamo costruire un'approssimazione migliore facendo la media:

$$\frac{1}{2}\left(\frac{1}{1}+\frac{2}{1}\right)=\frac{3}{2}$$

Proseguiamo applicando di nuovo questo metodo a $\frac{3}{2}$:

$$\frac{1}{2}\left(\frac{3}{2}+\frac{4}{3}\right)=\frac{17}{12}$$

Procediamo con $\frac{17}{12}$:

$$\frac{1}{2}\left(\frac{17}{12}+\frac{24}{17}\right)=\frac{577}{408}$$

Partendo dall'approssimazione (pessima, ma semplice) $\frac{1}{1}$, abbiamo trovato le migliori $\frac{3}{2}$, $\frac{17}{12}$, $\frac{577}{408}$, che apparivano già troncando la frazione continua di $\sqrt{2}$. Questo metodo di approssimazione si chiama *metodo babilonese*.

Si può osservare un'altra proprietà aritmetica di queste approssimazioni; torniamo alla nostra approssimazione $\frac{1}{1}$. Supponiamo di voler usare il metodo babilonese, ma ci siamo dimenticati come calcolare la media. Privi di idee, proviamo a sommare i numeratori e i denominatori delle frazioni[32]:

$$\frac{1+2}{1+1}=\frac{3}{2}$$

Il primo passo è uguale. Proviamo con $\frac{3}{2}$:

$$\frac{3+4}{2+3}=\frac{7}{5}$$

Proseguiamo con $\frac{7}{5}$:

$$\frac{7+10}{5+7}=\frac{17}{12}$$

Andiamo avanti ancora:

$$\frac{17+24}{12+17}=\frac{41}{29}$$

Abbiamo $\frac{1}{1}, \frac{3}{2}, \frac{7}{5}, \frac{17}{12}, \frac{41}{29}$, che sono precisamente le approssimazioni ottenute troncando la frazione continua. Ça va sans dire, procedendo su questa strada si ottengono precisamente le stesse approssimazioni date dalla frazione continua.

Esercizio. Qual è la frazione continua di $\frac{1+\sqrt{5}}{2}$ (la famosa *sezione aurea*)? Che numeri si ottengono troncandola?

[32]Per i nostri lettori più affezionati potrebbe scattare una sensazione di déjà vu: questa particolare operazione è a lungo discussa nell'articolo "Frazioni e fogli quadrettati" del numero 9, che include anche un accenno alle frazioni continue!

Riferimenti bibliografici

[1] G. H. Hardy, E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press.

[2] A. Khinchin, *Continued Fractions*, Dover Publications.

[3] C. Brezinski, *History of Continued Fractions and Padé Approximants*, Springer.

[4] D. Fowler, E. Robson, *Square Root Approximations in Old Babylonian Mathematics: YBC 7289 in Context*, Historia Mathematica, Volume 25, Issue 4.

[5] H. Davenport, *Aritmetica superiore*, Zanichelli.

21 The data Whisperers

Maria Christodoulou, n.11, Settembre 2020

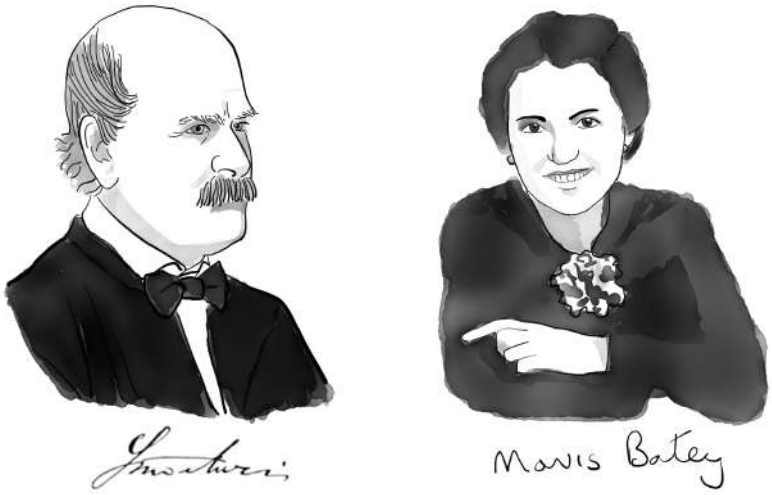
Traduzione italiana a seguire.

How do you find the epicentre of a cholera epidemic? How do you work out why women in one clinic die at a higher frequency than women in another one? How do you go about breaking a coded message? How do you decide which players will help you win the season?

The answer to all these questions is: you look at the data.

Let's set the scene, we're in London, and the year is 1854. It's the end of August, and the city is affected by yet another cholera outbreak. The people are scared, and the physicians have no concrete theories that can help them understand how the disease is transmitted and what they can do to stop it. It is a time of unsanitary conditions in this populous and impoverished part of the city. Enter our protagonist, physician John Snow, who started interviewing the locals and recording the patterns he observed. Through his meticulous data collection, he noticed something interesting: many of the affected residents had been using the water pump on Broad Street for their daily water needs. He collected some water from the pump and tried to study it, but as he was not yet sure what he was looking for, he found that line of investigation to be inconclusive. Yet his collected data, superimposed on hand-drawn local area maps, showed him an obvious pattern: the water pump on Broad Street had something to do with this. He took his findings to the local authorities and convinced them that the pump was somehow connected to the outbreak. The authorities removed its handle, effectively rendering it unusable, and the outbreak subsided. Or so the story goes. Dr Snow is now considered one of the founders of modern epidemiology. And it all started by collecting data.

Snow was not the only one doing so. A few years before him, in 1846 at a Viennese hospital, a young Hungarian physician called Ignaz Semmelweis was starting his new job. And he was greatly distressed by the rumours he was hearing. The hospital had two maternity clinics: one was staffed by physicians and medical students, and another staffed exclusively by midwives. The two



clinics had substantially different maternal mortality rates from postpartum infections: one with a 2% mortality rate, and the other with mortality ranging from 13% to 18%. That difference was staggering, and Semmelweis had absolutely no idea why this was happening. He started observing and collecting data, and the first thing he noted was that, contrary to what you may also be thinking, the clinic staffed by physicians was the one with the high mortality. The proverbial penny dropped under tragic circumstances – in March 1847, Semmelweis' friend and co-worker, the pathologist Jakob Kolletschka, died from an infection caused by a scalpel injury. Semmelweis noticed that Kolletschka had received his injury while performing an autopsy on a patient who had died from postpartum infection. That's when he made the connection: whatever it was that caused the death of the patient was transferred to the pathologist conducting the autopsy. His observation, together with his knowledge that almost all physicians had to conduct autopsies as opposed to midwives, led to one simple conclusion: something was transferred from the autopsies to the maternity ward leading to the increased maternal mortality in the physicians' clinic. He proposed an intervention: strict handwashing and disinfection protocols for all physicians after every autopsy. It was novel and radical. And it worked: suddenly mortality rates dropped dramatically in the physicians' clinic. When it comes to hygiene practices surrounding childbirth, Semmelweis was a trailblazer.

Let's move further forward in history, to the Second World War, a time when intel and data were valuable commodities. We find ourselves in Bletchley Park, the hub of British cryptanalysis. Let us not focus on Alan Turing right now – although he was undoubtedly a fascinating man - let us focus on a young linguist instead. Mavis Batey was 19 and studying German at University College London when the war started. Her language skills were absolutely crucial to the war effort. Breaking code after code, her work ensured the safe planning of D-Day. What does codebreaking have to do with data? Absolutely everything. Batey's edge as a codebreaker was provided by both her language skills, her talent in spotting patterns, and her familiarity with the enemy operators. She worked out that one of the operators had a girlfriend named "Rosa" and she used that knowledge to break the codes he was transmitting. She paid attention to the data. And here's another example: on one of the messages she was trying to decrypt, Batey noticed that the letter L was completely absent. Knowing well that Enigma machines never encoded a letter onto itself, she concluded that the sequence was a series of Ls. She used this to determine the setup for that day's machines, effectively helping decode crucial messages to the war effort.

Around sixty years later and across the Atlantic, looking for patterns in the data was used in a completely different setting – baseball. Although the study of baseball statistics was nothing new, a young economist named Paul DePodesta put his own spin on the whole thing. Working as an assistant to the General Manager of the Oakland Athletics, DePodesta decided to follow the data. He focused his analysis on the metrics that truly seemed to predict performance as opposed to those traditionally used by scouts in the past century. This ensured that the General Manager - Billy Beane - could use the Oakland A's limited budget to compose a team that was genuinely competitive, selecting players with complementary skills. After multiple strong seasons, their story was written up by financial journalist Michael Lewis and was also given the Hollywood treatment.

I could fill pages with stories of data collection and pattern recognition, but I would much rather show you just how intuitive and natural these processes are. Statistics help us detect patterns even when they are hiding in noisy data. They can also help us avoid the pitfall of biases, which we all inherently and unconsciously apply to everything we encounter.

Statistics can help us find the stories behind the numbers and they are the formalisation of our pattern recognition instincts.

Let me show you what I mean. This time I will put you in the shoes of somebody trying to uncover the story behind the numbers. Once upon a time there was a group of people, a little over 2000 of them, who found themselves experiencing an event. An event with devastating consequences. An event that led to the deaths of over 65% of that group. I will be upfront and tell you now that this event was not a disease. But something did happen, and people died. And to help us work out what happened we have data. Let us start by looking at the data.

	By economic status and sex					
	Population Exposed to Risk			Number of Deaths		
Economic Status	Male	Female	Total	Male	Female	Total
I (high)	180	145	325	118	4	122
II	179	106	285	154	13	167
III (low)	510	196	706	422	106	528
Other	862	23	885	670	3	673
Total	1731	470	2201	1364	126	1490

Maybe this is a little overwhelming at first glance but let us approach it systematically. The table is broken up into two parts: the left-hand side is telling us how many people were exposed to that particular event, and the right-hand side is telling us how many died because of that event. The first thing to note is that the number of dead is substantial. Whatever happened, it was terrible. But then... maybe you start noticing a little more detail. Like, for example, isn't it strange how men seem to have it much worse than women? Granted, there were fewer women in the group to begin with, but it still looks unusual. Let's look at the percentages of dead in the population to get a clearer idea.

	% of dead in each risk group		
Economic Status	Male	Female	Total
I (high)	65.56%	2.76%	37.54%
II	86.03%	12.26%	58.60%
III (low)	82.75%	54.08%	74.79%
Other	77.73%	13.04%	76.05%
Total	78.80%	26.81%	67.70%

The table shows the percentage of people who died in each risk group. Nearly 80% of all men in the group died. Contrast this to the women in the group, which is around 27% and you detect that this event impacts the two sexes differently. But the tables also give us some information on what is referred to as "Economic Status". This seems to be ranked from "I" which is the high level to "III" which is the low level. But there is something peculiar here: a category named "Other". This is unexpected because it is not labelled as missing data. It is not as if we simply do not know the economic status of that group: that group does not belong in any of these categories. But this is a substantial group. Out of the 2201 people in the population, 885 belong to that group. That's 40% of the population in the "Other" group. This is important.

So, if you are a statistician or a data analyst, or just a data explorer as we are now, what do you do?

When the data you have raise more questions than answers, then all you can do is ask more questions. And luckily for us, we have a relatively crude breakdown of the same dataset, but instead of splitting it into male and female, this time we split it in adults and children.

	By economic status and age group					
	Population Exposed to Risk			Number of Deaths		
Economic Status	Adults	Children	Total	Adults	Children	Total
I (high)	319	6	325	122	0	122
II	261	24	285	167	0	167
III (low)	627	79	706	476	52	528
Other	885	0	885	673	0	673
Total	2092	109	2201	1438	52	1490

This table should really stop us in our tracks. This population is extremely unusual. Not only the number of children in the population is surprisingly low, at around 5%, they also seem to be following an unexpected pattern in the various economic groups. More specifically, the number of children in each group is not proportional to the size of that group. If it were, there would be more children recorded in the mysterious "Other" category. It also appears that dying as a child in this event is a situation exclusively found in the low economic status group.

These two datasets give us some particularly valuable clues: it appears that this population is not a natural population - otherwise the number of children would follow some more expected patterns, and the chance of dying appears to be lower if you are a woman or a child in the medium and high economic groups.

And there lies the essence of the dataset: we are exploring some sort of disaster that impacts men more than women and children, and economic status affects the likelihood of survival. Maybe war is your first thought, but the presence of economic status itself should make you suspicious. If on the other hand "women and children first" is more what comes to mind, then you're onto something. This maritime code of conduct, common in the 19th and early 20th century, suggested than in life-threatening situations women and children should be given priority to lifeboats. This rule really entered popular consciousness with the RMS Titanic maritime disaster, when on her maiden voyage she hit an iceberg and sank in the icy waters of the North Atlantic Ocean. And when she sank, she took down with her almost 1500 souls. To be precise, 1490 people out of the 2201 total perished when RMS Titanic sank.

Let's look at the datasets again, now that we know that we're dealing with a maritime disaster. The "women and children" pattern is obvious for the two upper classes. And women are overall doing better than men, but again, economic status is crucial. To survive the RMS Titanic your best chance is to be one of the 6 rich children on board. Now with the knowledge of the provenance of our dataset the economic status groupings make more sense: "I" is for first class ticket holders, "II" is for second class ticket holders, and "III" is for third class. The crew is allocated as "Other" and as crew members could not bring their children on board, that explains the absence of any children in that group.

It is extremely powerful to be able to discover the story hidden in the noise of data. Small explorations, such as the one we just did can get us part of the way, but statistics and modern machine learning can take us much further. We can now find the consistent patterns in datasets with millions of entries, study the migration of populations through time, investigate ageing patterns in human society, peer into the human genome to work out genetic inheritance, or even find the right song recommendation for our latest playlist. Data surround us and modern statistics offer the tools for their exploration.

Come si trova l'epicentro di un'epidemia di colera? Come si fa a scoprire perché il tasso di mortalità in una clinica di maternità sia più alto di quello di una struttura analoga? Come si tenta di decifrare un messaggio segreto? Come si fa a decidere quali giocatori siano i più adatti per portare la squadra alla vittoria durante la prossima stagione del baseball?

La risposta a ciascuna di queste domande è: si analizzano i dati.

Siamo a Londra, anno 1854. È la fine di agosto e la città è teatro di un ennesimo focolaio di colera. La gente ha paura e i medici non hanno teorie concrete che aiutino a comprendere come si trasmette la malattia o come se ne possa arrestare la diffusione. Sullo sfondo di una parte della città povera e sovrappopolata, in un tempo di condizioni a dir poco antigieniche, ecco a voi il nostro protagonista, il medico John Snow, mentre si aggira interrogando gli abitanti alla ricerca di pattern di cui prendere nota. Grazie alla sua meticolosa raccolta dati, si rende conto di un fatto interessante: molti dei residenti toccati dall'epidemia si servono della pompa dell'acqua di Broad Street per l'approvvigionamento giornaliero. Snow va a prendere dell'acqua alla pompa incriminata e tenta di studiarla, ma dato che non sa bene cosa cercare le sue analisi rimangono prive di frutti. Tuttavia i dati raccolti, riportati su mappe della zona da lui stesso disegnate, lasciano pochi dubbi: la pompa dell'acqua di Broad Street deve avere qualcosa a che fare con l'epidemia. Le autorità londinesi, su indicazione di Snow, rimuovono il manico della pompa rendendola inutilizzabile, e il focolaio si spegne, o così si racconta. Il Dottor Snow è considerato oggi uno dei fondatori dell'epidemiologia moderna. E tutto cominciò con una raccolta dati.

*Ma Snow non era l'unico a utilizzare questo approccio. Alcuni anni prima, nel 1846, un ospedale di Vienna assumeva un giovane medico ungherese di nome Ignaz Semmelweis. Semmelweis era preoccupato dalle voci che aveva sentito: l'ospedale aveva due reparti maternità, uno gestito da dottori e studenti di medicina, mentre l'altro impiegava esclusivamente ostetriche. I due reparti avevano tassi di mortalità da infezioni post-parto radicalmente diversi: in uno il tasso era del 2%, nell'altro si aggirava fra il 13% e il 18%. Una differenza sconcertante, che Semmelweis non aveva idea di come spiegare. Tanto più che, al contrario di quanto anche voi potreste pensare, il reparto con un alto tasso di mortalità era quello che impiegava i medici. L'*eureka *arrivò purtroppo in circostanze tragiche: nel marzo del 1847 un amico e collega di Semmelweis, il patologo Jakob Kolletschka, morì per un infezione causata da una ferita procuratasi col bisturi, e Semmelweis notò che l'amico si era procurato la ferita nell'effettuare un'autopsia su una donna deceduta di infezione post-parto. Fu allora che nella sua mente scattò un collegamento: qualunque cosa fosse che causava la morte delle pazienti poteva essere trasmessa al patologo che effettuava l'autopsia. Questa osservazione, combinata con la consapevolezza che quasi tutti i medici effettuavano autopsie, al contrario delle ostetriche, portava a un'inevitabile e semplice conclusione: qualcosa veniva trasmesso tramite le autopsie nel reparto maternità dei medici e provocava così il tasso di mortalità più alto. Semmelweis propose un intervento immediato: stabilire severi protocolli igienici – con lavaggio delle mani e disinfezione – per i medici dopo ogni autopsia. Era un provvedimento innovativo e radicale per i tempi. E funzionava: la mortalità delle pazienti subì un improvviso e consistente calo nel reparto dei medici. Semmelweis fu un apripista*

in fatto di pratiche igieniche in connessione al parto.

Adesso facciamo un salto in avanti verso l'epoca della Seconda Guerra Mondiale; come durante ogni guerra, l'informazione era una merce preziosa. Ci ritroviamo a Bletchley Park, il centro della crittoanalisi britannica. Non stiamo per concentrarci su Alan Turing, nonostante l'innegabile fascino della sua storia, ma su una giovane linguista. Mavis Batey aveva 19 anni e studiava tedesco al University College di Londra quando scoppiò la guerra. Le sue abilità linguistiche furono cruciali per lo sforzo bellico. Ruppe codice dopo codice e fu anche grazie al suo lavoro che fu possibile pianificare lo sbarco in Normandia. Cosa c'entra la crittoanalisi con in dati? Moltissimo. L'eccezionalità di Batey come crittoanalista è frutto delle sue abilità linguistiche, ma anche del suo talento nel riconoscere pattern e della sua familiarità con gli operatori nemici. Ad un certo punto dedusse che uno degli operatori aveva una fidanzata di nome "Rosa" e usò questa informazione per decodificare le sue trasmissioni. Era sempre attenta ai dati. Ecco un altro esempio: si accorse che uno dei messaggi che tentava di decifrare non conteneva nessuna lettera L. Ben sapendo che le macchine Enigma non codificavano mai una lettera con se stessa, immaginò che il messaggio fosse in realtà una sequenza di L. Grazie a questo fu in grado di determinare la configurazione delle macchine cifranti per quella giornata, portando un aiuto decisivo allo sforzo bellico alleato.

Circa sessant'anni dopo, dall'altra parte dell'Atlantico, la ricerca di pattern veniva impiegata in un contesto completamente diverso: il baseball. Nonostante lo studio delle statistiche sportive non fosse ovviamente nuovo, il giovane economista Paul DePodesta ne fece un uso tutto suo. Nel suo ruolo di assistente del Manager Generale della Oakland Athletics, DePodesta decise di seguire i dati. Concentrò la sua analisi sui parametri che sembravano davvero essere predittivi rispetto alle performance dei giocatori piuttosto che su quelli tradizionalmente utilizzati dai talent scout nell'ultimo secolo. Fu così che il Manager Generale — Billy Beane — poté utilizzare le limitate risorse a disposizione dell'Oakland A per mettere insieme una squadra veramente competitiva, selezionando giocatori con abilità complementari. Dopo il successo di numerose stagioni, la storia di DePodesta venne raccontata dal giornalista finanziario Michael Lewis, e nel frattempo ne è stato tratto addirittura un film.

Potrei riempire pagine e pagine di storie sulla raccolta e analisi dei dati, ma quello che invece voglio fare è mostrarvi quanto intuitivi e naturali possano essere questi processi. La statistica ci consente di riconoscere pattern anche quando si nascondono in una collezione di dati "rumorosi"; ci aiuta inoltre a evitare le trappole tese dai nostri "bias", che applichiamo inevitabilmente e inconsciamente a tutto ciò che ci viene messo di fronte.

La statistica, insomma, ci aiuta a far emergere le storie dietro i numeri e non è altro che una formalizzazione dei nostri istinti di riconoscimento di pattern.

Lasciate che vi mostri cosa voglio dire, questa volta mettendo voi nei panni di chi cerca di svelare la storia nascosta dietro i numeri. C'era una volta un gruppo di persone — poco più di 2000 — che si trovarono coinvolte in un avvenimento. Un avvenimento dalle conseguenze devastanti. Un avvenimento che avrebbe condotto alla morte di oltre il 65% di questo gruppo. Vi rivelo subito che questo avvenimento non fu il diffondersi di una malattia. Ma qualcosa accadde, e delle persone morirono. Per aiutarci a ricostruire

che cosa, abbiamo dei dati. E la prima cosa da fare è questa: guardare i dati.

	By economic status and sex					
	Population Exposed to Risk			Number of Deaths		
Economic Status	Male	Female	Total	Male	Female	Total
I (high)	180	145	325	118	4	122
II	179	106	285	154	13	167
III (low)	510	196	706	422	106	528
Other	862	23	885	670	3	673
Total	1731	470	2201	1364	126	1490

Magari a prima vista ci ritroviamo un po' confusi, ma vediamo di usare un approccio sistematico. La tabella è divisa in due parti: la metà sinistra ci dice quante persone siano state coinvolte nell'avvenimento in questione, la parte di destra quante ne siano morte. La prima cosa che possiamo osservare è che il numero di morti è notevole. Questo avvenimento, qualunque esso sia, dev'essere stato terribile. Ma poi... magari state cominciando a notare qualcosa in più. Per esempio, non è strano che che la situazione appaia peggiore per gli uomini che per le donne? È vero che le donne all'interno del gruppo sono di meno fin dall'inizio, ma la cosa sembra comunque particolare. Diamo un'occhiata alle percentuali dei morti all'interno della nostra popolazione per farci un'idea più chiara.

	% of dead in each risk group		
Economic Status	Male	Female	Total
I (high)	65.56%	2.76%	37.54%
II	86.03%	12.26%	58.60%
III (low)	82.75%	54.08%	74.79%
Other	77.73%	13.04%	76.05%
Total	78.80%	26.81%	67.70%

La tabella mostra la percentuale di persone decedute in ciascuna categoria. Mentre quasi l'80% degli uomini finisce col morire, questo vale solo per circa il 27% delle donne: il nostro avvenimento, evidentemente, ha avuto un impatto diverso sui due sessi. In più, le tabelle ci danno informazioni su quello che è definito "Economic status" (diremo "ceto"), e che appare classificato da "I" (il livello più alto) a "III" (il più basso). Ma c'è qualcosa di imprevisto: una categoria denominata "Altro". È un fatto inaspettato, tanto più che non sembra trattarsi di dati mancanti: non è che semplicemente non conosciamo

il "ceto" della categoria, ma che le persone al suo interno non appartengono a nessuno dei "ceti" della lista; per di più, la categoria contiene un numero non indifferente di persone. Delle 2201 persone nella nostra popolazione, 885 sono dichiarate "altro": circa il 40%.

Dunque, se foste uno statistico o un analista di dati, o semplicemente in quanto esploratore di dati, cosa fareste?

Quando i dati che abbiamo producono più domande che risposte, non resta che porre ancora più domande. In questo caso, per fortuna, abbiamo a disposizione un'ulteriore suddivisione abbastanza grezza dei nostri dati: invece che i dividere la popolazione in uomini e donne, possiamo dividerla in adulti e bambini.

	By economic status and age group					
	Population Exposed to Risk			Number of Deaths		
Economic Status	Adults	Children	Total	Adults	Children	Total
I (high)	319	6	325	122	0	122
II	261	24	285	167	0	167
III (low)	627	79	706	476	52	528
Other	885	0	885	673	0	673
Total	2092	109	2201	1438	52	1490

Questa nuova tabella dovrebbe immediatamente farci fare un passo indietro. La nostra popolazione è estremamente inusuale; non solo la percentuale di bambini — circa il 5% — è sorprendentemente bassa, ma è distribuita in modo inaspettato nei vari ceti. In particolare, il numero di bambini non è proporzionale al numero di appartenenti a ciascun ceto: se così fosse, dovrebbero esserci più bambini nella misteriosa categoria "altro". Inoltre, sembra che la morte sia toccata solamente ai bambini di "ceto" più basso.

Queste due collezioni di dati ci danno alcuni indizi molto utili: anzitutto si direbbe che la popolazione non sia naturale (altrimenti il numero di bambini seguirebbe pattern più usuali), e inoltre la probabilità di morte sembra essere più bassa se si è una donna o un bambino di ceto medio o alto.

Ecco qui l'essenza di questi dati: stiamo esplorando un qualche tipo di disastro le cui conseguenze hanno un impatto maggiore per gli uomini che per le donne e i bambini, e nell'ambito del quale il "ceto" influenza le probabilità di sopravvivenza. Potrete pensare per prima cosa a una guerra, ma già la presenza di queste categorie "ceto" dovrebbe darvi dei sospetti. Se invece quello a cui state pensando è il detto "prima le donne e i bambini", siete sulla buona strada. Il codice di condotta marittimo, diffuso nel diciannovesimo e ventesimo secolo, prevede che in una situazione di pericolo le donne e i bambini abbiano accesso prioritario alle scialuppe di salvataggio. Questa regola entrò a far parte della coscienza popolare in occasione del disastro marittimo della nave RMS Titanic, che nel corso del proprio viaggio inaugurale entrò in collisione con un iceberg e affondò nelle acque gelate dell'Oceano Atlantico settentrionale. Quasi 1500 anime furono perdute assieme

alla nave. Ad essere precisi, 1490 delle 2201 persone a bordo morirono nel naufragio del RMS Titanic.

Ora che sappiamo che si tratta di un naufragio, guardiamo di nuovo i nostri dati. Il pattern "donne e bambini" è evidente nelle due classi superiori. In generale le donne si salvano più degli uomini, ma di nuovo il ceto è un aspetto cruciale. Per avere le migliori chances di sopravvivere al naufragio del Titanic, conviene essere uno dei sei bambini ricchi presenti sulla nave. Sapendo da dove provengano i dati, quelle strane categorie per "ceto" hanno finalmente più senso: "I" sta per i viaggiatori in prima classe, "II" per quelli in seconda, "III" per quelli in terza classe. L'equipaggio è classificato come "altro", il che spiega perché la categoria "altro" non contenga bambini.

Quello di saper scoprire le storie nascoste dal rumore nei dati è un potere notevole. Piccole esplorazioni come quella che abbiamo appena fatto possono farci fare un pezzo di strada, ma la statistica moderna e il machine learning ci conducono molto, molto oltre. Oggi possiamo trovare le regolarità in collezioni di dati con milioni di punti, studiare le migrazioni di popoli interi nel tempo, andare alla scoperta dei processi di invecchiamento nelle società della storia, svelare i segreti del genoma umano e dell'ereditarietà, perfino trovare il brano giusto da suggerire per la nostra ultima playlist. I dati sono ovunque attorno a noi e la statistica moderna ci fornisce gli strumenti adatti per la loro esplorazione.

22 I Teoremi di Impossibilità di Arrow-Gibbard-Satterthwaite

Lucio Tanzini e Cristofer Villani, n.12, Aprile 2021

Supponiamo di avere a cena un gruppo di amici, e dover scegliere velocemente che cosa ordinare da mangiare. Decidiamo, allora, di metterla ai voti. A prima vista, il problema è risolto. In effetti, se abbiamo due sole possibilità, e ne preferiamo una, la votazione sarà rapida: a tutti è chiaro cosa votare.

Supponiamo però di averne varie: se pensiamo di essere gli unici a voler ordinare una pizza, ma il giapponese non ci va poi tanto male, potrebbe darsi che sia meglio votare sushi al posto di pizza, in modo che, almeno, si decida per la nostra seconda scelta. O magari, se votiamo indiano, potremmo almeno evitarci di ordinare il cibo che ci piace meno.

In altre parole, potremmo dover *strategizzare*, non votando quello che è effettivamente il nostro cibo preferito, per ottenere, alla fine dei conti, un miglior risultato.

Il problema sarebbe risolto se esistesse un sistema elettorale che non sfavorisca mai un elettore quando esprime onestamente la sua preferenza.

Mostreremo però che, purtroppo per noi, un tale sistema in generale non esiste.

Più nello specifico,

($\star$) *un qualsiasi sistema con almeno tre candidati che permetta la vittoria a ciascuno di essi e per il quale ad ogni elettore convenga votare seguendo onestamente le proprie preferenze è una dittatura,*

nel senso che c'è un elettore che determina interamente l'esito dell'elezione. Nel seguito, vedremo come arrivare a questo risultato, dovuto al filosofo Allan Gibbard (1973), che generalizzò una versione precedente, più debole: il celebre *Teorema di impossibilità di Arrow*, provato nel 1950 dall'economista Kenneth Arrow.

22.1 Il teorema di Gibbard-Satterthwaite

Vediamo prima un caso particolare: supponiamo che ogni elettore esprima il suo voto stilando la propria classifica dei candidati.

22.1.1 L'enunciato del teorema

Come tradurre questa situazione in un contesto matematico? Supponiamo di avere n elettori e m candidati, dove n, m sono numeri naturali: possiamo indicare gli elettori coi numeri da 1 a n, e i candidati con le lettere dell'alfabeto $A, B, C, \dots$. Chiamiamo $\mathcal{C} = \{A, B, C, \dots\}$ l'insieme degli m candidati.

Ogni *preferenza* di voto è una sequenza $P = (X_1, \dots, X_m)$, che rappresenta il fatto che il candidato X_i occupi l'i-esimo posto in classifica.

Quando tutti avranno votato, otterremo un *profilo* di voti, cioè una sequenza

$$\boldsymbol{P} = (P_1, \dots, P_n),$$

dove P_i è la preferenza di voto dell'elettore i. Ad esempio, con tre elettori $1, 2, 3$ e quattro candidati A, B, C, D potremmo avere il profilo (P_1, P_2, P_3), dove $P_1 = (A, B, C, D)$, $P_2 = (B, A, C, D)$, $P_3 = (C, A, D, B)$, che possiamo visualizzare con la tabella sotto.

1	**2**	**3**
A	B	C
B	A	A
C	C	D
D	D	B

Con tali informazioni, il sistema elettorale deve proclamare un vincitore tra gli elementi di $\mathcal{C}$. Pertanto possiamo dire che un sistema elettorale è una funzione

$$f : \mathcal{P} \to \mathcal{C},$$

dove $\mathcal{P}$ è l'insieme di tutti i profili possibili, che associa ad ogni profilo $\boldsymbol{P}$ un vincitore $f(\boldsymbol{P})$.

Dato un sistema elettorale f, diremo che

1. f è *dittatoriale* se esiste un elettore i, che chiamiamo un *dittatore*, tale che, comunque dato un profilo di voto $(P_1, \ldots, P_n)$, il vincitore è il candidato preferito da i in P_i.
2. f è *manipolabile* se esiste un profilo $\boldsymbol{P} = (P_1, \ldots, P_n)$ in cui un elettore i, sostituendo P_i con un'altra preferenza P_i', ottenga un vincitore che preferisce (in P_i) a $f(\boldsymbol{P})$. (In altre parole, se ogni elettore votasse in maniera onesta ci sarebbe un profilo in cui ad un elettore sarebbe invece covenuto mentire).

Vogliamo allora dimostrare il

Theorem 3 (Gibbard-Satterthwaite). Se $\mathcal{C}$ ha almeno tre candidati, un sistema elettorale $f : \mathcal{P} \to \mathcal{C}$ non manipolabile che permette a tutti i candidati di vincere è dittatoriale.

Procederemo in tre passi consecutivi:

i) prima, otterremo alcune proprietà "ragionevoli" di f, che ci aspettiamo da un sistema democratico;

ii) poi, vedremo che queste proprietà implicano l'esistenza di un elettore sospetto, il *pivot*;

iii) infine, dedurremo che il pivot è proprio il dittatore che stavamo cercando.

22.1.2 Un sistema (in apparenza) democratico

Iniziamo deducendo, dalle ipotesi, le seguenti proprietà.

Monotonia. Supponiamo che in un fissato profilo $\boldsymbol{P} = (P_1, \ldots, P_n)$ vinca il candidato A. Supponiamo poi che un elettore i cambi la sua preferenza *scalando in alto* un candidato B, cioè scambiandolo con il candidato immediatamente sopra di lui un certo numero di volte. In altre parole, se P_i' è la nuova preferenza di i, il passaggio $P_i \rightsquigarrow P_i'$ può ad esempio essere

$\boldsymbol{i}$		$\boldsymbol{i}$
$\vdots$		$\vdots$
C		$\boldsymbol{B}$
A	$\rightsquigarrow$	C
D		A
$\boldsymbol{B}$		D
$\vdots$		$\vdots$

Se $\boldsymbol{P'} = (P_1, \dots, P_{i-1}, P'_i, P_{i+1}, \dots, P_n)$, chi ci aspettiamo vinca in $\boldsymbol{P'}$? Dato che l'unica preferenza a essere cambiata è quella di i, e in essa sono conservate le posizioni relative tra i candidati diversi da B, che è salito in classifica, è ragionevole che vinca A (come in $\boldsymbol{P}$) oppure B.

Proposizione 2. *Sia f non manipolabile. Se due profili $\boldsymbol{P}, \boldsymbol{P'}$ differiscono solo per la preferenza P_i, P'_i di un unico elettore, e P'_i è ottenuta da P_i scalando verso l'alto un candidato B, $f(\boldsymbol{P'})$ è $f(\boldsymbol{P})$ oppure B.*

Dimostrazione. Supponiamo per assurdo che per il profilo $\boldsymbol{P'}$ vinca un candidato C diverso da A e B. Osserviamo che allora o i preferisce A a C sia in P_i che in P'_i oppure i preferisce C ad A sia in P_i che in P'_i.

Nel primo caso, in $\boldsymbol{P'}$, a i converrebbe cambiare la propria preferenza in P_i così da far vincere A (dato che preferisce A a C). Nel secondo caso, in $\boldsymbol{P}$, gli converrebbe cambiare la propria preferenza in P'_i così da far vincere C.

In entrambi i casi, f è manipolabile. □

Unanimità. Supponiamo che tutti gli elettori preferiscano, tra tutti, lo stesso candidato A, come in

1	**2**	**3**	$\cdots$	$\boldsymbol{n}$
A	A	A	$\cdots$	A
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Ricordiamo che, nelle ipotesi sul sistema elettorale f, abbiamo supposto che ogni candidato *possa* effettivamente vincere.

Proposizione 3. *Se un candidato A è la prima scelta di ogni elettore, A vince.*

Dimostrazione. Prendiamo un profilo $\boldsymbol{Q}$ in cui vince A. Possiamo allora modificare una per volta le preferenze di ogni elettore $1, \dots, n$ scalando A in prima posizione: chiamiamo $\boldsymbol{P}$ il nuovo profilo così ottenuto. Applicando n volte la proprietà di monotonia, abbiamo che, in $\boldsymbol{P}$, vince ancora A.

Sia ora $\boldsymbol{R}$ un profilo in cui A è la prima scelta di tutti gli elettori. Costruiamo profili intermedi

$$\boldsymbol{P} \rightsquigarrow \boldsymbol{P}^{(1)} \rightsquigarrow \boldsymbol{P}^{(2)} \rightsquigarrow \dots \rightsquigarrow \boldsymbol{P}^{(n-1)} \rightsquigarrow \boldsymbol{R}$$

modificando, una dopo l'altra, le preferenze di $\boldsymbol{P}$, in modo che nel profilo $\boldsymbol{P}^{(i)}$ le prime i siano come in $\boldsymbol{R}$ (e le altre come in $\boldsymbol{P}$). In ognuno dei $\boldsymbol{P}^{(i)}$, A continua a essere il candidato preferito da tutti. Se quindi in $\boldsymbol{P}^{(1)}$ A non vincesse, a 1 converrebbe dichiarare la preferenza che aveva in $\boldsymbol{P}$ per far vincere A; perciò, A vince anche in $\boldsymbol{P}^{(1)}$. Analogamente, si ottiene che A vince in ognuno dei $\boldsymbol{P}^{(i)}$, e infine in $\boldsymbol{R}$. □

Ostracismo. Vale anche la versione speculare di unanimità.

Proposizione 4. *Se un candidato B è l'ultima scelta di ogni elettore, B non vince.*

Dimostrazione. Ragioniamo come prima. Sia $\boldsymbol{P}$ un profilo della forma

1	**2**	**3**	$\cdots$	$\boldsymbol{n}$
A	A	A	$\cdots$	A
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B	B	B	$\cdots$	B

in cui la prima scelta di ogni elettore è A e l'ultima scelta di ogni elettore è B: per unanimità, B non vince. Se $\boldsymbol{R}$ è un qualsiasi profilo in cui B è per tutti l'ultima scelta, costruiamo profili

$$\boldsymbol{P} \rightsquigarrow \boldsymbol{P}^{(1)} \rightsquigarrow \boldsymbol{P}^{(2)} \rightsquigarrow \cdots \rightsquigarrow \boldsymbol{P}^{(n-1)} \rightsquigarrow \boldsymbol{R}$$

come sopra. A ogni passo, B è l'ultima scelta di ogni elettore, e continua a non vincere (perché?), quindi non vince nemmeno in $\boldsymbol{R}$. □

22.1.3 Un tipo sospetto: il pivot

Troviamo il pivot. Finora, le nostre ipotesi su f non hanno avuto conseguenze strane. Vediamo però che succede se applichiamo quelle trovate a un fissato candidato, diciamo B. Partiamo da un profilo $\boldsymbol{P}^{(0)}$ della forma

1	**2**	**3**	$\cdots$	$\boldsymbol{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B	B	B	$\cdots$	B

in cui ogni elettore ha B come ultima scelta. Per ostracismo, sappiamo che B non vince: supponiamo allora vinca $K \neq B$.

Se ora 1 cambia la sua preferenza scalando B al primo posto, otteniamo un profilo $\boldsymbol{P}^{(1)}$

1	**2**	**3**	$\cdots$	$\boldsymbol{n}$
B	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	B	B	$\cdots$	B

in cui, per monotonia, la vittoria va a K o a B. Continuando a cambiare, in successione, le preferenze di $2, \ldots, n$ allo stesso modo, otteniamo una sequenza di profili

$$\boldsymbol{P}^{(0)} \rightsquigarrow \boldsymbol{P}^{(1)} \rightsquigarrow \boldsymbol{P}^{(2)} \rightsquigarrow \cdots \rightsquigarrow \boldsymbol{P}^{(n-1)} \rightsquigarrow \boldsymbol{P}^{(n)}$$

per cui, in $\boldsymbol{P}^{(n)}$, ogni elettore preferisce B agli altri candidati. In ognuno dei $\boldsymbol{P}^{(i)}$, per monotonia, continua a vincere o K o B; d'altra parte, per unanimità, nel profilo $\boldsymbol{P}^{(n)}$ è necessariamente B a vincere. Inoltre, se in un certo profilo $\boldsymbol{P}^{(k)}$ sappiamo che vince B, ancora per monotonia B dovrà vincere in ogni profilo successivo al k-esimo. Esisterà quindi un elettore r, che chiameremo *pivot* per B, tale che

i) in ogni profilo $\boldsymbol{P}^{(j)}$ con $0 \leq j < r$, vince K;

ii) in ogni profilo $\boldsymbol{P}^{(j)}$ con $r \leq j \leq n$, vince B.

I primi sospetti. Indaghiamo ora i poteri di r.

Consideriamo i profili

$\boldsymbol{P}^{(r-1)}$:

1	**2**	$\cdots$	$\boldsymbol{r-1}$	$\boldsymbol{r}$	$\boldsymbol{r+1}$	$\cdots$	$\boldsymbol{n}$
B	B	$\cdots$	B	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	B	B	$\cdots$	B

$\xrightarrow{f}$ K

$$\boldsymbol{P}^{(r)}:\quad \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & B & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & B & \cdots & B \end{array} \quad \xrightarrow{\ f\ } \quad B$$

e vediamo che cosa riusciamo a dire grazie alla non manipolabilità di f.

i) In $\boldsymbol{P}^{(r-1)}$, se uno degli elettori tra 1 e $r-1$, diciamo i, cambia la sua preferenza, B deve continuare a perdere: altrimenti, in $\boldsymbol{P}^{(r-1)}$, a i converrebbe dichiarare la nuova preferenza per far vincere B.

   Se invece un elettore j tra $r+1$ e n cambia la sua preferenza lasciando B in fondo, B deve ancora perdere: altrimenti, nel nuovo profilo, a j converrebbe usare la sua preferenza in $\boldsymbol{P}^{(r-1)}$ perché B non vinca.

ii) In $\boldsymbol{P}^{(r)}$, analogamente, comunque un elettore j compreso tra $r+1$ e n cambi la sua preferenza, B dovrà continuare a vincere; viceversa, se i è compreso tra 1 e $r-1$ e cambia la sua preferenza lasciando B in testa, B continua a vincere.

Se ne deduce che

Proposizione 5. *Se r è un pivot per B, valgono i seguenti fatti.*

MB. *In un qualsiasi profilo in cui $r,\dots,n$ votano B per ultimo, B non vince.*

WB. *In un qualsiasi profilo in cui $1,\dots,r$ votano B per primo, B vince.*

La prova schiacciante. Per quanto curiose, le proprietà ***MB***, ***WB*** ora trovate non sono eclatanti. Vedremo però ora che, da esse, possiamo dedurre un fatto ben più grave.

Proposizione 6. *Se r è un pivot per B, r ha il seguente potere.*

OB. *Se tutti votano B per ultimo, vince il candidato preferito da r.*

Dimostrazione. Consideriamo un profilo

$$\boldsymbol{P}:\quad \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline \vdots & \vdots & \vdots & \vdots & K & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B & B & \cdots & B & B & B & \cdots & B \end{array}$$

in cui ogni candidato mette B all'ultimo posto, e r preferisce un certo $K \in \mathcal{C}$ a tutti gli altri candidati, e supponiamo che, in $\boldsymbol{P}$, vinca un candidato $G \neq K$: vogliamo vedere che questo è impossibile. Per farlo, costruiremo un profilo $\boldsymbol{Q}$ e due sequenze di profili

$$\begin{array}{ccccc} & & \boldsymbol{R}^{(1)} & \rightsquigarrow & \boldsymbol{R}^{(2)} \\ \boldsymbol{P} & \nearrow\!\!\searrow & & & & \searrow\!\!\nearrow & \boldsymbol{Q} \\ & & \boldsymbol{S}^{(1)} & \rightsquigarrow & \boldsymbol{S}^{(2)} \end{array}$$

in modo che, seguendo la prima, si deduca che in $\boldsymbol{Q}$ vince K e, seguendo la seconda, si ottenga invece che in $\boldsymbol{Q}$ vince B: dato che questo è assurdo, dev'essere $K = G$.

Partiamo allora da $\boldsymbol{P}$: se tutti gli elettori $i \neq r$ scalano, uno dopo l'altro, K in cima, otteniamo il profilo

$$\boldsymbol{R}^{(1)}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline K & K & \cdots & K & K & K & \cdots & K \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B & B & \cdots & B & B & B & \cdots & B \end{array}$$

Per unanimità, sappiamo che in $\boldsymbol{R}^{(1)}$ vince K. Ora, supponiamo che gli elettori $1, \ldots, r-1$ scalino B in cima, in modo che il profilo diventi

$$\boldsymbol{R}^{(2)}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & K & K & \cdots & K \\ K & K & \cdots & K & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & B & B & \cdots & B \end{array}$$

Per monotonia, in $\boldsymbol{R}^{(2)}$ vince B o K; ma, per $\boldsymbol{\mathcal{MB}}$, è necessariamente K a vincere. Se, infine, anche r mette B al secondo posto, cioè il profilo diventa

$$\boldsymbol{Q}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & K & K & \cdots & K \\ K & K & \cdots & K & B & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & B & \cdots & B \end{array}$$

continua a vincere K; altrimenti, a r converrebbe non dichiarare il cambiamento, dato che continua a preferire K a B.

Ripartiamo adesso da $\boldsymbol{P}$: se $1, \ldots, r-1$ scalano, in sequenza, B in cima, otteniamo

$$\boldsymbol{S}^{(1)}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & K & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & B & B & \cdots & B \end{array}$$

in cui, per monotonia, possono vincere G o B; ma, per $\boldsymbol{\mathcal{MB}}$, vince G. Se, poi, r scala B in seconda posizione, si ha

$$\boldsymbol{S}^{(2)}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & K & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & B & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & B & \cdots & B \end{array}.$$

Ancora per monotonia, possono vincere G o B. D'altra parte, se r sposta B in testa, sappiamo che vince B (***WB***); ma r, in $\boldsymbol{S}^{(2)}$, preferisce sicuramente B a G. Se quindi vincesse G, a r converrebbe dichiarare di volere B al primo posto.

Infine, se gli elettori $1, \dots, r-1$ scalano K al secondo posto, e $r+1, \dots, r$ scalano K in cima, riotteniamo

$$\boldsymbol{Q}: \begin{array}{cccccccc} \mathbf{1} & \mathbf{2} & \cdots & \boldsymbol{r-1} & \boldsymbol{r} & \boldsymbol{r+1} & \cdots & \boldsymbol{n} \\ \hline B & B & \cdots & B & K & K & \cdots & K \\ K & K & \cdots & K & B & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & B & \cdots & B \end{array}.$$

Tuttavia, in entrambi $\boldsymbol{S}^{(2)}$ e $\boldsymbol{Q}$,

1. $1, \dots, r-1$ preferiscono B a K;
2. $r+1, \dots, n$ preferiscono K a B.

Pertanto, a ogni cambio di preferenza di $1, \dots, r-1$, deve continuare a vincere B, altrimenti a uno di questi elettori converrebbe non cambiare preferenza per veder vincere B; analogamente, a ogni cambio di preferenza di $r+1, \dots, n$, B continua a vincere (perché?). In definitiva, in $\boldsymbol{Q}$ vince B, e ciò conclude. □

La nuova proprietà ***OB*** smaschera definitivamente r: ha decisamente troppo potere. Per concludere resta solo da vedere che, in effetti, r è proprio un dittatore.

22.1.4 Il dittatore

Prendiamo un profilo $\boldsymbol{P}$ in cui r preferisce un candidato K a tutti gli altri. Vogliamo vedere che, in $\boldsymbol{P}$, vince K.

Supponiamo prima che $K \neq B$. Sappiamo, dalle ipotesi su f, che esiste un candidato C *distinto* da K e B: se ripetiamo il ragionamento che abbiamo fatto con r, troveremo un elettore s *pivot* per C, quindi tale che

i) ***MC.*** In un profilo in cui $s, \dots, n$ votano C per ultimo, C non vince.

ii) ***WC.*** In un profilo in cui $1, \dots, s$ votano C per primo, C vince.

iii) ***OC.*** Se tutti votano C per ultimo, vince il candidato preferito da s.

Che relazione c'è tra r e s? Prendiamo un profilo

$\mathbf{1}$	$\mathbf{2}$	$\cdots$	$\boldsymbol{r-1}$	$\boldsymbol{r}$	$\boldsymbol{r+1}$	$\cdots$	$\boldsymbol{n}$
B	B	$\cdots$	B	B	K	K	K
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C	C	$\cdots$	C	C	C	$\cdots$	C

Per ***WB***, vince B; d'altra parte, per ***OC***, vince il candidato preferito da s, che deve allora essere B. Di conseguenza, $s \leq r$. Viceversa, guardando

$\mathbf{1}$	$\mathbf{2}$	$\cdots$	$\boldsymbol{s-1}$	$\boldsymbol{s}$	$\boldsymbol{s+1}$	$\cdots$	$\boldsymbol{n}$
C	C	$\cdots$	C	C	K	K	K
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
B	B	$\cdots$	B	B	B	$\cdots$	B

otteniamo che $r \leq s$. Quindi, r e s sono la stessa persona, e r è un *pivot* sia che B che per C! Se allora prendiamo un *qualsiasi* profilo $\boldsymbol{P}$ in cui r preferisce $K \neq B$, possiamo, mantenendo le posizioni relative tra i candidati,

i) spostare B in fondo per tutti gli elettori, poi riportarlo dov'era;

ii) spostare C in fondo per tutti gli elettori, poi riportarlo dov'era.

Nel primo caso, per $\boldsymbol{\mathcal{O}B}$ e monotonia, vince B o K; nel secondo, per $\boldsymbol{\mathcal{O}C}$ e monotonia, vince C o K. Ma B, C, K sono distinti, e il vincitore è uno solo: in conclusione, è proprio K.

Resta il caso in cui r, che era un pivot per B, preferisce proprio B. Allora, esisteranno altri due candidati distinti A, C: è facile adattare il ragionamento per mostrare che r è anche un *pivot* per A; otterremo, come sopra, che l'unico caso possibile è che vinca B. Questo dimostra il teorema 3. □

22.2 Il teorema di Gibbard

Per ottenere una dimostrazione di quanto annunciato in ($\star$), dobbiamo ora eliminare l'ipotesi che il sistema sia *ordinale*, cioè che gli elettori si esprimano fornendo una classifica dei candidati.

Pensiamo allora a un sistema elettorale in questi termini. Abbiamo ancora il nostro insieme $\{1, \dots, n\}$ di elettori, con le loro personali preferenze sui candidati nell'insieme $\mathcal{C} = \{A, B, C, \dots\}$.

Stavolta, però, il loro voto non consiste *direttamente* in una preferenza P_i, ma più in generale ogni candidato i ha un insieme (finito) di *strategie* possibili per esprimere la propria preferenza: volutamente, non chiariamo *che cosa* effettivamente siano queste strategie, e pensiamo quindi ad esse come agli elementi di un insieme $\mathcal{S}_i$, l'insieme delle strategie dell'elettore i.

Dopo che gli elettori avranno votato, avremo quindi davanti un *profilo di strategie* $\boldsymbol{S}$, cioè una sequenza $(s_1, \dots, s_n)$ in cui ogni s_i è un elemento di $\mathcal{S}_i$, cioè è una strategia per l'elettore i. Il sistema elettorale sarà allora una funzione $f : \mathcal{S} \to \mathcal{C}$ dall'insieme di tutti i profili di strategie $\boldsymbol{S}$ possibili a quello dei candidati.

In questo contesto, che cosa vogliono dire le parole in ($\star$)? Se f è un sistema elettorale (nella nuova formulazione), diciamo che

1. f è *non manipolabile* se, ogni volta che un elettore i ha una preferenza P_i sui candidati, i ha a disposizione una *strategia dominante* $s_i^*(P_i)$, cioè una strategia per cui, comunque gli altri elettori scelgano strategie

$$s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n,$$

   e comunque i cambi $s_i^*(P_i)$ con un'altra strategia $s_i \in \mathcal{S}_i$, il candidato $f(s_1, \dots, s_i, \dots, s_n)$ non è preferito, in P_i, a $f(s_1, \dots, s_i^*(P_i), \dots, s_n)$;

2. f è *dittatoriale* se esiste un *dittatore* r, cioè un elettore tale che, per ogni candidato $A \in \mathcal{C}$, ha una strategia $s_A \in \mathcal{S}_r$ che, comunque gli altri elettori scelgano la propria strategia, assicura la vittoria ad A.

Theorem 4 (Gibbard). Se $\mathcal{C}$ ha almeno tre candidati, un sistema $f : \mathcal{S} \to \mathcal{C}$ non manipolabile che permette a tutti i candidati di vincere è dittatoriale.

Dimostrazione. Supponiamo che esista una funzione $f : \mathcal{S} \to \mathcal{C}$ che permetta la vittoria a ogni candidato e sia non manipolabile e non dittatoriale.

Possiamo allora costruire una funzione

$$g : \mathcal{P} \to \mathcal{C},$$

dove $\mathcal{P}$ è l'insieme dei profili di preferenze degli elettori, in questo modo. Dato un profilo di preferenze $\boldsymbol{P} = (P_1, \ldots, P_n)$, poniamo

$$g(\boldsymbol{P}) = f((s_1^*(P_1), \ldots, s_m^*(P_m)),$$

dove $s_i^*(P_i)$ è una strategia dominante per i rispetto alla preferenza P_i. Questa nuova funzione g rappresenta, in effetti, un sistema elettorale ordinale, come l'abbiamo definito sopra. Potete allora facilmente verificare che g viola il teorema di Gibbard-Satterthwaite, e pertanto f non può esistere. □

22.3 Esercizi

Per chi volesse cimentarsi, proponiamo qualche variante del teorema di Gibbard.

22.3.1 Un'elezione, più turni

Che succede ammettendo che le elezioni si possano svolgere in vari turni (come, ad esempio, nel caso in cui siano previsti ballottaggi)?

22.3.2 A pari merito

Ammettiamo che nelle preferenze due o più candidati possano essere a pari merito. Dimostrate che, se il sistema permette la vittoria a tutti i candidati, che sono almeno tre, e non è manipolabile, allora esistono un modo di ordinare gli elettori, diciamo $i_1, \ldots, i_n$, e uno di ordinare i candidati, diciamo $X_1, \ldots, X_m$ con la seguente proprietà.

Se associamo ad ogni candidato X una $(n+1)$-upla $(X^{(1)}, \ldots, X^{(n+1)})$ tale che, per $i \leq n$, $X^{(i)}$ indica la posizione del candidato X nella preferenza dell'elettore i, e $X^{(n+1)}$ è la posizione di X nell'ordine dei candidati, allora la prima di queste $(n+1)$-uple in ordine lessicografico è quella associata al candidato vincente.

22.3.3 Il teorema di Arrow

Si può richiedere che un sistema elettorale, invece di stabilire un vincitore, restituisca una classifica dei candidati. In tal caso vale il seguente risultato.

Theorem 5. Consideriamo un sistema elettorale con le proprietà seguenti.

i) *(Unanimità)* Dato un profilo $\boldsymbol{P}$ e due candidati A e B, se in $\boldsymbol{P}$ ogni elettore preferisce A a B, allora A è preferito a B per $\boldsymbol{P}$.

ii) *(Indipendenza dalle Alternative Irrilevanti)* Supponiamo di avere due profili $\boldsymbol{P}, \boldsymbol{Q}$ e due candidati A, B, tali che ogni elettore preferisce A a B in $\boldsymbol{P}$ se e solo se preferisce A a B in $\boldsymbol{Q}$; allora, A è preferito a B per $\boldsymbol{P}$ se e solo se A è preferito a B per $\boldsymbol{Q}$.

Allora, tale sistema è dittatoriale, nel senso che esiste un elettore i tale che, presi due candidati A e B e un profilo $\boldsymbol{P}$, A è preferito a B per $\boldsymbol{P}$ se e solo se A è preferito a B in P_i.

Questo è il teorema di Arrow (in una versione del 1963). Dimostratelo, usando il teorema di Gibbard.

Riferimenti bibliografici

[1] K. J. Arrow, *A difficulty in the concept of social welfare*, Journal of Political Economy Vol. 58 (1950).

[2] J.-P. Benoît, *The Gibbard-Satterthwaite theorem: a simple proof*, Economics Letters Vol. 69 (2000).

[3] A. Gibbard, *Manipulation of voting schemes: a general result*, Econometrica Vol. 41 (1973).

[4] M. A. Satterthwaite, *Strategy-proofness and Arrow's conditions: existence and correspondence theorems for voting procedures and social welfare functions*, Journal of Economic Theory Vol. 10 (1975).

23 Il principio di induzione e i numeri di Fibonacci

Alessandro Cordelli, n.12, Aprile 2021

Nelle dimostrazioni per induzione quasi sempre la maggior parte del lavoro consiste nel ricavare la validità della proposizione al passo successivo assumendo la sua validità a quello precedente; molto più semplice (spesso quasi banale) è il riconoscimento della validità della relazione in un caso particolare. Delle due condizioni comprese nel principio di induzione cioè, è la seconda quella su cui ci si concentra di più. Di fatto, non è facile trovare esempi significativi e non banali di proprietà estendibili da un intero al successivo che non sono tuttavia valide perché non verificate nemmeno in un caso particolare. Inoltre, l'analisi dei casi particolari è un momento fondamentale nel percorso di scoperta delle relazioni matematiche. Il presente articolo si concentra proprio sull'importanza del caso particolare, sia nel processo di scoperta di una relazione che come condizione necessaria per la sua dimostrazione, con un esempio in cui gli strumenti di calcolo richiesti non vanno oltre l'applicazione di semplici relazioni algebriche.

23.1 Introduzione

La successione di Fibonacci possiede un fascino difficilmente eguagliato da altri argomenti in matematica. Malgrado una definizione tutto sommato semplice, innumerevoli sono le sue applicazioni[33] all'algebra, alla geometria, all'analisi. Il presente contributo tratta proprio di una delle innumerevoli proprietà della successione di Fibonacci. Nella prima parte si va alla scoperta di tale proprietà, che poi però deve essere dimostrata e lo strumento adatto a questo scopo è il principio di induzione.[34]

Ora, come è noto, le dimostrazioni per induzione si basano su due passaggi logici: in primo luogo si verifica la proprietà in un caso particolare e poi si dimostra che se è valida all'ordine n allora è valida anche all'ordine $n+1$. Tanto nei casi più semplici come nelle dimostrazioni più complicate il grosso del lavoro consiste nel mostrare l'estensione della proprietà da n a $n+1$, mentre la verifica di un caso particolare è un compito quasi sempre banale.

Il termine "induzione" richiama il passaggio dal particolare all'universale, e in quanto tale non è solo uno strumento per la dimostrazione ma, prima ancora, per la scoperta di relazioni matematiche. In questo senso è particolarmente significativa la distinzione che fa George Polya [3] tra *induzione* (il riconoscimento della validità di una certa relazione) e *induzione matematica* (la sua dimostrazione). E poiché la scoperta di una relazione avviene (non sempre, ma molto spesso) osservando alcuni casi e riconoscendo una regolarità che si ripete senza eccezioni, quando si mette mano al processo dimostrativo il caso particolare di solito c'è già.

La relazione qui discussa è una proprietà della successione di Fibonacci che viene "scoperta" osservando alcuni casi particolari e dimostrata per induzione. Poiché i calcoli necessari per la dimostrazione si basano solo sulla relazione che definisce i numeri di Fibonacci (cioè che ogni elemento della successione è dato dalla somma dei due precedenti), sembrerebbe ragionevole aspettarsi che tale proprietà valga anche per qualsiasi altra successione generata con la stessa legge di quella di Fibonacci, anche se i valori iniziali non sono 1 e 1. Questo però non accade. Si dà cioè il caso notevole di una proprietà per la quale l'estensione da n a $n+1$ si può dimostrare utilizzando la relazione ricorsiva che definisce una qualsiasi successione analoga a quella di Fibonacci, ma la cui validità dipende in maniera stringente dai valori iniziali della successione stessa.

[33] Per una rassegna esaustiva sui numeri di Fibonacci, vedi [1].

[34] Sul principio di induzione vedi ad esempio [2].

Il presente articolo è così strutturato: il prossimo paragrafo è dedicato ad alcune considerazioni sul principio di induzione, mentre il successivo descrive il processo euristico di scoperta di una relazione tra i numeri di Fibonacci la cui dimostrazione viene sviluppata e discussa nelle sezioni 4 e 5; il sesto paragrafo contiene le conclusioni.

23.2 Induzione per scoprire, induzione matematica per dimostrare

"*840 è un numero molto particolare. Esso gode infatti di una notevolissima proprietà: è divisibile per qualsiasi numero. Non ci credete? Proviamo. Come tutti i numeri è ovviamente divisibile per 1. È un numero pari e quindi è divisibile per 2. La somma delle sue cifre è 12, un multiplo di 3, quindi è divisibile per 3. La sua metà è ancora un numero pari, quindi è divisibile per 4. Finisce per zero, quindi è divisibile per 5. Essendo divisibile per 2 e per 3 è anche divisibile per 6. E poi per 7, e anche per 8... Insomma, sappiamo che le coincidenze in matematica non esistono, e qui sembra proprio che questa proprietà abbia un po' troppe conferme per essere un caso...*". Ecco quale potrebbe essere l'inizio di una lezione un po' teatrale sul principio di induzione matematica.

Induzione e deduzione sono schemi di ragionamento complementari sia per quel che riguarda gli ambiti che gli obiettivi. L'induzione è il metodo di indagine principe nelle scienze della natura, secondo cui da un insieme di casi particolari che presentano tutti la stessa caratteristica si estrapola la presunta universalità di tale caratteristica sotto le medesime condizioni. Tuttavia tale universalità non potrà mai essere confermata (ciò richiederebbe la verifica sperimentale di un'infinità di casi particolari), cosicché una teoria fisica non può essere verificata, mentre può sempre essere falsificata [4]. Viceversa la deduzione, fondata sulla discesa dall'universale al particolare secondo lo schema del sillogismo aristotelico [5], garantisce l'assoluta certezza della tesi dimostrata a partire dalle ipotesi fatte. Ma anche qui c'è un prezzo da pagare, il fatto cioè che il punto di partenza di ogni ragionamento soffre di una ineliminabile arbitrarietà, per cui la verità non è tanto da intendersi come corrispondenza alla realtà quanto piuttosto come coerenza delle conclusioni con le premesse. Per dirla con le parole di un grande filosofo e matematico [6], *làmatematica è unàscienza nella quale non si sa di cosa si parla e non si sa se le affermazioni che vi si fanno sono vere o false.*

Da queste considerazioni sembrerebbe emergere una rigida separazione dei metodi: induzione nelle scienze della natura, deduzione in matematica. E tuttavia, come la fisica non si può ridurre alla ricerca euristica di leggi scollegate da una teoria generale, così il processo di scoperta in matematica non può prescindere dall'osservazione dei casi particolari alla ricerca di regolarità, che dovranno comunque essere successivamente dimostrate. Molto spesso nelle aule scolastiche la matematica viene presentata come una collezione di applicazioni di un certo numero di regole stabilite sulla base di nozioni e principi introdotti senza alcuna giustificazione: ciò non rende ragione né dello sviluppo storico di questa scienza né del lavoro dei matematici. Il momento deduttivo è la parte finale di un processo che molto spesso inizia proprio dall'osservazione dei casi particolari.

Ovviamente i casi particolari non bastano per asserire una legge, tutt'al più sostanziano una congettura. L'uso dei calcolatori negli ultimi decenni ha solo aumentato la quantità di casi particolari che possono essere presi in considerazione in un tempo ragionevole, ma non ha certo cambiato questo fondamentale limite logico. Anche con il numero impressionante di conferme che si possono ottenere con l'uso del calcolatore, una congettura è destinata a rimanere tale fino a che non viene prodotta una dimostrazione vera e propria. Questa è stata ad esempio la

storia dell'ultimo teorema di Fermat, enunciato nel 1637 ma dimostrato da Andrew Wiles[35] solo nel 1994 (tra l'altro utilizzando concetti e metodi che non potevano assolutamente essere nella cassetta degli attrezzi di un matematico del XVII secolo, cosicché un certo mistero intorno a questo teorema rimane). Un caso a parte è invece quello del teorema dei quattro colori (che asserisce che per colorare una qualsiasi mappa in modo che due regioni confinanti siano colorate diversamente bastano quattro colori), dimostrato analizzando per mezzo di un software tutte le numerose ma finite (1476) topologie non equivalenti che può avere una qualsiasi mappa piana.

L'analisi dei casi particolari può invece avere un valore probante quando si tratta di falsificare una congettura, basta infatti un solo controesempio per invalidare una proposizione universale. Un caso notevole in tal senso è quello di un altro teorema anch'esso legato al nome di Fermat. Già in un manoscritto cinese del 500 a.C. è presente questa congettura: l'espressione $2^n - 2$ è divisibile per n se e solo se n è un numero primo. In effetti, se facciamo qualche prova ci rendiamo conto della plausibilità di tale affermazione. Fermat riuscì a dimostrare il teorema diretto, cioè che se n è un numero primo, allora $2^n - 2$ è divisibile per n. Il teorema inverso, se dimostrato, costituirebbe un test per verificare se un certo numero è primo senza dover provare le divisioni per tutti i numeri primi fino alla radice quadrata del numero stesso. Di fatto però il teorema inverso non vale, e questo viene stabilito portando un controesempio. Non è tuttavia facile trovare un controesempio, dato che il più piccolo valore di n per cui la congettura cade è 341. Infatti $341 = 31 \cdot 11$, mentre $2^{341} - 2$ è divisibile per 341. Ora, il valore di $2^{341} - 2$ è un numero mostruosamente grande, dell'ordine di 10^{102}. È però degno di nota il fatto che il test di divisibilità di $\left(2^{341} - 2\right) : 341$ non è stato realizzato negli ultimi anni utilizzando un supercomputer, ma nel 1819 applicando in maniera ingegnosa le classi di resto [8].

Se l'induzione può essere una valida guida nella formulazione di una congettura, l'*induzione matematica* (o *principio di induzione*) è una cosa sostanzialmente diversa che permette la dimostrazione di enunciati universali. L'idea, come è noto, è che se una certa proprietà vale nel caso particolare $n = 0$ e, assumendone la validità nel caso n se ne può dedurre la validità nel caso $n + 1$, la proprietà è valida per ogni intero. Il principio di induzione si trova utilizzato già nel XVI secolo da Francesco Maurolico e, successivamente, da Fermat (che gli diede il nome di *principio della discesa infinita*) e Pascal, ma la formulazione moderna è quella del quinto assioma di Peano [9]: *Ogni sottoinsieme di numeri naturali che contenga lo zero e il successore di ogni proprio elemento coincide con l'intero insieme dei numeri naturali.* Naturalmente, se il più piccolo valore per cui una certa proprietà vale non è zero ma un qualche $k > 0$, tramite il principio di induzione viene stabilito che tale proprietà vale per tutti gli interi maggiori o uguali a k.

23.3 Giocando con la serie di Fibonacci

L'idea di prendere in considerazione la proprietà descritta in questo articolo nasce dal titolo accattivante di un video sui numeri di Fibonacci complessi [10], nel quale mi ero imbattuto casualmente. In realtà il video tratta dell'estensione a valori non interi della formula di Binet (che fornisce un'espressione esplicita anziché ricorsiva dei numeri di Fibonacci[36]) ma, prima ancora di vederlo, mi ero domandato come potrebbero essere dei numeri di Fibonacci complessi, e la cosa più semplice venutami in mente era stata modificare i valori iniziali della serie: $1, i$ anziché $1, 1$. La successione di Fibonacci così modificata diventa:

$$1, i, 1 + i, 1 + 2i, 2 + 3i, 3 + 5i, 5 + 8i, 8 + 13i, \ldots$$

[35] Sull'ultimo teorema di Fermat e la dimostrazione di Andrew Wiles vedi [7].

[36] Sulla formula di Binet vedi ad esempio [11].

La prima cosa che colpisce di questa successione è che le parti reale e immaginaria di ogni termine sono a loro volta i numeri di Fibonacci. Questo non è difficile da capire, infatti i primi due valori reali che appaiono nella sequenza sono 1 e 1, mentre i primi due valori immaginari sono i e i, e l'algoritmo per la generazione della successione è lo stesso del caso reale, cioè ogni termine è dato dalla somma dei due precedenti. Indicando quindi con c_n l'n-esimo termine della successione complessa e con f_n l'n-esimo numero di Fibonacci, avremo per $n \geq 3$: $c_n = f_{n-2} + i f_{n-1}$ con $c_1 = 1$ e $c_2 = i$. La successione c_n gode però di un'altra proprietà, non evidente a prima vista, che balza all'occhio prendendo in esame la successione dei moduli quadri $|c_n|^2$: $1, 1, 2, 5, 13, 34, 89, 233, \ldots$ vale a dire i termini della successione $|c_n|^2$ sono tutti numeri di Fibonacci, anche se non sono tutti i numeri di Fibonacci (per esempio mancano l'8, il 21, il 55...). Ricordando poi che la parte reale e quella immaginaria sono due f_n consecutivi e osservando che i termini della sequenza di Fibonacci che compaiono nella successione dei moduli quadri $|c_n|^2$ sono solo quelli di posto dispari, possiamo scrivere: $f_n^2 + f_{n-1}^2 = f_{2n-1}$.

Assumendo $f_0 = 0$ e facendo un po' di prove si constata che la relazione è verificata per $n \geq 1$ senza eccezioni, almeno per i primi valori di n. L'induzione ci ha così condotto a formulare una congettura, che però è destinata a rimanere tale fino a che non venga prodotta una dimostrazione.

Sembrerebbe quindi di essere giunti al momento di passare il testimone all'induzione matematica, ma a questo punto sorge un problema: nel passaggio $n \to n+1$ il secondo membro dell'uguaglianza diventa f_{2n+1}, che è uguale a $f_{2n} + f_{2n-1}$, cioè occorrerebbe anche il $2n$-esimo elemento della successione, mentre la formula che stiamo cercando di dimostrare fornisce solo gli elementi di posto dispari. In altri termini, sfruttando la definizione dei numeri di Fibonacci, bisognerebbe dimostrare che $f_n^2 + f_{n-1}^2 = f_{2n-1}$ implica $f_{n+1}^2 + f_n^2 = f_{2n+1}$, ma f_{2n+1} è a sua volta definito come $f_{2n-1} + f_{2n}$; ecco perché serve anche una relazione per i termini di posto pari. Apparentemente siamo finiti in un vicolo cieco, a meno che non si riesca a trovare una analoga formula per f_{2n}: il lavoro dell'induzione non è ancora terminato.

Facendo un po' di tentativi osserviamo che f_{2n} risulta sempre divisibile per f_n: $f_4 = 3$ è divisibile per $f_2 = 1$; $f_6 = 8$ è divisibile per $f_3 = 2$; $f_8 = 21$ è divisibile per $f_4 = 3$; $f_{10} = 55$ è divisibile per $f_5 = 5$; ecc... Inoltre il valore di $\frac{f_{2n}}{f_n}$ è esprimibile in maniera semplice in termini degli stessi numeri di Fibonacci. Infatti (a partire da $n = 3$):

$$\frac{f_6}{f_3} = 4 = 2 \cdot 2 + 0, \quad \frac{f_8}{f_4} = 7 = 2 \cdot 3 + 1, \quad \frac{f_{10}}{f_5} = 11 = 2 \cdot 5 + 1, \quad \ldots$$

sembra quindi che sia: $\frac{f_{2n}}{f_n} = 2 \cdot f_n + f_{n-3}$. Siamo così arrivati alla seconda relazione richiesta, quella per gli elementi di posto pari che, scritta esplicitamente è: $f_{2n} = 2f_n^2 + f_n \cdot f_{n-3}$.

Osserviamo che, se confermate, queste due relazioni permetterebbero di ottenere il valore di un particolare elemento nella sequenza di Fibonacci senza dover calcolare tutti i precedenti. Così, ad esempio, il calcolo di f_{50} mediante $f_{50} = f_{49} + f_{48}$ richiede i 49 valori precedenti, mentre con $f_{50} = 2f_{25}^2 + f_{25} \cdot f_{22}$ ne occorrono soltanto 25. L'applicazione delle due formule può anche essere iterata riducendo ulteriormente il numero di calcoli richiesti. Nel nostro esempio: $f_{25} = f_{13}^2 + f_{12}^2$ e $f_{22} = 2f_{11}^2 + f_{11} \cdot f_8$. A questo punto si possono ancora applicare le due formule, oppure considerare che sotto $n = 15$ i valori di f_n sono sufficientemente piccoli da essere calcolati con la formula ricorsiva:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, 610, \ldots$$

Quindi $f_{22} = 2 \cdot 89^2 + 89 \cdot 21 = 17711$ e $f_{25} = 233^2 + 144^2 = 75025$, da cui: $f_{50} = 2 \cdot 75025^2 + 75025 \cdot 17711 = 12586269025$.

23.4 Una dimostrazione per induzione

Avendo trovato due relazioni che riguardano i numeri di Fibonacci osservando un certo numero di casi particolari, la dimostrazione attraverso il principio di induzione si riduce a verificare il passaggio da n a $n+1$. Cioè, assumendo valide entrambe le relazioni della coppia:

$$\begin{cases} f_{2n-1} = f_n^2 + f_{n-1}^2 \\ f_{2n} = 2 \cdot f_n^2 + f_n \cdot f_{n-3} \end{cases}$$

bisognerà mostrare che valgono altresì entrambe le relazioni della nuova coppia:

$$\begin{cases} f_{2n+1} = f_{n+1}^2 + f_n^2 \\ f_{2n+2} = 2 \cdot f_{n+1}^2 + f_{n+1} \cdot f_{n-2} \end{cases}$$

Il punto di partenza, nonché ingrediente chiave della dimostrazione, è la regola che definisce i numeri di Fibonacci: $f_n = f_{n-1} + f_{n-2}$, che applicheremo anche all'elemento $n-1$, cioè:

$$f_{n-1} = f_{n-2} + f_{n-3}.$$

Partiamo quindi dall'espressione $f_{n-1} + f_{n-2} + f_{n-3}$, che può essere interpretata in due modi diversi a seconda dell'ordine in cui viene eseguita la somma:

$$(f_{n-1} + f_{n-2}) + f_{n-3} = f_n + f_{n-3},$$

ma anche:

$$f_{n-1} + (f_{n-2} + f_{n-3}) = 2f_{n-1}.$$

Quindi, uguagliando i secondi membri delle due precedenti uguaglianze:

$$f_n + f_{n-3} = 2f_{n-1}.$$

Moltiplichiamo adesso ambo i termini dell'uguaglianza per f_n e sommiamo f_n^2; in tal modo nel membro di sinistra rimane $2 \cdot f_n^2 + f_n \cdot f_{n-3}$ che per ipotesi è f_{2n}. Abbiamo quindi:

$$f_{2n} = 2f_n f_{n-1} + f_n^2.$$

A questo punto sommiamo ad ambo i termini l'espressione $f_n^2 + f_{n-1}^2$. In tal modo a sinistra rimane:

$$f_{2n} + \left(f_n^2 + f_{n-1}^2\right) = f_{2n} + f_{2n-1} = f_{2n+1}.$$

A destra, invece:

$$\left(f_{n-1}^2 + 2f_n f_{n-1} + f_n^2\right) + f_n^2 = f_{n+1}^2 + f_n^2.$$

Risulta così dimostrata la prima delle due relazioni: $f_{2n+1} = f_{n+1}^2 + f_n^2$.

Si passa adesso alla dimostrazione della relazione per i termini di posto pari. In questo caso si parte dall'espressione $2f_n f_{n+1}$ che svilupperemo in due modi diversi.

Primo sviluppo:

$$2f_nf_{n+1} = 2f_n \cdot (f_n + f_{n-1}) = f_n \cdot (2f_n + 2f_{n-1}) =$$
$$= f_n \cdot (2f_n + f_{n-1} + f_{n-1}) .$$

Esprimiamo il secondo f_{n-1} come $f_{n-2} + f_{n-3}$. In tal modo:

$$2f_nf_{n+1} = f_n \cdot [2f_n + (f_{n-1} + f_{n-2}) + f_{n-3}] =$$
$$= f_n \cdot (3f_n + f_{n-3}) = 3f_n^2 + f_nf_{n-3}.$$

Si arriva così a scrivere:

$$2f_nf_{n+1} = f_n^2 + \left(2f_n^2 + f_nf_{n-3}\right) = f_n^2 + f_{2n}.$$

Secondo sviluppo:

$$2f_nf_{n+1} = (f_n + f_{n-1} + f_{n-2})\, f_{n+1} =$$
$$= (f_{n+1} + f_{n-2})\, f_{n+1} = f_{n+1}^2 + f_{n+1}f_{n-2}$$

in cui si è posto $2f_n = (f_n + f_{n-1} + f_{n-2})$ e, nella parentesi, $f_n + f_{n-1} = f_{n+1}$.

Uguagliando i due sviluppi si ottiene:

$$f_n^2 + f_{2n} = f_{n+1}^2 + f_{n+1}f_{n-2}$$

In questa equazione basta sommare f_{n+1}^2 ad entrambi i membri per ottenere:

$$f_{n+1}^2 + f_n^2 + f_{2n} = 2f_{n+1}^2 + f_{n+1}f_{n-2}$$

Cioè $f_{2n+1} + f_{2n} = f_{2n+2} = 2f_{n+1}^2 + f_{n+1}f_{n-2}$, che è proprio la seconda relazione, quella che vale per i termini di posto pari.

23.5 Il ruolo del caso particolare

Se seguiamo in dettaglio tutti i passaggi della dimostrazione di entrambe le relazioni noteremo come l'unica proprietà utilizzata per trasformare una scrittura nell'altra sia stata, oltre all'ipotesi che le due relazioni valgano all'ordine n, la relazione che definisce i numeri di Fibonacci: $f_n = f_{n-1} + f_{n-2}$. Sembrerebbe quindi ragionevole aspettarsi che tali relazioni valgano in qualsiasi sequenza definita con la stessa legge di ricorrenza dei numeri di Fibonacci. La celebre successione $1, 1, 2, 3, 5, 8, 13, 21, \ldots$ non è infatti unica: utilizzando la stessa legge di ricorrenza ma cambiando i valori iniziali se ne possono ottenere quante se ne vogliono. Ad esempio, godono di un certo interesse i cosiddetti *numeri di Lucas* [12] L_n definiti esattamente mediante la stessa relazione di ricorrenza $L_n = L_{n-1} + L_{n-2}$ ma con valori iniziali $L_0 = 2$ e $L_1 = 1$. La successione di Lucas è quindi: $2, 1, 3, 4, 7, 11, 18, 29, \ldots$ Se si prova però ad applicare alla sequenza dei numeri di Lucas le relazioni dimostrate nella precedente sezione, si osserva che esse non valgono. Perché? Semplicemente, in questo caso è venuta meno una delle due condizioni della dimostrazione per induzione: il fatto cioè che la proprietà valga in un caso particolare.

Dal punto di vista strettamente logico, il fatto che le due relazioni dimostrate per i numeri di Fibonacci non valgano per *tutti* i numeri di Lucas non significa ancora che non valgono per

nessun numero di Lucas. Potrebbe infatti darsi il caso che tali relazioni siano verificate per un certo valore $n = \overline{n}$, magari anche molto grande. In tal caso la dimostrazione sviluppata per i numeri di Fibonacci ci garantirebbe la validità delle stesse relazioni per tutti i numeri di Lucas con $n \geq \overline{n}$. È però possibile dimostrare per altra strada[37] che, di fatto, le due relazioni che abbiamo dimostrato per i numeri di Fibonacci non possono valere in nessun caso per i numeri di Lucas.

Infatti, tra le molte proprietà dei numeri di Lucas ce ne sono due che li legano direttamente alla sequenza di Fibonacci [12]: $L_n^2 = 5f_n^2 + 4 \cdot (-1)^n$ e $L_n = f_{n-1} + f_{n+1}$. Supponiamo quindi (per assurdo) che le due relazioni che abbiamo dimostrato per i numeri di Fibonacci valgano anche per quelli di Lucas, per esempio quella relativa ai termini di posto dispari, cioè che: $L_{2n-1} = L_n^2 + L_{n-1}^2$. Applicando le due relazioni viste sopra a entrambi i termini dell'uguaglianza otteniamo: $f_{2n}+f_{2n-2} = 5\cdot\left(f_n^2 + f_{n-1}^2\right) = 5\cdot f_{2n-1}$ (nell'ultimo passaggio si è sfruttata la proprietà dimostrata per la sequenza di Fibonacci). Otteniamo così una ulteriore relazione per i numeri di Fibonacci: $f_{2n} + f_{2n-2} = 5 \cdot f_{2n-1}$ da cui, sviluppando f_{2n}, si ha: $f_{2n-1} + 2 \cdot f_{2n-2} = 5 \cdot f_{2n-1}$, cioè: $f_{2n-2} = 2 \cdot f_{2n-1}$, che possiamo anche scrivere come $f_{2n-2} = 2 \cdot f_{2n-2} + 2 \cdot f_{2n-3}$, da cui infine: $f_{2n-2} + 2 \cdot f_{2n-3} = 0$, che è manifestamente assurdo in quanto i numeri di Fibonacci sono interi positivi.

23.6 Conclusioni

Una coppia di relazioni ricorsive che permette di ottenere l'n-esimo numero di Fibonacci senza dover preventivamente calcolare tutti i precedenti ma solo i primi $\frac{n}{2}$ è stata congetturata dall'osservazione di un certo numero di casi particolari e dimostrata utilizzando la tecnica dell'induzione matematica.

Analizzando la dimostrazione, si vede chiaramente che l'unico ingrediente utilizzato è stata la relazione di ricorrenza che definisce i numeri di Fibonacci, il che farebbe supporre che le relazioni trovate si possano ugualmente applicare a qualsiasi altra successione generata con la stessa regola. Se però prendiamo in esame i numeri di Lucas (o anche altre successioni generate con la stessa regola dei numeri di Fibonacci), vediamo che questo non accade. Il motivo è che in tutti questi casi viene a mancare la prima delle condizioni che una dimostrazione per induzione deve rispettare: la validità in un caso particolare.

Le implicazioni epistemologiche sono interessanti. Se ci venisse chiesto se preferiamo dormire su un mucchio di paglia secca in una stanza con pareti, pavimento e soffitto in legno, nella quale non ci sono però fiammiferi, scintille, ecc., oppure su un divano in materiale ignifugo in una stanza in muratura in cui un paio di candele sono accese sopra un tavolo di marmo, molto probabilmente ci sentiremmo più sicuri nella seconda situazione, eppure il rischio di incendio è nullo nel primo come nel secondo caso. È come se l'estendibilità da n a $n + 1$ denotasse in qualche modo la possibilità per la validità di una certa relazione e il caso particolare l'evento contingente che ne innesca l'esistenza. Ma in matematica esistono le contingenze?

Riferimenti bibliografici

[1] K. J. Devlin, D. Didero, *I numeri magici di Fibonacci. L'avventurosa scoperta che cambiò la storia della matematica*, BUR Rizzoli, Milano (2013).

[2] A. Cordelli, *I fondamenti della matematica, la logica e gli insiemi*, e-book Kindle (2013), URL: `http://www.amazon.it/gp/product/B00CKY2G1U`.

[37] Sono debitore di questa dimostrazione all'amico prof. Giovanni Gaiffi.

[3] G. Polya, *Come risolvere i problemi di matematica*, Feltrinelli, Milano (1967).

[4] K. Popper, *Logica della scoperta scientifica. Il carattere autocorrettivo della scienza*, Einaudi, Torino (2010).

[5] Aristotele, *Analitici Primi (a cura di Milena Bontempi) in Organon (coord. Maurizio Migliori)*, Bompiani, Milano (2016).

[6] B. Russel, *Misticismo e Logica*, Longanesi, Milano (1964).

[7] S. Singh, *L'ultimo teorema di Fermat*, Rizzoli, Milano (1999).

[8] A. H. Beiler, *Recreations in the theory of numbers*, Dover Publications inc., New York (1966).

[9] G. Peano, *Arithmetices Principia (testo italiano e latino)*, Aragno, Torino (2001).

[10] Stand-up Maths, *Complex Fibonacci Numbers?*, `https://youtu.be/ghxQA3vvhsk`.

[11] E. W. Weisstein, *Binet's Fibonacci Number Formula*, from MathWorld – A Wolfram Web Resource: `https://mathworld.wolfram.com/BinetsFibonacciNumberFormula.html`.

[12] E. W. Weisstein, *Lucas Number*, from MathWorld – A Wolfram Web Resource: `https://mathworld.wolfram.com/LucasNumber.html`.

24 Poliedri equiscomponibili e teorema di Dehn

Lucio Tanzini e Cristofer Villani, n.13, Ottobre 2021

Consideriamo due poligoni, P e Q, nel piano, e supponiamo di suddividerli in pezzi poligonali più piccoli. Ci chiediamo quando P e Q siano *equiscomponibili*, cioè quando riusciamo a suddividerli in modo che i pezzi di P siano congruenti a quelli di Q.

Sicuramente, perché P e Q siano equiscomponibili, è necessario che abbiano la stessa area. Sorprendentemente, ciò è anche sufficiente: se cioè scegliamo due poligoni con la stessa area, riusciamo sempre a decomporli in pezzi poligonali a due a due congruenti. Una dimostrazione interattiva di questo risultato, il *teorema di Wallace-Bolyai-Gerwien*, si trova in [2].

Sorge allora la domanda: è vero anche per i poliedri? Più precisamente, dato un poliedro P, una *decomposizione* (poliedrica) di P (Figura 65) è un insieme di poliedri $P_1, \dots, P_n$ tali che

i) $P_1 \cup \cdots \cup P_n = P$;

ii) per ogni $1 \leq i, j \leq n$, P_i e P_j sono *quasi disgiunti*, cioè l'intersezione $P_i \cap P_j$ è contenuta nell'unione delle loro facce.

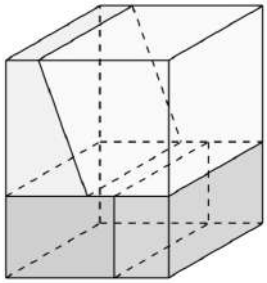


Figura 65: Una decomposizione di un parallelepipedo in quattro poliedri.

Ancora una volta, si ha certamente che, se due poliedri P e Q sono *equiscomponibili*, cioè ammettono decomposizioni $P_1, \dots, P_n$ e $Q_1, \dots, Q_n$ in cui P_i e Q_i sono congruenti per ogni i, allora P e Q hanno lo stesso volume.

Stavolta, però, scopriremo che il viceversa non è vero. In particolare, servendoci di uno strumento algebrico introdotto dal matematico Max Dehn nel 1901, che prende appunto il nome di *invariante di Dehn*, mostreremo che *un cubo e un tetraedro non sono mai equiscomponibili.*

24.1 Angoli diedrali

Sia P un poliedro, e sia l un suo spigolo. Oltre a considerare la sua lunghezza ℓ, possiamo associare a l un angolo θ, nel seguente modo.

Consideriamo le due facce di P, diciamo F, G, che si intersecano in l, e scegliamo un piano π perpendicolare a l, in modo che π intersechi F, G in due segmenti, r e s. Chiamiamo *angolo diedrale* associato a l l'angolo (piano) θ tra i segmenti r e s.

Nel seguito, misureremo gli angoli diedrali guardandoli come frazioni dell'angolo giro: pertanto, se ad esempio P è un cubo di volume unitario ed l è un qualsiasi suo spigolo, vale $\ell = 1$, e l'angolo diedrale associato a l è retto, per cui poniamo $\theta = 1/4$.

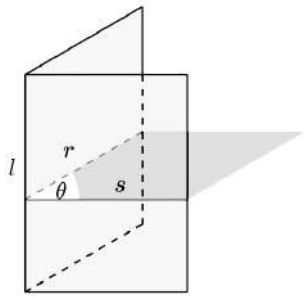


Figura 66: Angolo diedrale θ associato allo spigolo l.

24.2 L'invariante di Dehn

Possiamo adesso associare a uno spigolo l di un poliedro P due oggetti: la sua lunghezza ℓ e il suo angolo diedrale θ.

Per ottenere il nostro invariante, abbiamo bisogno di un modo di legarli.

Presi due numeri reali $\ell, \theta \in \mathbb{R}$, diciamo che il simbolo formale $\ell \otimes_{\mathbb{Q}} \theta$ (o, più semplicemente, $\ell \otimes \theta$) è un *tensore*. Stabiliamo inoltre che, se $q \in \mathbb{Q}$ è un numero razionale e T, T' sono tensori, anche i simboli

$$0, \qquad T + T', \qquad q \cdot T$$

sono tensori. Quindi, sono ad esempio tensori

$$\ell_1 \otimes \theta_1 + \ell_2 \otimes \theta_2, \qquad q_1 \cdot (\ell_1 \otimes \theta_1) + \ell_2 \otimes \theta_2 + q_3 \cdot (\ell_3 \otimes \theta_3),$$

e così via, per $\ell_i, \theta_i \in \mathbb{R}$, e $q_i \in \mathbb{Q}$.

Riferendoci ai tensori come simboli formali, intendiamo che non pensiamo a un tensore come a un preciso oggetto matematico, ma come a una stringa di caratteri che possiamo manipolare algebricamente secondo regole precise. Le mosse possibili sono le seguenti.

S) Se ℓ, ℓ', θ sono numeri reali, valgono

$$\begin{aligned}(\ell + \ell') \otimes \theta &= \ell \otimes \theta + \ell' \otimes \theta,\\ \ell \otimes (\theta + \theta') &= \ell \otimes \theta + \ell \otimes \theta';\end{aligned}$$

P) Se $q \in \mathbb{Q}$ e $\ell, \theta \in \mathbb{R}$, vale

$$q \cdot (\ell \otimes \theta) = (q \cdot \ell) \otimes \theta = \ell \otimes (q \cdot \theta);$$

R) Se $\ell, \theta, \theta' \in \mathbb{R}$ e $\theta - \theta'$ è un numero razionale, allora

$$\ell \otimes \theta = \ell \otimes \theta'$$

e, in particolare, $\ell \otimes \theta = \ell \otimes 0$ se $\theta \in \mathbb{Q}$;

U) Vale $\ell \otimes \theta = 0$ se e solo se $\ell = 0$ oppure $\theta \in \mathbb{Q}$.

Rimarchiamo che la possibilità di sommare numeri razionali senza cambiare il tensore (mossa (R)) vale solo per il termine *a destra* di $\otimes$ (l'angolo diedrale), e che è possibile spostare un coefficiente q tra i termini del tensore e fuori da esso (mossa (P)) *solo se q è razionale.*

Osserviamo inoltre che, con le mosse che abbiamo stabilito, non possiamo decidere se due tensori qualsiasi sono uguali; d'altra parte, possiamo in effetti cercare di ridurre un tensore T nella forma $\ell \otimes \theta$ usando le regole (S), (P) e (R); se ci riusciamo, sappiamo dire se è uguale a zero oppure no grazie alla regola (U). Vedremo che, per i nostri scopi, questo è sufficiente.

Possiamo ora definire l'invariante di Dehn.

Definizione 2. Sia P un poliedro con n spigoli, diciamo $l_1, \dots, l_n$, e siano $\ell_1, \dots, \ell_n$ e $\theta_1, \dots, \theta_n$ le lunghezze e gli angoli diedrali rispettivi. L'*invariante di Dehn* associato a P è il tensore

$$\begin{aligned}\langle P \rangle &= \ell_1 \otimes \theta_1 + \cdots + \ell_n \otimes \theta_n \\ &= \sum_{i=1}^{n} \ell_i \otimes \theta_i.\end{aligned}$$

Supponiamo, per esempio, che P sia un cubo di lato ℓ, con $\ell \in \mathbb{R}$. Si ha allora $\ell_i = \ell$ e $\theta_i = 1/4$ per ogni $i = 1, \dots, 12$, da cui

$$\begin{aligned}\langle P \rangle &= \ell \otimes 1/4 + \cdots + \ell \otimes 1/4 && \text{(S)} \\ &= (\ell + \cdots + \ell) \otimes 1/4 \\ &= (12 \cdot \ell) \otimes 1/4 && \text{(R)} \\ &= (12 \cdot \ell) \otimes 0 && \text{(U)} \\ &= 0.\end{aligned}$$

24.3 L'invariante di Dehn del tetraedro

Consideriamo adesso un tetraedro regolare Q, tale che uno spigolo l abbia lunghezza ℓ_Q e angolo diedrale θ_Q. L'invariante di Dehn di Q è allora

$$\langle Q \rangle = \ell_Q \otimes \theta_Q + \cdots + \ell_Q \otimes \theta_Q = (6 \cdot \ell_Q) \otimes \theta_Q$$

che, per la mossa (U), è uguale a zero se e solo se $\theta_Q = 0$, cioè θ_Q è un numero razionale. Vogliamo vedere che, in effetti, $\theta_Q \notin \mathbb{Q}$, da cui $\langle Q \rangle \neq 0$.

Iniziamo calcolando θ_Q. Chiamiamo F, G le due facce del tetraedro che si intersecano in l, e fissiamo una delle due, diciamo F, come base di Q. Consideriamo ora il triangolo rettangolo T i cui lati sono, rispettivamente,

i) l'altezza h del tetraedro rispetto a F;

ii) il raggio r di F su l;

iii) l'apotema a di Q su G.

Per le proprietà elementari dei triangoli, e dato che F e G sono congruenti, sappiamo che $r = \frac{1}{3}a$. Inoltre, dalla definizione segue che $2\pi\theta_Q$ è l'angolo tra r e a in T (il fattore 2π viene dal fatto che stiamo misurando gli angoli come frazioni dell'angolo giro). Pertanto,

$$\cos(2\pi\theta_Q) = r/a = \frac{1}{3}.$$

Per mostrare che $\theta_Q \notin \mathbb{Q}$, supporremo che sia un numero razionale e, usando l'uguaglianza sopra, vedremo come questo conduca a un assurdo.

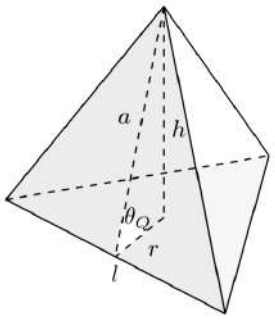


Figura 67: Il tetraedro Q. Le facce ombreggiate sono la faccia F (quella inferiore) e la faccia G.

Supponiamo allora che $\theta_Q = m/n$, con m/n una frazione ridotta ai minimi termini. Ne segue che, se chiamiamo $\alpha = 2\pi\theta_Q$,

$$n\alpha = 2\pi m$$

è un multiplo intero dell'angolo giro. Di conseguenza, deve essere

$$\cos(n\alpha) = 1.$$

Per vedere che questo è impossibile, cerchiamo una relazione che leghi $\cos(n\alpha)$ a $\cos(\alpha)$: come saprete, se $n = 2$, tale relazione è data dalla formula di duplicazione del coseno, vale a dire

$$\cos(2\alpha) = 2\cos^2(\alpha) - 1.$$

Se quindi poniamo $\cos(\alpha) = x$ e $\cos(2\alpha) = T_2(x)$, otteniamo $T_2(x) = 2x^2 - 1$. Vorremmo ottenere un risultato analogo in generale, ponendo $T_n(x) = \cos(n\alpha)$. Per farlo, osserviamo che, per $n > 2$, le formule di addizione e sottrazione danno

$$\cos((n+1)\alpha) = \cos(n\alpha + \alpha) = \cos(n\alpha)\cos(\alpha) - \sin(n\alpha)\sin(\alpha)$$
$$\cos((n-1)\alpha) = \cos(n\alpha - \alpha) = \cos(n\alpha)\cos(\alpha) + \sin(n\alpha)\sin(\alpha)$$

da cui, sommando membro a membro,

$$\cos((n+1)\alpha) + \cos((n-1)\alpha) = 2\cos(\alpha)\cos(n\alpha),$$

ovvero

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Non ci interessa conoscere esplicitamente la formula per $T_n(x)$, ma da questa relazione si deducono facilmente, usando un ragionamento induttivo, due cose:

i) per ogni $n \geq 1$, $T_n(x)$ è un polinomio (detto *polinomio di Čebyšev di prima specie*) di grado n, diciamo
$$T_n(x) = c^{(n)}x^n + \text{ termini di grado più basso;}$$

ii) per ogni $n \geq 1$, $c^{(n)} = 2^{n-1}$.

Se avete familiarità con l'induzione, non avrete problemi a verificare (i) e (ii); altrimenti, potete assumere che valgano, limitandovi a controllare che siano vere nel caso $n = 2$.

Torniamo adesso alla nostra uguaglianza, che possiamo riscrivere come

$$T_n(x) = 1,$$

dove $x = \cos(\alpha)$ è uguale, nel nostro caso, a 1/3. Usando le proprietà di T_n appena viste, otteniamo

$$2^{n-1}\left(\frac{1}{3}\right)^n + \text{ termini di grado più basso} = 1.$$

Tuttavia, moltiplicando i due membri dell'uguaglianza per 3^n, si ha

$$2^{n-1} + \text{ termini divisibili per } 3 = 3^n,$$

il che vorrebbe dire che 2^{n-1} è uguale alla differenza di due multipli di 3, e quindi è a sua volta divisibile per 3. Dato che questo non è possibile, ne deduciamo che la nostra assunzione $\theta_Q \in \mathbb{Q}$ è assurda, e θ_Q è un numero irrazionale.

In conclusione, per quanto osservato all'inizio del paragrafo, l'invariante di Dehn di un tetraedro regolare è diverso da 0.

24.4 Il teorema di Dehn

Abbiamo quindi ottenuto che, se P e Q sono rispettivamente un cubo e un tetraedro regolare, $\langle P\rangle \neq \langle Q\rangle$. Per concludere, ci resta allora da provare il seguente risultato.

Theorem 6 (Dehn). Se P e Q sono due poliedri equiscomponibili, allora

$$\langle P\rangle = \langle Q\rangle.$$

L'ingrediente fondamentale per la dimostrazione è la risposta alla seguente domanda: se decomponiamo un poliedro P nei pezzi $P_1, \ldots, P_n$, come si comporta l'invariante di Dehn $\langle P\rangle$ rispetto a $\langle P_1\rangle, \ldots, \langle P_n\rangle$? Per i nostri scopi, possiamo limitarci a considerare decomposizioni più "pulite" di quelle definite inizialmente. Nello specifico,

Definizione 3. Diciamo che una decomposizione $P_1, \ldots, P_n$ di un poliedro P è *netta* se, per ogni $1 \leq i, j \leq n$, l'intersezione $P_i \cap P_j$ è un vertice, uno spigolo o una faccia di entrambi.

Per decomposizioni nette, l'invariante di Dehn è *additivo*, nel senso che

Proposizione 7. *Se $P_1, \ldots, P_n$ è una decomposizione netta di un poliedro P, allora*

$$\langle P\rangle = \langle P_1\rangle + \cdots + \langle P_n\rangle = \sum_{i=1}^{n} \langle P_i\rangle.$$

Dimostrazione. Consideriamo uno spigolo l, di lunghezza $\ell(l)$, di uno dei poliedri P_i. Supponiamo che l sia contenuto in $P_{i_1}, \ldots, P_{i_k}$ (e non contenuto nei restanti P_j) con angoli diedrali $\theta_{i_1}, \ldots, \theta_{i_k}$ rispettivamente, e poniamo

$$\Theta_l = \theta_{i_1} + \cdots + \theta_{i_k}.$$

Dal fatto che la decomposizione è netta, abbiamo che i poliedri P_i che intersecano l in un suo segmento sono esattamente $P_{i_1}, \ldots, P_{i_k}$, e pertanto

i) se l è interno a P, cioè non contenuto nell'unione delle sue facce, i P_{i_j} descrivono complessivamente attorno a l un angolo giro, cioè $\Theta_l = 1$;

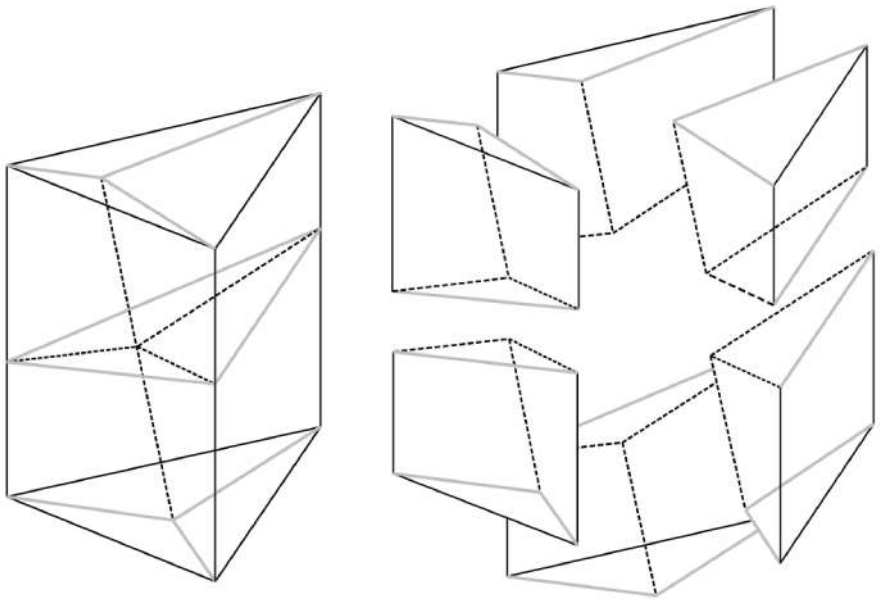


Figura 68: Decomposizione netta di un prisma a base triangolare, in cui gli spigoli di tipo (i), (ii), (iii) sono tratteggiati, grigi e neri rispettivamente.

ii) se l è contenuto nell'unione delle facce di P, ma non in un suo spigolo, i P_{i_j} descrivono complessivamente attorno a l un angolo piatto, e $\Theta_l = 1/2$;

iii) se l è contenuto in uno spigolo L di P, Θ_l è proprio l'angolo diedrale θ_L associato a L in P.

Guardiamo ora $\langle P_1\rangle + \cdots + \langle P_n\rangle$: per definizione, ognuno dei $\langle P_i\rangle$ è somma di tensori $\ell(l)\otimes\theta$, dove l è uno spigolo di P_i e θ è il suo angolo diedrale (in P_i).

Se esplicitiamo $\langle P_1\rangle + \cdots + \langle P_n\rangle$ e raccogliamo i tensori che condividono lo stesso spigolo, otteniamo

$$\sum_{i=1}^{n}\langle P_i\rangle = \sum_{l}\left(\ell(l)\otimes\theta_{i_1} + \cdots + \ell(l)\otimes\theta_{i_k}\right),$$

dove l varia tra gli spigoli distinti dei P_i e i θ_{i_j} (che dipendono da l) sono come sopra.

Di conseguenza, se l è come in (i) e ha lunghezza ℓ, otteniamo

$$\ell\otimes\theta_{i_1} + \cdots + \ell\otimes\theta_{i_k} = \ell\otimes(\theta_{i_1} + \cdots + \theta_{i_k}) = l\otimes\Theta_l = 0,$$

dato che $\Theta_l = 1 \in \mathbb{Q}$. Analogamente, se l è come in (ii),

$$\ell\otimes\theta_{i_1} + \cdots + \ell\otimes\theta_{i_k} = \ell\otimes 1/2 = 0,$$

mentre per l come in (iii) abbiamo

$$\ell\otimes\theta_{i_1} + \cdots + \ell\otimes\theta_{i_k} = \ell\otimes\Theta_l = \ell\otimes\theta_L.$$

Quindi, gli unici tensori che contano nella somma sopra sono quelli in cui l è contenuto in uno spigolo L di P. D'altra parte, poiché i P_i decompongono P, gli spigoli $l_1,\ldots,l_h$ dei P_i contenuti in uno *stesso* L sono tali che $L = l_1\cup\cdots\cup l_h$, e inoltre

i) se L è lungo ℓ_L e l_i è lungo ℓ_i,
$$\ell_L = \ell_1 + \cdots + \ell_h;$$

ii) se θ_L è l'angolo diedrale associato a L in P, per ogni i si ha $\Theta_{l_i} = \theta_L$.

Ma allora, se nella somma sopra raccogliamo ulteriormente i tensori in cui l è contenuto in L, abbiamo addendi della forma

$$\ell_1 \otimes \theta_L + \cdots + \ell_h \otimes \theta_L = (\ell_1 + \cdots + \ell_h) \otimes \theta_L = \ell_L \otimes \theta_L.$$

In definitiva, otteniamo
$$\sum_{i=1}^{n} \langle P_i \rangle = \sum_{L} \ell_L \otimes \theta_L$$
che, per definizione, è $\langle P \rangle$. □

Prendiamo ora una decomposizione $P_1, \ldots, P_n$ *qualsiasi* di un poliedro P: se, per ogni faccia di ciascuno dei P_i, tracciamo il piano in cui tale faccia è contenuta, otteniamo che ogni P_i è decomposto dai piani tracciati in (finiti) poliedri $P_i^{(1)}, P_i^{(2)}, \ldots$ in modo che

i) per ogni i, $P_i^{(1)}, P_i^{(2)}, \ldots$ è una decomposizione *netta* di P_i;

ii) complessivamente, i $P_i^{(j)}$ danno una decomposizione *netta* di P.

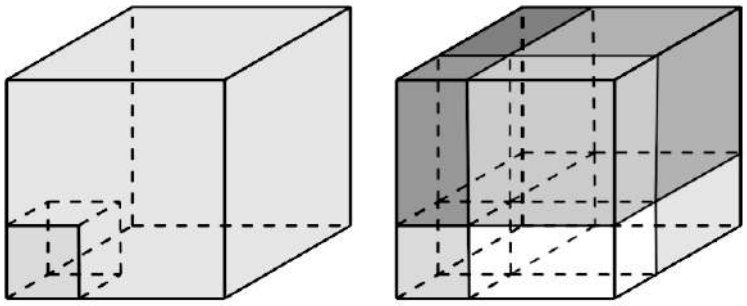


Figura 69: La decomposizione di cubo in un cubo più piccolo e nel suo complementare, resa netta tagliando il cubo con i piani associati alle facce del cubo piccolo.

Dimostrazione (del teorema). Supponiamo che due poliedri P e Q siano equiscomponibili, e prendiamo decomposizioni (qualsiasi) $P_1, \ldots, P_n$ e $Q_1, \ldots, Q_n$ di P e Q rispettivamente, in modo che il poliedro P_i sia congruente a Q_i.

Costruiamo quindi decomposizioni $P_i^{(j)}$ e $Q_i^{(j)}$ di P e Q come sopra: notiamo che, in generale, i pezzi delle decomposizioni non saranno a due a due congruenti.

Per ogni i, però, i $P_i^{(j)}$ danno una decomposizione netta di P_i: grazie alla proposizione 7, possiamo quindi scrivere
$$\langle P_i \rangle = \langle P_i^{(1)} \rangle + \langle P_i^{(2)} \rangle + \cdots,$$
e complessivamente otteniamo
$$\sum_i \langle P_i \rangle = \sum_{i,j} \langle P_i^{(j)} \rangle.$$

Ma, ancora per la proposizione 7, il membro a destra dell'uguaglianza è proprio $\langle P \rangle$, dato che i $P_i^{(j)}$ forniscono una decomposizione netta di P.

D'altra parte, dato che P_i è congruente a Q_i, abbiamo in particolare che $\langle P_i \rangle = \langle Q_i \rangle$, e perciò

$$\sum_i \langle P_i \rangle = \sum_i \langle Q_i \rangle.$$

Applicando lo stesso discorso a Q, otteniamo allora

$$\sum_i \langle Q_i \rangle = \sum_{i,j} \langle Q_i^{(j)} \rangle = \langle Q \rangle,$$

e unendo le uguaglianze trovate arriviamo a $\langle P \rangle = \langle Q \rangle$, che era quello che volevamo. □

Abbiamo perciò mostrato che esistono poliedri con lo stesso volume che non sono equiscomponibili, usando il fatto che i loro invarianti di Dehn sono distinti.

Concludiamo notando che, nel 1965, J.P. Sydler ha dimostrato che vale anche il viceversa del teorema 6: se due poliedri P e Q hanno lo stesso volume e *lo stesso invariante di Dehn* allora sono, effettivamente, equiscomponibili.

24.5 Esercizi

1. Mostrate che un cubo P di lato ℓ e un prisma Q di altezza ℓ la cui base è un poligono regolare di area ℓ^2 sono equiscomponibili (potete usare il teorema di equiscomponibilità dei poligoni). Quanto vale $\langle Q \rangle$?
2. Calcolate gli angoli diedrali dei poliedri regolari, mostrando che hanno i valori seguenti.

Poliedro	**Angolo diedrale** $(2\pi\theta)$
Tetraedro	$\arccos(1/3)$
Cubo	$\pi/2$
Ottaedro	$\arccos(-1/3)$
Dodecaedro	$\arccos(-\sqrt{5}/5)$
Icosaedro	$\arccos(-\sqrt{5}/3)$

3. Usando l'esercizio precedente, mostrate che un cubo e un poliedro regolare sono equiscomponibili se e solo se sono congruenti.

Riferimenti bibliografici

[1] M. Dehn, *Uber Dem Rauminhalt*, Math Ann. 55 (1902), pp. 465-478.

[2] D. Smirnov and Z. Epstein, *An interactive demonstration of the Wallace–Bolyai–Gerwien theorem*, `https://dmsm.github.io/scissors-congruence/`.

[3] R. Schwartz, *Dehn's Dissection Theorem*, `http://www.math.brown.edu/reschwar/Papers/dehn.pdf`.

[4] J.-P. Sydler, *Conditions necessarires et suffisiantes pour l'équivalence des polyedres de l'espace euclidean a trois dimensions*, Comment Math Helv. 40 (1965).

25 Teoria dei campi e costruzioni con riga e compasso

Antonio Di Nunzio, n.14, Aprile 2022

25.1 Introduzione

Sin dalle scuole medie, abbiamo avuto a che fare con la riga e il compasso: abbiamo costruito triangoli equilateri, quadrati, pentagoni, esagoni... ma anche l'asse di un segmento o la bisettrice di un angolo. Queste costruzioni si realizzano seguendo una procedura in cui si tracciano linee rette e archi di circonferenza, determinando, passo dopo passo, i punti che individuano la figura da costruire (come ad esempio i vertici di un poligono). I *problemi di costruzione con riga e compasso* risalgono all'antica Grecia e hanno stimolato l'intero sviluppo della geometria euclidea. Questi richiedono di costruire un certo oggetto geometrico a partire da un dato insieme di punti, facendo uso soltanto della riga e del compasso.

Nella loro moltitudine, ve ne sono alcuni che sono stati studiati per secoli dai più illustri matematici, ma che non hanno mai trovato soluzione. I più importanti sono la **duplicazione del cubo** (cioè la costruzione di un cubo di volume doppio rispetto a un cubo assegnato), la **quadratura del cerchio** (cioè la costruzione di un quadrato avente la stessa area di un cerchio assegnato) e la **trisezione dell'angolo** (cioè la determinazione di un metodo generale che permetta di dividere un angolo assegnato in tre parti congruenti). Nel corso dei secoli, il continuo fallimento nella ricerca di una soluzione spinse la comunità matematica a ritenere che questi problemi fossero impossibili da risolvere. Nacque quindi un nuovo problema: come si fa a dimostrare che effettivamente queste costruzioni non sono realizzabili?

La matematica, come si sa, è ricca di sorprese: un problema che risiede in una certa branca può essere interpretato e tradotto in termini di un'altra dove lo studio ne risulta più semplice o efficace. Questo è proprio quello che accade nel nostro caso: i problemi di costruzione risiedono nella geometria, ma possono essere tradotti in linguaggio algebrico ed essere affrontati con metodi e strumenti propri dell'algebra; strumenti in grado di formalizzare e dimostrare il fatto che i tre problemi precedenti non abbiano soluzione.

In questo articolo scopriremo il processo di traduzione in linguaggio algebrico di un problema di costruzione con riga e compasso, addentrandoci tacitamente nella *teoria dei campi*: un ramo dell'algebra moderna sviluppato principalmente per indagare la risolubilità delle equazioni polinomiali, ma che oggi riveste un ruolo centrale in diverse aree della matematica. Il suo linguaggio, come vedremo, si adatta anche al contesto delle costruzioni con riga e compasso, ed è in grado di offrire una risposta a questi tre famosi problemi.

Iniziamo a dare qualche dettaglio sui primi due, dando per buono che partendo da un segmento è sempre possibile costruire un quadrato sfruttando solo riga e compasso (a breve, ciò potrà essere facilmente dedotto).

Assegnato un cubo avente spigoli di misura ℓ, il suo volume è $V = \ell^3$. Si vuole costruire un cubo di volume $V' = 2V = 2\ell^3$. Questo avrà come spigolo un segmento di lunghezza $\ell' = \ell\sqrt[3]{2}$. Il problema si riduce a determinare se, dato un segmento di lunghezza ℓ, sia possibile costruire con riga e compasso un segmento di lunghezza $\ell\sqrt[3]{2}$ (Figura 70, sinistra).

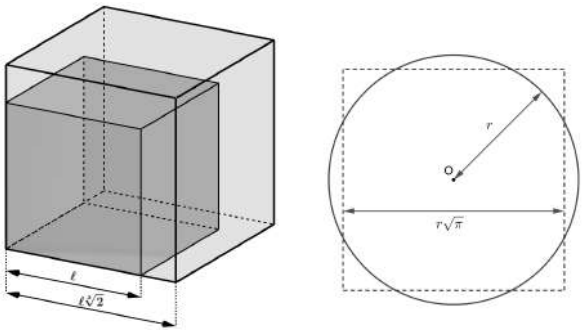


Figura 70: Duplicazione del cubo e quadratura del cerchio.

Assegnato un cerchio di raggio r, la sua area è $A = \pi r^2$. Si vuole costruire un quadrato di area A. Questo avrà come lato un segmento di lunghezza $\ell = r\sqrt{\pi}$. Il problema si riduce a determinare se, dato un segmento di lunghezza r, sia possibile costruire con riga e compasso un segmento di lunghezza $r\sqrt{\pi}$ (Figura 70, destra).

25.2 Costruzioni con riga e compasso e numeri costruibili

Cominciamo richiamando i concetti matematici che sono alla base del nostro studio, in particolare specifichiamo cosa vuol dire effettuare una costruzione con riga e compasso. Innanzitutto precisiamo che in questo contesto gli strumenti riga e compasso si considerano Â«idealiÂ», cioè non graduati e quindi senza la possibilità di effettuare misurazioni. Questi vincoli discendono dalla struttura assiomatica della geometria euclidea, infatti le operazioni "lecite" sono quelle stabilite dai primi tre **postulati di Euclide**, così enunciati:

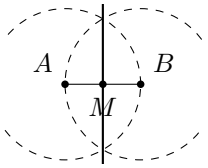
1. *È possibile condurre una linea retta da un qualsiasi punto ad ogni altro punto.*
2. *È possibile prolungare illimitatamente una retta finita (un segmento) in linea retta.*
3. *È possibile descrivere un cerchio con qualsiasi centro e distanza (raggio) qualsiasi.*

Quindi, la riga e il compasso devono essere considerati come strumenti che permettono di effettuare *solo* queste tre operazioni (e non altre derivanti dalla loro struttura fisica, come ad esempio misurare e riportare la misura di segmenti o conservare l'apertura del compasso).

Nella formulazione classica quindi, una *costruzione con riga e compasso* è una sequenza finita di queste tre operazioni di base, partendo da almeno due punti.

Vediamo alcuni esempi, a cui nel seguito ci riferiremo come *costruzioni elementari*.

1. **L'asse di un segmento**



Dato un segmento di estremi A e B, si costruiscono la circonferenza di centro A e passante per B, e la circonferenza di centro B e passante per A; si individuano i punti di intersezione tra le due circonferenze e si traccia la retta che li congiunge.

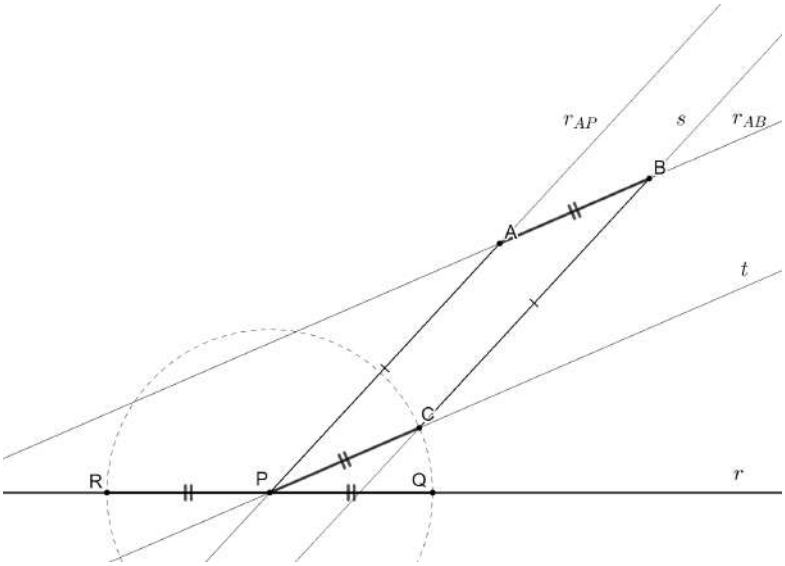
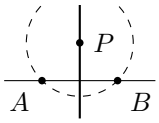


Figura 71: Dati un segmento di estremi A e B, una retta r e un punto $P \in r$, si costruiscono la retta r_{AP} passante per A e P, la retta r_{AB} passante per A e B, la retta s parallela a r_{AP} passante per B e la retta t parallela a r_{AB} passante per P. Si individua il punto C di intersezione tra s e t e si costruisce la circonferenza di centro P e passante per C. Questa individua due punti Q e R su r, ciascuno dei quali determina con P un segmento congruente ad AB.

2. **La perpendicolare ad una retta data in un punto dato**



Data una retta r e un punto P, si costruisce una circonferenza di centro P che interseca r in due punti A e B e si traccia l'asse del segmento AB.

3. **La parallela ad una data retta in un punto dato**
   Data una retta r e un punto P, si costruisce la perpendicolare s a r passante per P e si costruisce la perpendicolare r' a s passante per P.
4. **Un segmento della stessa lunghezza di un segmento dato su una retta data partendo da un punto dato**
   Riuscite a effettuare da soli questa operazione (ovvero quella di "applicare un segmento a una retta") tramite una costruzione con riga e compasso che sfrutti le precedenti costruzioni elementari? Per la soluzione, vedete la Figura 71 e la sua didascalia.

Il primo passo da compiere è quello di gettare una base analitica: a partire da due punti, possiamo fissare un'unità di misura (la loro distanza) e possiamo costruire un sistema di assi cartesiani con origine in uno dei due (possiamo tracciare la retta per i due punti e costruire la sua perpendicolare per uno dei due). Questo è quello che facciamo ad esempio su un foglio a quadretti: possiamo scegliere un punto come origine e scegliere la lunghezza del lato di un quadretto come unità di misura.

Consideriamo allora un sistema di assi cartesiani con origine O e un punto U di coordinate $(1, 0)$. A partire da O e U possiamo costruire rette e circonferenze, inoltre possiamo individuare i punti di intersezione degli oggetti costruiti e sfruttarli per fare altre costruzioni. Definiamo quindi i **punti**, le **rette** e le **circonferenze costruibili** secondo le seguenti regole ricorsive:

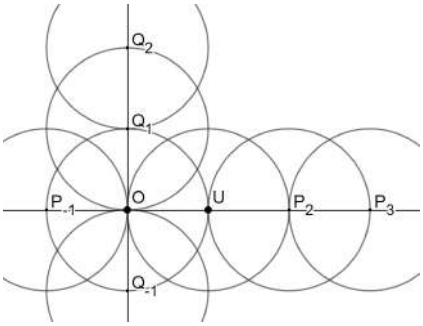


Figura 72: Costruzione dei punti di coordinate $(m, 0)$ e $(0, n)$.

- I punti O e U sono costruibili.
- Una retta tra punti costruibili è costruibile.
- Una circonferenza con centro un punto costruibile e passante per un punto costruibile è costruibile.
- Un punto di intersezione di due rette costruibili o di una retta e una circonferenza costruibili o di due circonferenze costruibili è costruibile.

Il nostro obiettivo è quello di determinare l'insieme dei punti costruibili, che indichiamo con Π, cioè l'insieme di tutti quei punti che sono raggiungibili mediante un numero finito di costruzioni a partire da O e U. Vediamo ad esempio che partendo da O e U è possibile costruire tutto il reticolo dei punti del piano di coordinate intere:

nel nostro sistema di assi cartesiani, possiamo costruire la circonferenza di centro O e passante per U, individuando sugli assi i punti P_{-1} , Q_{-1} , Q_1 di coordinate rispettivamente $(-1, 0), (0, -1), (0, 1)$.

Possiamo costruire la circonferenza di centro U e passante per O, individuando sull'asse delle ascisse un punto P_2 di coordinate $(2, 0)$, e possiamo costruire la circonferenza di centro P_2 e passante per U, individuando il punto P_3 di coordinate $(3, 0)$. Procedendo in questo modo sia sull'asse delle ascisse che su quello delle ordinate, è facile rendersi conto del fatto che, se $m, n \in \mathbb{Z}$, i punti di coordinate $(m, 0)$ e $(0, n)$ sono costruibili. Ora, tracciando le perpendicolari agli assi cartesiani per questi punti e considerando le loro intersezioni, possiamo individuare tutti i punti di coordinate intere (m, n) (vedi la Figura 72).

Come si può notare dalla figura, in realtà già con queste prime costruzioni abbiamo individuato molti altri punti del piano. Come possiamo caratterizzare i punti di Π?

Per rispondere, ci svincoliamo dell'approccio sintetico della geometria euclidea per fare spazio a un approccio algebrico, che si focalizzi sulle proprietà - algebriche - delle coordinate dei punti costruibili. A partire dal nostro sistema di riferimento, diamo la seguente:

Definizione. *Un numero $\alpha \in \mathbb{R}$ si dice **costruibile (con riga e compasso)** se è possibile costruire un segmento di misura $|\alpha|$ facendo uso soltanto della riga e del compasso. Indichiamo con $\mathcal{C}$ l'insieme dei numeri costruibili.*

Questa prima definizione sembra non tenere conto delle coordinate dei punti costruibili, ma solo delle misure dei segmenti che li congiungono. In realtà vale la seguente caratterizzazione dei punti del piano costruibili (provate a dimostrarla sfruttando le costruzioni elementari):

Un punto di coordinate (x, y) è costruibile se e solo se le sue coordinate x e y sono numeri costruibili.

Questo significa che, fissato un sistema di riferimento, determinare l'in-sieme Π dei punti del piano costruibili equivale a determinare l'insieme $\mathcal{C}$ dei numeri costruibili.

Cerchiamo allora di capire come è fatto $\mathcal{C}$. Abbiamo già visto che tutti i punti di coordinate (m, n) con $m, n \in \mathbb{Z}$ sono costruibili, quindi sicuramente $\mathbb{Z} \subseteq \mathcal{C}$. Inoltre possiamo costruire il segmento che congiunge O al punto di coordinate (m, n), e questo ha lunghezza $\sqrt{m^2 + n^2}$. Allora per definizione di numero costruibile $\sqrt{m^2 + n^2} \in \mathcal{C}$ per ogni $m, n \in \mathbb{Z}$. Possiamo anche individuare i punti medi dei segmenti, dimezzandone la lunghezza. Da ciò segue che se $\alpha \in \mathcal{C}$, anche $\frac{\alpha}{2^n} \in \mathcal{C}$ per ogni $n \in \mathbb{N}$.

Vediamo ora come possono essere tradotti i due problemi dell'antichità che abbiamo approfondito alla fine della Sezione 1 in termini di numeri costruibili:

- Nel caso della duplicazione del cubo, possiamo scegliere i punti O e U ad esempio su due vertici adiacenti di una faccia del cubo, e scegliere quindi la lunghezza di uno spigolo come unità di misura. In questo sistema di riferimento, duplicare un cubo equivale a costruire un segmento di misura $\sqrt[3]{2}$. Quindi il problema è equivalente alla domanda: " $\sqrt[3]{2} \in \mathcal{C}$? "
- Assegnato un cerchio, possiamo sempre individuare il suo centro con riga e compasso (basta tracciare due corde non parallele della sua circonferenza e intersecare i loro assi). Possiamo quindi scegliere un sistema di riferimento dove O è il centro e U un punto arbitrario della circonferenza, fissando quindi come unità di misura il raggio del cerchio. Allora quadrare un cerchio equivale a costruire un segmento di misura $\sqrt{\pi}$. Quindi il problema è equivalente alla domanda: " $\sqrt{\pi} \in \mathcal{C}$? "

Anziché approfondire qui come si possa formalizzare il problema della trisezione dell'angolo in termini di numeri costruibili, invitiamo il lettore a riflettere sulla questione indipendentemente. Avremo qualcosa da dire in proposito alla fine di questo articolo!

Il prossimo passo è quello di indagare le proprietà algebriche dell'insieme $\mathcal{C}$. Ad esempio dati due numeri costruibili a e b, cosa possiamo dire sulla somma $a + b$ e sul prodotto $a \cdot b$? Sono ancora numeri costruibili? E la differenza $a - b$ e il quoziente a/b sono ancora numeri costruibili?

Vediamo subito che la risposta è affermativa grazie alle costruzioni rappresentate in Figura 73:

- **Chiusura per somma e differenza**: se $a, b \in \mathcal{C}$, allora $a \pm b \in \mathcal{C}$.
  Possiamo costruire i punti A e P di coordinate rispettivamente $(a, 0)$ e (a, b). Possiamo costruire la circonferenza di centro A e passante per P che individua sull'asse delle ascisse un punto Q_1 di coordinate $(a-b, 0)$ e un punto Q_2 di coordinate $(a+b, 0)$, quindi $a \pm b \in \mathcal{C}$.
- **Chiusura per quoziente**: se $a, b \in \mathcal{C}$ e $a \neq 0$, allora $b/a \in \mathcal{C}$.
  Considerando ancora i punti A e P, possiamo tracciare la retta r per O e P, tracciare la retta s parallela all'asse delle ordinate passante per U e, intersecando r e s, individuare il punto R. I triangoli OUR e OAP sono simili e quindi i loro lati sono i proporzione: $UR : OU = AP : OA$. Ne consegue che $\overline{UR} = \frac{b}{a} \in \mathcal{C}$.
- **Chiusura per prodotto**: se $a, b \in \mathcal{C}$, allora $a \cdot b \in \mathcal{C}$.
  Poiché $1 \in \mathcal{C}$, per quanto appena visto si deve avere $\frac{1}{a} \in \mathcal{C}$, ma quindi anche $\frac{b}{1/a} = a \cdot b \in \mathcal{C}$.

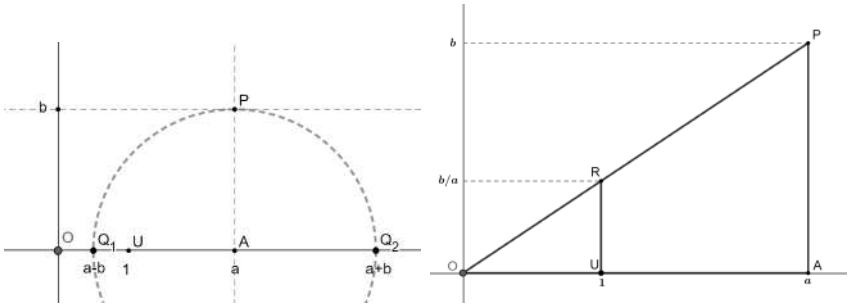


Figura 73: Somma e differenza di numeri costruibili (a sinistra); rapporto di numeri costruibili (a destra).

Come diretta conseguenza di questa proprietà abbiamo che ogni numero razionale $\frac{m}{n}$, con $m, n \in \mathbb{Z}$ e $n \neq 0$, è costruibile. Dunque otteniamo che $\mathbb{Q} \subsetneq \mathcal{C} \subseteq \mathbb{R}$.

L'insieme dei numeri costruibili $\mathcal{C}$, quindi, è *chiuso* rispetto alle quattro operazioni razionali, esattamente come l'insieme $\mathbb{Q}$ dei numeri razionali o come l'insieme $\mathbb{R}$ dei numeri reali: è sempre possibile sommare, sottrarre, moltiplicare e dividere due numeri (eccetto dividere per 0) restando all'*interno* di $\mathcal{C}$. Gli insiemi di numeri (ma non solo) che prevedono queste regolarità rispetto alle quattro operazioni sono molto importanti nello studio dell'algebra, e si chiamano **campi**. La nozione di campo sarà cruciale nello studio delle costruzioni con riga e compasso.

Un'ulteriore proprietà che caratterizza in modo decisivo l'insieme $\mathcal{C}$ è che questo è *chiuso per estrazione di radice quadrata* di numeri non negativi:

Sia $\alpha \in C$ tale che $\alpha \geq 0$. Allora $\sqrt{\alpha} \in \mathcal{C}$.

Dimostrazione. Se $\alpha \geq 0$ è un numero costruibile, allora possiamo costruire sull'asse delle ascisse un segmento di lunghezza α avente come estremi i punti O e il punto Q di coordinate $(\alpha, 0)$. Inoltre sappiamo che il punto P di coordinate $(-1, 0)$ è costruibile, per cui è costruibile anche il punto medio M del segmento PQ.

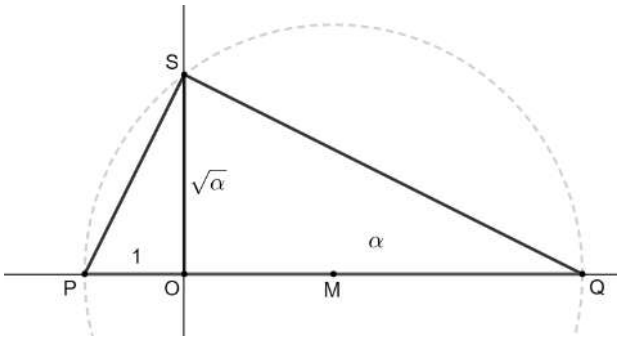


Figura 74: Costruzione della radice quadrata di un numero costruibile.

Costruiamo la circonferenza di centro M e passante per P e Q (Figura 74). Questa individua sul semiasse positivo delle ordinate un punto S. Il triangolo PQS è rettangolo in S perché un angolo inscritto in una semicirconferenza è retto. Per il secondo teorema di Euclide vale

$$\overline{OS}^2 = \overline{PO} \cdot \overline{OQ} = \alpha$$

da cui $\overline{OS} = \sqrt{\alpha}$, e quindi $\sqrt{\alpha} \in \mathcal{C}$. □

Abbiamo quindi che all'interno di $\mathcal{C}$ oltre che sommare, sottrarre, moltiplicare e dividere due numeri, è possibile eseguire l'operazione di estrazione di radice quadrata di ogni numero non negativo. Nella prossima sezione, vedremo che queste proprietà forniscono una descrizione completa dell'insieme $\mathcal{C}$, nel senso che ogni numero costruibile si può ottenere a partire dai numeri razionali combinando le quattro operazioni razionali e l'estrazione di radice quadrata.

25.3 Costruzioni con riga e compasso dal punto di vista algebrico

Per definizione, se α è un numero costruibile, allora è possibile costruire un segmento di lunghezza $|\alpha|$ per mezzo di un numero finito di costruzioni a partire da due punti. Il nostro intento è quello di analizzare i singoli passi della costruzione con riga e compasso di α, interpretandoli algebricamente uno alla volta.

Nella nostra trattazione, abbiamo scelto O e U come punti di partenza per poter effettuare costruzioni e, da questi, abbiamo osservato che è possibile costruire l'intero insieme dei punti del piano di coordinate razionali, che indichiamo con $\Pi_0 \subsetneq \Pi$. Per semplificare le cose ed evitare di partire da due soli punti, possiamo supporre di avere già a disposizione Π_0 come insieme di Â«punti baseÂ» per costruire α.

Partendo da Π_0, le costruzioni possibili sono rette passanti per punti di Π_0 e circonferenze di centro un punto di Π_0 e passanti per un punto di Π_0. Nello specifico, siano $P_1,\ P_2 \in \Pi_0$ punti di coordinate $(x_1, y_1),\ (x_2, y_2)$ rispettivamente. Indichiamo $r = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ la distanza tra P_1 e P_2. Ricordiamo che:

- La retta passante per P_1 e P_2 ha equazione

  $$(y_2 - y_1)\, x - (x_2 - x_1)\, y + x_2 y_1 - x_1 y_2 = 0$$

  ossia è della forma $ax + by + c = 0$ con $a = y_2 - y_1,\ b = x_1 - x_2,\ c = x_2 y_1 - x_1 y_2 \in \mathbb{Q}$.
- La circonferenza di centro P_1 e passante per P_2 ha equazione

  $$x^2 + y^2 - 2x_1\, x - 2y_1\, y + {x_1}^2 + {y_1}^2 - r^2 = 0$$

  ossia è della forma $x^2 + y^2 + ax + by + c = 0$ con $a = -2x_1,\ b = -2y_1,\ c = {x_1}^2 + {y_1}^2 - r^2 \in \mathbb{Q}$

In particolare, le equazioni delle rette e circonferenze costruibili a partire da Π_0 hanno coefficienti in $\mathbb{Q}$. Osserviamo che questo è conseguenza del solo fatto che $\mathbb{Q}$ è chiuso rispetto alle operazioni razionali. Le coordinate dei punti di intersezione di due di questi oggetti sono determinate risolvendo il sistema algebrico formato dalle loro equazioni.

- Il caso di **due rette**: se le equazioni di due rette (non parallele) hanno coefficienti in $\mathbb{Q}$, allora sono in $\mathbb{Q}$ anche le coordinate del loro punto di intersezione (provate a verificarlo). Geometricamente, questo significa che l'uso della sola riga non ci permette di *uscire fuori* da Π_0, e ciò è conseguenza del solo fatto che $\mathbb{Q}$ è un campo.

- Il caso di **due circonferenze** si riduce a quello di una retta e una circonferenza: basta considerare il sistema (equivalente) ottenuto sottraendo la seconda equazione alla prima.
- Il caso di **una retta e una circonferenza**: consideriamo un sistema di secondo grado della forma

$$\begin{cases} x^2 + y^2 + a_1 x + b_1 y + c_1 = 0 \\ a_2 x + b_2 y + c_2 = 0 \end{cases}, \qquad \text{con} \quad a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{Q}$$

Osserviamo che almeno uno tra a_2 e b_2 è non nullo. Ad esempio se $b_2 \neq 0$ (ma un discorso analogo vale se $a_2 \neq 0$), sostituendo nella prima equzione $y = -\frac{1}{b_2}(a_2 x + c_2)$, si ottiene

$$\underbrace{(a_2^2 + b_2^2)}_{=A\in\mathbb{Q}} x^2 + \underbrace{(2a_2c_2 + a_1{b_1}^2 - a_2b_1b_2)}_{=B\in\mathbb{Q}} x + \underbrace{{a_2}^2 - b_1b_2c_2 + c_1{b_2}^2}_{=C\in\mathbb{Q}} = 0$$

Sia $\Delta = B^2 - 4AC$. Nel caso in cui valga la condizione $\Delta \geq 0$ (che geometricamente si traduce con l'esistenza di punti di intersezione tra le curve, condizione che nel seguito supporremo sempre vera in virtù degli scopi del nostro studio) la soluzione dell'equazione, che individua le ascisse dei punti di intersezione, è data dalla formula:

$$x = -\frac{B}{2A} \pm \frac{\sqrt{\Delta}}{2A}, \quad \text{da cui} \quad y = \frac{B - 2Ac_2}{2Ab_2} \mp \frac{a_2\sqrt{\Delta}}{2Ab_2}$$

Entrambe le coordinate sono quindi della forma $a + b\sqrt{\Delta}$, con $a, b \in \mathbb{Q}$.

Nel caso in cui Δ non sia il quadrato di un numero razionale (ossia $\sqrt{\Delta} \notin \mathbb{Q}$), l'uso del compasso ci premette di *uscire fuori* da Π_0, introducendo punti di coordinate irrazionali.

Quando introduciamo un punto nuovo $\widetilde{P} \notin \Pi_0$, possiamo passare alla fase successiva: costruire rette e circonferenze sfruttando anche questo punto. Come possiamo esprimere questo in termini algebrici?

Vogliamo scrivere le equazioni di queste rette e circonferenze per caratterizzare le coordinate dei loro punti di intersezione, in modo del tutto analogo a quanto fatto partendo da Π_0. Come già sottolineato, il punto di forza di questa prima analisi è stata la struttura di campo di $\mathbb{Q}$. Per replicare lo studio precedente, possiamo allora pensare di *estendere* il campo $\mathbb{Q}$ a un nuovo campo $\mathbb{K}_1$ di numeri costruibili, che contiene $\mathbb{Q}$ e le coordinate di $\widetilde{P}$. Allora come prima, potremmo considerare l'insieme dei punti di coordinate in $\mathbb{K}_1$, che indichiamo con Π_1, e ripartire considerando rette e circonferenze di equazioni a coefficienti in $\mathbb{K}_1$. Chi può essere $\mathbb{K}_1$?

Abbiamo visto che $\widetilde{P}$ ha coordinate irrazionali $(x_{\widetilde{P}}, y_{\widetilde{P}})$ che sono della forma $a + b\sqrt{\Delta}$, con $a, b, \Delta \in \mathbb{Q}$ e $\sqrt{\Delta} \notin \mathbb{Q}$. Poiché $\mathcal{C}$ è un campo, allora possiamo combinare a piacimento le quattro operazioni tra numeri costruibili restando all'interno di $\mathcal{C}$. Quindi in particolare $\frac{a+b\sqrt{\Delta}}{b} - \frac{a}{b} = \sqrt{\Delta} \in \mathcal{C}$, ma anche $p + q\sqrt{\Delta} \in \mathcal{C}$ per ogni $p, q \in \mathbb{Q}$. Ma allora l'insieme $K = \left\{ p + q\sqrt{\Delta} \mid p, q \in \mathbb{Q} \right\}$ è tutto contenuto in $\mathcal{C}$ e per di più è un campo (provate a verificare che K è chiuso rispetto alle quattro operazioni razionali).

Poniamo allora $\mathbb{K}_1 = K$ e ripartiamo dai punti di Π_1 (che sappiamo essere tutti costruibili). Iteriamo il processo finché non raggiungiamo un segmento di misura $|\alpha|$ (cioè, a meno di una costruzione elementare, il punto di coordinate $(\alpha, 0)$), quindi finché non raggiungiamo un campo

$\mathbb{K}_n$ che contenga il numero α.
In termini algebrici, la sequenza delle costruzioni può essere tradotta in una sequenza di *estensioni di campi*:

$$\mathbb{Q} \subseteq \mathbb{K}_1 \subseteq \mathbb{K}_2 \subseteq \mathbb{K}_3 \subseteq ... \subseteq \mathbb{K}_n$$

dove ogni contenimento è il dato algebrico di ciascun passo della costruzione geometrica: si ha $\mathbb{K}_{i-1} = \mathbb{K}_i$ se nel passaggio i-esimo si è individuato un punto di coordinate in $\mathbb{K}_{i-1}$, cioè si è rimasti all'interno di Π_{i-1}; si ha invece $\mathbb{K}_{i-1} \subsetneq \mathbb{K}_i$ se nel passaggio i-esimo si è usciti al di fuori di Π_{i-1}. In questo secondo caso, abbiamo $\mathbb{K}_i = \left\{p + q\sqrt{\Delta_i} \mid p, q \in \mathbb{K}_{i-1}\right\}$ per qualche $\Delta_i \in \mathbb{K}_{i-1}$ positivo tale che $\sqrt{\Delta_i} \notin \mathbb{K}_{i-1}$.

Facciamo un esempio concreto: supponiamo di voler costruire il numero $1+\sqrt{2}+\sqrt[4]{3}$. Partendo da Π_0, una strada può essere la seguente:

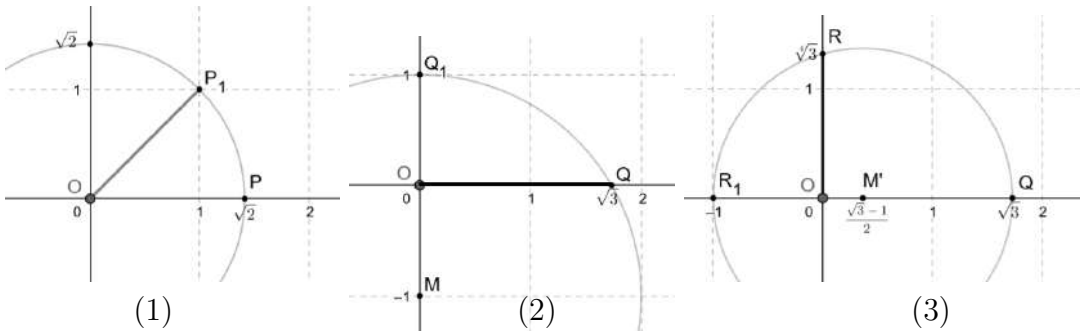


Figura 75: Tre costruzioni geometriche per ottenere il numero $1 + \sqrt{2} + \sqrt[4]{3} \in \mathcal{C}$.

	Geometria	*Algebra*
(1)	Costruiamo un segmento di misura $\sqrt{2}$ congiungendo l'origine con il punto P_1 di coordinate $(1, 1)$. Puntando il compasso in O con apertura in P_1, tracciamo una circonferenza che individua sull'asse delle ascisse il punto P di coordinate $(\sqrt{2}, 0)$.	Estendiamo $\mathbb{Q}$ a $\mathbb{K}_1 = \left\{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\right\}$
(2)	Costruiamo un segmento di misura $\sqrt{3}$ tracciando la circonferenza di centro nel punto M di coordinate $(0, -1)$ e passante per il punto Q_1 di coordinate $(0, 1)$. Individuiamo quindi il punto Q di intersezione con il semiasse positivo delle ascisse, che ha coordinate $(\sqrt{3}, 0)$.	Estendiamo $\mathbb{K}_1$ a $\mathbb{K}_2 = \left\{p + q\sqrt{3} \mid p, q \in \mathbb{K}_1\right\}$
(3)	Costruiamo un segmento di misura $\sqrt[4]{3}$ con la classica costruzione della radice quadrata: a partire da Q, dal punto R_1 di coordinate $(-1, 0)$ e dalla circonferenza di centro nel loro punto medio M' e passante per Q, si individua il punto R di coordinate $(0, \sqrt[4]{3})$.	Estendiamo $\mathbb{K}_2$ a $\mathbb{K}_3 = \left\{x + y\sqrt[4]{3} \mid x, y \in \mathbb{K}_2\right\}$ e ci fermiamo perché $1 + \sqrt{2} + \sqrt[4]{3} \in \mathbb{K}_3$

A questo punto è facile costruire un segmento di misura $1 + \sqrt{2} + \sqrt[4]{3}$: applicando il segmento OR all'asse delle ascisse a partire da P, troviamo un punto S di coordinate $(\sqrt{2} + \sqrt[4]{3}, 0)$. Questo individua con R_1 un segmento di misura $1 + \sqrt{2} + \sqrt[4]{3}$.

25.4 Problemi impossibili

Abbiamo visto come ciascuno dei problemi si riduca a un test di appartenenza di un certo numero irrazionale α all'insieme $\mathcal{C}$. Abbiamo anche visto che se $\alpha \in \mathcal{C}$, allora esiste una successione di estensioni di campi:

$$\mathbb{Q} = \mathbb{K}_0 \subseteq \mathbb{K}_1 \subseteq \mathbb{K}_2 \subseteq \mathbb{K}_3 \subseteq ... \subseteq \mathbb{K}_n$$

con $\alpha \in \mathbb{K}_n$. Possiamo assumere che tutti i contenimenti siano stretti, a meno di eliminare dalla successione quei contenimenti che sono uguaglianze. Possiamo anche assumere di fermarci non appena si raggiunge un campo contenente α, cioè possiamo assumere che $\alpha \in \mathbb{K}_n$ ma $\alpha \notin \mathbb{K}_{n-1}$. In questo caso $\mathbb{K}_n = \left\{ p + q\sqrt{\delta} \mid p, q \in \mathbb{K}_{n-1} \right\}$ per qualche numero positivo $\delta \in \mathbb{K}_{n-1}$.

25.4.1 È impossibile duplicare un cubo con riga e compasso ($\sqrt[3]{2} \notin \mathcal{C}$)

Supponiamo per assurdo $\sqrt[3]{2} \in \mathcal{C}$. Allora, con la notazione precedente si ha $\sqrt[3]{2} = a + b\sqrt{\delta} \in \mathbb{K}_n$ per qualche $a, b \in \mathbb{K}_{n-1}$ con $b \neq 0$ (perché $\sqrt[3]{2} \notin \mathbb{K}_{n-1}$). Elevando al cubo entrambi i membri dell'identità precedente, si ottiene:

$$2 = a^3 + 3ab^2\delta + b(3a^2 + b^2\delta)\sqrt{\delta}$$

Osserviamo che poiché $\delta > 0$, si ha $b(3a^2 + b^2\delta) \neq 0$. Dividendo entrambi i membri per questa quantità e isolando $\sqrt{\delta}$ troviamo

$$\sqrt{\delta} = \frac{\overbrace{2 - a^3 - 3ab^2\delta}^{\in \mathbb{K}_{n-1}}}{\underbrace{b(3a^2 + b^2\delta)}_{\in \mathbb{K}_{n-1}}} \in \mathbb{K}_{n-1}$$

e questo è assurdo, perché ad esempio implica che $\sqrt[3]{2} = a + b\sqrt{\delta} \in \mathbb{K}_{n-1}$, contro le ipotesi iniziali.

25.4.2 È impossibile quadrare un cerchio con riga e compasso ($\sqrt{\pi} \notin \mathcal{C}$)

Sia $\alpha \in \mathcal{C}$. Con la notazione iniziale abbiamo $\alpha = p + q\sqrt{\delta} \in \mathbb{K}_n$ per qualche $p, q \in \mathbb{K}_{n-1}$. Questo significa che $(\alpha - p)^2 = q^2\delta$ e quindi α soddisfa l'equzione di secondo grado $x^2 - 2px + p^2 - q^2\delta = 0$ a coefficienti in $\mathbb{K}_{n-1}$.
Inoltre, possiamo scrivere questi coefficienti nella forma $p_1 + p_2\sqrt{\delta'}$, con $p_1, q_1, \delta' \in \mathbb{K}_{n-2}$, da cui, isolando $\sqrt{\delta'}$ ed elevando al quadrato entrambi i membri, troviamo che α soddisfa un'equazione di grado 4 a coefficienti in $\mathbb{K}_{n-2}$. Procedendo in questo modo, è facile rendersi conto del fatto che ogni numero costruibile è radice di un polinomio a coefficienti in $\mathbb{Q}$ di grado una potenza di 2.

In matematica, i numeri che soddisfano un'equazione polinomiale a coefficienti razionali si dicono *algebrici*, tutti gli altri si dicono *trascendenti*. Nel 1882 Ferdinand von Lindemann ha dimostrato che π è un numero trascendente, e questo ha chiuso definitivamente il problema della quadratura del cerchio: $\pi \notin \mathcal{C}$ e in particolare $\sqrt{\pi} \notin \mathcal{C}$.

25.4.3 È impossibile trisecare l'angolo di $60°$ con riga e compasso ($\cos(\frac{\pi}{9}) \notin \mathcal{C}$)

Nel corso dell'intero articolo abbiamo sempre accantonato il terzo dei problemi menzionati nell'introduzione: quello della trisezione dell'angolo. Lasciamo sia il processo di formalizzazione e "algebrizzazione" del problema, sia la sua soluzione (in termini della non costruttibilità di un certo numero) come sfida per il lettore.

Qui di seguito proponiamo alcuni suggerimenti che potrebbero risultare utili!

a) Costruire un angolo di $60°$ avente vertice nell'origine e come lato il semiasse positivo delle ascisse, e provare che è trisecabile se e solo se $\cos(\pi/9) \in \mathcal{C}$.

b) Provare l'identità goniometrica $\cos(3t) = 4\cos^3(t) - 3\cos(t)$.

c) Sostituendo $3t = \pi/3 + 2k\pi$, con $k \in \mathbb{Z}$, dedurre che il polinomio $p(x) = 4x^3 - 3x - 1/2$ ammette le tre radici distinte

$$\cos\left(\frac{\pi}{9}\right), -\cos\left(\frac{2\pi}{9}\right), -\cos\left(\frac{4\pi}{9}\right).$$

d) Provare che le radici di $p(x)$ sono irrazionali (cosa succede imponendo $4(r/s)^3 - 3(r/s) - 1/2 = 0$, se $r, s \in \mathbb{Z}$ non hanno divisori primi in comune?)

A questo punto supponiamo per assurdo che $\cos(\frac{\pi}{9}) \in \mathcal{C}$. Allora possiamo scrivere $\cos(\frac{\pi}{9}) = a + b\sqrt{\delta} \in \mathbb{K}_n$ per qualche $a, b \in \mathbb{K}_{n-1}$ e $\cos(\frac{\pi}{9}) \notin \mathbb{K}_{n-1}$.

e) Verificare che se $a + b\sqrt{\delta}$ è radice di $p(x)$, allora lo sono anche $a - b\sqrt{\delta}$ e $-2a$.

f) Sfruttando i punti precedenti e le formule di duplicazione del coseno, dimostrare che se uno tra $\cos(\frac{2\pi}{9}), \cos(\frac{4\pi}{9})$ è in $\mathbb{K}_{n-1}$, allora anche $\cos(\frac{\pi}{9}) \in \mathbb{K}_{n-1}$, trovando la contraddizione cercata.

Riferimenti bibliografici

[1] R. Chirivì, I. Del Corso, R. Dvornicich, *Esercizi scelti di Algebra, Vol. 2*, pp. 47–49.

[2] R. Courant, H. Robbins, *Che cos'è la matematica?*, pp. 179–194.

[3] I tre postulati di Euclide dal web, vedi ad esempio `https://en.wikipedia.org/wiki/Axiom`

26 Stuzzicadenti buffi e π

Luigi Amedeo Bianchi, n.14, Aprile 2022

26.1 Introduzione

Per l'esperimento che vogliamo fare ci occorreranno: un foglio A4 con quadrettatura 1 cm, uno stuzzicadenti, un righello (per misurare lo stuzzicadenti) e qualcosa per tenere traccia dei dati dell'esperimento.

"Ma come, esperimento? Non parliamo di matematica? E da quando la matematica è una scienza sperimentale?" Fidatevi! Parleremo di matematica, ma prendendo una strada diversa.

26.2 L'esperimento

Come prima cosa dobbiamo misurare la lunghezza in centimetri dello stuzzicadenti. Il mio è lungo $L = 6.8$ cm. Fatto questo, mettiamo il foglio quadrettato su una superficie orizzontale liscia.

Lanciamo ora lo stuzzicadenti in modo che cada sul foglio. Contiamo quante righe interseca, sia orizzontali, sia verticali. Un modo comodo e veloce di farlo è contare quante righe orizzontali ci sono tra un estremo e l'altro dello stuzzicadenti e sommare questo numero a quello delle righe verticali comprese tra un estremo e l'altro. In questo modo non rischiamo di perderci qualche intersezione o di farci imbrogliare (non contandolo con la dovuta molteplicità) dal fatto che lo stuzzicadenti passi "esattamente" in un punto nel quale si intersecano due righe.

Segnamoci questo valore, nel mio caso 9. Ripetiamo questa procedura per un totale di 10 misurazioni, nel mio caso: $9, 9, 8, 9, 10, 7, 9, 10, 10, 9$.

Ora vogliamo calcolare il numero medio di intersezioni osservate $\overline{I}$, ossia sommare tutti i valori ottenuti e dividere il risultato per il numero delle osservazioni (nel nostro caso 10):

$$\overline{I} = \frac{9+9+8+9+10+7+9+10+10+9}{10} = \frac{90}{10} = 9.$$

Ora dividiamo la lunghezza L per il numero medio di intersezioni osservate $\overline{I}$ e moltiplichiamo questo risultato per 4: $\frac{L}{\overline{I}} \cdot 4 = \frac{6.8}{9} \cdot 4 \approx 3.022$. Il vostro risultato sarà quasi certamente diverso, ma, mi sento di scommettere, non troppo!

Ripetiamo l'esperimento da capo e raccogliamo altre 10 misurazioni: $7, 9, 8, 9, 9, 9, 8, 7, 9, 8$. In questo caso $\overline{I} = 8.3$, mentre $L = 6.8$, come prima. Quindi questa volta, per me, $\frac{L}{\overline{I}} \cdot 4 = \frac{6.8}{8.3} \cdot 4 \approx 3.277$.

Possiamo anche considerare le 20 misurazioni fatte come parte di un solo esperimento e calcolare il numero medio di intersezioni sommando sia i numeri nella prima lista, sia quelli nella seconda e dividendo il risultato (nel mio caso 173) per il numero di osservazioni, ossia 20. Questo ci dà, per $\frac{L}{\overline{I}} \cdot 4$, un valore all'incirca 3.145, intermedio tra i due ottenuti prima (ma non la loro media aritmetica, come mai?).

Potremmo continuare a fare lanci di stuzzicadenti, conteggi e usare la "formula" $\frac{L}{\overline{I}} \cdot 4$ e il valore ottenuto si avvicinerà a $3.14\ldots$ Sì, proprio a π. Ma perché? Da dove viene la "formula"?

26.3 Un modello matematico

Cerchiamo di descrivere in termini matematici quello che stiamo facendo. Possiamo descrivere lo stuzzicadenti come un segmento di lunghezza L. Come facciamo però a rappresentare il numero medio di intersezioni che ha con le righe?

Il *valore atteso* (o media) di una quantità casuale (come è il numero I di intersezioni di un segmento con una quadrettatura) è dato dalla media dei possibili valori pesata con le corrispondenti probabilità: se chiamiamo p_i la probabilità che il segmento intersechi i righe (per $i \in \mathbb{N}$), allora il numero medio di intersezioni è

$$E(I) = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + \dots \tag{17}$$

Questa notazione può sembrare imprecisa: cosa significano quei puntini? E cosa vuol dire sommare infiniti oggetti? In realtà possiamo osservare che c'è un numero massimo (legato alla lunghezza L) di intersezioni possibili, quindi esiste un N tale che per ogni $i \geqslant N$, $p_i = 0$. Come esercizio, potete provare a calcolarlo!

Osserviamo anche che i valori calcolati prima nel corso dell'esperimento, $\overline{I}$, sono approssimazioni sperimentali di questo valore $E(I)$ che viene dal modello. Si può dimostrare che al crescere del numero di ripetizioni dell'e-sperimento, $\overline{I}$ (detto anche *media campionaria*) converge a $E(I)$.

Torniamo al nostro modello matematico e, per continuare, proviamo a semplificarci un po' la vita considerando una versione speciale del problema: ci sono solamente righe orizzontali (sempre a distanza di 1 cm le une dalle altre) e il nostro segmento/stuzzicadenti ha una lunghezza $L < 1$. La ragione della prima semplificazione è abbastanza evidente: consideriamo separatamente le righe orizzontali e quelle verticali, anzi, magari riusciremo a usare qualche argomento di simmetria per passare gratuitamente o quasi dalle sole righe orizzontali alla quadrettatura. Per quanto riguarda la scelta della lunghezza ridotta, il motivo è legato alla fastidiosa somma con i puntini nella formula (17): se la lunghezza del segmento è minore di 1 (e abbiamo solamente righe orizzontali), possiamo avere o 0 intersezioni o 1 intersezione, quindi

$$E(I) = 0 \cdot p_0 + 1 \cdot p_1 = p_1 \tag{18}$$

e in particolare il numero atteso di intersezioni è uguale alla probabilità di averne una.

Non dobbiamo però farci prendere troppo dall'entusiasmo: le probabilità p_i che compaiono in (18) non sono le stesse scritte in (17), si riferiscono a un segmento diverso e quindi a un problema diverso! Ma la speranza è quella di riuscire, grazie a questa variante del problema, a capire qualcosa di più del problema originale.

Ripensando all'esperimento semplificato, possiamo osservare che la probabilità p_1 che un segmento di lunghezza minore di 1 intersechi o meno una riga orizzontale dipende dall'angolo φ che il segmento ha rispetto alle righe orizzontali. Sapendo che il segmento forma un angolo φ con le rette orizzontali, la probabilità infatti è uguale[38] alla proiezione del segmento sulla perpendicolare alle rette orizzontali, $h = L \cdot \sin \varphi$. Essa è tanto più alta quanto più φ è prossimo a un angolo retto. Lo stuzzicadenti copre infatti sull'asse verticale una porzione pari alla sua proiezione sull'asse verticale stesso e siccome la distanza tra due rette orizzontali è 1, la probabilità che una riga cada all'interno della porzione verticale coperta dal segmento è pari ad h, come illustrato in Figura 76. In particolare, se φ è retto, la probabilità è uguale alla lunghezza L del segmento stesso.

Dobbiamo però calcolare questo valore al variare di tutte le possibili ampiezze φ dell'angolo. Per ogni φ, come detto, la probabilità di intersecare una retta è il rapporto tra la lunghezza

[38] In questo caso è uguale, perché la distanza tra le rette è 1, in generale sarebbe proporzionale, con la rinormalizzazione data dalla distanza tra le righe.

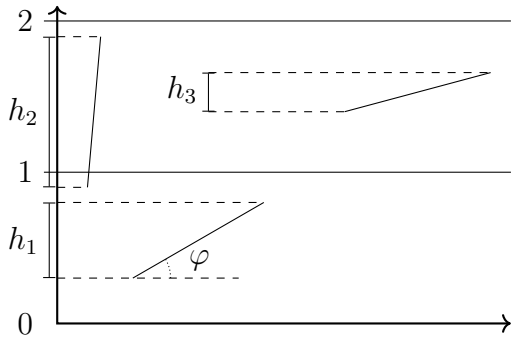


Figura 76: La proiezione verticale (e quindi la probabilità di intersezione) al variare dell'angolo φ.

$h = L \sin \varphi$ e la distanza tra le linee orizzontali, ossia 1. Tuttavia abbiamo un'infinità continua di possibili valori di φ, quindi possiamo pensare di trasformare il rapporto tra le lunghezze in un rapporto tra superfici.

Calcoliamo allora il valore tra l'area (in grigio) sotto la curva e il rettangolo rappresentati in Figura 77. Infatti, per simmetria, φ varia tra l'angolo nullo e l'angolo retto (ossia tra 0 e $\frac{\pi}{2}$, se misuriamo l'angolo in radianti), e per ogni valore di φ, come detto, la probabilità di intersecare una retta orizzontale è uguale al rapporto tra il valore della funzione $L \sin x$ calcolata in φ e la lunghezza del segmento unitario. Mettendo questi segmenti in verticale uno accanto all'altro in corrispondenza del valore di φ cui fanno riferimento abbiamo la Figura 77. La probabilità di successo è pari al rapporto tra l'area sotto la curva e quella del rettangolo di base $\frac{\pi}{2}$ e altezza 1. Per chi ha familiarità con gli integrali (chi non la ha può saltare direttamente al paragrafo successivo), si tratta di calcolare

$$\frac{\int_0^{\pi/2} L \sin x \mathrm{d}x}{\pi/2} = \frac{2L}{\pi}[-\cos x]_0^{\pi/2} = \frac{2L}{\pi}.$$

In particolare, possiamo osservare che questa probabilità (che come detto è anche il numero medio di intersezioni) è proporzionale alla lunghezza L con coefficiente $\frac{2}{\pi}$.

Abbiamo fatto comparire π nel nostro modello. Sarebbe però bello poter ricavare il medesimo risultato senza dover ricorrere al calcolo integrale, per poi passare a considerare il caso che ci interessa, ossia $L > 1$. Come vedremo ora, possiamo fare entrambe le cose in un colpo solo, grazie a una proprietà del valore atteso, la linearità.

26.4 Linearità del valore atteso e stuzzicadenti spezzati

Se abbiamo due quantità casuali A e B il valore atteso della loro somma è la somma dei valori attesi: $E(A+B) = E(A) + E(B)$. Cerchiamo di convincercene nel caso che stiamo considerando. Se abbiamo due segmenti di lunghezze $l_1, l_2 \leqslant 1$ e chiamiamo I_1 e I_2 il numero di intersezioni del primo e del secondo segmento, rispettivamente, abbiamo 4 possibilità:

A_1 Nessun segmento ha intersezioni, $A_1 = \{I_1 = 0\} \cap \{I_2 = 0\}$;

A_2 Il primo segmento ha un'intersezione e il secondo non ne ha, $A_2 = \{I_1 = 1\} \cap \{I_2 = 0\}$;

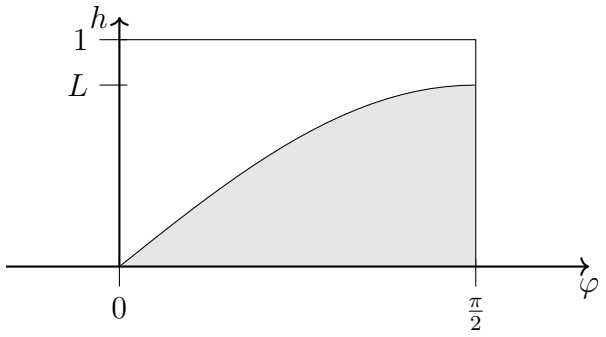


Figura 77: Quanto vale il rapporto tra l'area in grigio e quella del rettangolo?

A_3 Il secondo ha un'intersezione e il primo no, $A_3 = \{I_1 = 0\} \cap \{I_2 = 1\}$;

A_4 Entrambi i segmenti hanno un'intersezione, $A_4 = \{I_1 = 1\} \cap \{I_2 = 1\}$.

Questi eventi avranno ciascuno una propria probabilità $P(A_i)$. Osserviamo inoltre che $P(I_1 = 1) = P(A_2) + P(A_4)$. Infatti

$$\{I_1 = 1\} = \{I_1 = 1 \cap \{\{I_2 = 0\} \cup \{I_2 = 1\}\}\} =$$

$$= \{\{I_1 = 1\} \cap \{I_2 = 0\}\} \cup \{\{I_1 = 1\} \cap \{I_2 = 1\}\} = A_2 \cup A_4$$

perché $\{I_2 = 0\}$ e $\{I_2 = 1\}$ sono casi disgiunti e coprono tutte le possibilità, assieme al fatto che vale la proprietà distributiva dell'intersezione rispetto all'unione. Inoltre l'unione $A_2 \cup A_4$ è disgiunta, dal momento che i due insiemi non si intersecano, quindi la probabilità dell'unione è la somma delle probabilità. In modo simile abbiamo che $P(I_2 = 1) = P(A_1) + P(A_4)$.

Cerchiamo ora il valore atteso per $I_1 + I_2$:

$$\begin{aligned} E(I_1 + I_2) &= (0+0) \cdot P(A_1) + (0+1) \cdot P(A_2) + (1+0) \cdot P(A_3) \\ &\quad +(1+1) \cdot P(A_4) \\ &= P(A_2) + P(A_3) + P(A_4) + P(A_4) \\ &= (P(A_2) + P(A_4)) + (P(A_1) + P(A_4)) \\ &= 1 \cdot P(I_1 = 1) + 1 \cdot P(I_2 = 1) \\ &= E(I_1) + E(I_2). \end{aligned}$$

Abbiamo quindi mostrato la linearità in questo caso particolare con le intersezioni di due segmenti con le rette orizzontali. In modo simile possiamo estenderla alla somma del valore atteso del numero di intersezioni di più segmenti.

Tornando agli stuzzicadenti e alle righe orizzontali, la linearità ci aiuta chiaramente a passare dal caso $L \leqslant 1$ al caso $L > 1$: se il nostro segmento è lungo $L > 1$, possiamo spezzarlo in $m = \lceil L \rceil$ segmenti, ciascuno di lunghezza $l = \frac{L}{m} \leqslant 1$. Per linearità del valore atteso, il numero medio di intersezioni con righe orizzontali sarà allora, dal momento che tutti gli m segmenti hanno la medesima lunghezza l,

$$E(I_L) = mE(I_l) = m \cdot \frac{2 \cdot l}{\pi} = m \cdot \frac{2 \cdot \frac{L}{m}}{\pi} = \frac{2 \cdot L}{\pi},$$

dove con I_L indichiamo il numero di intersezioni di un segmento di lunghezza L e con I_l il numero di intersezioni di un segmento di lunghezza l.

Tuttavia, senza l'integrale calcolato prima non sapremmo che per un segmento di lunghezza $l < 1$ il numero medio di intersezioni è $\frac{2 \cdot l}{\pi}$. Possiamo ricavare anche questo risultato dalla linearità? Come?

Torniamo al valore atteso del numero di intersezioni. Per un segmento I_l di lunghezza l, esso dipenderà, secondo un'opportuna funzione f, dalla lunghezza l: $E(I_l) = f(l)$. Osserviamo anche che questa funzione è necessariamente lineare, dal momento che possiamo pensare di spezzare il segmento in due segmenti più corti, di lunghezza l_1 ed l_2 rispettivamente, che avranno I_1 e I_2 intersezioni rispettivamente:

$$f(l_1) + f(l_2) = E(I_1) + E(I_2) = E(I_l) = f(l) = f(l_1 + l_2).$$

Possiamo spingerci oltre: se spezziamo un segmento di lunghezza nl in $n \in \mathbb{N}$ segmenti di lunghezza l abbiamo $nf(l) = f(nl)$ e se lo spezziamo in $m \in \mathbb{N}$ segmenti di lunghezza $\frac{nl}{m}$ abbiamo

$$mf\left(\frac{n}{m}l\right) = f(nl) = nf(l),$$

da cui $f(ql) = qf(l)$ per ogni razionale q. Con qualche attenzione (appoggiandoci alla monotonia di questa funzione f) riusciamo a estendere al caso reale, concludendo che $f(x) = xf(1)$, con $f(1)$ una costante C da determinare.

Se però pensiamo alla proprietà di linearità enunciata sopra, non abbiamo chiesto che i segmenti fossero tra loro allineati: la linearità vale indipendentemente da questo fatto. Potrebbero benissimo formare una linea spezzata, aperta o chiusa, e il valore atteso del numero totale di intersezioni con le righe sarà la somma dei valori attesi per i singoli segmenti. Se la lunghezza della spezzata è L, il valore atteso del numero di intersezioni con le rette orizzontali sarà $C \cdot L$.

Possiamo ora osservare che, con le poligonali spezzate, possiamo dare approssimazioni successive (e sempre più precise) anche di curve che non sono rettilinee a tratti. In particolare possiamo approssimare delle circonferenze e, in particolare, la circonferenza di diametro 1. Essa ha lunghezza π e quindi ha un valore atteso di intersezioni con le rette orizzontali uguale a $C \cdot \pi$. Ma se ha diametro 1, avrà sempre due intersezioni (o due volte la stessa retta, oppure sarà tangente a due rette), quindi $2 = C \cdot \pi$, da cui $C = \frac{2}{\pi}$. Quindi il valore atteso delle intersezioni con le rette orizzontali a distanza 1 le une dalle altre di un segmento di lunghezza L è $\frac{2L}{\pi}$.

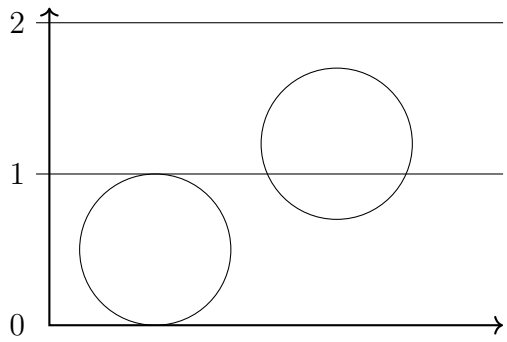


Figura 78: Stuzzicadenti circolari di diametro 1: hanno sempre esattamente due intersezioni.

In modo del tutto analogo possiamo procedere con le rette verticali, con le quali ci saranno in media $\frac{2L}{\pi}$ intersezioni. Complessivamente quindi in media ci sono $\frac{4L}{\pi}$ intersezioni I_q con la quadrettatura, $E(I_q) = \frac{4L}{\pi}$, da cui segue che $\frac{4L}{E(I_q)} = \pi$. Abbiamo quindi giustificato (quasi) rigorosamente il risultato dell'esperimento iniziale.

26.5 Conclusioni

Il problema del lancio dello stuzzicadenti (originariamente un ago) prende il nome dal Conte di Buffon, che lo pose nel 1777 nel caso delle sole rette orizzontali. La relazione venne usata nella seconda metà dell'Ottocento per stimare il valore di π. Naturalmente, volessimo fare oggi la stessa cosa sarebbe molto più pratico lanciare stuzzicadenti virtuali su pavimentazioni virtuali. Inoltre usare una quadrettatura invece delle sole rette parallele porta a una convergenza molto più rapida.

Sempre di fine Ottocento, per la precisione del 1860, è l'idea di usare uno stuzzicadenti circolare per evitare il calcolo dell'integrale, dovuta a E. Barbier [1]. Una trattazione molto più esaustiva del problema di Buffon, con alcune generalizzazioni, si può trovare in [2], che ha fortemente ispirato questa presentazione. In questo volume il problema dell'ago è il punto di partenza per lo studio della probabilità geometrica, chiamata dagli autori anche combinatoria continua, per via delle numerose analogie che ha con i conteggi della combinatoria.

Riferimenti bibliografici

[1] E. Barbier. Note sur le problème de l'aiguille et le jeu du joint couvert. *J. Mathématiques Pures et Appliquées*, 2(5):273–283, 1860.

[2] D. A. Klain and G.-C. Rota. *Introduction to Geometric Probability.* Lezioni Lincee. Cambridge University Press, Cambridge, UK; New York, 1997.

27 Quante configurazioni ha un cubo di Rubik?

Alessandro Iraci, n.15, Settembre 2022

Chi di noi non ha mai avuto a che fare con un cubo di Rubik? In quasi ogni casa ce n'è uno, che probabilmente giace mescolato su una mensola o in un cassetto. Ma quante configurazioni diverse ci sono? La risposta è sorprendente: ben 43.252.003.274.489.856.000, un numero comparabile con il numero di granelli di sabbia nel Sahara. Come si arriva a questo numero? È quello che scopriremo in questo articolo.

Se hai un cubo di Rubik in casa, è il momento di andarlo a cercare! Ti sarà utile per capire meglio questo articolo. Se non ne hai, scansiona il QR qui sotto per utilizzarne una versione online!

SCAN ME

Puoi trovare una versione interattiva del cubo online all'indirizzo `https://alg.cubing.net`.

27.1 Il cubo di Rubik

Un cubo di Rubik è un cubo con le facce colorate, in cui ciascuna faccia è divisa in 9 pezzi tramite 4 tagli paralleli ai lati del cubo (due in ciascuna direzione). Questo divide il cubo in 27 cubetti più piccoli, di quattro tipi diversi: abbiamo 8 angoli, ciascuno dei quali ha tre colori; 12 spigoli, di due colori; 6 centri, di un colore solo; un nucleo, corrispondente al pezzo interno al cubo, che non ha alcun colore, e che possiamo ignorare non essendo mai visibile. Una mossa su un cubo di Rubik consiste nel ruotare una faccia intorno al proprio centro di un multiplo intero di 90°: questo tiene fisso il centro corrispondente, e permuta in qualche modo 4 spigoli e 4 angoli (ma ovviamente gli spigoli vengono mandati in spigoli e gli angoli vengono mandati in angoli).

Per comodità, conviene fissare un po' di notazione. Lo schema di colori "standard" del cubo di Rubik ha la faccia in alto colorata di bianco, quella in basso di giallo, quella davanti di verde, quella dietro di blu, quella a destra di rosso, e quella a sinistra di arancione.

Per indicare una rotazione di 90° in senso orario di una certa faccia, useremo la lettera iniziale della corrispondente parola in inglese, scritta in maiuscolo: `U` per *up*, `D` per *down*, `F` per *front*, `B` per *back*, `R` per *right*, e `L` per *left*. Per indicare una rotazione in senso antiorario, apporremo un apice dopo la lettera, e per indicare una rotazione di 180° apporremo un 2: per esempio, `R'` denota una rotazione di 90° in senso antiorario della faccia a destra (quella rossa), mentre `F2` denota una rotazione di 180° della faccia davanti (quella verde).

Possiamo identificare un pezzo con l'insieme dei centri che lo toccano. Nello stato risolto, per esempio, chiamiamo `UF` il pezzo bianco-verde, e `FDL` il pezzo verde-giallo-arancione. Indicare un pezzo in questo modo è utile anche per specificare un'orientazione: se dico che il pezzo giallo-rosso

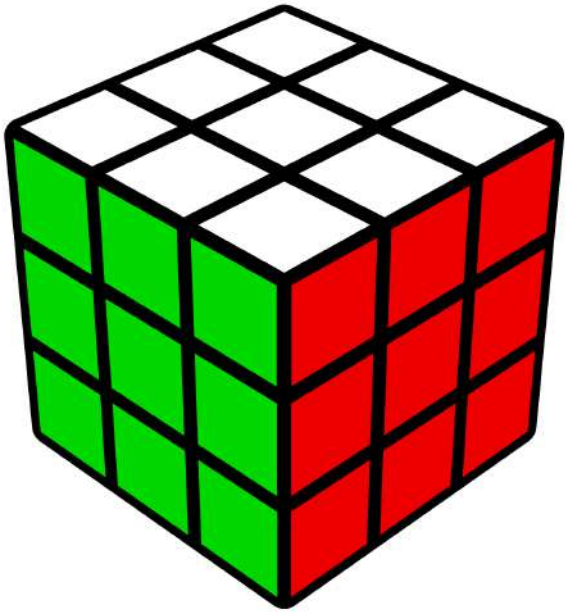


Figura 79: Un cubo di Rubik.

si trova in `UF`, intendo che lo sticker giallo è adiacente al centro bianco (quello in `U`), mentre lo sticker rosso è adiacente al centro verde (quello in `F`).

27.2 Struttura di gruppo

L'insieme delle possibili configurazioni del cubo di Rubik, che chiameremo $\mathfrak{R}$, è un gruppo. Cos'è un gruppo?

Definizione 4. Un gruppo è un insieme G dotato di un'operazione $G \times G \to G$, che a due elementi a, b di G associa un elemento $a \cdot b$, tale che

1. sia *associativa*, ossia che $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ per ogni $a, b, c \in G$;
2. ammetta un *elemento neutro*, ossia che esista un elemento $e \in G$ tale che $a \cdot e = e \cdot a = a$ per ogni $a \in G$;
3. ogni elemento ammetta un *inverso*, ossia che per ogni $a \in G$ esista un elemento $a' \in G$ tale che $a \cdot a' = a' \cdot a = e$.

Esercizio 1. Verificare che i seguenti insiemi con operazione rispettano le proprietà della Definizione 4, cioè sono *gruppi*:

1. i numeri interi con l'addizione $(\mathbb{Z}, +)$;
2. i numeri reali positivi con la moltiplicazione $(\mathbb{R}_+, \cdot)$;
3. le funzioni bigettive da un insieme X in se stesso con la composizione $(S(X), \circ)$.

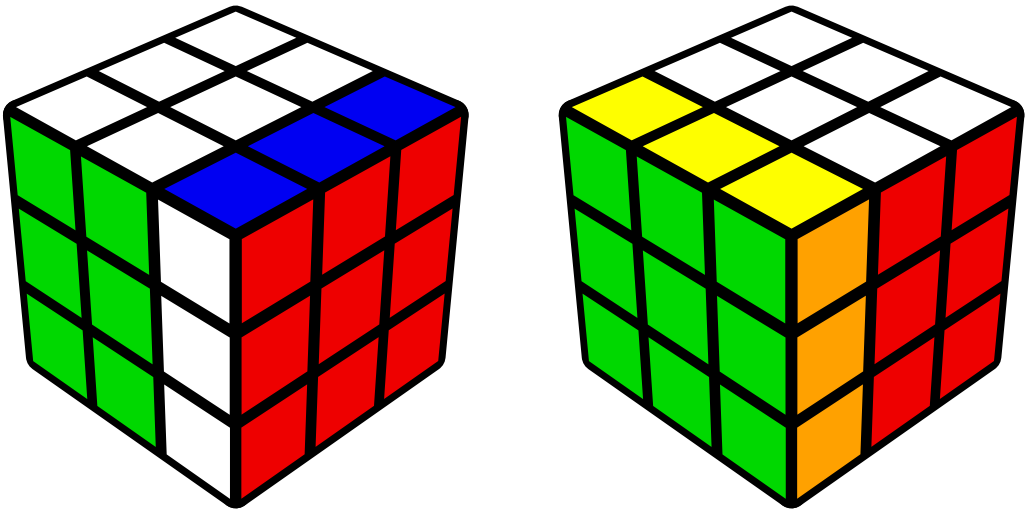


Figura 80: Le mosse `R'` (sinistra) e `F2` (destra) applicate ad un cubo risolto.

Le funzioni bigettive da $\{1, \dots, n\}$ in sé vengono dette *permutazioni*; il gruppo corrispondente viene chiamato *gruppo simmetrico su n elementi* e si denota con S_n. Questo gruppo tornerà utile in seguito.

Esercizio 2. I numeri reali con la moltiplicazione $(\mathbb{R}, \cdot)$ formano un gruppo? Perché?

In che modo l'insieme delle possibili configurazioni del cubo di Rubik è un gruppo? Quello che possiamo fare è identificare ogni configurazione con una (qualsiasi) sequenza di mosse che porta un cubo risolto in quella configurazione, e definire la moltiplicazione di due configurazioni come la configurazione ottenuta applicando le due sequenze di mosse una dopo l'altra. Se più sequenze di mosse distinte portano il cubo nella stessa configurazione, allora le due sequenze identificano lo stesso elemento del gruppo.

Esercizio 3. Verificare che $(\mathfrak{R}, \cdot)$ è un gruppo.

Per finire, una piccola osservazione che potrebbe esservi utile anche nel verificare le proprietà di gruppo dell'esercizio precedente:

Remark. In un gruppo, l'inverso del prodotto $a \cdot b$ è $b' \cdot a'$: infatti

$$(ab)(b'a') = a(bb')a' = aea' = aa' = e.$$

Nel caso del gruppo del cubo, l'inverso di una sequenza di mosse è la sequenza di mosse inverse applicata nell'ordine opposto.

27.3 Pillole di teoria dei gruppi

In questa sezione vogliamo elencare un po' di fatti sui gruppi di cui avremo bisogno in seguito.

Per studiare i gruppi, torna utile considerare alcune funzioni speciali fra gruppi, che rispettano le operazioni.

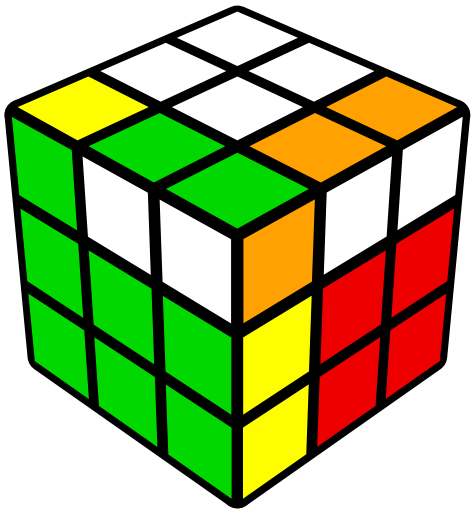


Figura 81: Le sequenze `U' R' F R F2` e `F2 L F L' U'` portano il cubo nello stesso stato, quindi sono uguali come elementi di $\mathfrak{R}$.

Definizione 5. Dati due gruppi $(G, \cdot), (H, *)$, una funzione $f\colon G \to H$ si dice *omomorfismo di gruppi* se, per ogni $g_1, g_2 \in G$, si ha $f(g_1 \cdot g_2) = f(g_1) * f(g_2)$.

Per esempio, la funzione esponenziale $\exp\colon \mathbb{Z} \to \mathbb{R}_+$ definita come $\exp(n) = e^n$ è un omomorfismo di gruppi fra $(\mathbb{Z}, +)$ e $(\mathbb{R}_+, \cdot)$: infatti $\exp(m+n) = e^{m+n} = e^m e^n = \exp(m) \cdot \exp(n)$, come volevamo.

Un omomorfismo di gruppi che sia anche una funzione bigettiva si dice *isomorfismo*: per esempio, se estendiamo la funzione exp dell'esempio precedente a tutto $\mathbb{R}$, otteniamo un isomorfismo di gruppi fra $(\mathbb{R}, +)$ e $(\mathbb{R}_+, \cdot)$.

Un'importante famiglia di gruppi è data dai *gruppi ciclici*.

Definizione 6. Un gruppo G si dice *ciclico* se esiste un elemento $g \in G$, detto *generatore*, tale che

$$G = \{g^n \mid n \in \mathbb{Z}\},$$

ovvero che, dato un qualsiasi elemento $h \in G$, esiste $n \in \mathbb{Z}$ tale che $h = g^n$.

Un gruppo ciclico è finito se e solo se esiste $n \neq 0$ tale che $g^n = e$. In questo caso, il minimo $n > 0$ con questa proprietà si dice *ordine* del generatore, e coincide con il numero di elementi del gruppo. Denotiamo con $(\mathbb{Z}/n\mathbb{Z}, +)$ il gruppo ciclico su n elementi.

Possiamo pensare a un gruppo ciclico finito come l'insieme dei numeri da 0 ad $n-1$ in un giorno da n ore: prendendo per esempio i nostri giorni da 24 ore, se ora sono le 17:00, fra 14 ore saranno le 7:00, e quindi $17 + 14 = 7 \in \mathbb{Z}/24\mathbb{Z}$; in notazione più compatta, questo si scrive anche $17 + 14 \equiv 7 \pmod{24}$.

Esercizio 4. Verificare che la funzione $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ che manda un intero a nel suo resto della divisione intera per n (ossia l'unico $0 \leq q < n$ tale che $a = nk + q$ per qualche $k \in \mathbb{Z}$) è un omomorfismo di gruppi fra $(\mathbb{Z}, +)$ e $(\mathbb{Z}/n\mathbb{Z}, +)$.

Così come possiamo moltiplicare fra loro gli elementi di un gruppo, possiamo anche moltiplicare fra di loro due gruppi!

Definizione 7. Dati due gruppi $(G, \cdot_G), (H, \cdot_H)$, possiamo definire il loro *prodotto* come il gruppo dato dall'insieme

$$G \times H = \{(g, h) \mid g \in G, h \in H\},$$

ossia l'insieme delle coppie formate da un elemento di G ed un elemento di H, con il prodotto termine a termine, ossia

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 \cdot_G g_2, h_1 \cdot_H h_2).$$

Esercizio 5. Verificare che $(G \times H, \cdot)$ è un gruppo.

Definire questa operazione ci permette di studiare i gruppi "scomponendoli" in pezzi più piccoli, un po' come scomporre i numeri interi in fattori. Per esempio, la funzione $f\colon \mathbb{Z}/24\mathbb{Z} \to \mathbb{Z}/3\mathbb{Z} \times \mathbb{Z}/8\mathbb{Z}$ che manda n nella coppia dei suoi resti per le divisioni per 3 e per 8 è un isomorfismo di gruppi, e questo ci dice che $\mathbb{Z}/24\mathbb{Z}$ si può "scomporre" in due pezzi.

Esercizio 6. La funzione $f\colon \mathbb{Z}/24\mathbb{Z} \to \mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/6\mathbb{Z}$ che manda n nella coppia dei suoi resti per le divisioni per 4 e per 6 è un omomorfismo di gruppi? Se sì, è anche un isomorfismo di gruppi? Perché?

Dato un omomorfismo di gruppi $f\colon G \to H$, l'insieme degli elementi di G che vengono mandati nell'elemento neutro di H si chiama *nucleo* dell'omomorfismo.

Esercizio 7. Verificare che il nucleo di un omomorfismo è a sua volta un gruppo (con l'operazione del gruppo che lo contiene).

Esercizio 8. Dato $f\colon G \to H$ omomorfismo di gruppi suriettivo con G finito, verificare che le controimmagini di un elemento di H hanno tutte lo stesso numero di elementi.

27.4 Il gruppo simmetrico

In precedenza abbiamo menzionato il gruppo simmetrico su n elementi, ovvero il gruppo di funzioni bigettive da $\{1, \dots, n\}$ in sé, che chiamiamo permutazioni. Quanti elementi ha questo gruppo? Facciamo un rapido conto: sia $\sigma \in S_n$ una permutazione. Abbiamo n possibilità per il valore di $\sigma(1)$, poi $n-1$ possibilità per il valore di $\sigma(2)$ (dev'essere diverso da $\sigma(1)$), poi $n-2$ possibilità per il valore di $\sigma(3)$ (dev'essere diverso da entrambi i precedenti) e così via, fino ad 1 solo valore possibile per $\sigma(n)$. Ognuna di queste scelte ci dà una permutazione diversa, e così in totale abbiamo

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

permutazioni. Questo numero è talmente ricorrente che vogliamo dargli un nome.

Definizione 8. Definiamo *fattoriale* di un numero intero positivo n il numero

$$n! := n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1$$

(si legge n *fattoriale*).

Ci sono vari modi di scrivere una permutazione. Quello forse più semplice è con la cosiddetta *one-line notation*, che consiste semplicemente nello scrivere i valori della funzione uno di fianco all'altro: per esempio, 74528316 indica la permutazione $\sigma \in S_8$ tale che $\sigma(1) = 7$, $\sigma(2) = 4$ e così via, fino a $\sigma(8) = 6$. Un altro modo, più utile ai nostri scopi, è con la *notazione in cicli*: la stessa permutazione $\sigma \in S_8$ si scriverebbe (17)(24)(3586), che significa che 1 va in 7, poi 7 va in 1 e si chiude il ciclo (che ha lunghezza 2); 2 va in 4, 4 va in 2 e si chiude il ciclo; 3 va in 5, che va in 8, che va in 6, che va in 3 chiudendo il ciclo.

Esercizio 9. Verificare che 74528316 e (17)(24)(3586) denotano la stessa permutazione.

Non è difficile verificare che ogni permutazione si può scrivere, anche non in modo unico, come prodotto (ovvero composizione) di *trasposizioni*, cioè di permutazioni composte da un solo ciclo di lunghezza 2. Per esempio,

$$(17)(24)(3586) = (17) \cdot (24) \cdot (68) \cdot (58) \cdot (35),$$

dove il "prodotto" a destra denota la permutazione ottenuta scambiando prima 1 con 7, poi 2 con 4, poi 6 con 8, poi 5 con 8 e infine 3 con 5.

Una permutazione si dice *pari* se è prodotto di un numero pari di trasposizioni, e *dispari* altrimenti. Così come la somma di due numeri interi con la stessa parità è pari, e la somma di due numeri interi con parità diverse è dispari, così il prodotto di due permutazioni con la stessa parità è pari, e il prodotto di due permutazioni con parità diverse è dispari.

Esercizio 10. Verificare che un ciclo è pari se la sua lunghezza è dispari, ed è dispari se la sua lunghezza è pari.

Esercizio 11. Verificare che la funzione $f\colon S_n \to \mathbb{Z}/2\mathbb{Z}$ che manda le permutazioni pari in 0 e quelle dispari in 1 è un omomorfismo di gruppi. Dedurre che esattamente metà delle permutazioni di n elementi sono pari, e metà sono dispari.

27.5 Smontare e rimontare il cubo

Siamo pronti per un conto intermedio. Prendiamo un cubo di Rubik, smontiamolo e rimontiamolo a caso. In quanti modi possiamo farlo? Dunque, abbiamo detto che il nucleo e i centri non si muovono, quindi quelli li teniamo fermi. Ci restano 12 spigoli e 8 angoli, e dobbiamo mettere gli spigoli al posto degli spigoli e gli angoli al posto degli angoli. Per ogni spigolo, dobbiamo scegliere fra quali due centri metterlo, e in che modo: per esempio, se decidiamo di mettere lo spigolo verde-rosso tra il centro bianco e quello blu, possiamo mettere il verde vicino al bianco e il rosso vicino al blu, o viceversa. Per il primo spigolo abbiamo 12 possibilità per la posizione, e 2 per l'orientazione; per il secondo, 11 per la posizione e 2 per l'orientazione, e così via, fino all'ultimo spigolo che ha 1 solo posto disponibile, con 2 possibilità per l'orientazione. In totale abbiamo quindi

$$12 \cdot 2 \cdot 11 \cdot 2 \cdot \ldots \cdot 1 \cdot 2 = 12! \cdot 2^{12}$$

modi di montare gli spigoli. Similmente, dato che ognuno degli 8 angoli ha invece 3 possibili orientazioni, abbiamo

$$8 \cdot 3 \cdot 7 \cdot 3 \cdot \ldots \cdot 1 \cdot 3 = 8! \cdot 3^8$$

modi di montare gli angoli.

In totale, quindi, smontando e rimontando un cubo abbiamo $12! \cdot 2^{12} \cdot 8! \cdot 3^8$ possibili modi di rimontarlo. Se fosse possibile raggiungere ciascuna di queste configurazioni semplicemente ruotando le facce del cubo, allora avremmo risolto il nostro problema! Dobbiamo controllare se sia così oppure no.

Per risolvere problemi di combinatoria come questo, spesso si fa ricorso agli invarianti. Un *invariante* è una quantità che non cambia, o che cambia in modo controllato, quando si eseguono delle mosse in certi giochi. Prendiamo per esempio gli scacchi: il colore della casella su cui si trova un alfiere non cambia mai, quindi non importa in che situazione ci si trovi o come ci si sia arrivati, noi sappiamo per certo che l'alfiere che all'inizio si trovava sulla casella `f1`, che è

bianca, se non è stato catturato si troverà ancora su una casella bianca; il colore della casella su cui si trova un cavallo, invece, cambia ad ogni mossa, quindi per esempio il cavallo che all'inizio si trovava sulla casella `g1`, che è nera, se è stato mosso un numero pari di volte si troverà ancora su una casella nera, e se è stato mosso un numero dispari di volte allora si troverà su una casella bianca (sempre se non è stato catturato).

Così come non tutte le disposizioni dei pezzi su una scacchiera si possono ottenere durante una partita, non tutte le configurazioni di un cubo di Rubik che si possono ottenere smontandolo e rimontandolo si possono anche ottenere semplicemente ruotando le facce. Cerchiamo di trovare degli invarianti, e cerchiamo di capire se sono tutti quelli che ci interessano.

27.6 Permutazione

Per trovare il nostro primo invariante, assegniamo un numero da 1 a 8 ad ogni angolo ed un numero da 1 a 12 ad ogni spigolo, e ad ogni configurazione associamo due permutazioni, $\sigma_a \in S_8$ per gli angoli e $\sigma_s \in S_{12}$ per gli spigoli, che ci dicono in che posizioni si trovano i vari pezzi; per il momento, ignoreremo l'orientazione. Per esempio, se lo spigolo bianco-blu (a cui associamo per esempio il numero 1) si trova tra il centro verde e quello rosso, e allo spigolo verde-rosso abbiamo assegnato il numero 7, allora avremo che $\sigma_s(1) = 7$. Questa funzione è un *omomorfismo di gruppi* fra $\mathfrak{R}$ e $S_8 \times S_{12}$, ossia (come spiegato alla Sezione 3) una funzione, che chiamiamo f, fra $\mathfrak{R}$ e il prodotto cartesiano di S_8 e S_{12} (che è un gruppo con la composizione componente per componente) tale che $f(e) = e$ (l'immagine dell'elemento neutro è l'elemento neutro), e che $f(a \cdot b) = f(a) \cdot f(b)$.

Con questa associazione, ruotare una singola faccia corrisponde a un ciclo di lunghezza 4 (o, più comodamente, 4-ciclo) sia per gli angoli che per gli spigoli. Ricordiamo che un 4-ciclo è una permutazione dispari. Data una configurazione qualsiasi scriviamo una qualsiasi sequenza di mosse che porta un cubo risolto in quella configurazione: la coppia di permutazioni associata alla configurazione è il prodotto delle coppie di permutazioni associate alle singole mosse, e dato che per ogni mossa le due permutazioni hanno la stessa parità, allora anche le due permutazioni associate alla configurazione finale hanno la stessa parità. Questo però ci dice che non tutte le configurazioni che possiamo ottenere smontando un cubo le possiamo anche ottenere ruotando le facce: per esempio, smontando il cubo posso scambiare di posto due spigoli e nient'altro, ma questo ci dà un 2-ciclo di spigoli (dispari) e la permutazione identica degli angoli (pari), che hanno parità differente; noi abbiamo però dimostrato che ruotando le facce possiamo ottenere solo permutazioni di angoli e spigoli che hanno la stessa parità, quindi questa specifica configurazione non si può ottenere.

Quante sono le coppie di permutazioni in $S_8 \times S_{12}$ con la stessa parità? Fissata la prima delle due, se è pari ($8!/2$ casi) allora anche l'altra deve essere pari ($12!/2$ possibilità), mentre se è dispari ($8!/2$ casi) allora anche l'altra deve essere dispari ($12!/2$ possibilità). In totale abbiamo quindi

$$\frac{8!}{2} \cdot \frac{12!}{2} + \frac{8!}{2} \cdot \frac{12!}{2} = \frac{1}{2}(8! \cdot 12!)$$

permutazioni possibili. Dobbiamo ancora verificare di poterle ottenere tutte, ma lo faremo in seguito.

27.7 Orientazione

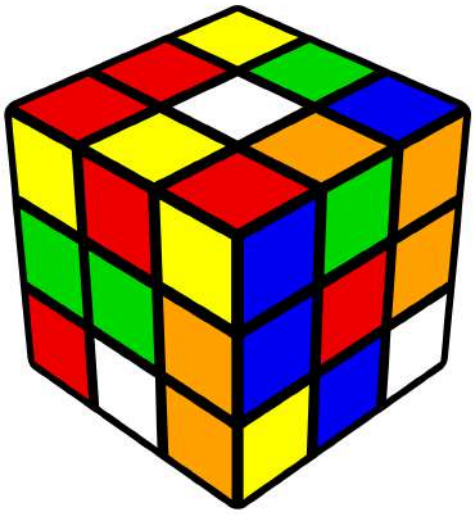
Abbiamo bisogno di altri due invarianti, uno per gli angoli ed uno per gli spigoli. Entrambi riguardano l'orientazione dei pezzi, ovvero guardano in che modo sono disposti i colori all'interno

di ciascun pezzo, ma non dove si trova il pezzo stesso.

27.7.1 Orientazione degli angoli

Per tenere traccia dell'orientazione degli angoli, dobbiamo fissare il colore della faccia in alto; finora abbiamo usato il bianco, quindi continueremo con questa scelta. È importante anche tenere traccia del colore della faccia opposta, che nel nostro cubo è gialla. Osserviamo che ogni angolo ha esattamente un adesivo bianco o un adesivo giallo, ma non entrambi. Adesso, ad ogni angolo associamo un numero fra $-1, 0, 1$ in questo modo: se l'adesivo bianco o giallo si trova in `U` oppure `D`, assegniamo 0; se occorre ruotare l'angolo di 120° su sé stesso in senso orario per portarlo su `U` o `D`, assegniamo 1; se occorre ruotare l'angolo di 120° su se stesso in senso antiorario per portarlo su `U` o `D`, assegniamo -1. Questo numero ci dice di quale multiplo di 120° dobbiamo ruotare l'angolo per portarlo nell'orientazione corretta: ovviamente assegnare -2 è la stessa cosa che assegnare 1, assegnare 3 è la stessa cosa che assegnare 0, e così via.

Nell'esempio in Figura 82, gli angoli `URF` e `FRD` vanno ruotati in senso orario e quindi assegniamo loro il valore 1; gli angoli `UFL`, `UBR`, `DRB` e `DLF` vanno ruotati in senso antiorario e quindi assegniamo loro -1; l'angolo `ULB` non va ruotato e quindi gli assegniamo 0. Il totale fa -2, quindi l'invariante ci garantisce che anche l'angolo `DBL`, che in figura non è visibile, va ruotato in senso antiorario. Controllate con il vostro cubo o scansionando il QR!



SCAN ME

https://alg.cubing.net/?setup=B_F2_L-_U-_F-_U2_L2_U2_F2_U-_B_D_L-_D_U-_B-_U2_R-_B-_D2_U_B-_D2_U_L-_R_D_L-_R2_U-

Figura 82: Un cubo mescolato con la sequenza `B F2 L' U' F' U2 L2 U2 F2 U' B D L' D U' B' U2 R' B' D2 U B' D2 U L' R D L' R2 U'`.

Cosa succede applicando una mossa? Applicare una mossa `U` o `D` non cambia nessuno dei valori assegnati agli angoli, mentre applicare una mossa `F`, `R`, `B` o `L` aumenta due valori di 1 e diminuisce due valori di 1; ricordiamo che 2 o -1 è la stessa cosa, quindi se su un angolo con valore 1 viene eseguita una mossa che aumenta il suo valore di 1, gli assegneremo invece -1. Questo ci dice che, effettuando una mossa qualsiasi, la somma dei valori assegnati agli angoli non cambia, se non che ogni tanto dobbiamo aggiungere o sottrarre 3 per mantenere tutti i valori uguali a -1, 0, o 1; quindi alla fine la somma resta sempre un multiplo di 3.

Quante orientazioni possibili abbiamo per gli angoli che siano compatibili con questa regola? Possiamo mettere i primi 7 come vogliamo, e per ciascuno abbiamo 3 possibilità; per l'ultimo invece abbiamo una sola scelta, dovendo per forza avere l'unica possibile orientazione che rende il totale un multiplo di 3: abbiamo quindi al massimo 3^7 orientazioni possibili. Come prima, dobbiamo ancora verificare di poterle ottenere tutte, ma lo faremo in seguito.

27.7.2 Orientazione degli spigoli

Per quanto riguarda l'orientazione degli spigoli, dobbiamo invece fissare il colore della faccia di fronte; finora abbiamo usato il verde, quindi continueremo con questa scelta. È importante anche tenere traccia del colore della faccia opposta, che nel nostro cubo è blu.

Definizione 9. Uno spigolo è *orientato* rispetto all'asse `F/B` se è possibile portarlo nella posizione corretta con l'orientazione corretta utilizzando solo mosse `U`, `D`, `R`, `L`, `F2` e `B2`, ed è *non orientato* altrimenti.

Usando solo le mosse date sopra, è sempre possibile portare uno spigolo nella posizione corretta: la sua orientazione a quel punto ci dirà se lo spigolo è orientato o no.

Sempre con riferimento all'esempio in Figura 82, lo spigolo `UF` è orientato, perché possiamo portarlo nella posizione e orientazione corretta con `F2 D`; lo spigolo `UR` non è orientato, perché con `U2 L` lo portiamo nella posizione corretta ma con l'orientazione sbagliata; neanche lo spigolo `FR` è orientato, perché con `F2 L2` lo portiamo nella posizione corretta ma con l'orientazione sbagliata.

Come prima, dobbiamo capire cosa succede applicando una mossa. Per definizione, applicare una mossa `U`, `D`, `R` o `L` non cambia l'orientazione di nessuno spigolo; applicare una mossa `F` o `B`, invece, cambia l'orientazione di tutti gli spigoli su quella faccia. In particolare, il numero di spigoli orientati cambia sempre di un numero pari: di 4 se sulla faccia `F` (o `B`) ci sono 0 o 4 spigoli orientati; di 2 se ce ne sono 1 o 3; oppure di 0 se ce ne sono 2.

Di nuovo, quante sono le possibili orientazioni degli spigoli compatibili con questa regola? Comunque mettiamo i primi 11 spigoli, esiste sempre un'unica scelta per l'orientazione dell'ultimo tale per cui il numero totale di spigoli orientati sia pari. Questo ci dice che abbiamo al massimo 2^{11} possibilità; come al solito, dobbiamo ancora verificare di poterle ottenere tutte.

27.8 Verifica finale

È finalmente arrivato il momento di verificare che gli invarianti che abbiamo trovato sono *completi*, cioè che ogni configurazione che soddisfa i requisiti che abbiamo trovato si può effettivamente ottenere ruotando le facce del cubo. Come bonus, impareremo una tecnica (non molto efficiente, ma funzionale) per risolvere il cubo a partire da qualsiasi configurazione ottenibile!

Per fare ciò, abbiamo bisogno di quello che in gergo si chiama *algoritmo*, cioè una sequenza di mosse che sposta solo alcuni pezzi in modo controllato. Useremo quella che i cuber chiamano *T-perm*, che si può ottenere per esempio con le mosse `R U R' U' R' F R2 U' R' U' R U R' F'`.

Questa sequenza scambia fra di loro gli spigoli `UL` e `UR`, e gli angoli `URF` e `UBR`. Vedremo come usare dei *coniugati* di questa sequenza per risolvere il cubo a partire da una qualsiasi configurazione compatibile con gli invarianti trovati in precedenza; l'inverso della soluzione ci permette poi di ottenere la configurazione di partenza partendo da un cubo risolto.

Definizione 10. Dato un gruppo G ed un elemento $a \in G$, si dice *coniugato di a tramite b* l'elemento bab'.

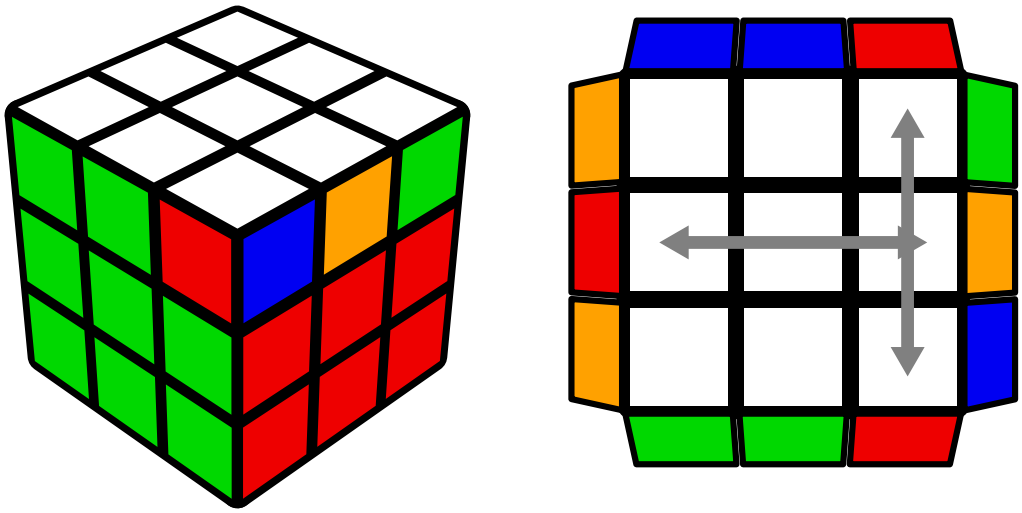


Figura 83: La T-perm da due angolazioni diverse.

Sebbene non ci sia una regola generale, spesso il coniugato di un elemento ha delle caratteristiche in comune con l'elemento di partenza; per esempio, il coniugato di una permutazione ha lo stesso numero di cicli di ciascuna lunghezza della permutazione di partenza. Nel nostro caso, un qualsiasi coniugato della T-perm sarà una qualche sequenza che scambia fra di loro due angoli e due spigoli.

27.8.1 Risolvere gli spigoli

Utilizzeremo la seguente strategia: guardiamo lo spigolo `UR` e vediamo dove dovrebbe andare (per esempio, se è rosso-verde, in `RF`); utilizzando una sequenza di mosse che lascia fissi lo spigolo `UR` e gli angoli `URF` e `UBR`, portiamo il pezzo in questione in `UL` (nel nostro esempio, possiamo usare la sequenza `U' F' U`); dopodiché, applichiamo la T-perm; infine, applichiamo l'inverso della sequenza precedente (ovvero `U' F U`). In questo modo, abbiamo scambiato lo spigolo `UR` e lo spigolo `RF`, e gli angoli `URF` e `UBR`.

Ripetiamo la procedura fino a che nel posto `UR` ritroviamo lo spigolo bianco-rosso (in qualsiasi orientazione). Se a questo punto gli spigoli sono risolti, passiamo agli angoli; altrimenti, scambiamo `UR` con un qualsiasi spigolo non risolto; continueremo così fino a quando tutti gli spigoli tranne al più `UR` saranno risolti. A questo punto siamo certi che anche `UR` sarà risolto: se tutti gli altri spigoli sono al loro posto, allora lo spigolo bianco-rosso deve essere per forza nell'ultimo posto rimasto (cioè `UR`), e se tutti gli altri spigoli sono orientati correttamente, dato che il numero di spigoli non orientati deve essere pari, allora anche lo spigolo bianco-rosso deve essere orientato correttamente.

Quello che resta da fare è verificare che, per ogni spigolo (contando anche l'orientazione), esiste almeno un modo di portarlo in `UL` senza toccare lo spigolo `UR` e gli angoli `URF` e `UBR`. Procediamo nel modo seguente:

1. se lo spigolo (che denoteremo genericamente con `XY`) è nello strato equatoriale (cioè non è né su `U` né su `D`), allora possiamo ruotare la faccia `U` in modo che gli sticker `LU` e `YX` si

trovino sulla stessa faccia (cioè Y), dopodiché ruotiamo quella faccia di 90° in senso orario o antiorario, in modo da portare lo spigolo che ci interessa sullo strato U, e infine ruotiamo la faccia U in modo da portarlo in UL. Abbiamo visto prima un esempio con RF;

2. se lo spigolo è nello strato D, a prescindere dall'orientazione, possiamo portarlo in DL con delle mosse D, dopodiché portarlo nello strato equatoriale con una mossa L, e ricondurci al caso 1;

3. se lo spigolo è già in UL (orientato correttamente), allora non dobbiamo fare nulla; se è il LU (cioè nello stesso posto ma non orientato correttamente), possiamo invece portarlo nello strato equatoriale con una mossa L, riconducendoci nuovamente al caso 1;

4. se lo spigolo è in UF (rispettivamente UB), a prescindere dall'orientazione, la sequenza R F' R' (rispettivamente R' B R) lo porta nello strato equatoriale senza influenzare lo spigolo UR e gli angoli URF e UBR, riconducendoci ancora al caso 1.

La casistica è esaustiva, e quindi questo ci permette di risolvere gli spigoli a partire da qualsiasi configurazione compatibile con gli invarianti trovati.

27.8.2 Risolvere gli angoli

Anche per gli angoli useremo una strategia simile: guardiamo l'angolo URF e vediamo dove dovrebbe andare (per esempio, se è rosso-giallo-verde, in RDF); utilizzando una sequenza di mosse che lascia fissi gli spigoli UR e UL e l'angoli URF, portiamo il pezzo in questione in UBR (nel nostro esempio, possiamo usare la sequenza D2 R D' R'); dopodiché, applichiamo la T-perm; infine, applichiamo l'inverso della sequenza precedente (ovvero R D R' D2). In questo modo, abbiamo scambiato l'angolo URF e l'angolo UBR, e gli spigoli UR e UL.

Ripetiamo la procedura fino a che nel posto URF ritroviamo l'angolo bianco-rosso-verde (in qualsiasi orientazione). Se a questo punto gli angoli sono risolti, abbiamo finito; altrimenti, scambiamo URF con un qualsiasi angolo non risolto; continueremo così fino a quando tutti gli angoli tranne al più URF saranno risolti. A questo punto, il cubo è interamente risolto: se tutti gli altri angoli sono al loro posto, allora l'angolo bianco-rosso-verde deve essere per forza nell'ultimo posto rimasto (cioè URF); se tutti gli altri angoli sono orientati correttamente, dato che il numero totale di rotazioni di 120° da fare deve essere un multiplo di 3, allora anche l'ultimo angolo deve essere orientato correttamente. Infine, dato che partivamo da una configurazione con gli spigoli risolti, che è pari, allora anche la permutazione degli angoli da fare doveva essere pari; questo vuol dire che abbiamo applicato la T-perm un numero pari di volte, il che ci dice che abbiamo scambiato gli spigoli UR e UL un numero pari di volte (e non abbiamo spostato nessun altro spigolo), e pertanto alla fine anche gli spigoli saranno risolti.

Come prima, dobbiamo verificare che, per ogni spigolo (contando anche l'orientazione), esiste almeno un modo di portarlo in UL senza toccare l'angolo URF e gli spigoli UR e UL. Procediamo nel modo seguente:

1. se l'angolo è in DBL (con questa orientazione), eseguiamo B2;

2. se l'angolo è in BDL (con questa orientazione), eseguiamo D' B;

3. se l'angolo è in LDB (con questa orientazione), eseguiamo R D' R';

4. se l'angolo è nello strato D, possiamo portarlo in DBL (senza considerare l'orientazione) con delle mosse D, riconducendoci ad uno dei casi precedenti;

5. se l'angolo è già il `UBR` (orientato correttamente), allora non dobbiamo fare nulla; se invece non è orientato correttamente, possiamo portarlo in `LDB` eseguendo `R D R'`, riconducendoci al caso 3;

6. se l'angolo è in `UFL` (rispettivamente `ULB`), a prescindere dall'orientazione, la sequenza `L D L'` (rispettivamente `L' D' L`) lo porta nello strato `D` senza influenzare l'angolo `URF` e gli spigoli `UR` e `UL`, riconducendoci al caso 4.

In ogni caso, ricordatevi sempre di fare l'inverso di tutte le mosse che avete usato per portare l'angolo nella posizione giusta, inclusi tutti i setup!

Anche per gli angoli la casistica è esaustiva, e quindi questo ci permette di risolvere il cubo a partire da qualsiasi configurazione compatibile con gli invarianti trovati.

27.8.3 Conclusione

Abbiamo trovato una (lunga) sequenza di mosse che risolve il cubo a partire da una qualsiasi configurazione compatibile con gli invarianti trovati; applicando l'inverso di questa sequenza ad un cubo risolto, otteniamo la configurazione di partenza. Questo dimostra che qualsiasi configurazione compatibile con gli invarianti si può ottenere ruotando le facce del cubo, senza doverlo smontare. Abbiamo quindi $\frac{1}{2}(8! \cdot 12!)$ possibili permutazioni, 3^7 possibili orientazioni per gli angoli, e 2^{11} possibili orientazioni per gli spigoli, tutte indipendenti fra di loro, per un totale di

$$\frac{1}{2}(8! \cdot 12!) \cdot 3^7 \cdot 2^{11} = 43.252.003.274.489.856.000$$

configurazioni, come avevamo detto all'inizio.

Riferimenti bibliografici

[1] Come risolvere il cubo di Rubik? `https://rubiks.com/solve-it`

[2] Guida alla Fewest Moves Challenge: `https://fmcsolves.cubing.net/`

[3] In quante mosse è sempre possibile risolvere un cubo di Rubik? `http://kociemba.org/moves20.htm`

28 Julia e Mandelbrot

Samuele Mongodi, n.16, Febbraio 2023

28.1 Noia al quadrato

Consideriamo un numero reale a; che succede se ne facciamo ripetutamente il quadrato, senza fermarci? Otterremo una successione di valori:

$$a, a^2, a^4, a^8, \dots, a^{2^n}, \dots$$

di cui è semplice studiare il comportamento. Facciamolo però nel dettaglio; per farlo, utilizziamo la disuguaglianza di Bernoulli, che lasciamo come esercizio di induzione allo studente studioso:

Siano $x > -1$ un numero reale e $n \in \mathbb{N}$ un numero naturale, allora si ha

$$(1+x)^n \geq 1 + nx \ .$$

Vediamo ora cosa fa la nostra noiosa successione:

- se $a = 0$, tutti gli elementi sono 0
- se $a = \pm 1$, tutti gli elementi dal secondo in poi sono 1
- se $1 < a$, scriviamo $a = 1 + (a-1)$ ed applichiamo Bernoulli: $a^{2^n} = (1 + (a-1))^{2^n} \geq 1 + 2^n(a-1)$ che ci fa capire subito che la nostra successione diventa sempre più grande, cioè, *tende a* $+\infty$
- se $0 < a < 1$, $a = 1/b$ con $b > 1$ e dunque $a^{2^n} = 1/b^{2^n}$; se b^{2^n} diventa sempre più grande, $1/b^{2^n}$ diventa sempre più piccolo, più di ogni reale positivo, e dunque *tende a* 0
- se $a < 0$, $a^2 > 0$ e $|a| < 1$ se e solo se $a^2 < 1$, il che ci dice che abbiamo concluso i casi possibili.

La retta reale è dunque divisa in tre[39]:

1. l'intervallo aperto $(-1, 1)$ "va in 0" (cioè, partendo da lì, la successione tende a 0)
2. l'unione delle due semirette $(-\infty, -1) \cup (1, +\infty)$ "va a $+\infty$"
3. i punti -1, $+1$ danno una successione costantemente uguale a 1 dal secondo termine in poi.

Dunque, in un certo senso, la retta è divisa in due "grossi pezzi", ognuno dei quali contiene[40] un valore (0 o $+\infty$) che attrae gli altri. In più, il "confine" tra questi due pezzi contiene un terzo valore limite, che però è molto più brutale degli altri due: le successioni che partono sulla zona di confine si stabilizzano sul valore limite (cioè 1) in un numero finito di passi (in realtà, in uno o due passi); notiamo che, se siamo in uno dei due pezzi grossi, un punto reale si comporta come quelli che gli sono immediatamente vicini, mentre, al confine tra i due pezzi, basta muoversi di pochissimo (da 1 o da -1) per trovare comportamenti radicalmente diversi.

La noia di un simile esempio ci spinge a porci alcune domande.

[39] Recta est omnis divisa in partes tres.

[40] ok, dire che l'insieme $\{a \in \mathbb{R} \ : \ |a| > 1\}$ "contiene" ∞ è un po' un abuso di linguaggio, ma capite cosa voglio dire...

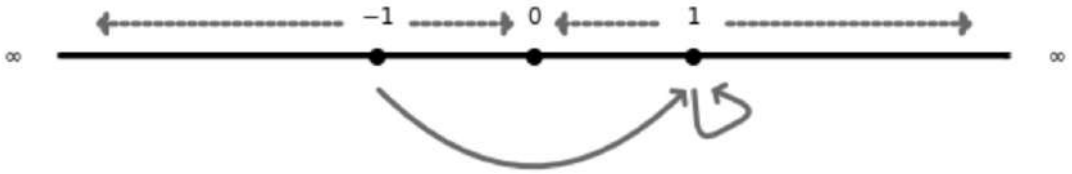


Figura 84: Dinamica della funzione $f(x) = x^2$.

Da dove arrivano i valori limite? I valori limite devono essere valori *fissi* per la nostra successione, cioè se partiamo da lì dobbiamo restare lì. Ovvero, devono soddisfare l'equazione $L^2 = L$, che, come sappiamo tutti, ha tre soluzioni, giusto? No, ne ha 2 nei numeri reali, cioè $L = 0$ e $L = 1$; da dove possiamo vedere che anche $+\infty$ sarà un possibile valore limite?

Che succede se cambiamo la regola della successione? Ad esempio, sarà chiaro a tutti che prendere le quarte potenze invece dei quadrati non cambia di fatto nulla, o che prendere i cubi invece dei quadrati aggiunge punti limite[41], ma che succede con regole più complicate?

Ci sono solo questi due tipi di comportamento? Cioè, regioni "attratte" da un loro punto fisso, oppure valori che diventano fissi dopo un numero finito di passi?

Come distinguere i valori limite tra loro? Nel nostro esempio noioso, il valore 0 attira a sé i valori vicini, come fa, dando opportuno significato al termine "vicini", il valore $+\infty$. D'altra parte il punto 1 non funziona affatto così. Ci sono indizi che possono rivelare questa differenza di comportamento?

28.2 Disegnetti

Generalizzando con foga l'esempio da cui siamo partiti, potremmo considerare una funzione $f : \mathbb{R} \longrightarrow \mathbb{R}$ e definire, dato $a \in \mathbb{R}$, la successione per ricorrenza

$$\begin{cases} a_n = f(a_{n-1}) & \text{se} \quad n \geq 1 \\ a_0 = a \end{cases}$$

Cioè, $a_1 = f(a)$, $a_2 = f(f(a))$, etc. etc., ogni a_n è ottenuto applicando n volte f al valore iniziale a.

La nostra successione noiosa era ottenuta con $f(x) = x^2$. Interpretiamo il problema in termini del grafico di $f(x)$ (vedi la Figura 85).

Innanzitutto, i punti fissi (cioè i candidati valori limite, a parte $+\infty$) dati dall'equazione $f(L) = L$ sono le intersezioni del grafico di $f(x)$ con la bisettrice del primo e terzo quadrante, $y = x$. Inoltre, notiamo che la coppia (a_{n-1}, a_n) rappresenta, per ogni n, un punto sul grafico di $f(x)$, poiché $a_n = f(a_{n-1})$.

Se dunque partiamo dal punto $(a, 0)$ sull'asse delle x, possiamo muoverci in verticale fino ad incontrare il grafico di $f(x)$ nel punto $(a, f(a)) = (a_0, a_1)$; ora, muovendosi in orizzontale

[41] la semiretta $(-\infty, -1)$ va a $-\infty$, -1 e 1 sono valori fissi, $(-1, 1)$ va a 0, $(1, +\infty)$ va a $+\infty$

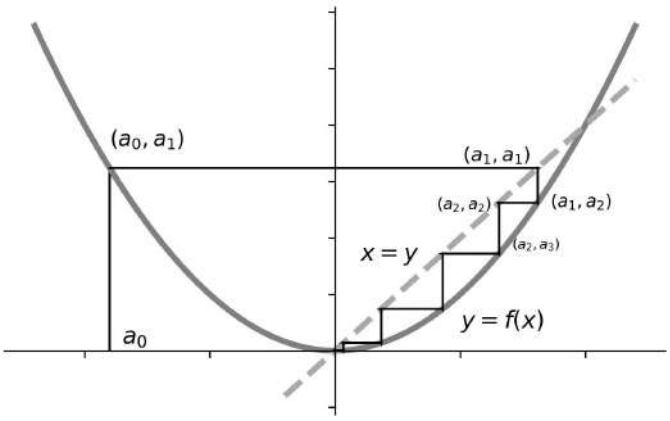


Figura 85: Iterate di $f(x) = x^2$, a partire da $a_0 = -0.9$. E' chiara la convergenza a 0.

fino ad incontrare la bisettrice del primo e del terzo quadrate, otteniamo il punto (a_1, a_1), dal quale possiamo muoverci in verticale nuovamente fino ad incontrare il grafico di $f(x)$ nel punto $(a_1, f(a_1)) = (a_1, a_2)$.

Iterando questa costruzione, otteniamo dei punti sul grafico di $f(x)$ le cui ascisse sono i valori della successione a_n. Provate pure a farlo da soli, variando il punto di partenza, usando il grafico di $f(x) = x^2$, per riottenere i risultati del punto precedente. Ovviamente, può essere divertente provare a farlo anche per altri grafici di funzione.

28.3 Proiezione stereografica

Un modo per far entrare l'infinito nei giochi, almeno graficamente, è "avvolgere" la retta reale su una circonferenza; ci avanzerà, se lo avremo fatto per bene, un punto, che conterà come infinito. In pratica, possiamo fare così: sul piano cartesiano consideriamo la circonferenza unitaria $x^2 + y^2 = 1$ e l'asse delle ascisse; per ogni punto $(a, 0)$ dell'asse, costruiamo la retta per esso e per il punto $(0, 1)$, cioè $ay = a - x$, ed intersechiamola con la circonferenza unitaria, ottenendo il punto

$$\left(\frac{2a}{a^2+1}, \frac{a^2-1}{a^2+1}\right) .$$

Chi sa di trigonometria, riconoscerà queste formule!

In questo modo, i punti dell'asse x corrispondono ai punti della circonferenza, tolto $(0, 1)$, che conta come infinito. All'inverso, dato un punto (b, c) sulla circonferenza, diverso da $(0, 1)$, la stessa costruzione gli associa il punto di ascissa $b/(1 - c)$ sull'asse x.

Ora proviamo a vedere cosa diventa $f(x) = x^2$ sulla circonferenza: partiamo da un punto (b, c) sulla circonferenza, otterremo il punto $(b/(1 - c), 0)$ sull'asse x; applicando la funzione $f(x)$ elevando al quadrato l'ascissa, otteniamo il punto $(b^2/(1 - c)^2, 0)$, che sulla circonferenza è

$$\left(\frac{2b^2(1-c)^2}{b^4+(1-c)^4}, \frac{b^4-(1-c)^4}{b^4+(1-c)^4}\right)$$

e, tenendo conto che $b^2 = 1 - c^2$, possiamo semplificarlo in $\left(\frac{1-c^2}{1+c^2}, \frac{2c}{1+c^2}\right)$.

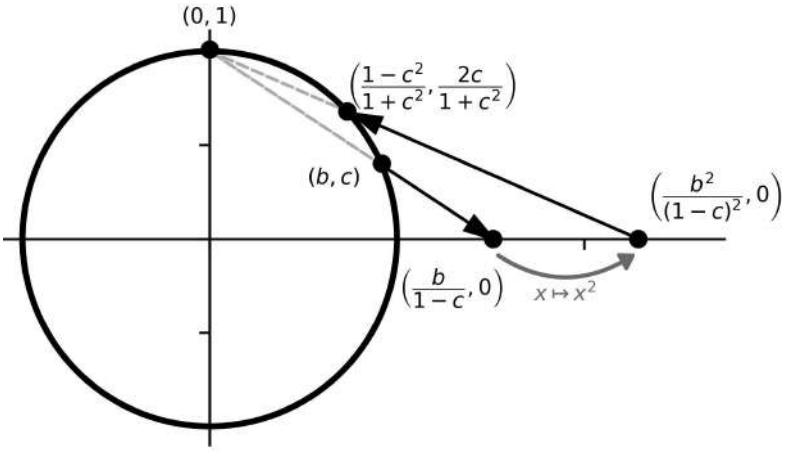


Figura 86: Proiezione stereografica e funzione indotta da $x \mapsto x^2$ sulla circonferenza.

Per quali valori di c questo è uguale al punto iniziale? Beh, dobbiamo innanzitutto avere

$$\frac{2c}{1+c^2} = c$$

che implica $c = 0, 1, -1$. Sostituendo in

$$\frac{1-c^2}{1+c^2} = b$$

otteniamo che b deve essere, rispettivamente $1, 0, 0$. Dunque qui otteniamo tre punti fissi: $(1, 0)$, $(0, 1)$, $(0, -1)$ che corrispodono ai valori 1, ∞ e 0.

Così, al prezzo di complicare orrendamente le formule, abbiamo capito da dove arriva il terzo punto fisso!

28.4 Vicino ai punti fissi

Torniamo ora nel più confortevole mondo della retta reale e cerchiamo di capire cosa succede "vicino" ad un punto fisso. In corrispondenza di quel punto, abbiamo detto, il grafico della funzione interseca la bisettrice; tutto dipenderà da "come" si intersecano.

Supponiamo, per ora, che la nostra funzione, a destra del punto fisso, cresca. Ci sono, allora, tre possibilità[42]: la retta tangente al grafico di $f(x)$ può essere più, meno o ugualmente inclinata rispetto alla bisettrice.

Se la tangente di $f(x)$ è meno inclinata della bisettrice, partendo da un punto sul grafico di $f(x)$ a destra del punto fisso e muovendoci in orizzontale, per trovare la bisettrice dovremo andare a sinistra; da qui, per trovare il grafico muovendoci in verticale, dovremo proseguire in basso.

Ci muoveremo dunque sempre a sinistra e in basso, restando compresi tra il grafico e la bisettrice e dunque tendendo al punto fisso.

Se invece la tangente è più inclinata, con lo stesso tipo di considerazione grafica è chiaro che andremo sempre verso destra e in alto, fintantoche la funzione resta sopra alla bisettrice.

[42]Ok, assumendo che la nostra funzione $f(x)$ sia abbastanza bella da avere una retta tangente nel punto...

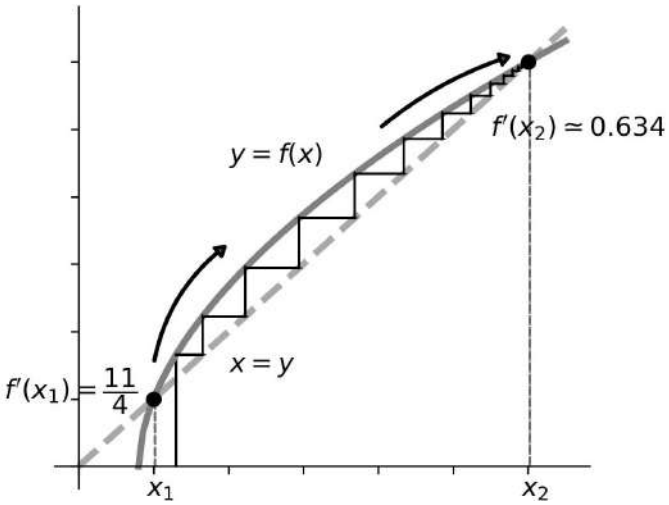


Figura 87: Funzione con punto fisso repulsivo (x_1) e attrattivo (x_2): pur partendo vicino a x_1, la successione tende a x_2.

Analoghe analisi si potrebbero condurre per i casi in cui la funzione è decrescente.

Per chi sa di derivate, possiamo riassumere il tutto e dimostrarlo, tenendo in mente che il coefficiente angolare della retta tangente al grafico di $f(x)$ in un punto è la derivata di $f(x)$ calcolata nel valore dell'ascissa di quel punto.

Consideriamo un punto fisso di $f(x)$, x_0 tale che $f(x_0) = x_0$; supponiamo che $f(x)$ sia derivabile con continuità "vicino"[43] a x_0.

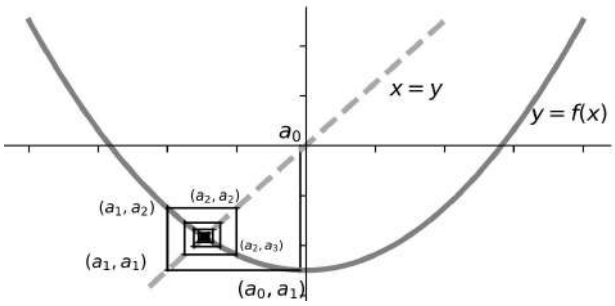


Figura 88: Punto fisso attrattivo con $f'(x) < 0$.

Se $|f'(x_0)| < 1$, allora vale $|f'(x)| \leq A < 1$ "vicino"[44] a x_0.
Ora ci viene in aiuto un classico teorema di analisi matematica.

Il *teorema di Lagrange* ci dice che, se una funzione $f(x)$ è continua su $[a, b]$ e derivabile su (a, b), allora esiste $\xi \in (a, b)$ tale che

$$f(b) - f(a) = f'(\xi)(b - a) \ .$$

Quindi, per il *teorema di Lagrange*, se x è abbastanza vicino a x_0, si ha

$$|f(x) - x_0| = |f(x) - f(x_0)| = |f'(c)(x - x_0)|$$

[43] cioè in un intervallo del tipo $(x_0 - \varepsilon, x_0 + \varepsilon)$
[44] cioè in un intervallo del tipo $(x_0 - \delta, x_0 + \delta)$

con c più vicino a x_0 di x e dunque $|f'(c)| \leq A$. Applicando questo ragionamento alla nostra successione per ricorrenza, se L è un punto fisso e a è il valore di partenza, abbastanza vicino ad L, si ha

$$|a_n - L| = |f(a_{n-1}) - f(L)| \leq A|a_{n-1} - L| \leq \ldots \leq A^n|a - L|$$

e poiché $A < 1$, avremo che tale valore tende a 0 al crescere di n, cioè a_n tende a L (poiché la loro distanza tende a 0).

Allo stesso modo, se $|f'(x_0)| > 1$, potremo dimostrare che la distanza $|a_n - L|$ aumenta con n.

Dunque, possiamo classificare i punti fissi in base alla derivata di $f(x)$:

1. se $|f'(x_0)| < 1$, x_0 attira a sé i punti vicini (è un punto fisso attrattivo)
2. se $|f'(x_0)| > 1$, x_0 respinge da sé i punti vicini (è un punto fisso repulsivo)
3. se $|f'(x_0)| = 1$, boh!

Nel caso noioso di $f(x) = x^2$, notiamo che $f'(0) = 0$ (e dunque 0 è un punto fisso attrattivo – anzi, super-attrattivo!) e $f'(1) = 2$ (e dunque 1 è un punto fisso repulsivo). Per capire il comportamento di ∞, dovremmo sporcarci un po' le mani con le proiezioni stereografiche: ad esempio, portare la retta sulla circonferenza con la proiezione stereografica di prima e poi usare una seconda proiezione stereografica, centrata in $(0, -1)$ stavolta, per far finire il punto $(0, 1)$ – che arrivava da ∞ nel punto 0; poi potremmo calcolare una nuova $f(x)$ seguendola attraverso queste trasformazioni e calcolarne la derivata in 0.[45]

28.5 Una famiglia di esempi

Per variare un poco, ma non troppo, consideriamo la funzione $f : \mathbb{R} \longrightarrow \mathbb{R}$ data da $f(x) = x^2 + c$.

Non tireremo più in ballo le proiezioni stereografiche, ma ∞ è un punto fisso attrattivo per ogni c. I suoi punti fissi reali sono dati dall'equazione $x^2 + c = x$, che ha soluzioni

$$x_\pm = \frac{1 \pm \sqrt{1 - 4c}}{2}$$

che sono reali se e solo se $c \leq 1/4$.

Inoltre, $f'(x) = 2x$ per ogni valore di c e dunque, nel caso $c < 1/4$, abbiamo che

$$f'(x_+) = 1 + \sqrt{1 - 4c} > 1 \qquad f'(x_-) = 1 - \sqrt{1 - 4c}\,.$$

Dunque, $|f'(x_-)| < 1$ se e solo se $\sqrt{1 - 4c} < 2$, ovvero se e solo se $1 - 4c < 4$ ovvero se $c > -3/4$. Riassumendo:

1. se $c > 1/4$, per ogni valore di partenza, iterando f si tende all'infinito
2. se $c = 1/4$ abbiamo un unico punto fisso reale di tipo boh (vale $f'(1/2) = 1$)
3. se $c \in (-3/4/, 1/4)$ abbiamo un punto fisso attrattivo e un punto fisso repulsivo
4. se $c = -3/4$ abbiamo un punto fisso repulsivo e un punto fisso boh
5. se $c < -3/4$ abbiamo due punti fissi repulsivi.

[45] Per chi voglia farsi il conto, la nuova $f(x)$ sarà sempre $f(x) = x^2$ e la sua derivata in 0 sarà sempre 0, dunque anche ∞ è un punto fisso super-attrattivo!

Nel caso (1) abbiamo già capito cosa succede e, nel caso (3), ci aspettiamo che succeda quel che succedeva, noiosamente, per $c = 0$: l'intervallo $(-x_+, x_+)$ verrà attratto da x_-, per $|a| > x_+$ tenderemo all'infinto, x_+ sarà fisso e $-x_+$ finirà su x_+ dopo un'iterazione.

Qualche esperimento grafico ci convince che il caso (2) si comporta come il caso (3): l'intervallo $(-1/2, 1/2]$ si contrae in $1/2$, il punto $-1/2$ diventa $1/2$ alla seconda iterazione e se $|a| > 1/2$ la successione va all'infinito.

Il risultato finale, nel caso $c = -3/4$, è ancora simile, dividendo la retta tra l'intervallo $(-3/2, 3/2)$, i suoi estremi e il suo complementare.

Prima però di lanciarci ad analizzare il caso (5), vediamo di formalizzare i primi 4. Consideriamo quindi la successione per ricorrenza

$$\begin{cases} a_n = a_{n-1}^2 + c \quad \text{se} \quad n \geq 1 \\ a_0 = a \end{cases}$$

28.6 Il caso $c > 1/4$

Si ha $a_{n+1} = a_n^2 + c > a_n$ (e ciò è equivalente al fatto che non vi siano punti fissi: la retta e la parabola non si intersecano!). Quindi la successione a_n è sempre crescente; dove può fermarsi? Beh, se esiste un numero L tale che a_n tende ad L, si dovrà anche avere che a_n^2 tenda a L^2 e dunque che $L = L^2 + c$, ma tale equazione non ha soluzioni! Dunque a_n cresce senza mai fermarsi e dunque arriva a ∞.

28.7 Il caso $0 \leq c < 1/4$

Come sopra, chiamiamo x_-, x_+ le due soluzioni di $x^2 - x + c = 0$; avremo che $0 \leq x_- < x_+$ e dunque $f(x)$ è crescente sull'intervallo (x_-, x_+). Di conseguenza, se $x_- \leq a_n < x_+$, allora $f(x_-) < f(a_n) < f(x_+)$, cioè $x_- < a_{n+1} < x_+$. Inoltre, su tale intervallo il grafico di $f(x)$ è sotto la bisettrice, dunque $f(x) < x$, ovvero $a_{n+1} < a_n$.

In questo caso, a_n decresce strettamente e deve rimanere tra x_- e x_+; non ha altre possibilità se non quella di avvicinarsi sempre più a x_-. Partendo invece nell'intervallo $[0, x_-)$, si avrà che per gli stessi motivi si deve rimanere entro tale intervallo, ma la successione sarà crescente e dunque come sopra a_n tenderà a x_-.

Per parità della funzione $f(x)$, partendo tra $-x_+$ e 0 si finirà, dopo uno step, tra 0 e x_+, potendo ripetere il ragionamento sopra. Dunque per $|a| < x_+$, la successione tende a x_-, per $|a| = x_+$, la successione si stabilizza su x_+.

Applicando poi l'idea del caso $c > 1/4$, si vede che per $|a| > x_+$ la successione tende all'infinito.

28.8 Il caso $-3/4 < c \leq 0$

Qui abbiamo che se $x_- < a_n < x_+$, allora $f(0) \leq a_{n+1} < x_+$ (poiché la funzione $x^2 + c$ è decrescente su $(x_-, 0)$ e crescente $(0, x_+)$). E si ha che $|c| \leq x_+$ se e solo se $-2 \leq c \leq 1/4$, dunque anche nel caso in esame. Qui però non possiamo sempre dire che partendo dall'intervallo (x_-, x_+) vi restiamo dentro; infatti, sappiamo solo che resteremo all'interno di $(-x_+, x_+)$. In tale intervallo, a sinistra di x_- otteniamo che $a_{n+1} > a_n$ e a destra di x_- otteniamo che $a_{n+1} < a_n$; si può dimostrare che, da un certo punto in poi, la nostra successione si dividerà in due, con i termini pari da una parte di x_- e i termini dispari dall'altra, entrambe le *sottosuccessioni* che tendono a x_-.

Come sopra, quindi, l'intervallo $(-x_+, x_+)$ si contrarrà su x_- e via dicendo.

28.9 Casi $c = 1/4$, $c = -3/4$

Lasciamo questi casi allo studente studioso, spoilerando il finale: succederà ancora che l'intervallo $(-x_+, x_+)$ si contrarrà su x_-, mentre al suo esterno tutto fuggirà verso l'infinito.

28.10 Hic sunt leones: $c < -3/4$

Dai conti che già abbiamo fatto, sappiamo che, se $-2 \le c \le 1/4$, allora $-x_+ \le c = f(0)$ e dunque, se la successione cade nell'intervallo $(-x_+, x_+)$, vi rimane.

Però, se $-2 \le c < -3/4$, non vi sono punti fissi attrattivi e dunque, non appena la successione si avvicina ad un punto fisso, ne viene respinta, aumentando la propria distanza da esso. Quindi è ragionevole supporre che non potrà mai raggiungere tali punti dai punti vicini... e dunque, che succede? Beh, ad esempio, per $c = -1$ ci possiamo accorgere che, se $a_0 = 0$, si ha $a_1 = -1$, $a_2 = 0$, $a_3 = -1$, etc e dunque c'è un *periodo* o *ciclo* (in questo caso, con periodo pari a 2: $0, -1$). Non è detto che il periodo si manifesti immediatamente, ad esempio per $a = \sqrt{1+\sqrt{2}}$ abbiamo la seguente successione

$$\sqrt{1+\sqrt{2}},\ \sqrt{2},\ 1,\ 0,\ -1,\ 0,\ -1, \dots$$

e dunque c'è un antiperiodo prima del periodo vero e proprio. Notiamo però che i punti da cui può iniziare un periodo lungo k sono tra le soluzioni di

$$f(f(\cdots(f(x))\cdots)) = x$$

dove a sinistra f è composta con se stessa k volte; questa equazione è un polinomio di grado 2^k e dunque questo è il massimo numero possibile di cicli. A loro volta, le partenze degli antiperiodi prima di un dato periodo sono soluzioni di un'equazione polinomiale di grado 2^h (dove h è la lunghezza dell'antiperiodo) e dunque ancora sono al più 2^h. Quindi i punti di periodi e antiperiodi sono al più tanti quanti i numeri razionali, cioè non possono comporre tutto l'intervallo $(-x_+, x_+)$. Per capire cosa succede agli altri punti, almeno nel caso $c = -1$, studiamo un attimo la funzione $g(x) = f(f(x))$, per la quale $0, -1$ sono punti fissi: si ha $g(x) = x^2(x^2-2)$ e $g'(x) = 4x(x^2-1)$, dunque $0, -1$ sono punti fissi (super-)attrattivi. Il che vuol dire che, se guardiamo solo le iterate pari o le iterate dispari, esse tenderanno a uno e all'altro, mentre la successione, in totale, oscillerà tra l'uno e l'altro. Ad esempio, partendo da $a = 0.1$ otteniamo questi primi valori:

$$0.1,\ -0.99,\ -0.01\dots,\ -0,9996\dots,\ -0.0007\dots,\ -0.9999993\dots,\ -0.000001\dots$$

e come si vede, si alternano due successioni, una che converge a 0 ed una che converge a -1. Da notare che, invece, la funzione $f(f(f(x)))$ non ha punti fissi a parte x_- e x_+ e dunque non vi sono cicli periodici di lunghezza 3.

Non è nemmeno detto che succeda sempre questo: consideriamo $c = -2$. Ci sono due punti fissi, $x_- = -1$ e $x_+ = 2$, e ci sono due punti di periodo due

$$\beta_- = (-1-\sqrt{5})/2$$
$$\beta_+ = (-1+\sqrt{5}))/2$$

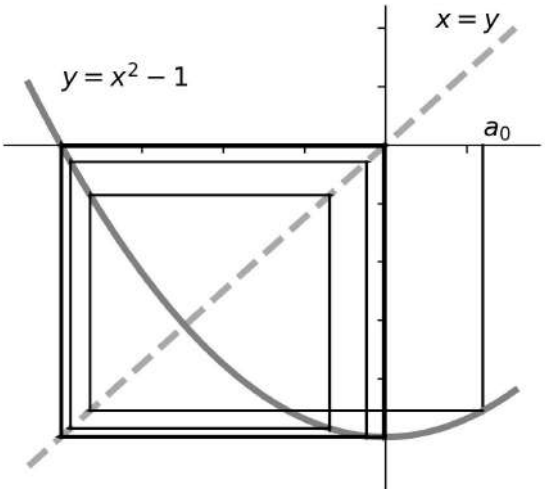


Figura 89: Orbita che si avvicina al ciclo $0, -1$.

ottenuti come punti fissi di $g(x) = f(f(x)) = x^4 - 4x^2 + 2$.

Tali punti fissi sono però repulsivi (la derivata di g in $\beta_\pm$ è -4) e dunque non avremo la situazione di prima in cui i valori della successione "convergono al ciclo periodico". Ad esempio, per $a = \beta_- + 0.01$ otteniamo i seguenti primi valori (troncati alle prime due cifre dopo la virgola)

$$-1.60, 0.58, -1.66, 0.74, -1.44, 0.09, -1.99, 1.97, 1.88, \ldots$$

e come si vede non sembra esserci nessuna intenzione di avvicinarsi al ciclo β_-, β_+ (-1.62, 0.62 le prime cifre dopo la virgola). Inoltre, sempre per $c = -2$, possiamo trovare cicli arbitrariamente lunghi: per $k \geq 3$, dal valore

$$2 - 4\sin^2\left(\frac{2\pi}{2^k - 1}\right)$$

parte un periodo lungo k, fatto di punti fissi repulsivi per la funzione ottenuta componendo f con se stessa k volte.

Dunque, partendo vicino a uno di questi valori, si osserverà come prima un comportamento *caotico*, in tutto l'intervallo $(-2, 2)$ (da cui però ancora una successione non può uscire).

Infine, vediamo brevemente che per $c = -6$ otterremo ad esempio due punti fissi repulsivi ($x_- = -2$ e $x_+ = 3$), un periodo lungo 2 (fatto dai punti $(-1 \pm \sqrt{21})/2$, anch'essi repulsivi), 6 punti periodici di periodo 3 (e quindi 2 cicli lunghi 3) di cui già non è possibile trovare una espressione esatta. Vi sono anche cicli di lunghezza maggiore, altrettanto impossibili da scrivere esplicitamente. Osserviamo però che, ad esempio, se $|a_n| < \sqrt{3}$, otterremo $|a_{n+1}| > 3 = x_+$, e dunque, iterando, avremo che a_n tende all'infinito, per tutti i valori di partenza in $(\sqrt{3}, \sqrt{3})$. In questo caso, succede una cosa molto particolare: proviamo a trovare i punti che *non vanno all'infinito*. Notiamo intanto che se $|a_n| > 3$, allora $|a_{n+1}| = a_n^2 - 6 > 3$ quindi anche in questo caso, se si parte con $|a| > 3$, si tende a infinito. Dunque, una successione che non va all'infinito deve sempre restare in $[-3, 3]$, ma se $a_1 \in [-3, 3]$, si può calcolare facilmente che a_0 poteva stare solo in $[-3, -\sqrt{3}] \cup [\sqrt{3}, 3]$; se poi a_2 sta in $[-3, 3]$, allora a_1 doveva stare in $[-3, -\sqrt{3}] \cup [\sqrt{3}, 3]$ e dunque a_0 doveva stare in

$$\left[-3, -\sqrt{6+\sqrt{3}}\right] \cup \left[-\sqrt{6-\sqrt{3}}\right] \cup \left[\sqrt{3}, \sqrt{6-\sqrt{6}}\right] \cup \left[\sqrt{6+\sqrt{3}}, 3\right] .$$

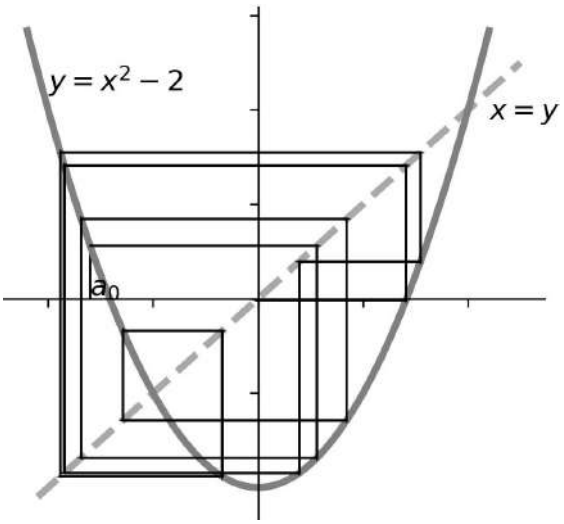


Figura 90: Orbita caotica da $a_0 = -1.6$.

Figura 91: Primi passi della costruzione dell'insieme dei punti che non vanno all'infinito, per $c = -6$, cioè di un insieme di Cantor.

Provando a proseguire, notiamo la costruzione è la seguente: partiamo dall'intervallo $I = [-3, 3]$ e togliamo un pezzo centrale, lungo $1/\sqrt{3}$ volte I; il risultato sono due intervalli e ad ognuno di essi togliamo un segmento centrale di lunghezza $1/\sqrt{3}$ volte la loro lunghezza, ottenendo 4 intervalli; ad ognuno di essi ancora togliamo la stessa cosa. Il *limite* di questo procedimento è un insieme, detto *insieme di Cantor*[46], che non contiene nessun intervallo, ma non è vuoto. È l'insieme dei punti che non vanno all'infinito per $c = -6$!

Insomma, per $c < -3/4$, succedono davvero cose strane...

28.11 Vera complessità

Ma le vere cose strane succedono quando si abbandona quell'innaturale costrizione che sono i numeri reali.

Consideriamo l'insieme $\mathbb{C}$ dei numeri complessi e definiamo la successione

$$\begin{cases} z_n = z_{n-1}^2 + (c_0 + ic_1) & \text{se} \quad n \geq 1 \\ z_0 = a_0 + ia_1 \end{cases}$$

[46]Quello originale, proprio di Cantor, sarebbe quello che ogni volta toglie un intervallo centrale lungo 1/3 di quello prima, ma per ottenerlo avremmo dovuto guardare a $c = -45/16$...

Ad ogni numero complesso z possiamo associare un punto del piano (x, y) di modo che $z = x + iy$, dove x e y sono numeri reali; se abbiamo due numeri complessi, $z_1 = x_1 + iy_1$ e $z_2 = x_2 + iy_2$, le operazioni di somma e prodotto hanno questo aspetto:

$$z_1 + z_2 = (x_1 + x_2) + i(y_1 + y_2) \qquad z_1 z_2 = (x_1 x_2 - y_1 y_2) + i(x_1 y_2 + x_2 y_1)$$

Allora, l'operazione di elevamento al quadrato sembra già più minacciosa qui, infatti $z^2 = (x^2 - y^2) + i(2xy)$ e dunque stiamo di fatto studiando la trasformazione $(x, y) \mapsto (x^2 - y^2, 2xy)$.

Notiamo però una cosa: se $z = r\cos\alpha + ir\sin\alpha$, per $r \in \mathbb{R}$, $r > 0$ e α un angolo (quindi inteso a meno di multipli di 2π) allora

$$z^2 = r^2(\cos^2\alpha - \sin^2\alpha) + ir^2 2\sin\alpha\cos\alpha = r^2\cos(2\alpha) + ir^2\sin(2\alpha)\ .$$

Ad esempio, partendo da $1 + i$ ed iterando la funzione $f(z) = z^2$, otteniamo

$$1 + i \mapsto (1 + i)^2 = 2i \mapsto -4 \mapsto 16 \mapsto \dots$$

oppure, partendo da $0.1 + 0.2i$, otteniamo

$$0.1 + 0.2i \mapsto -0.03 + 0.04i \mapsto -0.0007 - 0.0024i \mapsto \dots\ .$$

Anche qui, possiamo pensare (formalmente allo stesso modo di $\mathbb{R}$, quindi con le stesse formule), di "chiudere" $\mathbb{C}$ aggiungendo un punto all'infinito (∞) e ottenendo una sfera, di solito chiamata $\mathbb{CP}^1$, tramite la proiezione stereografica. Dunque, ha senso dire che una successione "tende all'infinito" senza aggiungere altro, quando il *modulo*[47] di z_n tende all'infinito come numero reale positivo.

Per comodità, continueremo a scrivere $a = a_0 + ia_1$ per il punto iniziale e $c = c_0 + ic_1$ per il parametro.

28.12 Il caso noioso

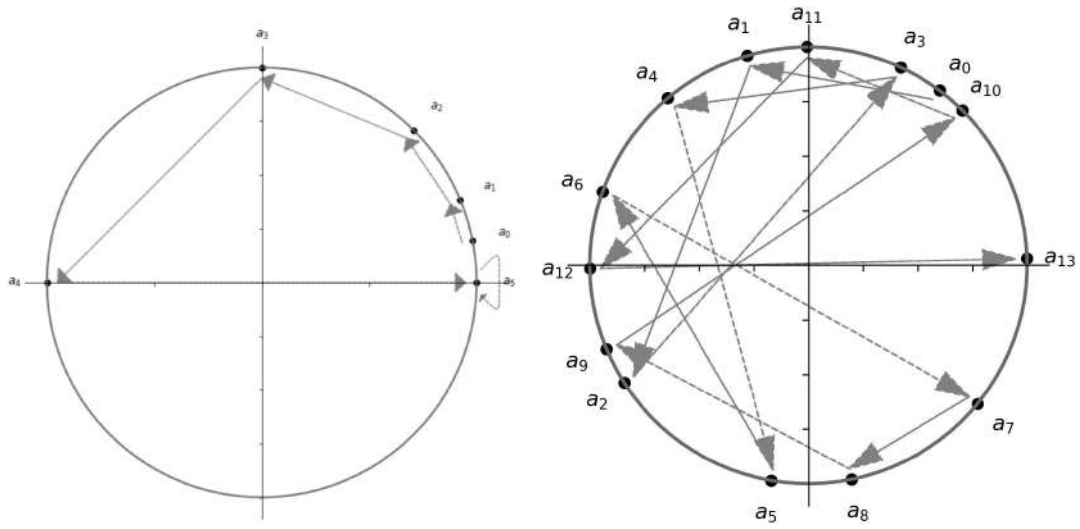
Se poniamo $c = 0$, stiamo, come all'inizio, facendo i quadrati successivi; guardando i moduli, possiamo ragionare come all'inizio: se $|a| < 1$, la successione z_n tende a 0 (perché il suo modulo tende a 0), se invece $|a| > 1$ la successione tende a ∞. Se però $|a| = 1$, allora $|z_n| = 1$ per ogni n... ancora avremo che $a = 1$ è un punto fisso e che $a = -1$ va su un punto fisso, ma, ad esempio, avremo dati iniziali che ci mettono quanto vogliamo ad arrivare su un punto fisso:

$$a = \cos\left(\frac{2\pi}{2^k}\right) + i\sin\left(\frac{2\pi}{2^k}\right)$$

si comporta così: ogni volta che eleviamo al quadrato, otteniamo un numero della stessa forma dove però c'è $k - 1$ al posto di k; in k passi si arriva a $k = 0$, che è 1. E poi ci sono un sacco di punti il cui quadrato continuerà ad avere modulo 1, ma non diventeranno mai 1. Ad esempio, provate a partire da $(3/5) + i(4/5)$ e vedete cosa succede...

La situazione è dunque la seguente: fuori dalla circonferenza di raggio 1, tutti i punti vanno verso ∞, dentro la circonferenza di raggio 1, tutti i punti vanno verso 0, sulla circonferenza i punti si muovono in modo arbitrariamente complicato (caotico!) senza, in generale, andare da alcuna parte.

[47] $|z| = \sqrt{x^2 + y^2}$, cioè la distanza di (x, y) dall'origine



(a) $\cos(\pi/16) + i\sin(\pi/16)$ **(b)** $(3/5) + i(4/5)$

Figura 92: Orbite di punti della circonferenza unitaria tramite $z \mapsto z^2$.

Una regione *connessa*[48] di piano in cui se la successione che parte da un punto ha un certo comportamento[49], le successioni che partono dai punti vicini si comportano allo stesso modo [50], si dice *componente di Fatou* (della successione o, più semplicemente in questo caso, della funzione $f(z)$ usata per creare la successione); l'insieme complementare dell'unione tutte le componenti di Fatou, cioè l'insieme in cui c'è un comportamento "non uniforme" delle successioni rispetto ai punti di partenza[51], si dice *insieme di Julia*.

Nel caso (noioso?) di $f(z) = z^2$, abbiamo due componenti di Fatou, l'esterno e l'interno della circonferenza unitaria, che è il nostro insieme di Julia. Notate che (in questo e in ogni altro caso), l'insieme di Julia è il "bordo comune" delle componenti di Fatou; il che ha senso: per passare da una zona dove vige un certo comportamento al limite ad una dove ne vige uno diverso, bisogna attraversare una transizione "caotica".

Va infine detto che spesso si considera un terzo insieme, il cosiddetto *insieme di Julia pieno*, cioè il complementare della componente di Fatou che è attratta dall'infinito, cioè l'insieme di tutti i punti per i quali la successione resta limitata, non tende all'infinto. Nel nostro caso, sarebbe il cerchio di raggio 1.

28.13 Art attack!

Un ovvio vantaggio del lavorare sui numeri complessi è il poter fare disegni nel piano e non sulla retta. Ecco dunque i disegni di alcuni insiemi di Julia pieni, per i valori di c reali che

[48] cioè fatta di un pezzo solo

[49] ad esempio: tende ad un dato valore, tende a un ciclo periodico,...

[50] tendono allo stesso valore o allo stesso ciclo periodico, ...

[51] Spesso si utilizza l'aggettivo *caotico*, riferendosi al fatto che il caos è una grande variazione di risultati dovuta a una piccola variazione di situazioni iniziali

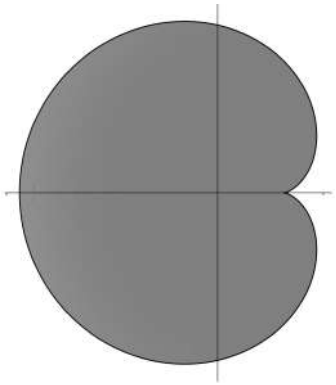


Figura 93: Una cardioide.

avevamo studiato prima. Riuscite a ritrovare alcune caratteristiche (sui reali) delle successioni associate?

In ognuno di questi disegni, l'insieme di Julia vero e proprio è il "bordo" della zona nera (che, ad esempio, per $c = -2, -6$ è tutta la zona nera), mentre l'interno è fatto da componenti di Fatou in cui i punti non tendono all'infinito, ma a un limite finito (un punto fisso attrattivo).

Poiché su $\mathbb{C}$ l'algebra funziona quanto su $\mathbb{R}$ (anzi, meglio![52]), per ogni valore di c possiamo calcolare i punti fissi z_+, z_- risolvendo $z^2 + c = z$ e possiamo determinare se sono attrattivi o repulsivi calcolando $|f'(z_\pm)|$, dove $f'(z) = 2z$. Quindi, a differenza di prima, avremo sempre due punti fissi (tranne che per $c = 1/4$, dove avremo due radici coincidenti), ma come prima avremo un punto fisso attrattivo solo se $|2z_-| < 1$ o $|2z_+| < 1$. Avere un punto fisso attrattivo corrisponde ad avere una componente di Fatou in cui i punti non vanno all'infinito e dunque corrisponde ad avere un "interno" per l'insieme di Julia pieno.

Come detto, $z_\pm = (1 \pm \sqrt{1-4c})/2$; sia w la radice quadrata di $1-4c$ con parte reale maggiore o uguale a 0. Allora z_- sarà il punto con derivata in modulo più piccola. Tale derivata sarà $1 - w$. L'insieme delle c per cui c'è un punto fisso attrattivo è perciò

$$\{c \ : \ |1-w| < 1\} = \{c = -(w^2-1)/4 \ : \ |1-w| < 1\};$$

in tale insieme, w varia in un cerchio di centro 1 (cioè $(1, 0)$ nel piano) e raggio 1 e la regione che ci interessa è quella ottenuta calcolando, per ogni tale w, il numero complesso $-(w^2-1)/4$. Il risultato è la curva di Figura 93, detta *cardioide*.

Abbiamo visto prima, però, che ci possono essere fenomeni più complicati, ad esempio un ciclo periodico attrattivo; supponiamo di avere un ciclo di lunghezza 2, fatto da $z_\circ$ e $z_\dagger$: $f(z_\circ) = z_\dagger$ e $f(z_\dagger) = z_\circ$, allora $z_\circ$ e $z_\dagger$ sono attrattivi per $g(z) = f(f(z))$ se $g'(z_\circ) = f'(f(z_\circ))f'(z_\circ) = f'(z_\dagger)f'(z_\circ) = 4z_\circ z_\dagger = g'(z_\dagger)$ ha modulo minore di 1.

I punti di periodo 2 sono le soluzioni di $f(f(z)) = z$, cioè di $z^4 + 2cz^2 - z + c^2 + c = 0$ che si fattorizza come $(z^2 + c - z)(z^2 + z + 1 + c) = 0$; le soluzioni del primo fattore sono i punti fissi e ci interessano quelle del secondo. Il loro prodotto è $1 + c$ e dunque formeranno un ciclo attrattivo se $|1 + c| < 1/4$. Questa condizione descrive un cerchio di centro $(-1, 0)$ e raggio $1/4$. Partendo da tale cerchio, otterremo che non ci sono punti con un limite finito, ma ci sono punti attratti da questo ciclo di lungehzza 2.

Ad esempio, se $c = 1$ e $|z_0| = |a| < 1/3$, allora $|z_1 + 1| = |a^2| < 1/9$ (e anche $|z_1| < 10/9$) e dunque $|z_2| = |z_1^2 - 1| < |z_1 - 1| \cdot |z_1 + 1| < 10/81 < 1/3$. Quindi se parto abbastanza vicino a

[52] Su $\mathbb{C}$, ogni polinomio di grado n ha esattamente n radici, se contate con molteplicità!

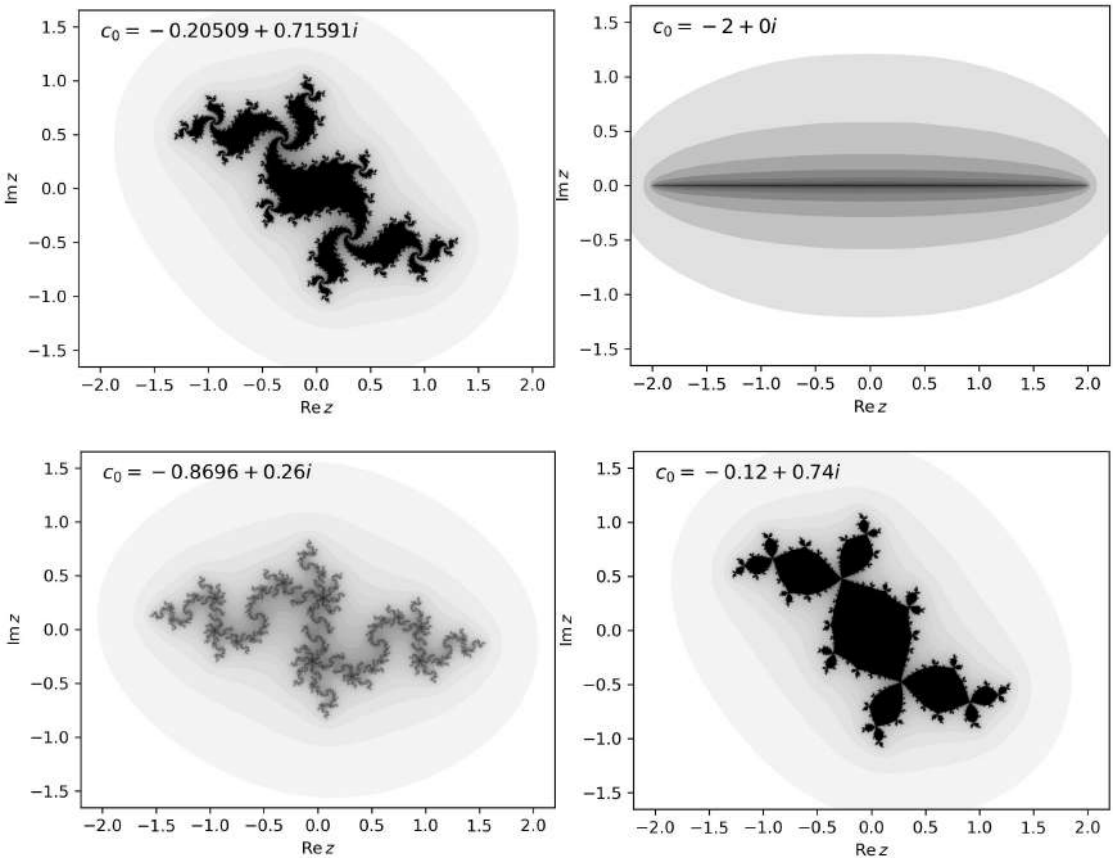


Figura 94: Insiemi di Julia

0, vengo attratto da 0 ai passi pari e da -1 ai passi dispari. Dunque, c'è una componente di Fatou che contiene 0 e c'è una componente di Fatou che contiene -11. Ve ne sono altre, come si può capire dal disegno per $c = 1$, anche se nessuna contiene un punto attrattivo, ma solo cicli periodici attrattivi.[53]

Se c è tale che non vi siano punti, né cicli attrattivi, possiamo ancora avere delle componenti di Fatou in cui i punti non vanno all'infinito, ad esempio succede per $c = 1/4$, dove l'unico punto fisso è neutro $f'(1/2) = 1$. In questo caso, il puunto fisso è nell'insieme di Julia, in quanto, da un lato i punti scappano all'infinito, dall'altro tendono verso di lui.

Inoltre, possiamo avere componenti di Fatou ancora più strane, in cui i punti "girano" senza mai ripassare da dove sono già stati, ma rimanendo dentro una certa regione.

Infine, abbiamo valori, come $c = i$ o $c = -2$ per cui l'insieme di Julia pieno non ha un interno e coincide con Julia, ma è ancora un pezzo unico (seppure non si riesca a dire "una linea", vista la sua natura poco disegnabile[54]), ma se prendiamo valori ancora più lontani da 0, come $c = -6$, troviamo solo *polvere*, insiemi che non contengono alcun intervallo e sembrano fatti di infiniti punti staccati.

28.14 Matrioske

Notiamo anche che, in molti casi, gli insiemi di Julia hanno una natura frattale, cioè contengon parti che sono una loro versione rimpicciolita.

Intuitivamente, questo è dovuto al fatto che la funzione $f(z) = z^2 + c$ è *conforme*; questo vuol dire che preserva gli angoli tra curve (tranne che in 0, dove $f'(z) = 0$). Questo può essere verificato solo per $c = 0$, visto che modificare c equivale a comporre con una traslazione nel piano complesso, che non modifica alcun angolo.

Consideriamo la funzione $f(z) = z^2$ e, a titolo di esempio, prendiamo una retta per l'origine e una circonferenza centrata nell'origine, che sono due curve ortogonali; una retta per l'origine si scirve come $\{tz_0 \ : \ t \in \mathbb{R}\}$ e dunque i quadrati dei suoi punti sono gli elementi di $\{t^2 z_1^2 \ : \ t \in \mathbb{R}\}$, cioè una semiretta. Una circonferenza centrata nell'origine di raggio r, tramite $f(z) = z^2$ viene mandata in una circonferenza centrata nell'origine di raggio r^2. Dunque, rimangono ortogonali. Questa non è una dimostrazione, ovviamente, ma cerca di farvi capire in che senso la mappa $f(z)$ "conserva gli angoli".

28.15 Keeping it all together

Sarebbe ottimo avere una sorta di mappa del piano complesso che descrive come cambiano gli insiemi di Julia a partire dal c che scegliamo.

Questa mappa è l'*insieme di Mandelbrot*. Tutto parte da un teorema dovuto a Julia e Fatou (eh sì, sono esistiti davvero): l'insieme di Julia di $f(z) = z^2 + c$ è connesso se e solo se 0 appartiene all'insieme di Julia pieno, cioè se e solo se la successione che parte da 0 è limitata. E dunque definiamo l'insieme di Mandelbrot proprio come l'insieme dei c per cui succede questo.

Una prima cosa da notare è che, come per i reali, possiamo subito dimostrare che se $|c| = 2 + h$ con $h > 0$, abbiamo che $|c^2 + c| \geq |c^2| - |c| = (4 + 4h + h^2) - (2 + h) = 2 + 3h + h^2 \geq 2 + 3h$; se

[53] Tutte le "zone piene" dell'insieme di Julia pieno vanno, dopo una iterazione, sulla componente di Fatou di 0 o su quella di -1 e tutti gli altri punti periodici sono sul bordo, cioè nell'insieme di Julita, poiché spostandosi poco da loro si cambia drasticamente comportamento, andando all'infinito o andando verso il ciclo attrattivo lungo 2.

[54] Gli insiemi di Julia come per $c = i$ sono detti *dendriti*

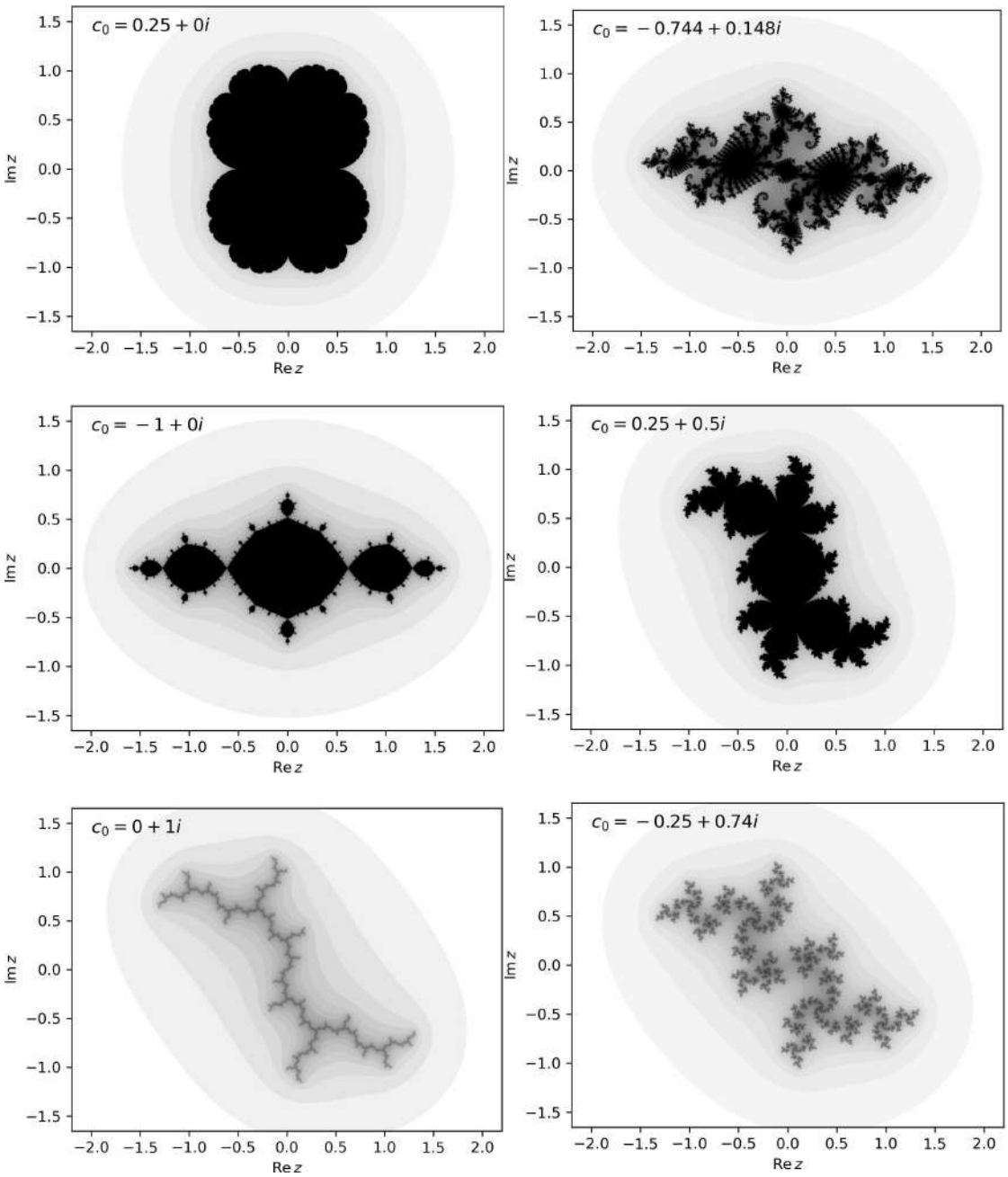


Figura 95: Insiemi di Julia - parte 2

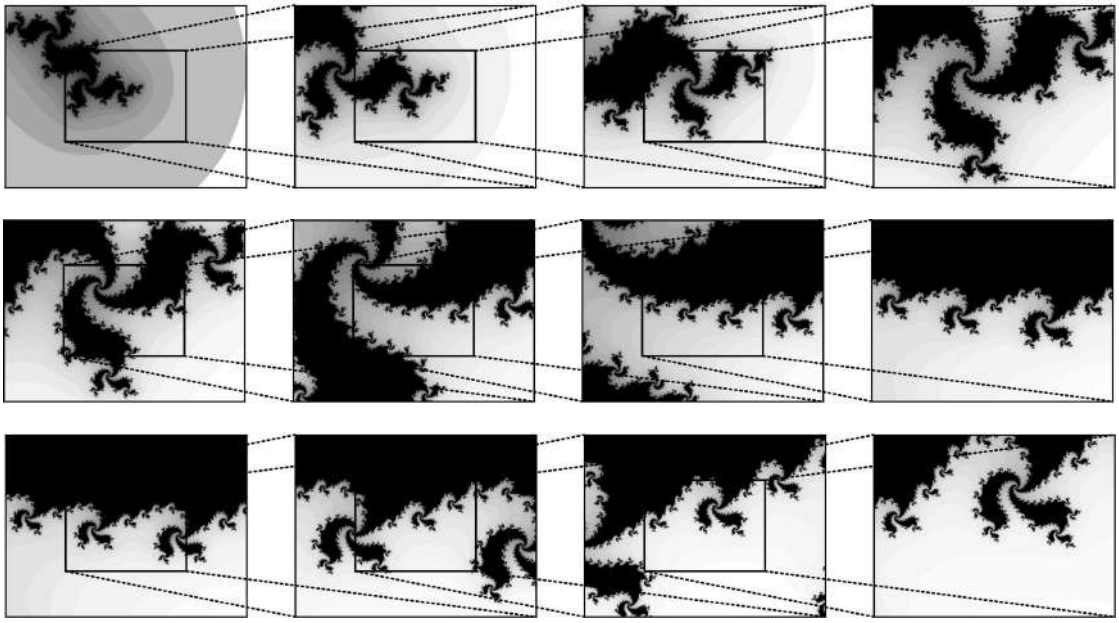


Figura 96: Ingradimenti successivi dell'insieme di Julia per $c = 0.205 + 0.716i$, centrati nel punto $1 - 0.67i$.

poi $|z_n| = 2 + mh$, allora $|z_{n+1}| \geq |z_n|^2 - |c|^2 = (2 + mh)^2 - (2 + h) \geq 2 + 2(m - 1)h$ e da qui è chiaro che se $|c| > 2$, allora c non sta nell'insieme di Mandelbrot, che dunque è contenuto nel disco di raggio 2.

Noterete poi che tale insieme contiene la cardioide delle c per cui c'è un punto fisso attrattivo e il disco delle c per c'è un ciclo di lunghezza 2 attrattivo. Gli altri pezzi "pieni" del Mandelbrot corrispodono a insiemi del parametro c per cui vi sono cicli attrattivi di lunghezza maggiore. Ad esempio i due pallini sopra e sotto la cardioide corrispondono a valori di c per cui c'è un ciclo attrattivo di lungehzza 3 (e questo dice che non vi sono valori reali per cui questo succede). Vi sono poi parti di tale insieme che non ci azzarderemmo a chiamare "piene", ad esempio tutti i tentacoli che si protendono verso l'alto o verso sinistra sull'asse reale; su di essi si trovano valori come $c = i$ o $c = -2$, per cui l'insieme di Julia pieno non ha un interno. Fuori dal Mandelbrot, infine, troviamo quei valori per cui l'insieme di Julia si disgrega in insiemi di Cantor, come $c = 6$. Molte sono le proprietà curiose dell'insieme di Mandelbrot: è fatto di un pezzo solo (nonostante i "filamenti" che si propagano dai bulbi), è frattale vicino a certi punti, il suo bordo è, in un certo senso, 2-dimensionale. Inoltre, non è ovvio che ci si debba limitare alle mappe quadratiche, si possono creare insiemi di Mandelbrot per altre famiglie dipendenti da un parametro, ad esempio $x^3 + c$ o cose più fantasiose.

Riferimenti bibliografici

[1] `https://e.math.cornell.edu/people/belk/dynamicalsystems/NotesJuliaMandelbrot.pdf`

[2] `https://complex-analysis.com/content/julia_set.html`

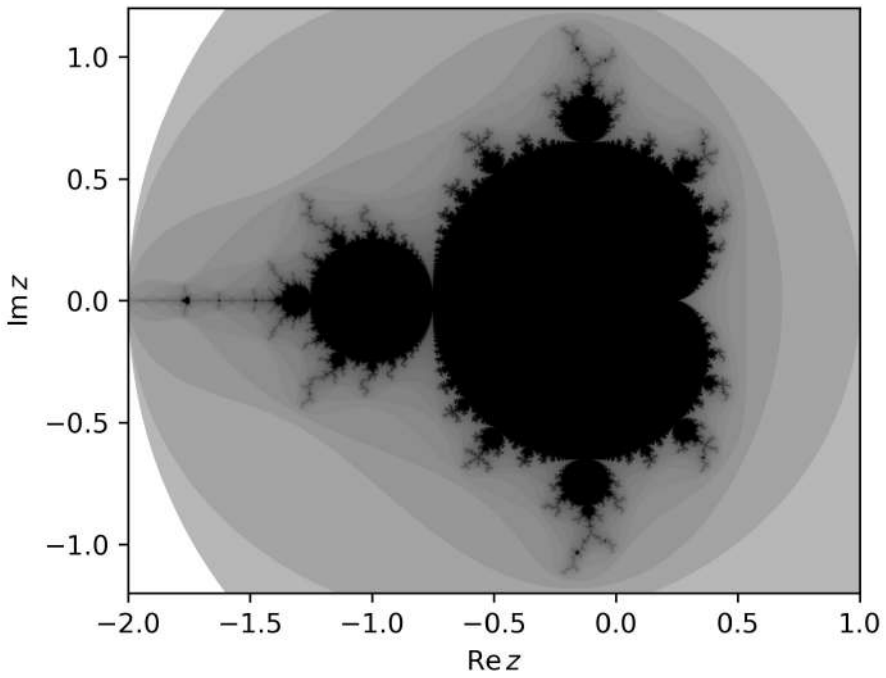


Figura 97: In nero, lo stupendo insieme di Mandelbrot

[3] https://images.math.cnrs.fr/L-ensemble-de-Mandelbrot/

[4] https://www.dynamicmath.xyz/mandelbrot-julia/

29 La matematica dietro le immagini: la mappa di Arnold

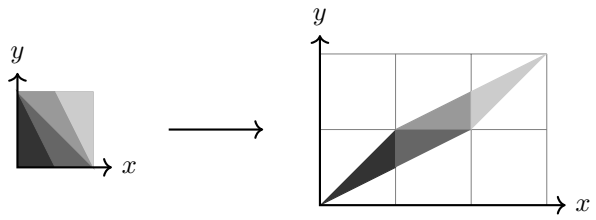
Chiara Gambicchia, n.17, Ottobre 2023

Tra i sistemi dinamici sul piano, una mappa piuttosto nota è la mappa di Arnold, dal nome del matematico Vladimir Arnold che negli anni sessanta ne studiò le proprietà per la prima volta. Questa stessa mappa è nota a molti come *Arnold's cat map*, "la funzione del gatto di Arnold", in quanto Arnold credeva molto nel potere delle immagini come strumento esplicativo, quindi egli stesso utilizzò l'immagine di un gatto per spiegare ai propri studenti l'azione della mappa.

Partiamo dalla definizione. Consideriamo la funzione f sul piano definita come segue:

$$\begin{array}{ccccc} f & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & & \begin{pmatrix} x \\ y \end{pmatrix} & \longmapsto & \begin{pmatrix} 2x+y \\ x+y \end{pmatrix} \end{array}$$

che manda il punto di coordinate (x, y) nel punto $(2x + y, x + y)$. Questo primo step si può visualizzare come nel disegno sottostante.



Osserviamo subito che valgono alcune proprietà:

1. tutti i punti che hanno coordinate intere vengono mappati in altri punti a coordinate intere: infatti sommando due numeri interi si ottiene ancora un intero e la stessa cosa vale se moltiplichiamo per 2 un numero intero;
2. la mappa preserva l'area: questo significa che, scegliendo una qualsiasi porzione di piano (che possiamo chiamare S) e misurandone l'area, l'area della parte di piano che corrisponde all'immagine di S tramite f è uguale all'area di S;
3. la mappa è bigettiva, cioè è sia iniettiva che suriettiva: quindi si può definire la funzione f^{-1}, che "riporta indietro" i punti alla loro posizione iniziale;
4. anche l'inversa f^{-1} gode delle stesse proprietà elencate sopra (come esercizio, provate a capire perchè!).

Un'altra cosa carina da notare è che ci sono due direzioni "privilegiate", chiamate *espandente* e *contraente*. Che vuol dire? Vuol dire che i punti sulla retta in direzione espandente vengono allontanati dall'origine degli assi, mentre quelli sulla direzione contraente vengono avvicinati all'origine. L'effetto è che gli insiemi vengono deformati allungandosi lungo la direzione espandente e accorciandosi lungo la direzione contraente.

Proviamo a rendere più visibile questa cosa. Immaginate di avere un fil di ferro chiuso a quadrato (come se fosse il bordo del quadrato disegnato sopra): ora prendete due vertici

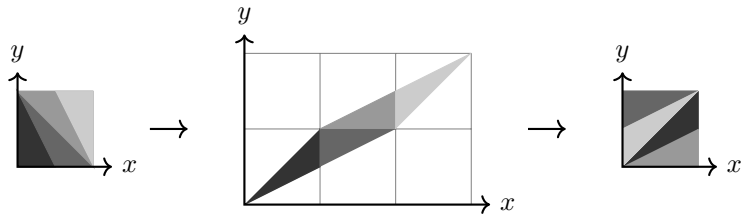
opposti e tirate fino a ottenere il parallelogramma che si vede nel secondo disegnino; come potete immaginare l'altra diagonale si accorcerà.

Continuiamo a comporre la nostra mappa. Grazie alle proprietà elencate sopra possiamo *passare la mappa f al quoziente*. Spieghiamo subito cosa significa questa cosa. Immaginate di disegnare una griglia sul piano, fatta di quadrati di lato 1, e scegliamo un quadrato principale (tipo quello che ha vertice nell'origine nel primo quadrante); applichiamo la nostra funzione f e poi tutto ciò che finisce fuori dal quadrato principale viene spostato nel quadrato principale senza cambiare posizione all'interno del singolo quadrato, come se ritagliassimo tutti i quadrati della griglia e li sovrapponessimo l'uno all'altro tutti sul quadrato principale. Questa cosa che abbiamo fatto, a livello delle coordinate, si dice *operare modulo 1*; cioè delle coordinate non ci interessa tutto quanto il numero, ma solo la parte decimale che si trova dopo la virgola.

Scritta formalmente la mappa risulta come segue

$$\begin{array}{rccc} \Gamma \quad : & [0,1]\times[0,1] & \longrightarrow & [0,1]\times[0,1] \\ & \begin{pmatrix} x \\ y \end{pmatrix} & \longmapsto & \begin{pmatrix} 2x+y \\ x+y \end{pmatrix} \quad \text{mod } 1 \end{array}$$

L'effetto finale della mappa si può rappresentare così.



La cosa interessante a questo punto avviene quando ripetiamo queste operazioni più volte iterando la mappa. Nella figuara sottostante per esempio si vede cosa succede a una figura dopo sole 6 applicazioni della mappa: del gatto non si vede nemmeno l'ombra, che caos!

C'è un termine matematico che descrive questo comportamento ed è "altamente mixing".

L'idea è quella di definire un parametro da studiare per capire quanto sia *ben sparpagliato* un insieme all'interno di un quadrato.

Facciamo un esempio per dirlo in termini più concreti. Immaginiamo un bicchiere pieno per metà d'acqua e per metà di olio; i due liquidi sono immiscibili, tuttavia possiamo "scomporre" l'olio in goccioline piccole a piacere in modo tale che ad occhio non si distingua l'olio dall'acqua. In questo caso possiamo dire che quanto più sono piccole queste goccioline, tanto più l'olio è *ben sparpagliato* nell'acqua.

Lo stesso ragionamento potremmo farlo con il bianco e il nero al posto dell'olio e dell'acqua: quanto più si vede grigio, tanto più si può dire che bianco e nero siano *ben sparpagliati* l'uno nell'altro.

Quest'idea intuitiva è espressa in termini analitici e geometrici da un parametro che si chiama *scala di mixing*.

Si può dimostrare che la mappa di Arnold che abbiamo analizzato ha la proprietà di essere *altamente mixing*, cioè un qualsiasi insieme nel quadrato a ogni iterazione della mappa viene "scomposto e sparpagliato" sempre di più, senza la possibilità di tornare alla disposizione di partenza.

Per capire meglio cosa voglia dire per questa mappa essere *altamente mixing*, provate a immaginare un disegno qualsiasi dentro il quadrato iniziale: stiratelo una prima volta nel

parallelogramma, poi ritagliate le parti che si trovano fuori dal quadrato principale e incollatele come abbiamo detto sopra. Già così l'immagine che da cui siete partiti è stata deformata abbastanza, ma ancora si intuisce qualcosa; ma provate a rifare la stessa cosa – stirate, tagliate e incollate – più volte: nel giro di poche iterazioni non si capisce più nulla!

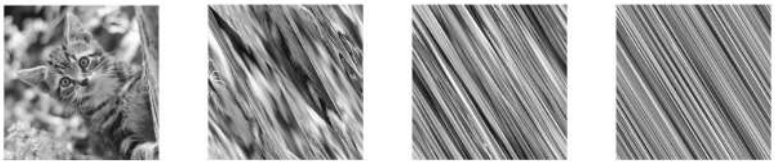


Figura 98: Rappresentazione di un'immagine a cui viene applicata la mappa 3 volte

Quindi per vedere questo comportamento *mixing* della mappa basta un computer che applichi la mappa infinite (o, più realisticamente, moltissime...) volte ad una foto? La risposta purtroppo è no. C'è un piccolo problema che impedisce al computer di riprodurre il comportamento della mappa all'infinito. Provate a fare uno scan dei QR code qui sotto, che mostrano la mappa applicata più volte a immagini di varie dimensioni.

Figura 99: Se per qualsiasi motivo avete difficoltà con i qr code, di seguito i link a cui rimandano:
`https://youtube.com/shorts/nofK0XjAo4E?feature=share`
`https://youtube.com/shorts/ckIgQQFdeUE?feature=share`
`https://youtu.be/3xnXUyXXwuI`

Fatto? Avete notato una cosa inaspettata (almeno rispetto a quel che abbiamo detto finora)?

Esatto! Ad un certo punto torna sempre l'immagine di partenza!

Per capire perchè si verifichi questo, è necessario formalizzare in termini matematici cosa sia una foto! Possiamo pensare a un'immagine digitale come se fosse una griglia di quadratini colorati, cioè i pixel, e deformarla vuol dire spostare e mischiarne i pixel.

Chiaramente la definizione della mappa va leggermente modificata per essere adattata al mondo digitale, in quanto stiamo parlando di un ambiente *discreto* e non *continuo* come quello del piano; cioè, mentre nel piano abbiamo un'infinità di punti, tutti "ammucchiati" dappertutto, nelle immagini digitali abbiamo una quantità finita di "punti" messi belli in ordine, perfettamente distinguibili tra loro. Un po' la stessa differenza che c'è tra colorare un disegno senza mai staccare il pennarello dal foglio e colorarlo con la tecnica puntinata.

Il principale cambiamento da fare è che tutte le operazioni vengono svolte *modulo N*, dove N è la dimensione dell'immagine. Operare *modulo N* vuol dire, un po' come prima, che ogni volta che abbiamo un numero non ci interessa il numero in sè per sè ma solo il resto della divisione con N: per esempio 7 modulo 3 diventa 1, mentre 44 modulo 6 diventa 2. Questo modo di vedere i numeri, che vi può sembrare così strano, in realtà già lo usate nella vita di tutti i giorni quando leggete l'orologio!

Ma torniamo alle nostre immagini digitali e soprattutto capiamo a che ci serve operare *modulo* N in questo caso.

Come abbiamo detto, un'immagine altro non è che una griglia di pixel. La posizione di un pixel è data dalle sue coordinate – orizzontale e verticale – nel quadrato, come se stessimo giocando a battaglia navale. Ora, se applichiamo la funzione iniziale, cioè quella che stira il quadrato, sicuramente usciamo dal quadrato e quindi ci ritroviamo con delle coordinate che non hanno senso: ma se leggiamo le coordinate *modulo* N ci ritroviamo di nuovo con dei numeri che hanno senso e individuano un punto dentro la foto!

Quindi la versione digitale della mappa si formalizza come segue:

$$\begin{array}{rccc} \Gamma & : & \left(\mathbb{Z}/N\mathbb{Z}\right)^2 & \longrightarrow & \left(\mathbb{Z}/N\mathbb{Z}\right)^2 \\ & & \begin{pmatrix} x \\ y \end{pmatrix} & \longmapsto & \begin{pmatrix} 2x+y \\ x+y \end{pmatrix} \mod N \end{array}$$

dove la scrittura $\left(\mathbb{Z}/N\mathbb{Z}\right)^2$ indica solamente che stiamo lavorando in una griglia deve le coordinate sono solo numeri compresi tra 1 ed N.

A questo punto dovrebbe essere ovvio il motivo per cui la mappa digitale è periodica, cioè ad un certo punto ritorna all'immagine di partenza. Infatti, visto che le foto sono composte da un numero finito di pixel, non si possono trovare infiniti modi di sistemarli, quindi non si possono proporre infinite immagini con quei pixel!

La cosa stupefacente però è che il numero di iterazioni da applicare per tornare alla foto di partenza è molto più basso di quanto ci si aspetti! Infatti se dovessimo passare per tutte le possibili permutazioni dei pixel dell'immagine avremmo a che fare con numeri spaventosi, ma è stato dimostrato che non è necessario fare tutti questi tentativi.

Chiamiamo $T(N)$ il periodo della mappa di Arnold applicata ad un'immagine di lato N, ossia il numero di iterazioni necessarie perchè si ritorni all'immagine iniziale. Si può dimostrare che valgono le seguenti stime:

- $T(N) = 3N$ per N della forma $2 \cdot 5^k$ con $k \geq 1$ intero;
- $T(N) = 2N$ per N della forma $6^a \cdot 5^k$ con $a = 0, 1$ e $k \geq 1$ intero;
- $T(N) \leq \frac{12}{7} N$ per gli altri valori di N.

Quindi, per fare degli esempi concreti, se consideriamo un'immagine di $50 = 2 \cdot 5^2$ pixel per lato, il periodo sarà di 150 iterazioni; se invece prendiamo una foto di $150 = 6 \cdot 5^2$ pixel per lato, il periodo sarà di 300 iterazioni.

Questo risultato ci dice in sostanza che in ogni caso siamo ben lontani dal dover passare da tutte le sitemazioni possibili dei pixel prima di riavere l'immagine iniziale, e siamo ancora più lontani dal comportamento *altamente mixing* della mappa da cui siamo partiti.

Letture consigliate

Sperando che questo articolo vi sia piaciuto, di seguito vi lascio alcuni spunti di lettura se siete interessati ad approfondire l'argomento trattato in questo articolo. **Disclaimer:** Alcuni di questi sono articoli non semplicissimi da comprendere; non vi scoraggiate se non ci capite nulla, anche io ho dovuto metterci molto impegno!

1. Link alla pagina Wikipedia sulla mappa di Arnold:
   `https://it.wikipedia.org/wiki/Gatto_di_Arnol'd`;

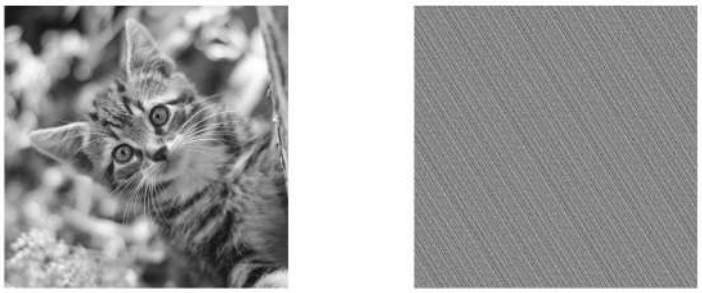


Figura 100: Immagine di partenza (a sinistra) e dopo 6 iterazioni della mappa (a destra)

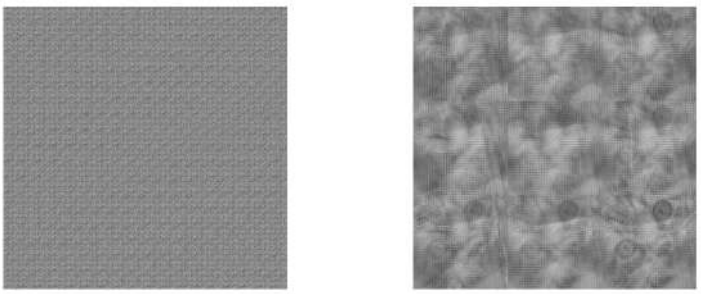


Figura 101: Immagine dopo 8 (a sinistra) e dopo 18 iterazioni della mappa (a destra).

Figura 102: Immagine dopo 25 (a sinistra) e dopo 36 iterazioni della mappa (a destra).

2. Articolo sul concetto di *scala di mixing*:
   A. Bressan – *A Lemma and a Conjecture on the Cost of Rearrangements*;
3. Articolo sul periodo della mappa discretizzata (il risultato riportato è nella sezione 2, ma potete leggere tutto l'articolo per vedere altre versioni della mappa):
   J. Bao, Q. Yang – *Period of the discrete Arnold cat map and general cat map.*

30 Contare con i polinomi

Davide Chionna, n.17, Ottobre 2023

30.1 Introduzione

Questo articolo nasce con uno duplice scopo: il primo, quello più immediato, è quello di fornire al lettore un'introduzione a delle potenti e importanti tecniche combinatoriche che permettono di enumerare un gran numero di famiglie di oggetti interessanti arrivando addirittura a fornire formule chiuse per il calcolo della loro cardinalità; il secondo scopo, vagamente più filosofico, è quello di mostrare al lettore la potenza della formalizzazione in matematica, ovvero come, semplicemente adottando una notazione intelligente per scrivere ciò di cui si parla, si arrivi subito a sviluppare degli strumenti algoritmici che ci facilitino nella soluzione di problemi altrimenti molto complessi. In questo caso specifico il ruolo della formalizzazione sarà giocato dalle serie generatrici con tutte le manipolazioni algebriche formali che si portano dietro.

30.2 Le funzioni generatrici

Immaginiamo che, per un qualche motivo, siamo interessati a contare esattamente un certo un numero di oggetti. Il primo istinto è certamente per tutti quello di elencare gli oggetti che ci interessano ad uno ad uno e contarli come farebbe un bambino; del resto, nel caso in cui gli oggetti in questione non siano accumunati da specifiche proprietà matematiche, questa appare l'unica soluzione (ad esempio io stesso non avrei in mente idee particolarmente furbe per contare quante siano di preciso le olive nella campagna di mio nonno). Però, facciamo un passo avanti nella matematica e portiamo qualche esempio in cui questi oggetti di cui ci interessa il numero abbiano anche qualche struttura in più.

Supponiamo di voler conoscere quanti siano gli anagrammi della parola "cane", (senza richiedere che questi abbiano un senso nella lingua). Si potrebbe tentare di elencarli tutti e presto ci si riuscirebbe (del resto sono solo quattro lettere!); magari li elencheremmo con un criterio per non sbagliarci: ad esempio prima quelli che iniziano per "C", poi per "A" e cosi via... ma se lettere aumentassero? Se fossero 10 lettere distinte? Ci andrebbe ancora di elencarli?

Prima di rispondere sì, teniamo a mente che qui gli anagrammi passerebbero da appena 24 a ben più di 3 milioni! Il matematico cerca quindi un metodo più furbo: all'inizio si chiede cosa vuole davvero contare e passa dagli anagrammi della parola cane agli anagrammi di una parola con 4 lettere distinte. A questo punto ha individuato la proprietà fondamentale che distingue la sua famiglia di oggetti! Poi riflette un po' (magari anche ripensando al criterio con cui li aveva elencati: "prima quelle che iniziano per C...") e si dice: "beh, ho quattro scelte per la prima lettera (in questo caso C,A,N o E), per ognuna di queste scelte potrò scegliere in 3 modi la seconda lettera tra le rimanenti e per ognuna di queste 12 coppie potrò scegliere in soli due modi la penultima lettera e quindi sarò obbligato a terminare con l'ultima restante, allora gli anagrammi sono $4 \cdot 3 \cdot 2 \cdot 1 = 24$".

Soddisfatto, torna a chiedersi: "e se le lettere fossero 10?" Come prima, stesso ragionamento, avremmo: $10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ anagrammi. Il matematico però non si accontenta. E se le lettere fossero una quantità qualsiasi, diciamo n? Il ragionamento non cambia e, ammesso che le n lettere siano tutte distinte, avremmo una quantità di anagrammi pari al prodotto di tutti i numeri interi da 1 a n. Si decide allora di dare un nome a questo prodotto e chiamarlo *fattoriale* dell'intero n ed indicarlo con $n!$.

La *funzione generatrice* è un nuovo passo verso la generalizzazione.

Supponiamo di avere una successione di famiglie di oggetti dipendenti da un parametro come l'intero n, nell'esempio precedente "gli anagrammi delle parole con n lettere distinte" (anche conosciute come *permutazioni di n elementi*). Questa successione di famiglie – o, come si dice in gergo matematico, "insiemi" – è di per se infinita (come i numeri naturali!). Decidiamo allora di chiamare il suo ennesimo termine A_n (ad esempio, A_5 sarà l'insieme delle permutazioni di 5 oggetti). È chiaro che una volta fissato il valore del parametro n abbiamo determinato un insieme ben preciso e dunque anche la sua cardinalità (ad esempio la cardinalità di A_5, che indichiamo con $|A_5|$, sarà 5!=120).

Le funzioni generatrici sono un modo per ricordare contemporaneamente tutte le cardinalità della nostra successione. Definiamo la serie generatrice di una generica successione di insiemi I_n come:

$$|I_0| + |I_1|x + |I_2|x^2 + \ldots = \sum_{n=0}^{\infty} |I_n|x^n,$$

dove la "sommatoria" all'ultimo termine non rappresenta altro che la scrittura prima dell'uguale: la cardinalità di I_0 per il monomio x^0, più la cardinalità di I_1 per il monomio x^1, e così via per $n = 2, 3, 4, \ldots$.

Questo oggetto, che assomiglia ad un polinomio ma ha infiniti termini, può (almeno per quanto ci riguarda) essere pensato semplicemente come un polinomio di grado infinito. Sull'insieme delle funzioni generatrici possiamo fare gran parte delle operazioni che faremmo sui polinomi: somme, moltiplicazioni, composizioni, etc... l'importante è assicurarsi sempre che il risultato di queste operazioni sia di nuovo un "polinomio" – magari di grado infinito, ma con la proprietà che per calcolare un preciso suo termine bastino sempre un numero finito di operazioni algebriche.

Ad esempio, nel caso degli insiemi A_n degli anagrammi di una parola di n lettere diverse, si può scrivere la funzione generatrice

$$1 + x + 2x^2 + 6x^3 + 24x^4 + \ldots = \sum_{n=0}^{\infty} n! \cdot x^n.$$

30.3 Funzioni generatrici, famiglie ricorsive e numeri di Fibonacci.

Quasi tutti conosceranno la sequenza di Fibonacci, ovvero quella successione di numeri che inizia con 0 ed 1 e in cui ogni termine successivo si determina sommando i due precedenti. Qui ci occuperemo di approfondire un aspetto della conoscenza di questi numeri, arrivando, attraverso lo strumento delle serie generatrici, ad una "formula chiusa". Detto F_n l'ennesimo numero di Fibonacci, questo si determina sommando i due precedenti, quindi notazionalmente scriveremo

$$F_{n+2} = F_{n+1} + F_n.$$

Quando siamo davanti a una formula ricorsiva di questo tipo, e ne vogliamo trovare un analogo chiuso, è sempre una buona idea ricorrere alle funzioni generatrici: proviamo!

Consideriamo la funzione generatrice associata alla successione F_n, supponendo che si abbia $F_0 = 0$ ed $F_1 = 1$; possiamo scrivere la funzione generatrice cercata come

$$F_0 + F_1x + F_2x^2 + F_3x^3 + F_4x^4 + \ldots = 0 + x + x^2 + 2x^3 + 5x^4 + \ldots$$

Consideriamo ora la legge:

$$F_{n+2} = F_{n+1} + F_n$$

e moltiplichiamo entrambi i membri per x^{n+2}, ottenendo così

$$F_{n+2}x^{n+2} = F_{n+1}x^{n+2} + F_n x^{n+2}.$$

Sommando ora l'identità qui sopra per $n = 0, 1, 2, \ldots$, si ottiene:

$$\sum_{n=0}^{\infty} F_{n+2}x^{n+2} = \sum_{n=0}^{\infty} F_{n+1}x^{n+2} + \sum_{n=0}^{\infty} F_n x^{n+2}; \tag{19}$$

Possiamo poi raccogliere un fattore x e un fattore x^2 e riscrivere l'uguaglianza come

$$\sum_{n=0}^{\infty} F_{n+2}x^{n+2} = x\sum_{n=0}^{\infty} F_{n+1}x^{n+1} + x^2\sum_{n=0}^{\infty} F_n x^n.$$

Notiamo ora che la funzione generatrice che vorremmo calcolare, che da ora in poi chiameremo $F(x)$ e che continuiamo a immaginare come un polinomio infinito nella variabile x, è:

$$F(x) = \sum_{n=0}^{\infty} F_n x^n = 0 + \sum_{n=1}^{\infty} F_n x^n = 0 + x + \sum_{n=2}^{\infty} F_n x^n.$$

Da qui segue guardando il primo e l'ultimo membro e portando la x dall'altra parte:

$$F(x) - x = \sum_{n=2}^{\infty} F_n x^n. \tag{20}$$

Fermiamoci per un attimo e osserviamo che possiamo scrivere, risistemando gli indici della sommatoria,

$$F(x) = t\sum_{n=1}^{\infty} F_n x^n = \sum_{n=0}^{\infty} F_{n+1}x^{n+1}. \tag{21}$$

Analogamente

$$\sum_{n=2}^{\infty} F_n x^n = \sum_{n=0}^{\infty} F_{n+2}x^{n+2}. \tag{22}$$

Se non credete a queste ultime due uguaglianze provate ad elencare, per ognuno, i termini delle somme a sinistra e a destra dell'uguale...

Pertanto la formula 19 trovata si trasforma in:

$$F(x) - x = \sum_{n=0}^{\infty} F_{n+2}x^{n+2} = x\sum_{n=0}^{\infty} F_{n+1}x^{n+1} + x^2\sum_{n=0}^{\infty} F_n x^n = xF(x) + x^2F(x),$$

dove la prima uguaglianza segue da 20, la seconda dall'equazione 19 e l'ultima dalla definizione di serie generatrici dall'equazione 21.

Possiamo ora trattare questa come un' equazione di primo grado nella variabile $F(x)$ (dove le x sono semplici coefficienti, a differenza di ciò a cui siamo abituati) e risolverla facilmente, ottenendo:

$$F(x) = \frac{x}{1 - x - x^2}.$$

Abbiamo dunque trovato la nostra funzione generatrice dei numeri di Fibonacci.

Chi abbia qualche familiarità con l'analisi saprà che ad una funzione del genere corrisponde uno sviluppo in serie (ovvero polinomio di grado infinito) e che questo può essere calcolato in molti

modi puramente algoritmici (teniamo presente che in questa circostanza siamo interessati solo al significato formale dell'espressione e quindi non ci importa di alcun problema di convergenza). Operato questo sviluppo otteniamo una formula per il coefficiente ennesimo dell'espansione, che in base a quanto detto sarà la formula chiusa per i numeri di Fibonacci: vediamo allora come procedere concretamente allo sviluppo. Le radici $x^2 + x - 1$ sono $r_\pm = (-1 \pm \sqrt{1+4})/2 = (-1 \pm \sqrt{5})/2$. Quindi abbiamo:

$$1 - x - x^2 = (1 - r_- x)(1 - r_+ x),$$

da cui

$$\frac{x}{1-x-x^2} = \frac{x}{(1-r_-x)(1-r_+x)} = \frac{1}{r_+ - r_-}\left(\frac{1}{1-r_+x} - \frac{1}{1-r_-x}\right) =$$

$$= \frac{1}{\sqrt{5}} \sum_j (r_+^j - r_-^j) x^j.$$

Ne abbiamo dunque dedotto la formula chiusa per i numeri di Fibonacci, conosciuta come *Formula di Binet*.

$$F_n = \frac{1}{\sqrt{5}}(r_+^n - r_-^n) = \frac{1}{\sqrt{5}}\left(\left(\frac{1+\sqrt{5}}{2}\right)^n - \left(\frac{1-\sqrt{5}}{2}\right)^n\right).$$

Questa formula, a mio avviso, è interessante per due motivi: il primo è che nonostante la sua apparente complessità di scrittura, restituisce sempre valori interi; il secondo è che spiega il rapporto tra la tanto decantata costante aurea $\frac{1+\sqrt{5}}{2}$ e la serie di Fibonacci che è tanto nota e tanto romanzata (se avete familiarità con l'analisi provate a considerare il valore del limite del rapporto di due termini successivi della sequenza quando n va ad infinito, resterete piacevolmente sorpresi!).

Abbiamo quindi qui avuto un primo assaggio della potenza delle funzioni generatrici. Mi auguro che al lettore, più dei singoli passaggi formali di sviluppo di espressioni, rimanga impressa la generalità e la riproducibilità dei metodi usati, tenendo a mente che questi diventano tanto più potenti tanto più lo sono gli strumenti di cui ci avvaliamo per avanzare nei calcoli (in questo senso la formula di inversione di Lagrange o l'utilizzo dello sviluppo in serie di Taylor aprono un mondo nuovo). In ogni caso, l'idea alla base resta sempre, semplicemente, quella mostrata qui.

Riferimenti bibliografici

[1] Herbert S. Wilf, *Generatingfunctionology*, `https://www2.math.upenn.edu/~wilf/DownldGF.html`.

31 La derivata aritmetica

Federico Allegri, n.18, Ottobre 2024

Sono le 7:00 del mattino, ti svegli spaventato perché sai che avrai l'interrogazione di matematica sulle derivate. Sul bus verso scuola ripassi tutte le derivate fondamentali: «La derivata di x^2 è $2x$, la derivata di $\sin(x)$ è $\cos(x)$...». Finché esclami ad alta voce: «La derivata di 15 è 0».

In quel momento, un signore calvo e con un vago accento spagnoleggiante ti guarda e ti dice: «Ragazzo, ma che cosa sta dicendo? La derivata di 15 è 8». Stupito, gli chiedi chiarimenti. Inizia così la magnifica avventura all'interno di un mondo nel quale la derivata di un numero non è sempre 0: scopriamo insieme la derivata aritmetica!

31.1 Introduzione

La derivazione di un numero è un concetto nato per la prima volta nel primo decennio del secolo scorso e fu definita dal matematico spagnolo José Mingot Shelly, che la presentò al Congresso nazionale della *Sociedad Matemática Española* di Granada nel 1911. Purtroppo, le sue idee vennero presto dimenticate e l'unica fonte che abbiamo su di lui e sulla sua definizione di derivata aritmetica è quella data da Dickson nel suo *History of the Theory of Numbers*.

Il concetto di derivata di un intero rinasce solo nel 2001 con la *Online Encyclopedia of Integer Sequences* (OEIS), un enorme database di sequenze di numeri fondato da Neil J. A. Sloane: la definizione di Shelly di derivata aritmetica venne inserita dallo stesso Sloane nell'OEIS e venne poi proposta nel 2002 alla quattordicesima *Summer Conference of International Tournament of Towns*.

Sorge spontaneo chiedersi quale sia lo scopo di questo nuovo strumento: la risposta è che l'idea di associare un concetto di derivata a un numero intero offre molti utili agganci nello studio della Teoria dei Numeri. In particolare, può essere vista come un mezzo di traduzione di alcuni problemi di natura numerica in una forma differente, permettendo così nuovi approcci risolutivi.

31.2 La derivata aritmetica

31.2.1 La derivata “classica"

Consideriamo una funzione reale $f : I \subset \mathbb{R} \longrightarrow \mathbb{R}$ con I intervallo aperto e supponiamo che $x_0 \in I$ sia un punto in cui f è definita. Vorremmo trovare un modo per determinare, se esiste, quale sia il coefficiente angolare della retta tangente al grafico di f nel suo punto di ascissa x_0. La derivata ci permette di determinare precisamente questo coefficiente angolare: fare la derivata della funzione f in x_0 significa determinare la pendenza della retta tangente al grafico di f nel punto di coordinate $(x_0, f(x_0))$. Questo concetto si può estendere a ogni punto di I: in questo modo possiamo definire una funzione f', detta *funzione derivata*, che ha la proprietà che $f'(x_0)$ coincide con il coefficiente angolare della retta tangente al grafico di f nel punto $(x_0, f(x_0))$. Dato che le uniche funzioni che ci serviranno nel corso della trattazione saranno i polinomi, vediamo come fare per derivare questi ultimi.

Cominciamo dai monomi: sia ax^n un monomio, dove x è una **indeterminata**, a è un numero (naturale, intero, irrazionale, come vi pare!) e n è un numero naturale (cioè 0, 1, 2, 3,...). Come si definisce la derivata rispetto all'indeterminata x del monomio ax^n? È facilissimo! L'idea è quella

di moltiplicare a per l'esponente n e diminuire di 1 l'esponente stesso. Scriviamolo in formule:

$$(ax^n)' = nax^{n-1},$$

dove con l'apostrofo $'$ indichiamo la derivata. Ad esempio la derivata di x^4 è $4x^{4-1} = 4x^3$, oppure la derivata di $(54x^8)' = 54 \cdot 8x^{8-1} = 432x^7$ e così via. Osserviamo ora una cosa importantissima: con la definizione data sopra, la derivata di un qualsiasi numero è 0. Infatti, detto a un numero reale, derivare la funzione $f(x) = a$ vuol dire determinare una funzione $f'(x)$, tale che per ogni punto $x_0 \in \mathbb{R}$, $f'(x_0)$ coincida con il coefficiente della retta tangente a $f(x) = a$ nel punto (x_0, a). Tuttavia, dato che la funzione $f(x) = a$ è rappresentata da una retta orizzontale nel piano cartesiano, in ogni punto la retta tangente coincide con la funzione stessa, e in particolare ha coefficiente angolare nullo.

Veniamo ora ai polinomi, che sono "somme di monomi": sia

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 + a_0$$

un polinomio, dove x è una indeterminata e $a_0, ..., a_n$ sono numeri. La derivata di $p(x)$, che si indica con $p'(x)$, è semplicemente:

$$p'(x) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + ... + a_1$$

ovvero è la somma delle derivate dei singoli monomi che compongono $p(x)$. Con un esempio cerchiamo di capire meglio: vogliamo derivare il polinomio $p(x) = 3x^4 + x^2 - 2$; basta derivare ogni singolo addendo: $(3x^4)' = 12x^3$, $(x^2)' = 2x$, e $(2)' = 0$, ottenendo così che

$$p'(x) = 12x^3 + 2x.$$

Potete provare a fare tutti gli esempi che volete.

31.2.2 La derivata aritmetica

Passiamo allora adesso alla definizione tanto attesa: come si definisce la *derivata aritmetica* di un numero? L'idea chiave per definire la derivata aritmetica di un numero è quella di adattare agli interi le classiche regole di derivazione per polinomi.

[width=boxrule=0.3mm, sharp corners] **Attenzione!** Questa è una **nuova** derivata, definita solo sui numeri, quindi non aspettiamoci come sopra che faccia sempre zero, tutto questo scritto sarebbe altrimenti poco interessante.

Nell'articolo di Shelly del 1911 viene presentato il seguente paragone: dato che la derivata di

$$p(x) = x^\alpha$$

è

$$p'(x) = \alpha \cdot x^{\alpha-1},$$

perché non definire la derivata aritmetica considerando come indeterminata i numeri primi? Definiamo prima cosa sono questi numeri.

Definizione 11 (Numero primo). Un numero intero p è detto *numero primo* se ha esattamente 4 divisori distinti.

Quindi, brutalmente, i monomi che consideriamo adesso sono i numeri del tipo p^n, dove p è un numero primo e n un numero naturale: avremo che la derivata aritmetica di un numero del tipo p^n è np^{n-1}, ancora una volta seguendo le regole riportate sopra. Facciamo degli esempi per scaldarci un po'.

- Qual è la derivata di 25? Dato che $25 = 5^2$, avremo che

$$25' = 2 \cdot 5 = 10.$$

- Qual è invece la derivata di 81? Dato che $81 = 3^4$, si avrà che

$$81' = 4 \cdot 3^3 = 4 \cdot 27 = 108.$$

- Qual è poi la derivata di $3 = 3^1$? Sarà

$$3' = 1 \cdot 3^0 = 1$$

Questa ultima regola vale in generale:

Proposition 1. La derivata aritmetica di un numero primo è sempre 1.

Come facciamo invece a derivare aritmeticamente un numero che non è una potenza di un primo, tipo 10 o 15 o 18? Per prima cosa definiamo una nuova regola di derivazione valida sui prodotti: se un numero si scrive come un prodotto tra due numeri primi, ad esempio $p \cdot q$ con p e q primi non necessariamente distinti, la derivata si fa in questo modo:

$$(pq)' = p' \cdot q + p \cdot q' = q + p,$$

ovvero spezziamo il prodotto in una somma di due addendi, nel primo scriviamo la derivata del primo numero che compare e lo moltiplichiamo per il secondo numero non derivato, mentre nel secondo addendo facciamo le cose al contrario (piccolo accorgimento per i più esperti: la derivata classica sui prodotti tra funzioni agisce proprio come sopra!). Questa regola vale anche per prodotti di 3, 4, 5 primi: basterà aumentare coerentemente il numero di addendi; ad esempio:

$$(pqrs)' = p'qrs + pq'rs + pqr's + pqrs'.$$

[width=boxrule=0.3mm, sharp corners] **Attenzione!** Questa regola è inoltre concorde con la regola riportata sopra, infatti possiamo pensare ad esempio p^2 come $p \cdot p$ e utilizzare la formula, ottenendo $(p^2)' = (p \cdot p)' = p' \cdot p + p \cdot p' = 2p' \cdot p = 2p$.

E se ci troviamo davanti un numero a caso, tipo 2539276429409192, cosa possiamo fare? Per derivare un numero come questo, dobbiamo appellarci a un risultato fondamentale dell'aritmetica, che riportiamo senza dimostrare:

Theorem 7 (fondamentale dell'aritmetica). Ogni numero intero diverso da 0, da 1 e da -1 può essere scritto in modo unico, a meno del segno e dell'ordine dei fattori, come prodotto di numeri primi.

Grazie a questo teorema e alla regola sui prodotti riportata sopra, possiamo derivare qualsiasi numero vogliamo!

Sia infatti N un numero intero, allora possiamo scrivere N come prodotto di numeri primi grazie al teorema precedente: esisteranno (unici) dei numeri primi $p_1, ..., p_n$ distinti e dei numeri naturali positivi $a_1, ..., a_n$ tali che

$$N = p_1^{a_1} ... p_n^{a_n}.$$

Utilizzando ora la regola di derivazione sul prodotto abbiamo concluso! Facciamo degli esempi per capire bene come funziona la cosa.

- La derivata di 45 si calcola come segue:

$$(45)' = (3^2 \cdot 5)' =$$
$$(3 \cdot 3 \cdot 5)' = 3' \cdot 3 \cdot 5 + 3 \cdot 3' \cdot 5 + 3 \cdot 3 \cdot 5' = 15 + 15 + 9 = 39.$$

- La derivata di 900 invece è

$$900' = (2^2 \cdot 3^2 \cdot 5^2)' =$$
$$(2^2)' \cdot 3^2 \cdot 5^2 + 2^2 \cdot (3^2)' \cdot 5^2 + 2^2 \cdot 3^2 \cdot (5^2)' = 900 + 600 + 360 = 1860.$$

In quest'ultimo esempio abbiamo agito in maniera differente: ci siamo accorti che non ha molto senso scrivere la derivata di 2^2 come la derivata di $2 \cdot 2$ (anche se è perfettamente lecito), ma possiamo considerare questo come un unico blocchetto e applicare la prima regola di derivazione che abbiamo visto: sta a voi verificare che i due risultati coincidono.

Infine, per completezza, dobbiamo esplicitare la derivata aritmetica di 0 e 1, in quanto questi non sono né numeri primi, né possono essere fattorizzati come prodotto di numeri primi, e quindi non possiamo applicare le regole viste: per non generare contraddizioni, poniamo semplicemente $1' = 0' = 0$.

La tabella sottostante mostra le derivate aritmetiche dei numeri da 0 a 38.

n	0	1	2	3	4	5	6	7	8	9	10	11	12
n'	0	0	1	1	4	1	5	1	12	6	7	1	16

n	13	14	15	16	17	18	19	20	21	22	23	24	25
n'	1	9	8	32	1	21	1	24	10	13	1	44	10

n	26	27	28	29	30	31	32	33	34	35	36	37	38
n'	15	27	32	1	31	1	80	14	19	12	60	1	21

Tabella 4: Derivate aritmetiche dei numeri da 1 a 38

31.3 Derivate successive

La derivata aritmetica di un numero è ancora un numero, quindi ha perfettamente senso continuare a derivarlo. Possiamo allora estendere il concetto alla derivata seconda, terza, quarta... Ad esempio, la derivata seconda di 15 è:

$$(15)'' = (3 \cdot 5)'' = (3' \cdot 5 + 3 \cdot 5')' = 8' = (2^3)' = 3 \cdot 2^2 = 12.$$

È interessante cercare di capire quale sia il comportamento della derivata k-esima di un numero n all'aumentare di k: in generale purtroppo l'andamento è molto irregolare, cioè non si può dire a priori cosa succederà a un numero quando lo deriviamo tante volte. Ci sono casi in cui la derivata va a 0 dopo poche derivazioni, ad esempio considerando le derivate successive di 10 otteniamo la sequenza $10, 7, 1, 0, 0, 0, ...$, e casi in cui invece cresce rapidamente e non sappiamo dove andrà a finire, come si vede considerando per esempio la sequenza delle derivate successive di 15: $15, 8, 12, 16, 32, 80, 176, 368...$

Analizziamo un caso particolare: consideriamo i numeri della forma

$$n = p^p n_1$$

dove p è primo e n_1 è un numero naturale non divisibile per p. Se deriviamo n otteniamo

$$n' = p \cdot p^{p-1} n_1 + p^p n_1' = p^p n_1 + p^p n_1' = n + p^p n_1' \geq n,$$

con uguaglianza solo quando $n_1 = 1$. Si ha quindi che per interi n che posseggono un divisore proprio (ovvero diverso da n) del tipo p^p, la derivata k-esima tenderà a crescere sempre di più (in modo più matematico si direbbe che $\lim_k n^{(k)} = +\infty$; se invece $n = p^p$, allora $n' = n$ e quindi anche $n^{(k)} = n$ per ogni k numero naturale. Il primo fatto è facilmente verificabile: sia $n = p^p n_1$ con $n_1 \neq 1$, allora

$$n' = n + p^p n_1' > n$$

$$n'' = (n')' = (n + p^p n_1')' = n' + p^p n_1' + p^p n_1'' = n + 2p^p n_1' + p^p n_1'' > n'$$

$$n''' = (n'')' = (n' + p^p n_1' + p^p n_1'')' = n'' + p^p n_1' + 2p^p n_1'' + p^p n_1''' > n''$$

e così via. Il punto è che a ogni derivazione si aggiunge il termine $p^p n_1'$, che è positivo.

Esiste una congettura molto interessante sull'andamento delle derivate successive, che risulta ancora oggi indimostrata (per maggiori dettagli si veda [6]):

Conjecture 1 (Barbeau). Per ogni intero n esiste una costante k_0 tale che per ogni $k \geq k_0$ possono succedere due cose:

- $n^{(k)} = 0$,
- $n^{(k)} \neq 0$ ed esiste un primo p tale che $n^{(k)}$ è un multiplo di p (ovvero p divide tutte le derivate successive alla k-esima).

Nulla vieta a voi di provare a risolvere questo problema. Proponiamo un esempio per capire cosa dice la congettura: abbiamo già osservato che la successione delle derivate successive di 10 va a 0 dopo 3 derivazioni. Consideriamo nuovamente la successione delle derivate di 15: possiamo notare che già dalla prima derivazione le derivate cominciano ad essere sempre pari ed effettivamente se continuassimo a derivare tante volte continueremmo a ottenere numeri pari. Questo fatto è concorde con la congettura di Barbeau: in questo caso $k_0 = 1$ e $p = 2$.

31.4 Equazioni differenziali aritmetiche

Passiamo ora a una parte molto importante, che è quella che ci servirà a dare un senso a tutto questo cumulo di parole e simboli. Parliamo di equazioni differenziali aritmetiche!

Consideriamo un numero naturale k; vogliamo capire chi sono quei numeri naturali n tali per cui $n' = k$. Quella scritta sopra è un'**equazione differenziale**, cioè un'equazione in cui compare l'incognita n, accompagnata da una sua derivata. Possono esistere equazioni differenziali anche legate alle derivate successive, cioè possiamo chiederci quali siano i numeri naturali n tali per cui $n'' = k$, $n''' = k$ e così via... e possiamo anche considerare equazioni più complicate in cui intrecciamo derivate successive e numeri, ad esempio $n''' + n'' - n^5 = 8$.

Nel primo caso che abbiamo introdotto, dovrebbe risultare abbastanza chiaro che per certi valori di k esistono infinite soluzioni all'equazione differenziale: se per esempio $k = 1$, si ha che l'equazione $n' = 1$ ha infinite soluzioni, ad esempio, tutti i numeri primi sono soluzione (e sono anche le sole soluzioni, lasciamo a voi il compito di provare a dimostrarlo).

L'equazione differenziale $n' = n$ invece, abbiamo visto ad esempio avere soluzione $n = p^p$, con p numero primo. Questi sono casi abbastanza semplici di equazioni differenziali aritmetiche, la cui soluzione risulta quasi immediata. Generalmente, però, è molto difficile trovare la soluzione di un'equazione differenziale, e spesso neanche i computer riescono in questo intento. Ma perché allora fasciarci la testa su questo problema? Perché, oltre a poter avere un interesse intrinseco, la derivata aritmetica permette di tradurre molti problemi della Teoria dei Numeri in termini di equazioni differenziali, con lo scopo di approcciare il problema da un punto di vista differente e nuovo. Questo significa che, se riuscissimo a trovare un metodo per risolvere equazioni differenziali aritmetiche, si sarebbe allo stesso tempo riusciti a risolvere alcuni problemi aperti di teoria dei numeri! Facciamo degli esempi di queste traduzioni.

31.4.1 La congettura di Goldbach (forte)

Uno dei problemi della matematica più famosi al giorno d'oggi è senz'altro la congettura di Goldbach. Su questo problema, Hardy, uno dei più importanti teorici dei numeri del '900, si è espresso come segue:

> *"Goldbach's conjecture is not only the most famous and difficult problem in number theory, but the whole of mathematics"*

Vediamo come recita questa congettura:

Conjecture 2 (Goldbach). Ogni numero pari maggiore di 2 può essere scritto come somma di due numeri primi, non necessariamente distinti.

Effettivamente se si prova a cercare un controesempio a mano, non ci si riesce: $4 = 2 + 2$, $6 = 3+3$, $8 = 5+3$, $10 = 5+5$ e così via. Il fatto che valga per tutti i numeri che fino ad oggi sono stati provati, purtroppo non è una dimostrazione; dovremmo mostrare infatti che questo risultato vale per ogni numero pari e non solo per alcuni (e questi sono infiniti, quindi a mano proprio non ci si può riuscire!). Ci sono molti risultati parziali sul problema di Goldbach, che affermano cose molto interessanti, ma nessuno di questi riesce effettivamente a dimostrare la veridicità o la falsità della congettura. Vediamone alcuni: un primo risultato, dovuto a Vinogradov, afferma che *quasi tutti* gli interi pari sono esprimibili come somma di due primi, dove il significato di *quasi tutti* è il seguente: chiamiamo $N(n)$ il numero di interi pari minori di n che non si possono scrivere come somma di due numeri primi. Allora

$$\lim_{n\to+\infty} \frac{N(n)}{n} = 0.$$

Una scoperta di Landau afferma invece che, se la congettura di Goldbach è falsa, allora questa è falsa al più per lo 0% dei numeri pari, dove con "al più 0%" indica che possono esistere delle eccezioni (possibilmente anche infinite), ma sparse un po' a caso tra tutti i numeri pari.

Vediamo come possiamo riscrivere il problema di Goldbach sotto forma di equazione differenziale aritmetica. In simboli il problema riportato sopra si scrive come segue: per ogni numero naturale $n > 1$, esistono p e q numeri primi tali che $2n = p + q$. Basta adesso osservare che $p + q$ è la derivata aritmetica di $m = pq$ e otteniamo il seguente modo equivalente di formulare la congettura (ricordiamo che un numero è detto **semiprimo** quando è prodotto di due primi non necessariamente distinti, ad esempio 9, 10, 15 e 21 sono numeri semiprimi):

Conjecture 3. Sia $a > 1$ un numero naturale. L'equazione differenziale aritmetica $n' = 2a$ ha sempre almeno una soluzione semiprima.

Dovrebbe adesso risultare chiaro perché lo studio delle equazioni differenziali aritmetiche possa essere importante nella risoluzione di certi problemi!

31.4.2 I numeri primi gemelli

Un'altra importante congettura, che risale a Euclide, riguarda i numeri primi gemelli.

Definizione 12. Due numeri primi p e q sono detti *primi gemelli* se $q = p + 2$, ovvero se distano 2 l'uno da l'altro.

Per esempio 5 e 7 o 107 e 109 sono coppie di primi gemelli. Vediamo come recita la congettura.

Conjecture 4 (dei primi gemelli). Esistono infiniti numeri primi p tali che anche $p + 2$ sia primo.

Un importante risultato di Eulero ci dice che

$$\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \cdots = \sum_{p \text{ primi}} \frac{1}{p} = \infty,$$

ovvero che prendendo tutti i reciproci dei numeri primi e sommandoli "otteniamo una cosa grande quanto vogliamo e che cresce sempre di più". Nel 1915 il matematico norvegese Viggo Brun ha dimostrato invece che

$$\sum_{p,p+2 \text{ primi}} \frac{1}{p} < \infty,$$

cioè che limitando la somma ai primi "che hanno un primo gemello" si ottiene invece un numero finito, che indichiamo con $B_2 = 1,902...$ e che chiamiamo *costante di Brun* (per una dimostrazione si veda, ad esempio [4]). Questo, matematicamente, può essere tradotto in due modi diversi: o i primi gemelli sono in numero finito, o crescono talmente velocemente che la somma dei loro reciproci rimane limitata. A oggi è stata dimostrata da Chen Jingrun una variante "debole" del problema dei primi gemelli, che dice che esistono infiniti numeri primi p tali che $p + 2$ sia un numero primo o un semiprimo (si veda, ad esempio [5]).

Vediamo ora come tradurre il problema dei primi gemelli in termini di derivata aritmetica.

Proposition 2 (Sui primi gemelli). Condizione necessaria per la validità della congettura dei primi gemelli è che l'equazione differenziale $n'' = 1$ abbia infinite soluzioni tra i naturali del tipo $n = 2p$, con p primo.

Infatti, se la congettura dei primi gemelli è vera, allora esistono infiniti primi p tali per cui anche $p + 2$ è primo, ovvero $(p + 2)' = 1$. Ma $p + 2$ è la derivata aritmetica di $2p$, e quindi in particolare $(2p)'' = 1$, cioè $2p$ è soluzione dell'equazione differenziale $n'' = 1$.

Una riscrittura di questo genere, in termini di condizione necessaria, è un approccio che generalmente si usa per dimostrare la falsità di un problema: se infatti si dimostrasse che l'equazione differenziale $n'' = 1$ ha un numero finito di soluzioni o che ne ha infinite, ma quelle del tipo $n = 2p$ sono in numero finito, allora si sarebbe risolta in negativo la congettura dei primi gemelli. In realtà, ad oggi, si pensa che questa sia vera, anche se non si è ancora riusciti a dimostrarla... Però non si sa mai.

31.5 La congettura di Giuga e il legame con i primari pseudo-perfetti

Un altro importante esempio di applicazione delle derivate aritmetiche è legato ai numeri di Giuga. Nella prima metà del secolo scorso, il matematico Giuseppe Giuga ipotizzò la seguente congettura, che è ancora oggi un problema aperto:

Conjecture 5 (Giuga (1950)). Sia n un numero naturale; n divide

$$1^{n-1} + 2^{n-1} + ... + (n-2)^{n-1} + (n-1)^{n-1}$$

solo se n è primo.

Questa è stata verificata dal matematico brasiliano Paulo Ribenboim per i numeri primi $p < 10^{1700}$ (si veda, ad esempio, [4]).

Giuga dimostrò che l'esistenza di un numero n che soddisfa le due seguenti relazioni (ricordiamo che quando scriviamo che $p \mid n$, intendiamo che p è un divisore di n, ovvero che esiste un numero intero a tale che $n = a \cdot p$)

$$p - 1 \Big| \frac{n}{p} - 1 \tag{23}$$

$$p \Big| \frac{n}{p} - 1 \tag{24}$$

per ogni primo p divisore di n, sarebbe un controesempio alla Congettura 5 (chiaramente a oggi non si sono trovati numeri che soddisfano contemporaneamente le condizioni (23) e (24), altrimenti sapremmo già che la congettura risulterebbe falsa). Un numero che rispetta la condizione (24) si chiama **numero di Giuga**. A oggi sono conosciuti soltanto 13 numeri di Giuga: il più piccolo è 30, a seguire invece troviamo 858, 1722, 66198, 2214408306... Non sappiamo se i numeri di Giuga siano in numero finito o no, anche se si pensa che ne esistano infiniti. Per capire meglio il significato di cosa sia un numero di Giuga, facciamo vedere, usando la definizione, che 30 rispetta la proprietà (23): si ha che

$$30 = 2 \cdot 3 \cdot 5$$

e vale che

$$2 \Big| \frac{30}{2} - 1 = 14,$$
$$3 \Big| \frac{30}{3} - 1 = 9$$
$$5 \Big| \frac{30}{5} - 1 = 5$$

e quindi la condizione (23) è rispettata per ogni primo che divide 30, ovvero 30 è un numero di Giuga.

Ma qual è il legame che c'è tra i numeri di Giuga e il nostro studio sulla derivata aritmetica? È stato osservato che tutti e 13 i numeri di Giuga conosciuti sono soluzione dell'equazione differenziale $n' = n + 1$. Possiamo approcciare il problema della ricerca dei numeri di Giuga come segue:

Conjecture 6. I numeri di Giuga sono tutte e sole le soluzioni dell'equazione differenziale aritmetica $n' = a \cdot n + 1$ con a intero, ma tutte le soluzioni a oggi note hanno $a = 1$.

31.5.1 Primari pseudo-perfetti

Cerchiamo adesso di collegare la congettura di Giuga a una classe di numeri naturali, detti numeri primari pseudo-perfetti. Definiamo prima di tutto che cos'è un numero perfetto.

Definizione 13. Un numero intero n è perfetto se è somma dei suoi divisori propri, incluso 1.

Ad esempio, 6 è un numero perfetto, infatti $6 = 1 + 2 + 3$ e 1, 2 e 3 sono gli unici divisori propri di 6. Anche 28 è un numero perfetto, in quanto $28 = 1 + 2 + 4 + 7 + 14$.

Prendiamo ora in considerazione il numero $42 = 2 \cdot 3 \cdot 7$ e consideriamo i seguenti rapporti: $\frac{42}{2} = 21$, $\frac{42}{3} = 14$, $\frac{42}{7} = 6$, e vale che $42 = 1 + 21 + 14 + 6$, ovvero 42 è somma tra 1 e i rapporti tra 42 e i suoi divisori primi. I numeri di questo tipo sono detti **primari pseudo-perfetti** (PPP per semplicità). Più formalmente:

Definizione 14. Un numero n è un *PPP (primario pseudo-perfetto)* se rispetta la relazione

$$n = 1 + \sum_{p|n,\ p \text{ primo}} \frac{n}{p}.$$

I numeri PPP sono stati utilizzati per la prima volta nel tentativo di dimostrazione di una congettura di Erdös e Moser (che non è ancora stata provata), la quale recita come segue (per informazioni più dettagliate, si veda, ad esempio [2]):

Conjecture 7 (Erdös e Moser (1955)). L'equazione $1^k + 2^k + ... + n^k = (n+1)^k$, con k e n numeri naturali, non ha soluzione se non quella banale $1 + 2 = 3$.

Si può dimostrare che i PPP godono di un'importante proprietà legata alla derivata aritmetica, vale infatti il seguente teorema.

Theorem 8. L'equazione differenziale aritmetica $n' = n - 1$ ha come soluzione i numeri PPP.

Grazie a questo teorema possiamo analizzare il legame che sussiste tra i numeri di Giuga e i PPP.

Consideriamo n un numero naturale e scriviamo $n = m \cdot k$, con m e k naturali. Derivando entrambi i membri dell'uguaglianza si ottiene che $n' = m' \cdot k + m \cdot k'$. Supponiamo ora che m sia un PPP, allora, per il Teorema 8

$$n' = k' \cdot m + k \cdot (m - 1) = n + (k' \cdot m - k).$$

Imporre che la somma tra parentesi faccia 1 equivale a imporre che n sia un numero di Giuga, mentre imporre che sia uguale a -1 equivale a imporre che n sia un PPP. Inoltre, nel caso speciale in cui k è primo, si ha che $k' = 1$ e quindi $m - k = 1$ implica che n è un numero di Giuga, mentre $m - k = -1$ implica che n è un PPP. In altre parole stiamo affermando quanto segue: se $m - 1$ o $m + 1$ è un numero primo, allora, partendo da un PPP, possiamo, tramite un semplice prodotto, calcolare o un numero di Giuga o un PPP: se $m - 1$ è primo allora $m(m - 1)$ è un numero di Giuga, mentre se $m + 1$ è primo allora $m(m + 1)$ è ancora un PPP. Mostriamo con due esempi la veridicità di queste affermazioni: 6 è un numero PPP e $6 - 1 = 5$ è primo, vale che $5 \cdot 6 = 30$ che è un numero di Giuga. Anche $6 + 1 = 7$ è primo, e vale che $6 \cdot 7 = 42$, che abbiamo già visto essere un numero primario pseudo-perfetto.

Riferimenti bibliografici

[1] Giorgio Balzarotti, Paolo P. Lava, *La derivata aritmetica*, Hoepli, 2013.

[2] William Butske, Lynda M. Jaje, Daniel R. Mayernik, *On the equation* $\sum_{p|n} \frac{1}{p} + \frac{1}{n} = 1$, *pseudoperfect numbers and perfectly weighted graphs*, Math. Comp. **69** (2000), no. 229, 407–420, `https://doi.org/10.1006/jnth.1999.2451`.

[3] William D. Banks, C. Wesley Nevans, Carl Pomerance, *A remark on Giuga's conjecture and Lehmer's totient problem*, MSC Numbers **11A07, 11N25** (2009).

[4] Paulo Ribenboim, *The New Book of Prime Numbers Records*, Springer-Verlag, New York, 1996.

[5] J.R. Chen, *On the representation of a large even integer as the sum of a prime and the product of at most two primes*, Kexue Tongbao **11** (1966), no. 9, 385–386.

[6] E. J. Barbeau, *Remark on an arithmetic derivative*, Canad. Math. Bull. **4** (1961), no. 2.

32 Limited revisited: teoria delle categorie

Francesca Pratali, n.18, Ottobre 2024

32.1 Perché Teoria delle Categorie?

La *teoria delle categorie* è una branca relativamente giovane della matematica, nata dalla *topologia algebrica*[55] e progettata per descrivere in modo uniforme vari concetti strutturali provenienti da diversi campi matematici. È una disciplina trasversale che interseca molte branche della matematica oltre alla topologia algebrica, come la *geometria algebrica*, l'*informatica*, la *logica*... In effetti, la teoria delle categorie fornisce un bagaglio di concetti (e teoremi su essi) che costituiscono un'astrazione di molte idee concrete in diversi rami della matematica.

Da un certo punto di vista, la teoria delle categorie non è altro che un *linguaggio* con cui descrivere uniformemente fenomeni e costruzioni che sono trasversali nella matematica. Molti matematici e matematiche si sono espressi al riguardo:

- La matematica Emily Riehl scrive, nel suo libro *Category Theory in Context*[56]:

  > Atiyah ha descritto la matematica come "scienza dell'analogia". In questo senso, la teoria delle categorie si occupa di analogia matematica. La teoria delle categorie fornisce un linguaggio interdisciplinare per la matematica, progettato per delineare fenomeni generali, che consente il trasferimento di idee da un'area di studio all'altra. La prospettiva della teoria delle categorie può funzionare come un'astrazione semplificante, isolando le proposizioni che valgono per ragioni formali da quelle la cui dimostrazione richiede tecniche proprie di una determinata disciplina matematica.

- Citando Hoare in *Notes on an Approach to Category Theory for Computer Scientists*:

  > La teoria delle categorie è la branca più generale e astratta della matematica pura. [...] Il corollario di un alto grado di generalità e astrazione è che la teoria non fornisce quasi alcun aiuto per risolvere i problemi più specifici all'interno di una qualsiasi delle sottodiscipline a cui si applica. È uno strumento per il generalista, di scarso beneficio per il professionista [...].

- Citando A. Asperti e G. Longo in *Categories, Types, and Structures. Foundations of Computing Series*:

  > La teoria delle categorie è un gergo matematico. [...] Molti formalismi e strutture diverse possono essere proposte per quello che è essenzialmente lo stesso concetto; il linguaggio e l'approccio categoriale possono semplificare attraverso l'astrazione, mostrare la generalità dei concetti e aiutare a formulare definizioni uniformi.

[55] La topologia algebrica è lo studio di *spazi topologici* (ovvero degli spazi geometrici con una certa nozione molto debole di *distanza* tra i punti) per mezzo di *invarianti algebrici*, ovvero oggetti di natura algebrica che sono gli stessi per spazi *equivalenti* (dove equivalenti può voler dire *omeomorfi*, o *omotopi* o..... ma questa è un'altra storia!)

[56] Disponibile online sul sito dell'autrice in libero accesso

- Concludiamo citando D.S. Scott, che nel suo libro *Relating theories of the lambda calculus* scrive che

  > [La teoria delle categorie offre] teoria pura delle funzioni, non una teoria delle funzioni derivate dagli insiemi.

In questo articolo adotteremo precisamente questo ultimo approccio alla disciplina, e affronteremo un concetto fondamentale della teoria delle categorie, quello di *limite*.

32.2 Insiemi senza elementi

Cominciamo in modo originale: facciamo un gioco.

Quanto possiamo dire sugli insiemi *senza* parlare dei loro elementi?

Abbiamo ancora il diritto di parlare di *funzioni* da un insieme all'altro, inclusa la funzione *identità*, e persino *composizione* di queste funzioni, ma non possiamo parlare delle cose all'interno di un insieme, dei suoi elementi.

Potrebbe sembrare una restrizione sciocca, e va bene, non devo convincervi che ha senso. Se, però, dovessi provare a convincervi, potrei menzionare il fatto che la stessa idea di appartenenza ad un insieme è un'idea primitiva indefinibile in matematica, e che usare questa intuizione informalmente ha portato, in passato, a paradossi logici - il *paradosso di Russel* è forse uno dei più celebri esempi.

[width=boxrule=0.3mm, sharp corners] **Paradosso di Russel:**
Sia X l'insieme i cui elementi sono gli insiemi che non appartengono a se stessi, cioè $X = \{Y \text{ tale che } Y \notin Y\}$. Allora $X \in X$ oppure $X \notin X$?

Dunque ha senso chiedersi se, al posto di *assumere* (nel senso di porre come *assioma*) che l'appartenenza ad un insieme ha senso, possiamo iniziare con delle fondazioni diverse, come ad esempio le funzioni, e *dedurre* l'idea di appartenenza in modo logicamente coerente.

Ma, di nuovo, non devo convincervi. Prendetemi pure in giro. È solo un gioco, e quelle sono le regole. Esercitiamoci.

Possiamo dire

$$\text{L'insieme } X \text{ ha un solo elemento} \tag{25}$$

senza parlare di elementi?

Sorprendentemente, **sì**! Invece che parlare di quell'unico elemento, posso dire questo:

$$\text{Dato un qualsiasi altro insieme } Y, \text{ esiste } \textit{un'unica} \text{ funzione } f\colon Y \longrightarrow\!\!\!\!\!\rightarrow X. \tag{26}$$

Dimostriamo che queste due affermazioni sono equivalenti. Chiaramente, se X ha un solo elemento, allora (26) è vera. Supponiamo adesso che (26) sia vera, e dimostriamo che X ha un solo elemento.

[width=boxrule=0.3mm, sharp corners] **Reductio ad absurdum:**
per provare che una certa *proposizione* matematica è vera. L'idea consiste nell'osservare che l'implicazione logica $A \Rightarrow B$ ha lo stesso valore di verità di $\neg B \Rightarrow \neg A$, e dimostrare quest'ultima.

Lo facciamo procedendo *per assurdo*: se (26) è vera e X ha più di un elemento, possiamo prendere un insieme Y con un elemento solo e scegliere dove inviarlo: ci sono esattamente tante possibilità quanto gli elementi di X, e visto che per ipotesi X aveva più di un elemento, abbiamo più di una funzione $Y \longrightarrow\!\!\!\!\!\rightarrow X$, il che contraddice (26), assurdo. Quindi se (26) è vera, X deve avere un unico elemento, come volevamo.

Quindi l'esistenza di esattamente un'unica funzione $Y \longrightarrow\!\!\!\!\!\rightarrow X$, *per qualsiasi insieme* Y, è esattamente equivalente a dire che X ha un unico elemento.

Esercizio 12. Riuscite a trovare un modo di dire

$$\text{L'insieme } X \text{ è vuoto}$$

senza parlare dei suoi elementi?

Proviamo adesso a spingerci oltre, e capiamo come dire

$$\text{L'insieme } X \text{ ha } \textit{gli stessi elementi} \text{ dell'insieme } Y$$

senza parlare degli elementi di questi insiemi. Okay, questa domanda è malposta: visto che non possiamo nominare o descrivere questi elementi (non possiamo proprio parlare di loro), non possiamo dire esattamente questo. Ma possiamo riformulare la proposizione dicendo che

$$\text{L'insieme } X \text{ ha lo stesso } \textit{numero} \text{ di elementi dell'insieme } Y \tag{27}$$

Non solo questa formulazione è coerente con il nostro gioco, ma siamo anche capaci di esprimere (27) con le nostre restrizioni di linguaggio. La risposta non sarà certo una novità:

$$\text{C'è una funzione } f\colon X \longrightarrow Y, \text{ ed un'altra funzione } g\colon Y \longrightarrow X, \text{ tali per cui } g \circ f = \mathrm{id}_X \text{ e } f \circ g = \mathrm{id}_Y.$$

In matematica, questa si chiama *corrispondenza bigettiva*, e dimostra che due insiemi hanno lo stesso numero di elementi - anzi, per meglio dire, che hanno la stessa *cardinalità*, visto che potrebbero essere infiniti, e quindi non un numero in sé. Un altro modo per esprimere (27) è dire che gli insiemi X e Y sono lo stesso *a meno di una corrispondenza bigettiva*, o, più astrattamente, *a meno di isomorfismo.* Fissiamo questa terminologia in una definizione.

Definizione 15. Una funzione tra insiemi $f\colon X \longrightarrow Y$ è un *isomorfismo* se esiste $g\colon Y \longrightarrow X$ tale che $f \circ g = \mathrm{id}_Y$ e $g \circ f = \mathrm{id}_X$. Diciamo che due insiemi X e Y sono *isomorfi* se esiste un isomorfismo tra i due.

Bene. Queste sono quindi le regole del gioco, e questo il pattern che seguiremo. Passiamo adesso a cose più complicate, e scopriamo che cos'è un limite in teoria delle categorie.

32.3 Limiti... ma in teoria delle categorie

32.3.1 Decostruiamo il prodotto cartesiano

Continuiamo dunque con il nostro gioco. Prendiamo una coppia di insiemi, chiamiamoli X e Y. Ci hanno insegnato che il *prodotto cartesiano* di X e Y, che chiamiamo $X \times Y$, è l'insieme dato dalle coppie ordinate (x, y), dove x è un elemento di X e y un elemento di Y:

$$X \times Y = \{(x, y) \text{ tale che } x \in X, y \in Y\}.$$

E adesso poniamoci la seguente domanda:

Possiamo definire il *prodotto cartesiano* senza parlare di elementi?

Di nuovo, non possiamo veramente sperare di sapere *come* gli elementi dei nostri insiemi siano chiamati, ma **possiamo** dire che un insieme ha il *numero* giusto di elementi per essere il prodotto cartesiano, e possiamo anche fornire una coppia di funzioni che forniscono le componenti

della coppia ordinata, fornendo un modo per interpretare tutto ciò che è presente in quell'insieme come una coppia ordinata.

Vediamo quindi come si fa. Innanzitutto, il prodotto cartesiano di due insiemi è naturalmente munito di due funzioni $p\colon X \times Y \longrightarrow X$, $q\colon X \times Y \longrightarrow Y$, chiamate *proiezioni*. In un mondo in cui possiamo parlare di elementi, p e q sono definite come

$$p(x, y) = x \qquad q(x, y) = y$$

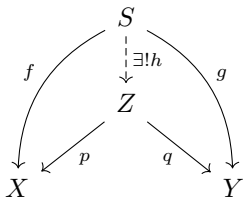
per ogni (x, y) in $X \times Y$.

Chiaramente, posta così abbiamo un problema: le proiezioni sono definite in termini di ciò che fanno sugli elementi, ma non possiamo parlare di elementi.

Quello che possiamo fare è osservare che le proiezioni sono caratterizzate da una certa **proprietà universale** che rende $X \times Y$ *finale* tra gli insiemi che godono di questa proprietà. Incredibilmente, questo basta a definire il prodotto cartesiano $X \times Y$!

Definizione 16. Il prodotto cartesiano di X e Y è dato da un insieme Z e due funzioni $p\colon Z \longrightarrow X$, $q\colon Z \longrightarrow Y$ con la seguente proprietà: per qualsiasi insieme S e per qualsiasi coppia di funzioni $f\colon S \longrightarrow X$ e $g\colon S \longrightarrow Y$ esiste un unico $h\colon S \longrightarrow Z$ tale che $f = p \circ h$ e $g = q \circ h$.

Possiamo esprimere quanto sopra con il seguente diagramma:



Esercizio 13. Se un tale insieme Z esiste, è unico a meno di isomorfismo. In particolare, possiamo chiamare Z direttamente $X \times Y$.

Svischeriamo la definizione:

- Le funzioni f e g servono a indicare che a ogni elemento di S è associata una certa coppia di elementi di X e Y, determinata da f e g.
- L'esistenza di h significa che, per qualsiasi scelta di elementi di X e Y, esiste un elemento in Z che vi corrisponde. Le proiezioni p e q ci dicono a quale X e quale Y corrisponde ogni elemento.
- Il fatto che h sia unico significa che esiste un solo elemento di questo tipo per ogni coppia, perché altrimenti si potrebbe scegliere a quale elemento h punta.

Insomma, tutto sommato questo dice che $X \times Y$ ha esattamente un elemento in esso per ogni scelta di una coppia di elementi di X e Y, e p e q ci dicono quale coppia è. Successo!

Esercizio 14. Il prodotto cartesiano di X e Y dipende solamente dalla *classe di isomorfismo* di X e Y. In altre parole, se X è isomorfo a X' e Y è isomorfo a Y', allora $X \times Y$ è isomorfo a $X' \times Y'$. Riuscite a dimostrarlo usando solo proprietà universali?

32.3.2 I limiti, per davvero

La caratterizzazione del prodotto cartesiano di due insiemi tramite una proprietà universale è istanza di un concetto più generale in teoria delle categorie, chiamato *limite*. (Attenzione: si tratta di un concetto solo lontanamente correlato ai limiti che potreste aver visto nel calcolo, quindi se è questo che vi è saltato in mente, fate pure e dimenticatelo di nuovo.)

La definizione rigorosa di limite prevede quella di *categoria* e di *funtore* tra due categorie, ma, per non rendere questo articolo troppo lungo, ci limiteremo a dare un'idea di cosa è un limite in generale e a dare una definizione precisa solo in casi molto particolari.

Definizione 17. Sia $n \geq 1$ un numero naturale, e consideriamo la collezione di insiemi $\{X_i\}_{i=1}^n$. Il *limite* di $\{X_i\}_{i=1}^n$ è un insieme Z e funzioni $p_i \colon Z \longrightarrow X_i$ tale che, per ogni altro insieme S e funzioni $f_i \colon S \longrightarrow X_i$, esiste un'unica funzione $h \colon S \longrightarrow Z$ tale che $f_i = p_i \circ h$ per ogni $i = 1, \dots, n$.

Remark. Il limite di $\{X_i\}_{i=1}^n$ non è altro che... il prodotto cartesiano $Z = X_1 \times X_2 \times \cdots \times X_n$, e le funzioni $p_i \colon Z \longrightarrow X_i$ non sono altro che le proiezioni. Come avevamo osservato in precedenza, il prodotto cartesiano è unico a meno di isomorfismo. In effetti, anche qua osserviamo che il prodotto cartesiano di insiemi non è strettamente *unico*: in particolare, ogni permutazione dell'ordine dei fattori in $X_1 \times \cdots \times X_n$ ha la stessa proprietà universale.

Finora abbiamo considerato il limite di una collezione di insiemi $\{X_i\}_i$, ma possiamo complicare la storia e aggiungere *relazioni* tra gli insiemi della nostra collezione, ovvero *funzioni* che collegano un certo X_i ad un altro X_j; chiamiamo *diagramma* una qualsiasi collezione di insiemi e di mappe tra questi. Vediamo un caso particolarmente semplice ma impo

Definizione 18. Siano X e Y due insiemi, e $f, g \colon X \longrightarrow Y$ due funzioni da X in Y. Consideriamo il diagramma dato da

$$X \underset{g}{\overset{f}{\rightrightarrows}} Y$$

Il *limite* di tale diagramma è il dato di un insieme Z, una funzione $\varphi \colon Z \longrightarrow X$ tale che $f \circ \varphi = g \circ \varphi$, e con la seguente proprietà universale: per ogni altro insieme S e funzione $k \colon S \longrightarrow X$ per cui $f \circ k = g \circ k$, esiste un unica mappa $h \colon S \longrightarrow Z$ tale per cui $\varphi \circ h = k$.

Diciamo che (Z, φ) è l'*equalizzatore* delle mappe $f, g \colon X \longrightarrow Y$.

Visivamente, possiamo illustrare quanto appena scritto come segue:

$$S \overset{\exists! h}{\dashrightarrow} Z \xrightarrow{\varphi} X \underset{f}{\overset{g}{\rightrightarrows}} Y \qquad (k : S \to X)$$

Consideriamo un caso particolare della Definizione 18, in cui $X = Y$, $f \colon X \longrightarrow X$ è una mappa qualsiasi e $g = \mathrm{id}_X$. Dunque il diagramma è semplicemente dato da

$$X \underset{\mathrm{id}_X}{\overset{f}{\rightrightarrows}} X$$

Esercizio 15. Quale è l'equalizzatore di $f \colon X \longrightarrow X$ e dell'identità di X ?

Sdipanando la definizione, arriviamo alla seguente caratterizzazione:

Il limite è dato da un insieme Z e una funzione $\varphi\colon Z \longrightarrow X$, tale per cui $f \circ \varphi = f$ e per ogni altro insieme S e funzione $k\colon S \longrightarrow X$ in cui $f \circ k = k$, esiste un'unica funzione $h\colon S \longrightarrow Y$ tale che $\varphi \circ h = k$.

Ma cosa significa veramente?

- $f \circ \varphi = f$ significa che φ non cambia nulla nel risultato di f. In altre parole, per qualsiasi x della forma $x = \varphi(y)$, si deve avere $f(x) = x$. Dunque l'immagine di φ è data dai valori che rimangono invariati dalla funzione f: questi sono chiamati **punti fissi** (o talvolta *punti stazionari*). Quindi φ può puntare solo ai punti fissi di f.

- Secondo la stessa logica, qualsiasi scelta di k può puntare solo ai punti fissi di k. In particolare, è possibile scegliere una funzione k che punti a qualsiasi sottoinsieme di punti fissi di X.

- Che h esista, quindi, significa che l'immagine di φ deve essere data da *tutti* i punti fissi di f. (Se per assurdo esistesse $x \in X$ punto fisso non nell'immagine di φ, allora potremmo scegliere $k\colon \{*\} \longrightarrow X$, $k(*) = x$, e k non potrebbe essere ottenuta componendo con φ, assurdo).

- Il fatto che h sia unico significa che φ deve essere *iniettiva*.

Quindi Z è in corrispondenza uno-a-uno con i punti fissi di f, e φ è quella corrispondenza uno-a-uno. In sostanza, *a meno di isomorfismo*, **l'equalizzatore di $f\colon X \longrightarrow X$ e dell'identità $\mathrm{id}_X\colon X \longrightarrow X$ è dato dai punti fissi di f, e la mappa $Z \longrightarrow X$ è l'inclusione.**

Più in generale, dato un diagramma qualsiasi, possiamo sempre chiederci quale sia il suo *limite*, ovvero quell'insieme Z e quelle mappe da Z negli insiemi del diagramma in questione che sono *universali*.

Per esempio, possiamo considerare il caso estremo dato dal diagramma *vuoto*, ovvero senza insiemi né funzioni tra di essi. Bene, allora il *limite* di questo diagramma deve essere solamente un insieme Z con la seguente proprietà universale: per qualsiasi altro insieme S, esiste un'unica funzione $f\colon S \longrightarrow Z$. Se questo vi suona familiare non è un caso: questa era la risposta alla prima domanda del nostro gioco in cui abbiamo descritto gli insiemi con un solo elemento! In teoria delle categorie, il limite del diagramma vuoto si chiama *oggetto terminale*.

Con questo nuovo linguaggio, dunque, possiamo dire che nella categoria degli insiemi l'oggetto terminale è l'insieme con un unico elemento, unico a meno di isomorfismo.

Esercizio 16. Quale è la proprietà universale del limite del diagramma di insiemi $X \xrightarrow{f} Y \xleftarrow{g} W$? Qual è il limite del diagramma di sopra quando scegliamo $X = \{*\}$ come l'insieme con un unico elemento?

Questi sono solo esempi di diagrammi di cui possiamo provare a calcolare il limite, ed è sorprendente quante costruzioni a noi ben note non siano altro che istanze della più generale nozione di limite. Sebbene i diagrammi possano diventare arbitrariamente complicati, un certo principio di semplicità sottende la nozione di limite di un diagramma, e scopriamo che prodotti ed equalizzatori sono *sufficienti* per calcolare il limite di un diagramma qualsiasi. Questo è il contenuto di un teorema fondamentale di teoria delle categorie.

Theorem 9. Il limite di un qualsiasi diagramma di insiemi può essere espresso come *equalizzatore* di due funzioni tra *prodotti* di insiemi.

32.4 Tirando le somme

La teoria delle categorie è un linguaggio per descrivere uniformemente costruzioni matematiche, e quella di oggi non era altro che la voce *limite* sul vocabolario, fatto di molte altre pagine. Studiarne e comprenderne la grammatica è un eccellente esercizio di astrazione, e permette di avere uno sguardo più ampio e trasversale nella matematica. Come tutti i linguaggi, d'altronde, la teoria delle categorie è un *mezzo*, e come concludere se non indicandovi alcune delle (numerose!) applicazioni che si possono fare di questa disciplina?

- provare a pensare in modo categorico la teoria degli insiemi, sostituendo la nozione primitiva di appartenenza con quella di *funzione*, non è semplicemente un gioco: portando avanti in modo rigoroso la teoria, si ottiene un tipo alternativo di teoria degli insiemi per la matematica, chiamata "Teoria Elementare della Categoria degli Insiemi", o in breve *ETCS* (per l'acronimo della traduzione in inglese). Se nella tradizionale teoria degli insiemi di Zermelo-Frenkel, si assume che esista una nozione di appartenenza e si lavora per costruire coppie ordinate, relazioni e funzioni. Nell'ETCS, invece, si assume una nozione di funzione e si lavora a ritroso per recuperare coppie ordinate e simili.

- questo stesso approccio, motivato stavolta da considerazioni inerenti alla branca della *logica matematica*, ha dato luogo alla disciplina della *teoria dei tipi*, sviluppatasi parallelamente alla teoria delle categorie, e, successivamente, della *teoria omotopica dei tipi* (*Homotopy Type Theory*, o *HoTT*). Queste teorie si collocano a cavallo tra logica matematica, informatica (precisamente la branca del *calcolo*) e, con l'avvento di HoTT, della topologia algebrica. Personal fact: lo studio di queste teorie è stato il centro focale della mia tesi triennale!

- riformulare costruzioni e proprietà di insiemi senza riferimento diretto agli elementi ha il vantaggio di permettere di dare le stesse definizioni in qualsiasi altra *categoria*, cioè un insieme di oggetti e di frecce tra di esse che si compongono in modo simile alla composizione di funzioni. Si può, ad esempio, considerare qualsiasi struttura algebrica (*gruppi*, *anelli*, ecc.) e gli omomorfismi tra di loro, o gli *spazi topologici* e le *funzioni continue* tra di loro. In particolare, in una qualsiasi categoria è possibile dare la nozione di *isomorfismo* tra due oggetti, e lavorare con gli oggetti di una categoria *a meno di isomorfismo*. Talvolta, però, può essere necessario considerare due oggetti di una categoria *equivalenti* non solo quando sono isomorfi. Per gli insiemi questo non è molto interessante, ma lo diventa non appena consideriamo oggetti con un po' più di struttura. Ad esempio, dati due *spazi topologici* (oggetti geometrici muniti di una certa nozione di *continuità*), possiamo domandarci se uno possa essere *deformato in modo continuo* fino ad arrivare a coincidere con l'altro. Se questo avviene, non diciamo più che i due spazi sono isomorfi (perché non c'è più una "corrispondenza uno-a-uno"), ma diciamo piuttosto che i due spazi topologici sono *omotopi*. Sono queste le considerazioni che hanno dato luogo alla *teoria dell'omotopia*, una branca della topologia algebrica che si occupa di studiare oggetti a meno di nozioni più deboli di isomorfismo. È una delle discipline della matematica in cui più sono applicati i metodi ed il linguaggio della teoria delle categorie!

Insomma, potremmo continuare ancora, dando più dettagli, espandendo definizioni, enunciando altri teoremi ed applicazioni del linguaggio che è la teoria delle categorie. Ci fermiamo qui, però, sperando di avervi incuriosito abbastanza da volerne sapere di più!

Riferimenti bibliografici

[1] The Univalent Foundations Program (2013) *Homotopy Type Theory: Univalent Foundations of Mathematics*, Institute for Advanced Study.

[2] Hatcher, Allen (2002) *Algebraic Topology*, Cambridge University Press.

[3] Riehl, Emily (2014) *Category Theory in Context*, `https://math.jhu.edu/~eriehl/context.pdf`

[4] Sito NLAB, ETCS, `https://ncatlab.org/nlab/show/ETCS`

33 Cosa sono i numeri di Betti

Luca Bruni, David Vencato, Jacopo Burelli, n.19, Febbraio 2025

33.1 Nota storica sui numeri di Betti

Enrico Betti nasce il 21 ottobre 1823 a Pistoia, dove compie gli studi classici al Liceo Forteguerri. Studia matematica e fisica all'Università di Pisa dove, nel 1846, si laurea in matematiche applicate, sotto la direzione scientifica di Giuseppe Doveri. Nei primissimi anni della vita scientifica di Enrico Betti, significativa sarà l'influenza di Mossotti. Questi, per esempio, scrivendogli da Viareggio nell'estate del 1847, scoraggia un iniziale interesse di Betti per la geometria descrittiva. Seguendo il severo (e forse un po' azzardato) giudizio del maestro, Betti lascia lo studio della geometria descrittiva per dedicarsi alla fisica matematica, in particolare ad un problema di idrodinamica, dato alle stampe nel 1850 negli "Annali di Scienze matematiche e fisiche", la nuova rivista fondata in quell'anno a Roma da Barnaba Tortolini. Questo primo lavoro di Betti doveva restare a lungo isolato nella sua produzione scientifica, fino al ben più significativo incontro con Riemann. Dal 1849 Betti insegna matematica al Liceo Forteguerri di Pistoia, sua città natale. Lontano dalla consuetudine di scambio quotidiano con gli antichi maestri, le sue ricerche acquistano un carattere decisamente più autonomo e originale. A questo periodo risalgono infatti i suoi primi studi di algebra, che lo occuperanno per tutto un decennio.

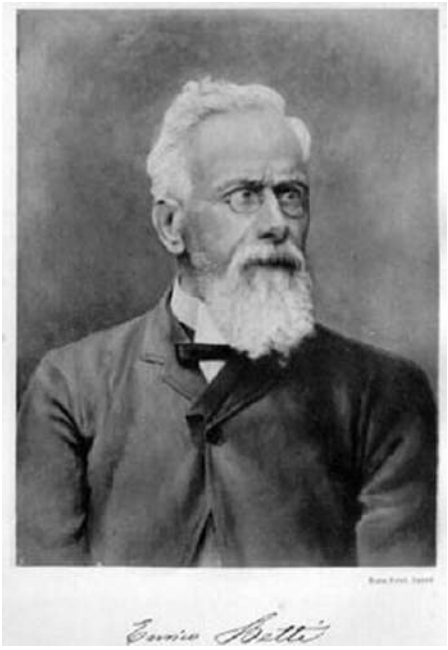


Figura 103: Enrico Betti

Punto di partenza di Betti sono i lavori di Galois sulla risolubilità per radicali delle equazioni algebriche, in quanto gli sviluppi promessi da Liouville sui lavori di Galois si lasciano inutilmente attendere. Così la prima effettiva ripresa delle idee di Galois è opera di Betti, nel corso del 1850. Betti è certamente consapevole della novità e dell'importanza dei lavori di Galois, alla cui chiarificazione sono destinate le sue ricerche.

All'estero gli studi di Betti non sembrano suscitare particolare interesse: l'argomento è nuovo e difficile e la forma dell'esposizione, se più chiara di quella di Galois, non è certamente trasparente. A ciò si aggiunga ancora il non banale inciampo costituito dalla lingua italiana in cui è redatto il lavoro. Di fatto, l'interlocutore scientifico di Betti capace di entrare con competenza nel suo ambito di ricerche è in quel periodo, in Italia, è Tardy ed è nella corrispondenza fra i due che si possono seguire gli sviluppi degli studi di Betti. Betti pubblica ancora negli "Annali" un lavoro sulla teoria delle sostituzioni, in cui si limita a chiarire alcuni punti della sua lontana memoria del 1852. Oltre alla intrinseca difficoltà di queste ricerche, un ulteriore ostacolo sembra essere di natura estranea ad esse: in quell'anno infatti Betti lascia l'insegnamento a Pistoia per assumere il ruolo più impegnativo di professore di Algebra Superiore al Liceo di Firenze. Connessa a questa nuova attività didattica, Betti intraprende la traduzione del volume di Algebra di Bertrand, che viene pubblicato con le sue aggiunte e note nel 1856. Nel 1857, Betti ottiene la cattedra di Algebra all'Università di Pisa.

In quegli anni, oltre alle lezioni universitarie, Betti tiene a casa propria, due volte la settimana, lezioni private per esporre a quattro dei suoi migliori studenti "le parti più elevate dell Algebra che non posso esporre nel corso che fò all'università". Alla primavera del 1855 risalgono le ultime ricerche di Betti nell'ambito della teoria delle equazioni: annunciate in una lettera a Sylvester, conosciuto allora a Firenze, appaiono negli "Annali" di Tortolini. Da questo momento in poi, fino verso la fine degli anni Cinquanta, Betti, pur continuando a lavorare nel campo dell'algebra, si dedica ad argomenti che rientrano nel più consolidato filone delle ricerche algebriche del tempo, dalla teoria delle serie alla teoria degli invarianti delle forme binarie alla teoria delle funzioni simmetriche delle radici di un equazione algebrica. Ed è proprio con una breve nota su quest'ultimo argomento che Betti raggiunge il primo effettivo riconoscimento all'estero, con la pubblicazione sulla prestigiosa rivista di Crelle. Ritornato da un soggiorno a Gottinga, nel quale insieme ai colleghi Brioschi e Casorati conosce Dedekind, Dirichlet e Riemann, Betti inizia ad interessarsi ai lavori di quest'ultimo.

Gli studi sulla teoria riemanniana delle funzioni di una variabile complessa si intrecciano allora, per Betti, con una temporanea ripresa dei suoi precedenti lavori sulla teoria delle equazioni. L'interesse prevalente di Betti è adesso rivolto all'analisi e, chiamato dalla fine del 1859 alla cattedra di Analisi Superiore, fa della teoria delle funzioni ellittiche l'argomento delle sue lezioni all'università. L'influenza di Riemann è accentuata dalla permanenza a Pisa di questi. Infatti Riemann soffriva di una forma acuta di tubercolosie negli ultimi anni della sua vita fece lunghi viaggi in Italia (e in particolare a Pisa) cercando sollievo nel mite clima mediterraneo.

Gli argomenti che animavano gli incontri e le discussioni tra Riemann e i matematici pisani dovevano abbracciare uno spettro estremamente ampio di questioni, di cui sono rimaste solo poche tracce nel Nachlass di Riemann e nella corrispondenza di Betti. Non è dunque un caso che proprio durante il soggiorno di Riemann a Pisa, quando il contatto fra i due è più stretto e fecondo, Betti riprenda a occuparsi di fisica matematica, che a partire da quegli anni fino alla morte egli fa oggetto principale delle proprie ricerche oltre che delle lezioni all'università accanto a quelle di Analisi Superiore. Tuttavia, anche se la presenza di Riemann può aver suggerito elementi di riflessione e di ricerca su aspetti particolari di questa o quella teoria, l'influenza più significativa e duratura esercitata dal matematico tedesco su Betti appare piuttosto essere stata di natura generale, sul modo di intendere la matematica come scienza unitaria, strettamente legata alla conoscenza del mondo fisico. In questa concezione riemanniana, che Betti fa propria, si trovano infatti i motivi ispiratori e le convinzioni più profonde che animano il terzo e più lungo periodo della sua attività scientifica, completamente dedicato alla fisica matematica, dopo le iniziali ricerche di carattere algebrico e il breve periodo (1859-1863) in cui egli si era prevalentemente occupato di questioni di analisi.

33.2 Topologia e Numeri di Betti per oggetti geometrici

33.2.1 Esempi introduttivi

Dopo aver esplorato la vita Di Enrico Betti, vogliamo ripercorrere i suoi studi fino alla definizione di uno degli oggetti più famosi della sua ricerca: *i numeri di Betti* (appunto). Ci lasceremo guidare da esempi pratici ed esercizi che ci introdurranno alla *topologia* e alla definizione di questi numeri. La trattazione è volutamente semplificata per fare uso di poche nozioni di matematica moderna, ma risulta essere completa anche se non abbraccia la topologia in tutti i suoi aspetti.

Partiamo con alcuni esempi:

Siano dati un disco pieno D^2 e un punto P. È facile capire perché questi due oggetti geometrici non sono *omeomorfi*[57]: se infatti esistesse un omeomorfismo tra i due, allora i due oggetti dovrebbero avere lo stesso numero di punti, ma D^2 ne ha infiniti, il punto P è costituito da un unico punto.

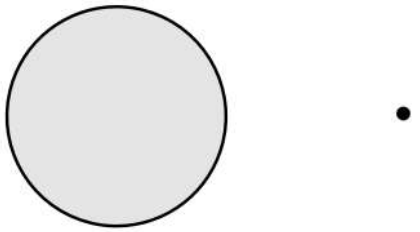


Figura 104: Un disco pieno D^2 e un punto P

Siano dati due oggetti geometrici X e Y con un numero differente di *componenti connesse*[58] che li distinguono. Allora questi non sono omeomorfi in quanto la deformazione che dovrebbe portare l'uno nell'altro dovrebbe strapparsi.

Sulla base di questi esempi invitiamo a riflettere sul seguente esercizio:

Esercizio 8. *Se prendiamo la retta* $\mathbb{R}$ *e la circonferenza* S^1 *come faccio a distinguerle?*

E se prendiamo una sfera S^2 e un toro (la superficie di una ciambella) $S^1 \times S^1$ come potremmo distinguerli? Intuitivamente percepiamo che sono oggetti diversi in quanto il toro, oltre a suddividere lo spazio $\mathbb{R}^3$, definisce anche dei cerchi sulla sua superficie che è impossibile "sciogliere" (formalmente contrarre a un punto).

L'obiettivo di questo articolo è ripercorrere le idee di Betti nell'introduzione di un invariante per omeomorfismo[59] che permetta di distinguere oggetti geometrici.

[57] Questa è la parola che useremo per dire che due oggetti sono uguali. In parole semplici, se posso portare un oggetto in un altro senza strapparlo. Vedremo la definizione formale più avanti.

[58] Una componente connessa è, come potrete immaginare, una parte di un oggetto che può essere percorsa continuando a rimanere sempre dentro l'oggetto

[59] Più precisamente per omotopia, ma lo vedremo più avanti

La questione di distinzione di oggetti geometrici è legata alla topologia. In questo articolo non daremo gli assiomi precisi della topologia, ci accontenteremo di dire che regolano le trasformazioni geometriche *continue*, ovvero, informalmente, trasformazioni senza strappi.

33.2.2 Oggetti geometrici e omotopia

Fino a questo momento abbiamo parlato di *oggetti geometrici*, senza aver dato una definizione formale di cosa stiamo parlando. Di seguito cerchiamo di chiarire di cosa stiamo parlando.

Definizione 19. Diremo che un sottoinsieme X di $\mathbb{R}^n$ è una *varietà* se per ogni punto $x \in X$ esiste un intorno di x che è omeomorfo a $\mathbb{R}^k$ per qualche $k \in \mathbb{N}$. Informalmente, una varietà è un sottoinsieme dello spazio tale che, se lo guardi da vicino, è come essere nello spazio $\mathbb{R}^k$

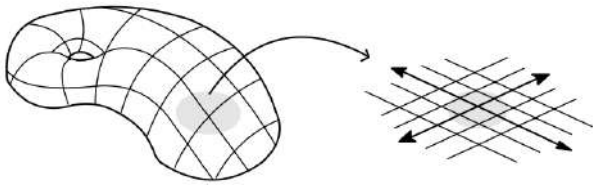


Figura 105: Una varietà 2-dimensionale

La Figura 105 rende bene l'idea, ma per avere un altro esempio potete pensare ad esempio al nostro pianeta Terra

Per interiorizzare questa nozione può essere utile pensare che se si sta camminando su una varietà che corrisponde a una superficie (varietà 2-dimensionale) significa che se mi trovo in un punto della varietà, ho l'impressione di essere su un piano. Di nuovo pensare al Pianeta Terra rende bene l'idea.

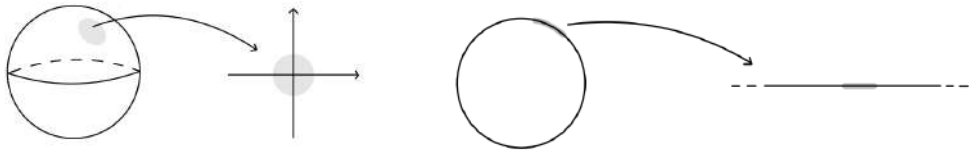


Figura 106: A sinistra una sfera, varietà 2-dimensionale. A destra una circonferenza, varietà 1-dimensionale.

Definizione 20. Diremo che un sottoinsieme X di $\mathbb{R}^n$ è un *oggetto geometrico* se è una unione finita di varietà che si intersecano a due a due in una varietà.

Lasciamo al lettore qualche semplice esercizio per prendere confidenza con la nozione appena introdotta

Esercizio 9. *Quali lettere dell'alfabeto (scritte in maiuscolo) sono delle varietà?*

Esercizio 10. *Qual è un oggetto geometrico che* non *è una varietà?*

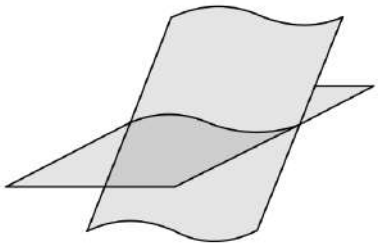


Figura 107: Un oggetto geometrico: l'intersezione tra due piani.

Vediamo adesso quali sono le giuste nozioni per il confronto tra oggetti geometrici.

Definizione 21. Diremo che due oggetti geometrici X, Y sono *omeomorfi* se $\exists f : X \longrightarrow Y$ funzione continua, bigettiva con inversa continua. Intuitivamente se posso deformare un oggetto in un altro senza strappi.

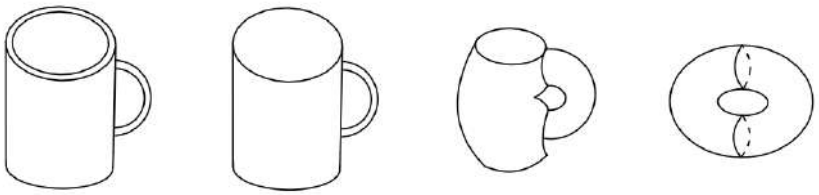


Figura 108: Un omeomorfismo tra una tazza ed un toro.

[width=boxrule=0.3mm, sharp corners] **Attenzione:**
Se la tazza avesse avuto il manico fatto solamente da una linea e non da un manico con volume, avremmo potuto concludere allo stesso modo?

Definizione 22. Siano $f, g : X \longrightarrow y$ due funzioni continue tra oggetti geometrici. Una *omotopia* tra f e g è una mappa continua $H \colon X \times [0,1] \longrightarrow Y$ tale che $H(x,0) = f$ e $H(x,1) = g$.

Analogamente diremo che una omotopia tra f e g è una famiglia di funzioni $f_t : X \to Y$ con $t \in [0,1]$ tali che $f_0 = f$ e $f_1 = g$.

Diremo che $\phi : X \longrightarrow Y$ è una *equivalenza omotopica* se esiste $\psi : Y \longrightarrow X$ tale che $\psi \circ \phi$ è omotopa a id_X e $\phi \circ \psi$ è omotopa a id_Y.

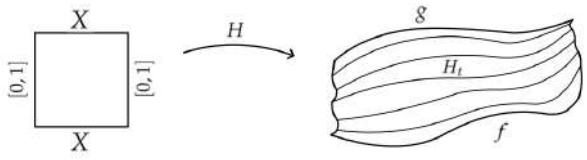


Figura 109: Omotopia come interpolazione tra una mappa f a una mappa g

La definizione risulta essere molto astratta, ma potete immaginarla come in Figura 109: una omotopia tra due funzioni è uno *scivolo continuo* di grafici tra il grafico della prima funzione e il grafico della seconda. Attenzione: questa idea è puramente intuitiva e serve come visualizzazione; ricordatevi che a volte il grafico non è possibile disegnarlo perché potrebbe essere più che 3-dimensionale

Definizione 23. Diremo che due oggetti geometrici X, Y sono *omotopicamente equivalenti* se esiste $f : X \longrightarrow Y$ equivalenza omotopica.

Remark. Il segmento $[0, 1]$ è omotopicamente equivalente a un punto. Il disco $\mathbb{D}^n$ è omotopicamente equivalente al punto.

Remark. Due oggetti omeomorfi sono omotopicamente equivalenti. La nozione di omeomorfismo è più rigida rispetto all'omotopia: per convincervi provate a pensare a una corona circolare e a una circonferenza: i due oggetti geometrici sono omotopicamente equivalenti, ma non omeomorfi

33.2.3 Complessi simpliciali e numeri di Betti

A questo punto vogliamo trovare un invariante per gli oggetti geometrici che sia ben definito e che sia facilmente calcolabile. Per farlo introduciamo i *complessi simpliciali*, che sono una approssimazione tramite segmenti, triangoli, tetraedri, etc. di oggetti geometrici.

Definizione 24. Il *simplesso $n-$dimensionale standard* è

$$\Delta^n = \left\{ (t_1, \cdots, t_{n+1}) \subseteq \mathbb{R}^{n+1} | \sum_{1 \leq i \leq n+1} t_i = 1, \quad 0 \leq t_i \leq 1 \right\} \tag{28}$$

e una sua *faccia* è un $n - k$ simplesso definito come:

$$F_{i_1,\ldots,i_k} = \left\{ (t_1, \cdots, t_{n+1}) \subseteq \Delta^n | t_{i_j} = 0 \text{ se } 1 \leq j \leq k \right\} \tag{29}$$

Un *simplesso* è un riscalamento delle facce che ne mantiene la forma essenziale.

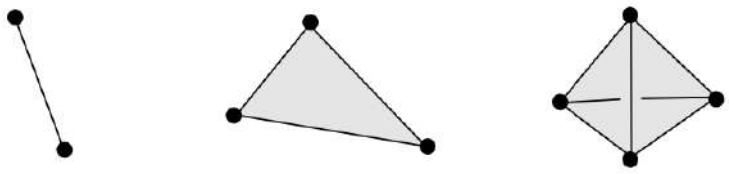


Figura 110: Simplessi di dimensione 1, 2 e 3

In Figura 110 è possibile osservare dei simplessi. In particolare:

- un simplesso $1-$dimensionale è un segmento.
- Un simplesso $2-$dimensionale è un triangolo pieno.
- Un simplesso $3-$dimensionale è un tetraedro pieno.

Quello che vogliamo fare adesso, è costruire oggetti attaccando tra di loro simplessi con la stessa logica con cui abbiamo definito gli gli oggetti geometrici.

Definizione 25. Un *complesso simpliciale* X è un sottoinsieme di $\mathbb{R}^N$ definito come una unione finita di simplessi che soddisfa le seguenti due proprietà:

1 Ogni faccia di un simplesso di X è ancora in X.

2 L'intersezione di due simplessi $X_1, X_2 \in X$ è una faccia sia di X_1 che di X_2.

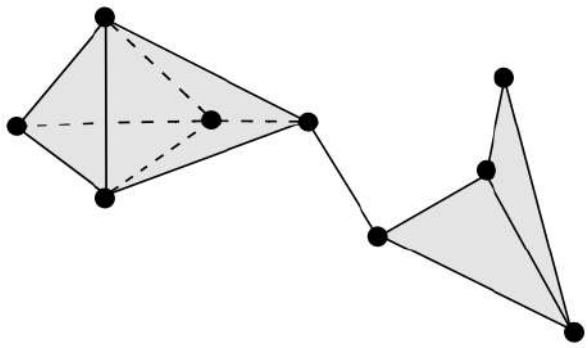


Figura 111: Complesso simpliciale

Siamo pronti a definire cosa sono i numeri di Betti per complessi simpliciali:

Definizione 26. Dato un complesso simpliciale X i *numeri di Betti* di X che indicheremo con $b_0(X), b_1(X), \cdots$ sono gli unici numeri che verificano le seguenti proprietà:

1 Se X è omotopicamente equivalente a Y allora $b_i(X) = b_i(Y)$ per ogni $i \geq 0$. In altre parole, i numeri di Betti sono invarianti per equivalenza omotopica.

2 $b_0(pt) = 1, b_i(pt) = 0$ per ogni $i \neq 0$; $b_i(\emptyset) = 0$ per ogni i e infine $b_0(X)$ conta il numero di componenti connesse di X.

3 Verificano la proprietà di Mayer-Vietoris: se $X = X_1 \cup X_2$ con X_1, X_2 complessi simpliciali, detta $Y = X_1 \cap X_2$, allora abbiamo una successione di numeri

$$\begin{aligned} \cdots \longrightarrow b_2(X) \longrightarrow b_1(Y) \longrightarrow b_1(X_1) + b_1(X_2) \longrightarrow b_1(X) \longrightarrow \\ \longrightarrow b_0(Y) \longrightarrow b_0(X_1) + b_0(X_2) \longrightarrow b_0(X) \longrightarrow 0 \end{aligned}$$

tale che

a) La somma alternata dei numeri tra due 0 deve fare 0.

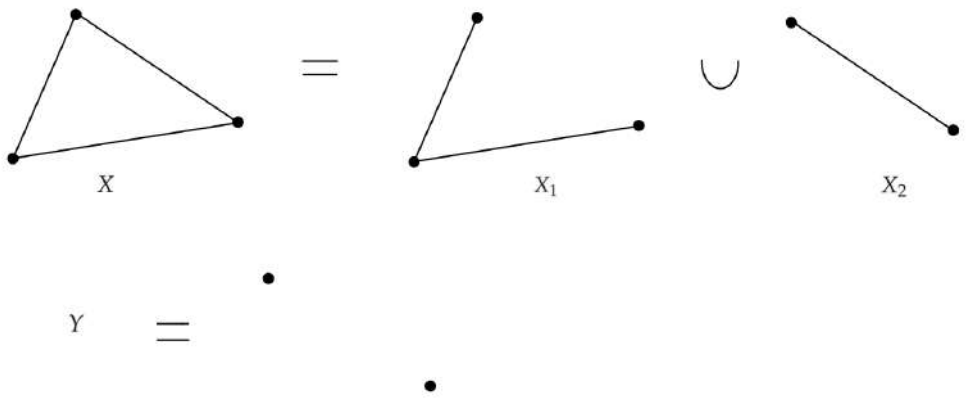
b) "Surgettività": se $a \longrightarrow b \longrightarrow 0$ allora $a \geq b$.

c) "Iniettività": se $0 \longrightarrow a \longrightarrow b$ allora $a \leq b$.

Usiamo la definizione per calcolare i numeri di Betti di un triangolo. Spezziamo il triangolo come due segmenti e un segmento che si intersecano in due punti. Chiamando l'intersezione Y si ha $b_0(Y) = 2$ e $b_i(Y) = 0$ per ogni $i \geq 1$.

Applicando adesso la proprietà di Mayer-Vietoris si ha:

$$\begin{aligned} b_2(Y) \longrightarrow \underbrace{b_2(X_1) + b_2(X_2)}_{0} \longrightarrow b_2(X) \longrightarrow \underbrace{b_1(Y)}_{0} \longrightarrow \underbrace{b_1(X_1) + b_1(X_2)}_{0} \longrightarrow \\ \longrightarrow \underbrace{b_1(X)}_{?} \longrightarrow \underbrace{b_0(Y)}_{2} \longrightarrow \underbrace{b_0(X_1) + b_0(X_2)}_{2} \longrightarrow \underbrace{b_0(X)}_{?} \longrightarrow 0 \end{aligned}$$



da cui $b_0(X) = b_1(X)$ e dunque $b_1(X) = 1$.

Per poter calcolare i numeri di Betti di un oggetto geometrico qualsiasi l'idea è quella di triangolarlo, ovvero di approssimarlo come fosse un complesso simpliciale.

Definizione 27. Una *triangolazione di* M è una suddivisione di M in $n-$simplessi (i.e. immagine di $n-$simplessi). Formalmente è una mappa $f : K \longrightarrow M$ omeomorfismo, con K complesso simpliciale.

Una *triangolazione di un oggetto geometrico* X è una triangolazione di ogni varietà di cui è composto e una triangolazione per ogni intersezione tra varietà.

Concretamente parlando, una triangolazione di una superficie è un ricoprimento di una superficie tramite triangoli che si intersecano a due a due o in un lato o in un vertice. Se invece voglia triangolare una linea, è un ricoprimento di tali linea travite segmenti che si intersecano in vertici.

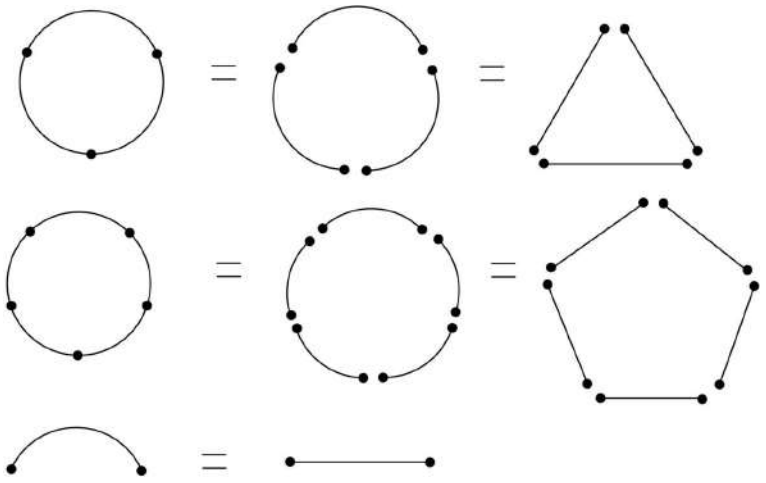


Figura 112: Possibili triangolazioni della circonferenza S^1

Siamo finalmente arrivati alla definizione dei numeri di Betti per oggetti geometrici.

Definizione 28. I numeri di Betti di un oggetto geometrico X sono i numeri di Betti di una triangolazione associata a X o a un oggetto geometrico Y omotopicamente equivalente a X, ovvero i numeri di Betti del complesso simpliciale con cui è stato triangolato lui o un suo rimpiazzamento Y.

[width=boxrule=0.3mm, sharp corners] **Attenzione:**
Non è detto che un oggetto geometrico sia triangolabile! Si pensi ad esempio al piano $\mathbb{R}^2$.

A priori la definizione non è ben posta perché potrebbe capitare che due triangolazioni diverse forniscono numeri di Betti differenti. Questo in realtà non accade: provare, per esercizio, a triangolare S^1 non con soli 3 segmenti, ma con 5 o con 2 e a calcolare ogni volta i numeri di Betti.

Il prossimo Teorema, di cui non diamo dimostrazione, fornisce la buona definizione dei numeri di Betti per oggetti geometrici

Theorem 10. I numeri di Betti sono ben definiti: ovvero scegliendo triangolazioni diverse per uno stesso oggetto geometrico, i numeri di Betti non cambiano.

Concludiamo la sezione con la seguente osservazione di cui il lettore è invitato a verificarne la veridicità (o semplicemente a convincersene) con gli esercizi proposti a fine articolo.

Remark. Geometricamente, il k-esimo numero di Betti conta il numero di buchi k-dimensionali di un oggetto.

Infine, per onestà intellettuale, facciamo presente che i numeri di Betti hanno una definizione molto più astratta legata all'omologia. Sono definiti come la dimensione di una ben definita struttura algebrica. Nel caso degli oggetti geometrici, la definizione data coincide con quella generale, ma nel caso di oggetti più complessi è necessario introdurre il (difficile) concetto di omologia.

33.3 Persistent Homology e Topological Data analysis

In questa sezione, vedremo una applicazione dei numeri di Betti per il riconoscimento di alcune caratteristiche di un oggetto a partire da una nuvola di punti che sappiamo far parte dell'oggetto.

Lo studio di dati e la determinazione di caratteristiche geometriche tramite tecniche topologiche prende il nome di *topological data analysis* e combina tecniche di geometria con la potenza dei calcolatori.

L'applicazione che presentiamo, che prende il nome di *persistent homology* ha come uno dei padri ancora un pistoiese, Patrizio Frosini.

Per fissare le idee immaginiamo di avere a disposizione un laser per individuare un meteorite in movimento. Non riusciamo a capire la forma completa del meteorite perché riusciamo solo ad ottenere le coordinate di alcuni punti. Cosa possiamo dedurre? Come potrebbe essere fatto il meteorite?

Tramite i numeri di Betti e il procedimento che andremo a vedere riusciremo ad ottenere informazioni sul meteorite anche se non è direttamente osservabile.

Per semplicità prendiamo un insieme di punti nel piano e non nello spazio.

Descriviamo operativamente in che cosa consiste lo studio di una nuvola di punti

1. Si prende una nuvola di punti nello spazio (o nel piano).
2. Fissato un raggio d si connettono i punti a distanza minore o uguale a d.
3. Si calcolano i numeri di Betti del complesso simpliciale ottenuto.

Le problematicità di questo metodo sono evidenti:

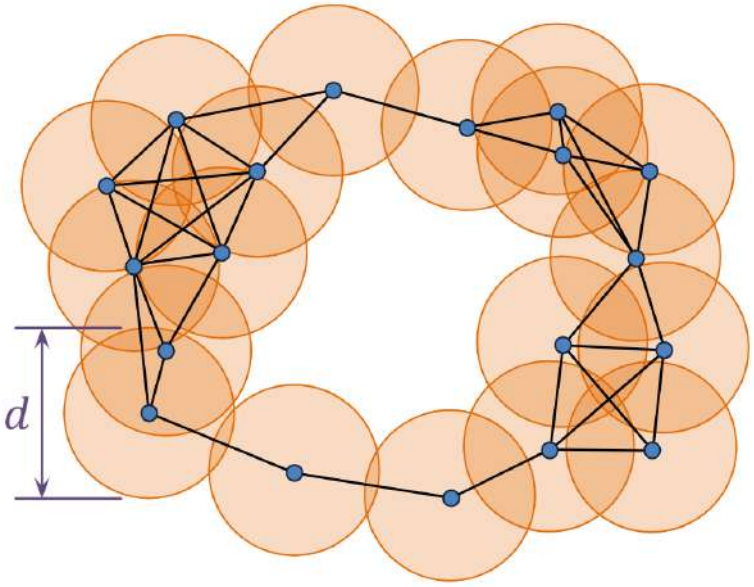


Figura 113: Algoritmo esposto per una generica nuvola di punti

- Se d viene scelto troppo grande o troppo piccolo non si riesce ad individuare la giusta geometria
- Otteniamo solamente un grafo e, come conseguenza dell'Esercizio 33.4, solamente lo 0-esimo e il primo numero di Betti diversi da 0.
- Se è presente del *rumore*, ovvero delle misurazioni erronee, tali misurazioni potrebbero risultare significative e portare a deduzioni non corrette.

[width=boxrule=0.3mm, sharp corners] **Idea chiave:**
Non guardo i numeri Betti fissata una singola distanza d, ma faccio variare d e cerco di monitorare la *persistenza* di alcuni numeri di Betti.

Per fare in modo che non si crei solamente un grafo, ogni qualvolta un insieme di punti formano un triangolo o un tetraedro (i.e. 3 o 4 punti sono distano a coppie almeno d) riempiamo il triangolo o il tetraedro da loro definito.

Se si aumenta la dimensione dello spazio in cui viviamo, ovvero abbiamo una nuvola di punti in uno spazio n- dimensionale, riempiamo ogni scheletro di simplesso n-dimensionale che si viene a formare.

Definizione 29. Fissato $d \in \mathbb{R}^+$ il *Complesso di Rips* è il complesso simpliciale R_d costruito tramite l'idea operativa definita sopra.

Vediamo come operativamente cambia il nostro algoritmo:

Fissiamo un $D \in \mathbb{R}^+$ (ad esempio $D = \max\{\text{dist}(x, y) \mid x, y \text{ punti della nuvola}\}$) e per ogni $0 < d < D$ eseguiamo le seguenti operazioni:

1. Costruiamo il complesso di Rips R_d.
2. Calcoliamo i numeri di Betti di R_d.

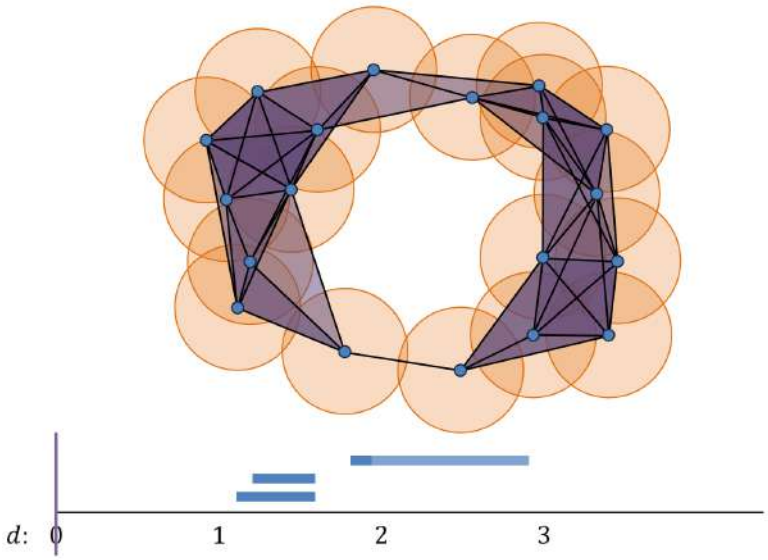


Figura 114: Codice a barre formatosi all'aumentare di d

Vogliamo dedurre da queste informazioni i numeri di Betti *reali* dell'oggetto di cui conosciamo solamente la nuvola di punti. Per poterlo fare, introduciamo il concetto di *codice a barre*.

Ogni buco (di qualsiasi dimensione), apparirà per un certo valore di d, che chiameremo d_1 e scomparirà (verrà riempito) per un secondo valore d_2 con $d_1 < d_2$. Dunque a ogni buco B, possiamo associare un intervallo $[d_1, d_2]$ che identifica il tempo per cui il buco è *sopravvissuto*. Tale intervallo lo possiamo rappresentare con una barra continua da d_1 a d_2 ed è detto la *persistenza del buco*. L'insieme di tutte le barre identificatrici dei buchi viene detto *codice a barre*. Una panoramica è data dalla Figura 114.

Remark. Fissato un tempo d, il numero di barre esistenti al tempo d per buchi n-dimensionali coincide esattamente con $b_n(R_d)$.

A partire dall'osservazione dei codici a barre è possibile intuire i numeri di Betti dell'oggetto codificato dalla nuvola di punti e dunque intuire la forma indicativa di tale oggetto: se infatti la persistenza di alcune barre è molto grande, questo permette di asserire che tale buco è effettivamente reale e non è soltanto frutto di rumore o errore di misurazione.

Possiamo enunciare i seguenti risultati, che rimangono volontariamente vaghi, ma che ci assicurano (sotto ipotesi che stiamo omettendo) l'affidabilità del metodo presentato

Proposition 3. Fissata una nuvola di punti associata alla misurazione di un oggetto, è possibile calcolare con una incertezza ε il numero di Betti $n-$esimo associato all'oggetto come il numero di barre che ha persistenza maggiore di un certo $\widetilde{d}$ dove $\widetilde{d}$ e ε sono numeri che dipendono dalla precisione della misurazione della nuvola di punti.

Theorem 11 (2007, Cohen-Steiner, Edelsbrunner,Hares). I codici a barre sono stabili per perturbazione dei dati, i.e. una minima perturbazione dei dati iniziali fornisce gli stessi risultati in termini di numeri di Betti.

33.4 Esercizi: Numeri di Betti

Di seguito trovate una serie di esercizi guidati e illustrati che possono aiutarvi alla comprensione dei numeri di Betti. Spero che saranno utili allo scopo!

Mostrare che S^1 non è omeomorfo ad $\mathbb{R}$.

Calcolare i numeri di Betti del nastro di Möbius.

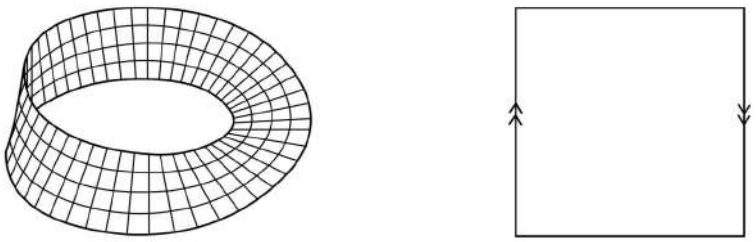


Figura 115: Il nastro di Moebius: visione tridimensionale e bidimensionale con identificazioni

Sia S^2 la sfera. Calcolare i numeri di Betti di S^2.

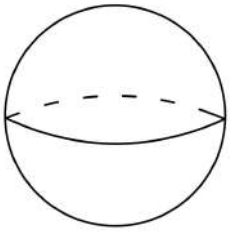


Figura 116: La sfera S^2

[*] Calcolare i numeri di Betti di S^n, con $n \in \mathbb{N}, n \geq 2$.

Risolvere i seguenti punti:

- Calcolare i numeri di Betti del wedge di due circonferenze, cioè di $S^1 \vee S^1$.

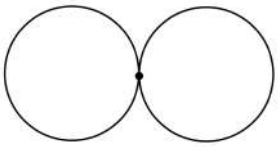


Figura 117: Un wedge di due circonferenze

- Calcolare i numeri di Betti di un wedge di $n \in \mathbb{N}$ circonferenze.

[*] Mostrare che $[a, b]$ è omomotopicamente equivalente ad un punto.

[**] Calcolare i numeri di Betti di un wedge di n circonferenze e m sfere, ovvero di $\underbrace{S^1 \vee \cdots \vee S^1}_{n} \vee \underbrace{S^2 \vee \cdots \vee S^2}_{m}$.

[***] Sia X un complessso simpliciale di dimensione n. Si dimostri che $b_k(X) = 0$ quando $k > n$.

[***] Calcolare i numeri di Betti del toro $T = S^1 \times S^1$.

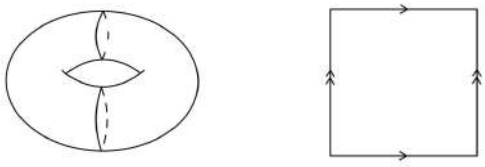


Figura 118: Il toro; visione tridimensionale e bidimensionale con identificazioni

Potrebbe essere utile seguire la seguente traccia risolutiva:

1. Spezzare lo spazio $\mathbb{R}^3$ come unione di un toro pieno X (un toro in cui i punti interni fanno parte del toro stesso) e il suo complementare in $\mathbb{R}^3$ che chiameremo Y. Tali oggetti geometrici hanno intersezione un oggetto geometrico emotopicamente equivalente al toro T se ingrassati.

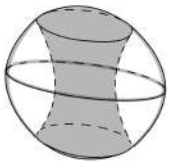


Figura 119: Un retratto per deformazione di Y

2. Calcolare i numeri di Betti di X e Y e mostrare che il primo numero di Betti di Y coincide con il primo numero di Betti del toro T.
3. Trovare una triangolazione per il Toro T e sfruttare le informazioni precedenti per calcolare i rimanenti numeri di Betti

34 L'integrale secondo Riemann

Margherita Zucchelli, n.19, Febbraio 2025

34.1 Vernice necessaria per muri parabolici e sinusoidali

Immaginiamo di dover dipingere il muro in figura:

Figura 120: Muro parabolico

Dobbiamo decidere quanta vernice acquistare: sappiamo che il profilo del muro si modellizza con la funzione $x^2 + 1$ in un intervallo che va da 0 ad un certo a in $\mathbb{R}$

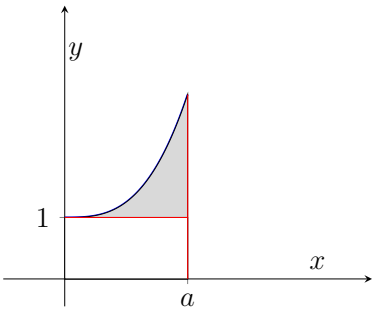


Figura 121: Matematizzazione del profilo: $f(x) = x^2 + 1$

Sicuramente, moltiplicando la base per l'altezza sappiamo calcolare l'area del rettangolo bianco, le difficoltà cominciano calcolando l'area scura.

Per capire quanta vernice ci serve per dipingere il muro vorremmo calcolare l'area sotto al grafico della funzione$f : [0, a] \longrightarrow \mathbb{R}$ definita da $f(x) = x^2$. L'obiettivo è calcolare l'area del sottografico di f utilizzando il metodo delle *somme di Riemann*. Procederemo stimando l'area per difetto e per eccesso in modo sempre più preciso, fino a che le due misure non coincidono.

34.1.1 Partizione e stima dell'area

Per precedere vorremmo suddividere l'intervallo $[0, a]$ in n sottointervalli di uguale lunghezza $\Delta x = \frac{a}{n}$:

$$\left[0, \frac{a}{n}\right], \left[\frac{a}{n}, \frac{2a}{n}\right], \ldots, \left[\frac{(n-1)a}{n}, a\right].$$

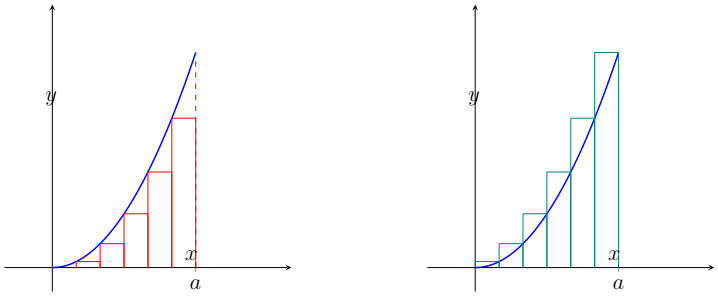


Figura 122: Approssimazione per rettangoli

Procediamo come avevamo detto, stimando l'area della figura per difetto. Per farlo, consideriamo n rettangoli, ognuno che ha come base inferiore un sottointervallo di lunghezza Δx. L'altezza di questi rettangoli è tale da permettergli di "toccare" il grafico della funzione dal basso. Calcoliamo la somma delle aree degli n rettangoli, sapendo comunque di perderci qualcosa dell'area che vogliamo trovare. L'altezza del rettangolo nel k-esimo intervallo è data da $f\left((k-1)\Delta x\right) = \left((k-1)\frac{a}{n}\right)^2$. L'area totale A_i dei rettangoli rossi (Figura 122) è quindi:

$$A_i = \sum_{k=1}^{n} \Delta x \cdot \left((k-1)\frac{a}{n}\right)^2 = \left(\frac{a}{n}\right)^3 \sum_{k=0}^{n-1} k^2.$$

Per la somma delle aree sovrastanti consideriamo i rettangoli che si "appoggiano" sulla funzione dall'alto. L'altezza del rettangolo nel k-esimo intervallo è $f\left(k\Delta x\right) = \left(k\frac{a}{n}\right)^2$: notiamo che questa coincide all'altezza del $k+1$-esimo intervallo nella somma inferiore, quindi effettivamente la differenza delle due somme coincide con l'area dell'n-esimo rettangolo, quindi:

$$A_s = \sum_{k=1}^{n} \Delta x \cdot \left(k\frac{a}{n}\right)^2 = \left(\frac{a}{n}\right)^3 \sum_{k=1}^{n} k^2 = A_i + \Delta x \cdot a^2.$$

Per ora, abbiamo dunque due stime dell'area, una per difetto ed una per eccesso, che dipendono da quanti rettangoli si utilizzano per misurare. Per calcolare le sommatorie utilizziamo una formula che non dimostriamo (se qualcuno volesse cimentarsi può procedere per induzione o cercare una via diretta):

$$\sum_{k=1}^{n} k^2 = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n.$$

Con questa possiamo esprimere le somme come:

$$A_i = \left(\frac{a}{n}\right)^3 \left[\frac{1}{3}(n-1)^3 + \frac{1}{2}(n-1)^2 + \frac{1}{6}(n-1)\right],$$

$$A_s = \left(\frac{a}{n}\right)^3 \left[\frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n\right].$$

Immaginiamo di stimare l'area con sempre più rettangoli: i "triangolini" di superficie che i rettangoli non coprono o coprono in eccesso diventano sempre più piccoli e quindi la misura più precisa. Esasperiamo questo concetto utilizzando infiniti rettangoli per coprire la superficie: questo è concretamente impossibile ma in "matematichese" si può formalizzare con un limite, portando il numero n delle suddivisioni $n \longrightarrow \infty$. Le nostre misure allora diventano:

$$\lim_{n \longrightarrow \infty} A_i = \lim_{n \longrightarrow \infty} A_s = \frac{a^3}{3}.$$

Per ogni n poi vale $A_i \leqslant A \leqslant A_s$, quindi per confronto $A = \frac{a^3}{3}$.

Benissimo! Adesso sappiamo che, se la lunghezza in metri del muro è a dovremo dipingere $\frac{a^3}{3}$ m^2 di superficie più la misura del rettangolo.

Prima di continuare vi voglio lasciare proposti i seguenti esercizi:

- Trovare una funzione la cui approssimazione con 2 rettangoli è più precisa di quella con 3.
- Dimostrare che, approssimando qualsiasi funzione con 4 rettangoli si ottiene sempre un'approssimazione più precisa che con 2 rettangoli.

Ci piacerebbe adesso dipingere un nuovo muro, quello in figura: L'architetto ci dice che il

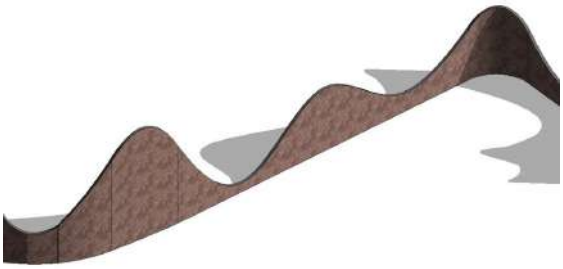


Figura 123: Muro sinusoidale

profilo del muro è $2\sin(x) + \cos(x)$, noi proviamo a suddividere la base in rettangoli ma non ci viene in mente nessun trucco che ci aiuti a calcolare tutte queste sommatorie con troppi fattori. Ci serve una formula più generale per stimare l'area dei sottografici.

34.2 Integrale secondo Riemann

34.2.1 Introduzione storica

Il concetto di integrale nasce nel contesto dello studio delle aree e delle somme infinite. Già nell'antichità, matematici come Archimede avevano sviluppato metodi per calcolare aree sotto curve, basati su processi di approssimazione e limiti. Tuttavia, il calcolo integrale come lo conosciamo oggi fu formalizzato nel XVII secolo grazie ai contributi di Isaac Newton e Gottfried Wilhelm Leibniz.

Nel XIX secolo, Karl Weierstrass e Bernhard Riemann posero le basi rigorose del calcolo integrale, introducendo il concetto di somme di Riemann. Queste somme permettono di definire formalmente l'integrale come limite di somme finite, fornendo uno strumento matematico fondamentale per l'analisi moderna.

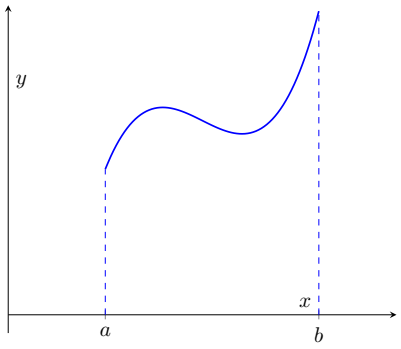
In questo articolo analizziamo il metodo di Riemann per calcolare l'area sotto una curva e dimostriamo alcune applicazioni significative, con esempi specifici.

34.3 Obiettivi

Data una $f : [a, b] \longrightarrow \mathbb{R}$ ci piacerebbe misurare l'area del sottografico di f, ovvero l'insieme

$$S = \big\{(x, y) \in \mathbb{R}^2 \,\big|\, 0 \leq y \leq f(x)\big\},$$

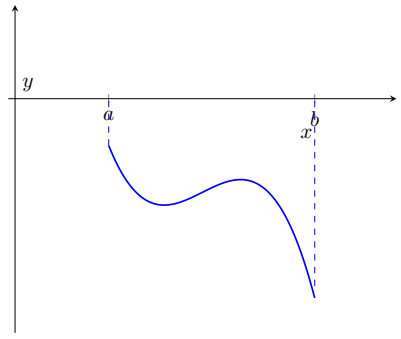
nel caso in cui $f \geq 0$. Un esempio è illustrato nella figura seguente.



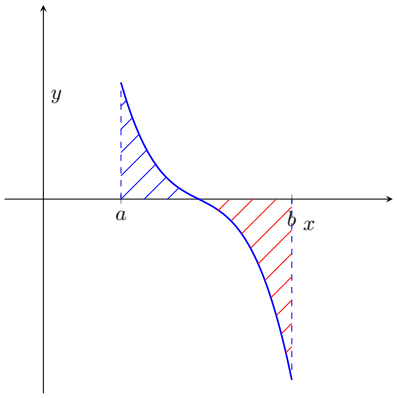
Se invece $f \leq 0$, cioè non positiva, l'area del sottografico viene considerata negativa:

$$S = \big\{(x, y) \in \mathbb{R}^2 \,\big|\, f(x) \leq y \leq 0\big\}.$$

Un esempio è mostrato nella figura seguente.



Infine, per una funzione f non esclusivamente positiva o negativa, come nell'esempio seguente, vogliamo che la misura sia data dall'area sopra l'asse x meno l'area sotto l'asse x.



Diamo adesso alcune definizioni per formalizzare le considerazioni:

Definizione 30. Una *suddivisione* di $[a,b]$ è un insieme $T = \{t_0, t_1, \dots, t_N\}$ tale che

$$t_0 = a,\ t_N = b, \quad t_i < t_{i+1}\ \forall\, i.$$

Cioè un insieme di $N+1$ punti in un intervallo.

Definizione 31. Sia $f : [a,b] \longrightarrow \mathbb{R}$ ed una suddivisione di $[a,b]$ che chiamiamo T. La *somma inferiore*[60] rispetto alla suddivisione T è la stima per difetto che abbiamo fatto prima, formalmente:

$$\mathcal{I}_-(f,T) := \sum_{i=1}^{N} \inf_{t\in(t_{i-1},t_i)} f(t)\cdot(t_i - t_{i-1}),$$

mentre la *somma superiore* è:

$$\mathcal{I}_+(f,T) := \sum_{i=1}^{N} \sup_{t\in(t_{i-1},t_i)} f(t)\cdot(t_i - t_{i-1}).$$

Non lo dimostriamo formalmente ma per ogni coppia di suddivisioni T, S di $[a,b]$ vale che la somma inferiore di una è minore della somma superiore dell'altra, cioè

$$\mathscr{I}_-(f,T) \leqslant \mathscr{I}_+(f,S).$$

Questo fatto è intuibile poiché abbiamo detto che tutte le somme inferiori stimano l'area per difetto e quelle superiori per eccesso.

Definizione 32. L'*integrale inferiore* di f è:

$$\mathcal{I}_-(f) := \sup\left\{\mathcal{I}_-(f,T)\,|\,T \text{ suddivisione di } [a,b]\right\},$$

cioè date tutte le possibili divisioni di $[a,b]$ e le rispettive somme inferiori chiamiamo integrale inferiore il sup di queste.

Questo valore sarà minore (non strettamente) dell'area che vogliamo calcolare, perché tutte le somme inferiori sono più piccole dell'area quindi anche il loro sup deve esserlo.

[60] Ricordiamo che dato E sottoinsieme di $\mathbb{R}$, il suo inf è il più grande degli elementi minori di tutti gli elementi di E, mentre il sup è il più piccolo di tutti gli elementi più grandi di tutti gli elementi di E

Definizione 33. L'*integrale superiore* è:

$$\mathcal{I}_+(f) := \inf\left\{\mathcal{I}_+(f,T)\,|\,T \text{ suddivisione di } [a,b]\right\}.$$

quindi l'inf delle somme superiori al variare della suddivisione di $[a,b]$ considerata.

L'integrale inferiore e l'integrale superiore sono quindi gli elementi più precisi che abbiamo per stimare l'area rispettivamente per difetto ed eccesso.

Definizione 34. Sia $f : [a,b] \longrightarrow \mathbb{R}$ una funzione. Se integrale inferiore e superiore coincidono diciamo che f è integrabile secondo Riemann e scriviamo

$$\int_a^b (f) = \mathcal{I}_-(f) = \mathcal{I}_+(f).$$

Remark. Può accadere che integrale inferiore e superiore non coincidano, cioè esistono funzioni non integrabili secondo Riemann. Vediamo un esempio. Si consideri la funzione $f : [0,1] \longrightarrow \mathbb{R}$ definita da

$$f(x) = 1 \quad \text{se } x \in \mathbb{Q}, \qquad f(x) = 0 \quad \text{se } x \notin \mathbb{Q}$$

Allora, per ogni suddivisione $T = \{t_0, t_1, \dots t_N\}$ di $[0,1]$ vale che ogni intervallino contiene almeno un elemento in $\mathbb{Q}$ ed un elemento in $\mathbb{R}$ (provare a dimostrarlo), quindi i rettangoli che toccano la funzione dal basso avranno altezza 0 e quelli che la toccano dall'alto altezza 1, in quanto

$$\inf_{t\in(t_{i-1},t_i)} f(t) = 0 \qquad \text{e} \qquad \sup_{t\in(t_{i-1},t_i)} f(t) = 1.$$

cioè il valore minimo della funzione in ogni intervallo è 0 ed il massimo è 1, poichè ogni intervallino in $[a,b]$ interseca sia $\mathbb{Q}$ che $\mathbb{R}\setminus\mathbb{Q}$ Per cui $\mathscr{I}_-(f,T) = 0$ e $\mathscr{I}_+(f,T) = 1$ per ogni suddivisione T di $[0,1]$ e quindi

$$\mathscr{I}_-(f) = 0 \qquad \text{e} \qquad \mathscr{I}_+(f) = 1.$$

Di conseguenza f non è integrabile.

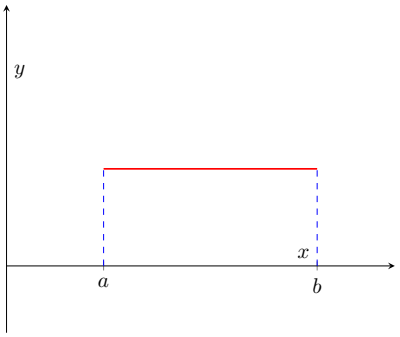
34.3.1 Alcuni fatti utili

Ricordiamo alcuni fatti noti che useremo nelle dimostrazioni più avanti:

- *Integrazione di una costante*: se $f : [a,b] \longrightarrow [m,M]$ è costante, ad esempio $f(x) = \alpha \in \mathbb{R}$ per ogni $x \in [a,b]$, si ha

  $$\int_a^b f(x)\,dx = \alpha(b-a)\,.$$

  Possiamo dedurlo direttamente dalla formula dell'area del rettangolo, ma notiamo che anche stimando l'area con il metodo di Riemann per qualsiasi partizione la somma interiore coincide con quella superiore che coincidono con l'area del retangolo.



- *Teorema dei valori intermedi*: se una funzione continua in un intervallo assume due valori distinti nell'intervallo allora assume tutti i valori compresi tra di essi
- *Teorema di Weierstrass*: se $f : [a, b] \longrightarrow \mathbb{R}$ è continua, l'immagine di f è un intervallo chiuso e limitato.
- *Derivata di una costante*: sia $f : [a, b]$ continua e derivabile. allora $f'(x) = 0$ per ogni $x \in [a, b]$ se e solo se f è costante.

34.4 Il teorema fondamentale del calcolo integrale

Vogliamo adesso enunciare e dimostrare un teorema che permette il calcolo dell'integrale in maniera più pratica rispetto alla somma di infinite aree di rettangoli.

Definizione 35. Data una funzione $f : [a, b] \longrightarrow [m, M]$, cioè una funzione definita su un intervallo $[a, b]$ la cui immagine è contenuta nell'intervallo $[m, M]$ definiamo la *media integrale di* f sull'intervallo $[a, b]$ come la quantità:

$$\frac{1}{b-a}\int_a^b f(x)\,dx,$$

cioè l'altezza che avrebbe il rettangolo di base $b-a$ e area uguale all'area compresa nel sottografico della funzione.

34.4.1 Il Lemma della media integrale

Lemma 2. Sia $f : [a, b] \longrightarrow [m, M]$. Si ha che:

$$m \leq \frac{1}{b-a}\int_a^b f(x)\,dx \leq M.$$

Inoltre, se f è continua, esiste un punto $c \in [a, b]$ tale che:

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(c).$$

Dimostrazione. Poiché $f : [a, b] \longrightarrow [m, M]$, si ha $m \leq f(x) \leq M$ per ogni $x \in [a, b]$. Integrando su $[a, b]$ e dividendo per $b - a$, otteniamo:

$$m = \frac{1}{b-a}\int_a^b m\,dx \leq \frac{1}{b-a}\int_a^b f(x)\,dx \leq \frac{1}{b-a}\int_a^b M\,dx = M,$$

dove per la prima uguaglianza abbiamo usato il primo dei fatti della sezione 34.3.1. Abbiamo così la tesi.

Se f è continua, l'immagine di f è un intervallo chiuso e limitato. Supponiamo che l'immagine sia $[m, M]$ (se l'immagine è un sottointervallo $[\ell, L] \subset [m, M]$, possiamo modificare il codominio per adattarlo). Per il teorema dei valori intermedi, $\frac{1}{b-a}\int_a^b f(x)\,dx$ appartiene all'immagine, dunque esiste $c \in [a, b]$ tale che:

$$\frac{1}{b-a}\int_a^b f(x)\,dx = f(c).$$

□

Remark. La continuità è essenziale affinché la media coincida con il valore di f in un punto.

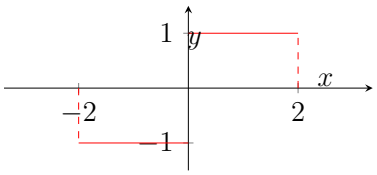


Figura 124: Funzione non continua per cui non esiste un punto c tale che $f(c)$ è la media integrale

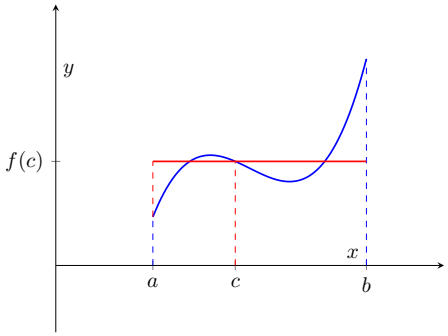
Consideriamo ad esempio $f : [-2, 2] \longrightarrow \mathbb{R}$ definita da:

$$f(x) = \begin{cases} -1 & \text{se } x \leq 0, \\ 1 & \text{se } x > 0. \end{cases}$$

La media di f su $[-2, 2]$ è 0, ma non esiste alcun punto in cui f assume il valore 0.

Il Lemma 2 della media integrale per funzioni continue può essere interpretato in due modi:

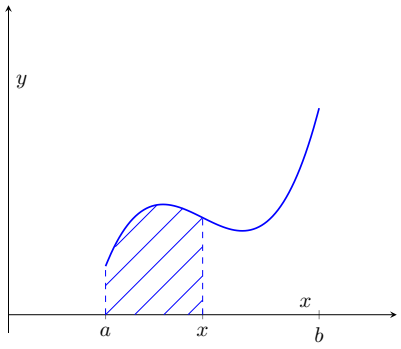
- Esiste almeno un punto $c \in [a, b]$ in cui $f(c)$ corrisponde alla media di f su $[a, b]$.
- Esiste un rettangolo, di base $[a, b]$ e altezza $f(c)$, con la stessa area del sottografico di f su $[a, b]$.



Definizione 36. Data una funzione $f : [a, b] \longrightarrow [m, M]$, definiamo la funzione integrale di f come:

$$F(x) := \int_a^x f(t)\, dt,$$

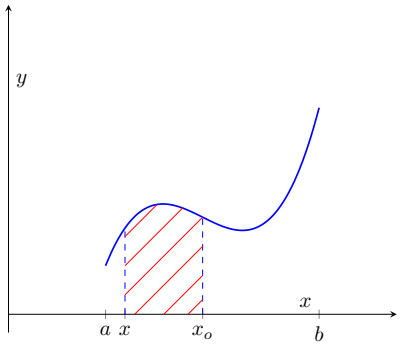
che rappresenta l'area (con segno) del sottografico di f sull'intervallo $[a, x]$.



La funzione integrale può anche essere definita rispetto a un generico $x_0 \in [a, b]$:

$$F_{x_0}(x) := \int_{x_0}^x f(t)\, dt,$$

dove x può essere anche minore di x_0, con il segno determinato dall'orientazione dell'intervallo di integrazione.



34.4.2 Teorema Fondamentale del calcolo integrale

Theorem 12 (Teorema fondamentale del calcolo integrale). Sia $f : [a, b] \longrightarrow \mathbb{R}$ una funzione continua. La sua funzione integrale $F(x)$ è derivabile e vale:

$$F'(x) = f(x) \qquad \text{per ogni } x \in [a, b].$$

Inoltre, se $G : [a, b] \longrightarrow \mathbb{R}$ è una funzione derivabile tale che $G' = f$, allora:

$$F(x) = G(x) - G(a).$$

Dimostrazione. Consideriamo il rapporto incrementale di F in $x_0 \in [a,b]$. Per h tale che $x_0 + h \in [a,b]$, abbiamo:

$$\frac{F(x_0+h) - F(x_0)}{h} = \frac{1}{h}\int_{x_0}^{x_0+h} f(t)\,dt.$$

Per il Lemma 2 della media integrale, esiste $c_h \in [x_0, x_0+h]$ tale che:

$$\frac{F(x_0+h) - F(x_0)}{h} = f(c_h).$$

Passando al limite per $h \longrightarrow 0$ e usando la continuità di f, si ottiene:

$$\lim_{h \longrightarrow 0} f(c_h) = f(x_0).$$

Quindi $F'(x_0) = f(x_0)$. Se G è tale che $G' = f$, allora $(F-G)' = 0$, e quindi $F(x) - G(x) = c$ per una costante $c \in \mathbb{R}$. Poiché $F(a) = 0$, segue $c = -G(a)$, da cui:

$$F(x) = G(x) - G(a).$$

□

34.5 Un'applicazione!

Siamo finalmente in grado di calcolare quanta vernice comprare per il nostro muro! A noi, per dipingere il tratto da a a b, interessava calcolare

$$\int_a^b 2\sin(x) + \cos(x)\,dx.$$

La funzione $2\sin(x) + \cos(x)$ è continua, quindi chiamata $F(x)$ una sua primitiva sappiamo che il valore che vogliamo calcolare coincide con $F(b) - F(a)$. Ci basta quindi trovare una primitiva di $2\sin(x) + \cos(x)$, valutarla in a ed in b e siamo a cavallo. Una primitiva della funzione è $-2\cos(x) + \sin(x)$, quindi dobbiamo comprare vernice per coprire un numero di metri quadri pari a

$$-2\cos(a) + \sin(a) + 2\cos(b) - \sin(b).$$

35 Storia di un esercizio

Jacopo Burelli, n.20, Gennaio 2026

35.1 L'importanza del contesto

[title=Problema, colback=blue!5, colframe=blue!40!black, coltitle=white] Trovare x tale che

$$(x-6)^3 = \sqrt[3]{x} + 6.$$

Probabilmente state già pensando a diversi modi per risolvere il problema e, con buona probabilità, almeno una delle soluzioni che troverete sarà corretta. Ma che cosa ci dice davvero questo sul problema? In realtà, molto poco. Da dove proviene questo esercizio? È tratto da un compito o da un'esercitazione? Aritmetica, Analisi, Analisi Numerica, o qualche altro ambito? Come dovrei affrontarlo? Quali strumenti posso utilizzare? E, inoltre, a quale insieme appartiene x?

Per dare una panoramica di come sono arrivato a formulare il problema, faccio un passo indietro e provo a motivare il processo creativo che sta dietro alla costruzione di questo esercizio. È proprio questo processo che mi ha interessato a tal punto da spingermi a scrivere questo articolo, il cui obiettivo vorrebbe essere, più che la soluzione in sé, **l'importanza del contesto**.

Ho pensato a questo problema mentre preparavo un compito per una terza di liceo scientifico e questo mi ha portato a riflettere su quanto sia importante il contesto in cui un problema viene proposto. Il modulo didattico legato alla sua risoluzione riguardava le funzioni, quindi vi suggerisco di provare a ragionare in questi termini, così da assecondare il mio processo creativo.

Se mi fossi trovato questo esercizio in un compito delle scuole superiori, sarei stato in grado di individuare una possibile soluzione? No, perché, nonostante il problema nasca dalla seguente osservazione, non lo avrei inquadrato nello stesso contesto di chi lo ha scritto.

Possiamo notare infatti che, se proviamo a invertire l'equazione

$$(x-6)^3 = y,$$

si ottiene

$$x = \sqrt[3]{y} + 6,$$

cioè le due funzioni sono una l'inversa dell'altra. In altre parole, sto cercando un valore di x che soddisfi

$$f(x) = f^{-1}(x).$$

Poiché avevamo visto il teorema della simmetria tra il grafico di una funzione f e quello della sua inversa rispetto alla bisettrice, potevo allora cercare le soluzioni proprio su $y = x$, andando a risolvere il problema semplificato

$$(x-6)^3 = x.$$

Tuttavia, anche in questo passaggio, ho dato per scontate alcune cose: per esempio il motivo per cui la funzione $(x-6)^3$ è invertibile, quale sia il dominio di f, e così via.

L'invertibilità di f è facilmente giustificabile. A meno di una traslazione (sulla quale, a sua volta, viene scaricata ulteriore conoscenza matematica) la funzione "è" x^3. In alternativa,

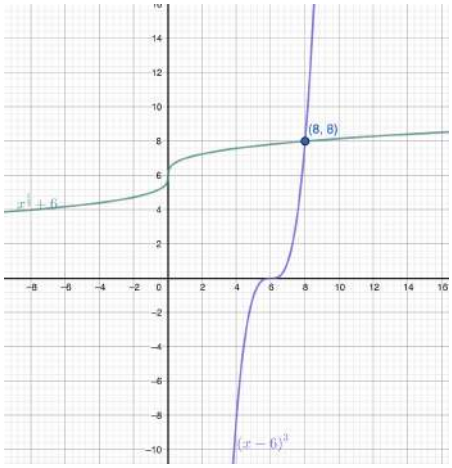


Figura 125: Intersezione tra i grafici di $\sqrt[3]{x}+6$ e $(x-6)^3$.

si possono utilizzare le derivate: insomma, dopo un corso di Analisi universitario questo non dovrebbe rappresentare un problema.

A scopi didattici, ne diamo comunque una dimostrazione a partire dalle definizioni, ricorrendo a strumenti efficaci e disponibili già dalle scuole superiori: le definizioni.

Presa $f : \mathbb{R} \longrightarrow \mathbb{R}$ definita da $f(x) = (x-6)^3$, per $x_1, x_2 \in \mathbb{R}$ e $y \in \mathbb{R}$ si ha

- Iniettività:

$$(x_1-6)^3 = (x_2-6)^3 \iff x_1-6 = x_2-6 \iff x_1 = x_2.$$

- Surgettività:

$$y = (x-6)^3 \iff \sqrt[3]{y} = x-6 \iff x = \sqrt[3]{y}+6.$$

Dopo aver inoltre preso coscienza del fatto che stavo cercando, quasi automaticamente, una soluzione reale, e degli impliciti che stavo assumendo lungo il percorso (si noti, ad esempio, la tranquillità che il *Teorema fondamentale dell'algebra* garantisce al solutore circa l'esistenza di tale x, interpretando $f(x)-x$ come un polinomio) ho scelto quella che mi è sembrata la soluzione migliore: non assegnare il problema in un compito scritto di terza scientifico, e mettere a dura prova l'ego di un caro amico, noto volto del dipartimento: Luca Bruni. A lui ho fornito il problema esattamente nella forma presentata all'inizio del paragrafo, in quanto particolarmente interessato a osservare il suo processo risolutivo.

> *"Sembra orribile elevando, però mentre ero in macchina e pensavo ai grafici, e ora forse ho fatto il conto male ma mi sembra che siano una la funzione inversa dell'altro, quindi stai cercando di risolvere $f(x) = f^{-1}(x)$, quindi stai cercando le intersezioni tra $f(x)$ e $f^{-1}(x)$ e i due grafici si intersecano nella retta $x = y$, quindi è sufficiente che risolvi un membro uguale a x. Bellissimo questo problema, bellissimo ti giuro sono troppo felice di averlo risolto così, fantastico, basta mettere la parte di sinistra uguale a x o la parte di destra uguale a x, e funziona.*
>
> *Comunque posso dire un esercizio bellissimo? Cioè ci ho dovuto pensare un pochino, ha una soluzione molto elegante, però non so il target, se è un esercizio sulle funzioni*

inverse è top, spettacolare però è molto difficile a meno che non l'abbiano visto, ho usato troppe competenze che loro non hanno, ho pensato al grafico ho ripensato al teorema.. però bellissimo problema."

Questa soluzione, porta a trovare $x = 8$. Come?

35.2 La nascita

Possiamo usare il criterio delle radici razionali, già presente sui testi dalle scuole superiori (se siamo abbastanza fortunati da averlo nel programmazione didattica e ricordarcelo)

[title=Criterio delle radici razionali [2]] Sia

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

un polinomio a coefficienti interi, con $a_n \neq 0$. Se $\dfrac{p}{q} \in \mathbb{Q}$, con $\gcd(p, q) = 1$, è una radice razionale di $p(x)$, allora

$$p \mid a_0 \quad \text{e} \quad q \mid a_n.$$

In altre parole, le possibili radici razionali di $p(x)$ sono della forma

$$\pm \frac{\text{divisore di } a_0}{\text{divisore di } a_n}.$$

Portando tutto a sinistra, l'equazione

$$(x-6)^3 = x$$

diventa

$$(x-6)^3 - x = 0.$$

Sviluppando il cubo:

$$(x-6)^3 = x^3 - 18x^2 + 108x - 216,$$

quindi

$$(x-6)^3 - x = x^3 - 18x^2 + 107x - 216.$$

A questo punto posso applicare il *Criterio delle radici razionali*: dato che il polinomio

$$p(x) = x^3 - 18x^2 + 107x - 216$$

è a coefficienti interi e ha coefficiente direttivo $a_n = 1$, ogni eventuale radice razionale deve essere un divisore del termine noto $a_0 = -216$, cioè un numero del tipo

$$\pm d \qquad \text{con } d \mid 216.$$

In particolare, l'elenco dei divisori positivi di 216 è

$$1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 27, 36, 54, 72, 108, 216,$$

e quindi i candidati sono

$$\pm 1, \pm 2, \pm 3, \pm 4, \pm 6, \pm 8, \pm 9, \pm 12, \pm 18, \pm 24, \pm 27, \pm 36, \pm 54, \pm 72, \pm 108, \pm 216.$$

Partendo dal più piccolo, con un po' di pazienza troviamo $x = 8$.

$$p(8) = 8^3 - 18 \cdot 8^2 + 107 \cdot 8 - 216 = 512 - 18 \cdot 64 + 856 - 216 = 0.$$

A questo punto ci siamo chiesti se potevamo dimostrare che $x = 8$ era l'unica soluzione reale. Applicando ad esempio la *Regola di Ruffini*[61] al polinomio

$$p(x) = x^3 - 18x^2 + 107x - 216$$

con radice $x = 8$, si ottiene la seguente scomposizione:

$$x^3 - 18x^2 + 107x - 216 = (x - 8)(x^2 - 10x + 27).$$

Effettivamente il polinomio quoziente di secondo grado non ha radici reali, in quanto una sua riscrittura è

$$x^2 - 10x + 27 = (x - 5)^2 + 2.$$

e finalmente concludiamo l'esercizio.

Questo potrebbe essere riformulato ed enunciato nel seguente modo:
[title=Problema, colback=blue!5, colframe=blue!40!black, coltitle=white] Dimostrare che $x = 8$ è l'unica soluzione reale di

$$(x - 6)^3 = \sqrt[3]{x} + 6.$$

Tuttavia, la gioia di Luca e il mio interesse per la didattica mi hanno portato a chiedermi quali potessero essere altri modi di risolvere un problema enunciato in questo modo, possibilmente coinvolgendo aree diverse della Matematica. Un obiettivo era quello di ampliare le conoscenze generali e trasversali nate attorno a questo esercizio e, allo stesso tempo, di riportare ai ragazzi una riflessione sui possibili diversi metodi di risoluzione di un problema, tema spesso trascurato.

Una delle frasi che più mi è rimasta impressa del mio professore di Analisi 1 e 2 è infatti:

"anche una patologia può diventare uno strumento."

35.3 Approfondire un problema

Per perseguire questo interesse, ho deciso di chiedere aiuto su Math Stack Exchange (MSE), un noto sito di domande e risposte frequentato da studenti e professori di matematica a ogni livello. Per completezza, rimando alla domanda originale tramite il qrcode a fianco.

Nel seguito dell'articolo, mi impegnerò a passare in rassegna e ad analizzare alcuni degli approcci segnalati.

[61] Ricordiamo che la regola di Ruffini è un metodo alternativo di effettuare la divisione tra polinomi nel caso in cui il polinomio divisore sia di primo grado.

[breakable, colback=yellow!10, colframe=yellow!60!black, title=John Bentin – Math Stack Exchange, fonttitle=] **It is easy to spot the solution $x = 8$ and verify it by substitution. The less trivial part is showing this solution to be unique. A sketch of the graphs of $y = (x-6)^3$ and $y = x^{1/3}+6$ shows clearly that they cross only once, at $y = x = 8$, while the line $y = x$ elsewhere lies between the two graphs (because they are inverse).**

However, this is not a mathematically rigorous method. We can prove that our solution is unique if we can show that $(x-6)^3 < x < x^{1/3}+6$ when $x < 8$, and $x^{1/3}+6 < x < (x-6)^3$ when $x > 8$. Thus there are four inequalities to verify.

The first inequality may be expressed as $(x-6)^3 - x < 0$ for $x < 8$. By expansion,

$$(x-6)^3 - x = x^3 - 18x^2 + 107x - 216 = (x-8)(x^2 - 10x + 27),$$

where the factor $x - 8$ is to be expected from our original spotted solution. The quadratic factor can be written as $(x-5)^2 + 2$, which is always positive, and the sign of the other factor $x - 8$ therefore establishes the inequality.

The second inequality can be put in the form $x - 6 < x^{1/3}$. Since cubing preserves order, we can write this as $(x-6)^3 < x$, and the proof proceeds similarly as before. The proofs of the remaining two inequalities are also similar.

John mette bene in luce il fatto che la semplificazione che riconduce alla ricerca delle soluzioni su $y = x$ non è affatto automatica, ma richiede un'argomentazione adeguata. Tale argomentazione risulta però particolarmente comprensibile nell'approccio proposto da Bentin (nonostante altri contributi, anche temporalmente precedenti come quello di *heropup* sempre sul forum).

[breakable, colback=yellow!10, colframe=yellow!60!black, title=Mari Strup - Math Stack Exchange, fonttitle=] **Here is what I feel is a relatively fun method. Begin by substituting $y^3 := x$ to obtain the equation**

$$(y^3 - 6)^3 = y + 6$$

and notice that if $y^3 - 6 = y$ then

$$(y^3 - 6)^3 = y^3 = y + 6$$

so we need only solve the equation $y^3 - 6 = y$. Rearrange to obtain

$$y(y^2 - 1) = 6$$

then factor using difference of two squares to find

$$y(y-1)(y+1) = 6.$$

In other words, we just need to find three consecutive numbers which have product 6. Famously $1 \times 2 \times 3 = 6$, so if $y = 2$ then

$$y(y-1)(y+1) = 2 \times 1 \times 3 = 6,$$

thus $y = 2$ is a valid solution. Since $y^3 = x$ it follows that $x = 8$.

Mari, per quanto fornisca a mio avviso una soluzione a posteriori, fa emergere una bellissima idea di sostituzione, che conduce a $y(y-1)(y+1)$, trasportando la ricerca delle soluzioni a un'osservazione aritmetica non avanzata e estremamente efficace. Certamente, la soluzione di Mari è influenzata dalla conoscenza dell'esistenza di soluzioni intere (nella domanda, infatti, è presente un commento su $x = 8$), che ha guidato la sua ricerca in una fattorizzazione del tipo proposto.

[breakable, colback=yellow!10, colframe=yellow!60!black, title=Dan - Math Stack Exchange, fonttitle=]

Converting to polynomial equation

Expand the cube to get:

$$x^9 - 54x^8 + 1296x^7 - 18162x^6 + 163944x^5 - 989496x^4$$
$$+3996972x^3 - 10429560x^2 + 15968015x - 10941048 = 0$$

Newton's Method

Let

$$f(x) = (x-6)^3 - (x^{\frac{1}{3}} + 6), \qquad f'(x) = 3(x-6)^2 - \frac{1}{3}x^{-2/3}.$$

$$x_{n+1} = x_n - \frac{(x-6)^3 - (x^{\frac{1}{3}} + 6)}{3(x-6)^2 - \frac{1}{3}x^{-2/3}}.$$

n	**0**	**1**	**2**	**3**	**4**	**5**
$x_n \approx$	**1**	**2.77**	**4.09**	**5.44**	**14.78**	**11.89**

n	**6**	**7**	**8**	**9**	**10**
$x_n \approx$	**10.00**	**8.84**	**8.22**	**8.02**	**8.00**

The sequence converges to the root $x = 8$.

Fixed-point iteration

Solving for the cube root gives

$$x = (x^{\frac{1}{3}} + 6)^{\frac{1}{3}} + 6.$$

n	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$x_n \approx$	**0**	**7.82**	**8.00**	**8.00**	**8.00**	**8.00**	**8.00**	**8.00**	**8.00**

The iteration converges rapidly to $x = 8$.

Bisection Method

A simple general-purpose algorithm for finding real roots.

[breakable, colback=gray!10, colframe=gray!70!black, title=Pseudocode, fonttitle=]

```
def bisect_solve(f, lo, hi):
    left_neg = f(lo) < 0
    right_neg = f(hi) < 0
    if left_neg == right_neg:
        raise ValueError('No sign change')
    for _ in range(100):
        x = (lo + hi) / 2
        y = f(x)
        if y == 0:
            return x
        elif (y < 0) == left_neg:
            lo = x
        else:
            hi = x
    return x
```

Applying the method on $[0, 10]$ **yields** $x = 8$**.**

Il contributo di Dan si distingue dagli altri per la sua natura computazionale. Dan esplora infatti tre approcci algoritmici: dal metodo di Newton, alla bisezione, fino all'iterazione di punto fisso [3], mostrando come tutti conducano coerentemente alla stessa soluzione reale. La capacità di mettere in evidenza i limiti pratici di alcuni approcci teorici è il motivo per il quale ho inserito tra le soluzioni riportate questa: l'espansione conduce a un polinomio di grado 9 con coefficienti che spesso non siamo abituati a maneggiare. Per quanto ami allo stesso tempo un'altra frase di un altro mio professore di Algebra 1 sulla Matematica:

"In Matematica vogliamo risolvere i problemi utilizzando meno conti possibile."

non posso fare a meno di essere soddisfatto dal gambetto dei metodi iterativi apprezzabili dalla risposta di Dan.

35.4 Commento finale

Sono presenti molte altre soluzioni che gli utenti stanno continuando a fornire: si va da metodi computazionali che non conoscevo, a diversi approcci algebrici basati su osservazioni e trucchi particolarmente interessanti. Per queste ulteriori soluzioni rimando quindi al collegamento riportato in bibliografia/sitografia [1].

Oltre agli approcci presentati in questo articolo, invito i lettori più navigati a cercare e a proporre metodi di risoluzione dell'esercizio che facciano uso della teoria dei corsi di Algebra 1 o 2 dell'Università di Pisa. La mancanza di un contributo di questo tipo in queste pagine lascia infatti, almeno a me, un leggero amaro in bocca.

Tra l'assegnazione dell'incarico, la fase di ideazione e la stesura finale di questo lavoro sono trascorsi all'incirca due mesi, durante i quali ho dedicato qualche ora, in modo sporadico e nei momenti liberi, alla ricerca di nuove soluzioni. Questo mi ha portato a riflettere su quanto sia importante non sottovalutare il tempo investito nel risolvere o nel creare un problema.

Nonostante conoscessi già un possibile metodo di risoluzione dell'esercizio (circostanza che probabilmente ne giustifica la nascita) è stato proprio il semplice atto di pormi una domanda a permettermi di entrare in contatto con idee e strumenti per me nuovi, sconosciuti: uno su tutti il metodo di Laguerre [1]. Strumenti che, se coltivati nel tempo, possono arricchire e diversificare il "terreno" concettuale su cui ciascuno di noi cammina durante un processo creativo o risolutivo in matematica.

Per concludere, desidero ringraziare tutte le persone che hanno contribuito a questo articolo: a partire da Luca, fino ai contributi forniti dagli utenti di MSE.

Riferimenti bibliografici

[1] Community di Math Stack Exchange, *Seeking different ways you can find* $x = 8$ *in* $(x-6)^3 = x^{1/3} + 6$, Math Stack Exchange, https://math.stackexchange.com/questions/5112197.

[2] P. Di Martino, *Algebra*, Pisa University Press, pisauniversitypress.it.

[3] D. A. Bini, *Appunti di Analisi Numerica*, Università di Pisa, poisson.phc.dm.unipi.it/~lbruni/Appunti/AnalisiNumerica.pdf.

36 Teoria dei grafi e la ricerca del cibo delle formiche

Luca Bruni, n.20, Gennaio 2026

Viviamo in un mondo di connessioni. Ogni giorno, senza rendercene conto, interagiamo con strutture che possono essere modellate come **grafi**. Intuitivamente, un grafo è una rappresentazione di oggetti e delle loro relazioni: è una struttura matematica formata da **nodi** (o *vertici*) e **collegamenti** (o *archi*) tra di essi. È uno strumento semplice ma potentissimo, che permette di modellare reti, connessioni, movimenti, interazioni. Obiettivo di queste pagine, è quello di introdurre alla teoria dei grafi e ai suoi algoritmi, mostrando come questa disciplina permetta la modellizzazione e la "spiegazione" del comportamento di sistemi complessi. In particolar modo ci occuperemo del problema della ricerca di cibo delle formiche.

36.1 Giochi e grafi

In questa prima sezione riproproniamo alcuni giochi che erano stati proposti in una vecchia rubrica del giornalino degli Open Days. Questi giochi, apparentemente semplici, nascondono in realtà una struttura di grafo che permette di risolverli in modo elegante e semplice.

36.2 Il gioco dei 4 cavalli

Immaginate di avere quattro cavalli e la seguente, insolita, scacchiera.

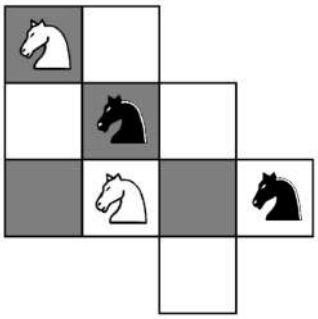


Figura 126: La scacchiera del gioco dei 4 cavalli.

Il problema che vogliamo porre è il seguente:

Data la scacchiera in Figura 126, è possibile scambiare di posizione i cavalli bianchi e quelli neri?

Le regole sono semplici: il cavallo si muove come nel gioco degli scacchi e può muoversi solo su una casella vuota.

Prima di andare avanti con la soluzione, è istruttivo provare a pensare da soli a una strategia. Ricordatevi che fare dei tentativi è a tutti gli effetti una strategia valida: spesso provando si scoprono regolarità inaspettate.

E se per caso riuscite nell'impresa, sapreste dire qual è il numero minimo di mosse necessarie a risolvere il rompicapo?

Spesso è utile capire per quale motivo il problema risulta difficile: in primis, il movimento degli scacchi a L può confonderci, inoltre, la ristrettezza della scacchiera non ci permette di muoverci liberamente. Cerchiamo in un colpo solo di risolvere entrambi i problemi. Eticchiamo le caselle come in Figura 127 e proviamo a pensare a una strategia.

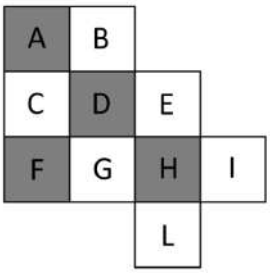


Figura 127: La scacchiera del gioco dei 4 cavalli con le caselle etichettate.

Costruiamo un grafo che catturi il movimento dei cavalli. I nodi del grafo sono le caselle della scacchiera e gli archi sono i movimenti possibili dei cavalli.

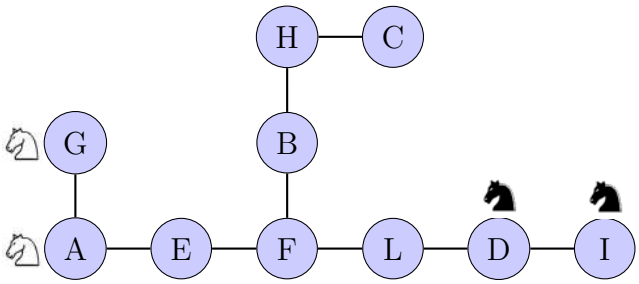


Figura 128: Il grafo che rappresenta i movimenti possibili dei cavalli sulla scacchiera.

La soluzione di questo problema è adesso immediata. I cavalli bianchi, ad esempio, si sposteranno nelle caselle B e H, permetteranno il passaggio dei cavalli neri e li sostituiranno nella loro posizioni. Anche la domanda per il numero minimo di mosse, che sembrava particolarmente complessa, è adesso facilmente approcciabile!

36.3 Il gioco degli smartphone

Il *gioco degli smartphone* è un interessante problema di disposizione in cui salta agli occhi la teoria dei grafi e la sua utilità.

Dati 4 smartphone, è possibile disporli in modo che ogni smartphone li tocca tutti tranne un altro?

Una possibile soluzione è quella di disporre i 4 smartphone in modo che formino i lati di un "quadrilatero" più grande.

La difficoltà aumenta notevolmente se si aumenta il numero di smartphone. Ad esempio provate a risolvere lo stesso quesito con 5 smartphone:

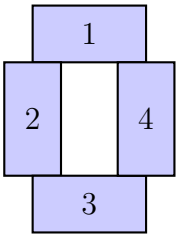


Figura 129: Una possibile soluzione del gioco dei 4 smartphone.

Dati 5 smartphone, è possibile disporli in modo che ogni smartphone li tocca tutti tranne un altro?

Dopo un po' di tentativi, ci rendiamo conto che qualcosa non va e che sembra che non sia possibile. Dobbiamo però trovare un modo per formalizzare questa impossibilità. La prima semplificazione che possiamo fare è quella di immaginare gli smartphone come nodi e i contatti tra di loro come dei collegamenti tra nodi. Ecco che, astraendo il problema dalla realtà fisica, abbiamo una visualizzazione molto più chiara di quello che sta accadendo.

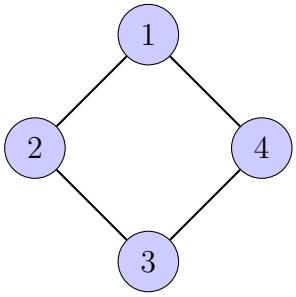


Figura 130: Il grafo associato alla soluzione del problema con 4 smartphone proposto sopra.

In questo modo, il problema si riduce a quello di trovare un grafo con 5 nodi in cui ogni nodo è collegato a esattamente 3 nodi.

Proviamo a rispondere alla seguente domanda: quanti collegamenti devono essere presenti nel grafo per soddisfare la condizione del problema?

Da ogni nodo partono 3 collegamenti, ma ogni collegamento collega due nodi. Questo vuol dire che il numero totale di collegamenti è uguale a 3 volte il numero di nodi diviso 2. In altre parole, il numero totale di collegamenti è uguale a $\frac{3 \cdot 5}{2} = \frac{15}{2}$ che non è un numero intero! Dunque questo grafo non può esistere e il problema dei 5 smartphone è impossibile da realizzare!

Un semplice ragionamento di teoria dei grafi ci ha permesso di risolvere elegantemente il problema che adesso, grazie all'idea introdotta, si presta alla seguente generalizzaione:

Dati n smartphone con $n \geq 3$, è possibile disporli in modo che ogni smartphone li tocca tutti tranne un altro?

Lasciamo al lettore la soluzione del problema generale.

36.4 Non solo giochi

Nei due giochi precedenti abbiamo visto come l'utilizzo dei grafi permette di modellizzare due giochi apparentemente complicati che si rivelano essere molto più comprensibili con il giusto formalismo.

Al di là dei giochi, i grafi sono uno strumento potentissimo per la modellizzazione di problemi complessi; nel seguito di queste note ci occuperemo di alcuni esempi di applicazione della teoria dei grafi a problemi reali. In particolare, dopo aver formalizzato il concetto di grafo, ci soffermeremo su due applicazioni e sulle implicazioni pratiche che ne derivano:

- La ricerca del cammino minimo in un grafo pesato;
- La ricerca di cibo delle formiche.

36.5 Elementi di teoria dei grafi

In questa sezione, ci occupiamo di dare le definizioni fondamentali per la teoria dei grafi.

Definizione 37. Un *grafo* è una coppia $G = (V, E)$, dove:

- V è un insieme finito di elementi chiamati *vertici* o *nodi*;
- $E \subseteq \{\{u, v\} \mid u, v \in V,\ u \neq v\}$ è un insieme di *archi*, ciascuno dei quali è una coppia non ordinata di vertici.

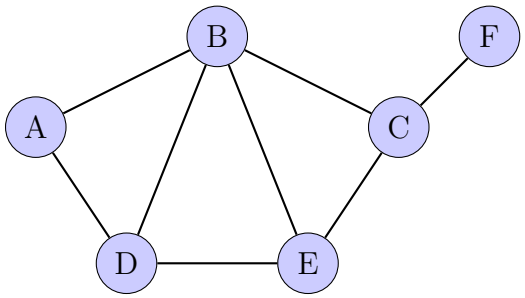


Figura 131: Esempio di grafo.

Remark. Nella definizione di grafo che abbiamo dato non ammettimo l'esistenza di *loop*, ovvero di nodi che hanno un arco che parte e arriva allo stesso nodo. Inoltre, non ammettiamo archi multipli, ovvero più archi che collegano la stessa coppia di nodi. In altre parole, ogni coppia di nodi può essere connessa da al massimo un arco.

Definizione 38. Un *grafo pesato* è un grafo in cui a ciascun arco è associato un valore numerico, detto *peso* o *costo*. Formalmente, un grafo ponderato è una terna $G = (V, E, w)$, dove:

- (V, E) è un grafo;
- $w : E \to \mathbb{R}$ è una funzione che assegna un peso reale a ciascun arco.

Definizione 39. Un *cammino* in un grafo $G = (V, E)$ è una sequenza di vertici $(v_0, v_1, \ldots, v_k)$ tale che per ogni $i = 0, \ldots, k-1$, l'arco $\{v_i, v_{i+1}\} \in E$.

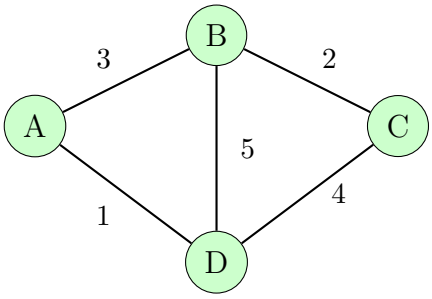


Figura 132: Esempio di grafo pesato con pesi sugli archi.

In altre parole, un cammino non è altro che una sequenza di vertici in cui ogni coppia consecutiva di vertici è connessa da un arco.

Definizione 40. Un *cammino semplice* o *cammino senza cicli* è un cammino in cui tutti i vertici sono distinti, cioè $v_i \neq v_j$ per ogni $i \neq j$.

Definizione 41. Un grafo non orientato è detto *connesso* se esiste un cammino tra ogni coppia di vertici.

Definizione 42. Un *albero* è un grafo connesso e aciclico, cioè che non contiene cicli.

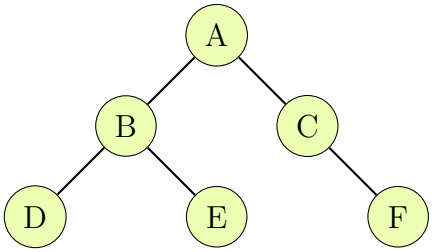


Figura 133: Un albero: grafo connesso senza cicli.

Nella prossima sezione, cominceremo a parlare di algoritmi sui grafi per l'esplorazione e per la ricerca di cammini. Siamo in particolare interessati ai cammini minimi.

Definizione 43. Dato un grafo pesato $G = (V, E, w)$, un *cammino minimo* tra due vertici è un cammino tale che la somma dei pesi degli archi attraversati è minima.

36.6 Dalle formiche ai grafi: ottimizzazione tramite intelligenza collettiva

36.6.1 Il problema della ricerca del cibo

Nel mondo naturale, uno dei comportamenti più affascinanti è quello delle formiche alla ricerca del cibo. Nonostante siano organismi molto semplici, prive di una vera intelligenza individuale, sono in grado di risolvere problemi complessi grazie alla cooperazione e alla comunicazione indiretta.

In particolare, le formiche riescono a trovare percorsi ottimali tra il nido e una fonte di cibo. Non possiedono mappe né un senso diretto delle distanze, ma riescono a scoprire e sfruttare cammini efficienti attraverso un meccanismo di **comunicazione chimica**, basato sui feromoni.

Questo comportamento ha ispirato un intero filone dell'intelligenza artificiale chiamato *Ant Colony Optimization* (ACO), in cui si cerca di riprodurre artificialmente il comportamento delle colonie di formiche per risolvere problemi di ottimizzazione su grafi.

Formalizzazione del problema

Il comportamento delle formiche può essere modellato attraverso un grafo, in cui:

- I **nodi** rappresentano posizioni fisiche (come il nido, i punti di passaggio e le fonti di cibo).
- Gli **archi** rappresentano i percorsi possibili tra le posizioni.
- Ogni arco ha un **peso**, che può rappresentare una distanza, un costo, o una difficoltà nel percorrerlo.

Il problema che ci poniamo è trovare **cammini convenienti** tra due nodi del grafo: il nodo sorgente (il nido) e il nodo obiettivo (una tra le fonti di cibo).

36.6.2 Algoritmo delle colonie di formiche - *Ant Colony Optimization* o ACO

Principi generali

Il modello ACO si basa su alcuni principi fondamentali, ispirati al comportamento reale:

1. **Deposizione di feromoni:** ogni formica, mentre percorre un cammino, deposita una quantità di feromone sugli archi attraversati.
2. **Evaporazione:** i feromoni evaporano col tempo, riducendo la loro intensità.
3. **Scelta probabilistica:** una formica decide quale nodo visitare in base a una probabilità che dipende:
    - dalla quantità di feromone presente sugli archi;
    - da un'informazione euristica (ad esempio, l'inverso della distanza). Prenderemo proprio questo esempio nel prosieguo[62].
4. **Rinforzo positivo:** i cammini migliori, ricevono più feromoni in tempi brevi, e diventano quindi più attraenti per altre formiche.

L'idea è quindi di simulare una colonia di formiche che esplora il grafo, depositando feromoni e aggiornando le probabilità di scelta dei cammini in modo iterativo, fino a convergere verso soluzioni ottimali o **quasi ottimali**.

[62]Si possono aggiungere parametri di pericolosità o affidabilità del cammino. L'implementazione è analoga a quanto viene fatto con la distanza

Formalizzazione matematica

Sia $\tau_{ij}(t)$ la quantità di feromone sull'arco (i, j) al tempo t (più precisamente all'iterazione t dell'algoritmo), e η_{ij} l'informazione euristica associata (ad esempio $\eta_{ij} = \frac{1}{d_{ij}}$, dove d_{ij} è la lunghezza dell'arco).

La probabilità che una formica attualmente in i scelga di muoversi verso il nodo j è data da:

$$P_{ij}(t) = \frac{[\tau_{ij}(t)]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{k \in N_i} [\tau_{ik}(t)]^\alpha \cdot [\eta_{ik}]^\beta}$$

dove:

- N_i è l'insieme dei nodi vicini a i ancora visitabili;
- α regola l'influenza del feromone;
- β regola l'influenza dell'euristica.

Remark. La probabilità è una media pesata tra il feromone e l'euristica. Il parametro α controlla quanto le formiche si affidano al feromone, mentre β regola l'importanza dell'euristica. Questo bilanciamento permette di esplorare nuovi cammini senza trascurare quelli già promettenti.

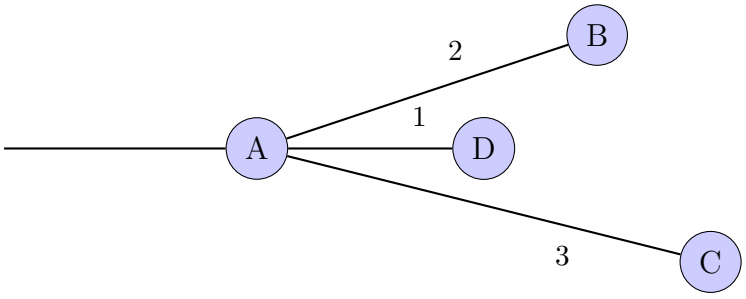


Figura 134: Un grafo con nodo A e tre archi in uscita. Supponendo feromoni $\tau_{AB} = 3$, $\tau_{AC} = 2$, $\tau_{AD} = 4$ e usando l'inverso della distanza come euristica ($\eta_{ij} = 1/w$), calcoliamo le probabilità. I parametri utilizzati sono $\alpha = 1$ e $\beta = 1$. Il denominatore è dato da $\sum_{k \in N_A} [\tau_{Ak}]^\alpha \cdot [\eta_{Ak}]^\beta = (3 \cdot 1/2) + (2 \cdot 1/3) + (4 \cdot 1/1) = 1.5 + 0.6667 + 4 = 6.1667$. Le probabilità risultano: $P_{AB} \approx 0.24$, $P_{AC} \approx 0.11$, $P_{AD} \approx 0.65$.

Aggiornamento dei feromoni:

Dopo che tutte le formiche hanno completato il loro cammino, la quantità di feromone su ogni arco viene aggiornata secondo:

$$\tau_{ij}(t+1) = (1 - \rho) \cdot \tau_{ij}(t) + \sum_{k=1}^{m} \Delta\tau_{ij}^{(k)}$$

dove:

- ρ è il tasso di evaporazione ($0 < \rho < 1$),
- $\Delta\tau_{ij}^{(k)}$ è la quantità di feromone depositata dalla k-esima formica.

La quantità depositata dalla k-esima formica è proporzionale alla qualità del percorso:

$$\Delta\tau_{ij}^{(k)} = \begin{cases} \frac{Q}{L_k} & \text{se l'arco } (i,j) \text{ è stato percorso dalla formica } k \\ 0 & \text{altrimenti} \end{cases}$$

dove L_k è la lunghezza totale del percorso della formica k, e Q è una costante positiva.

36.6.3 Implementazione e sperimentazione

Abbiamo sperimentato il funzionamento dell'algoritmo ACO su vari grafi pesati, utilizzando Python e librerie come NetworkX per la gestione dei grafi e Matplotlib per la visualizzazione. L'implementazione è stata testata su grafi di piccole e medie dimensioni.

Di seguito facciamo una panoramica dei dati che vengono raccolti ad ogni esecuzione dell'algoritmo. Le immagini riportate mostrano il grafo iniziale, il grafo con evidenziata la distribuzione del feromone sugli archi a una iterazione generica, la distribuzione percentuale dei feromoni nei vari archi a una iterazione generica e l'andamento del cammino di una formica. Le formiche partono dal nodo sorgente (evidenziato in rosso) e si muovono verso le fonti di cibo (evidenziate in verde), depositando feromone lungo il percorso. Durante le iterazioni, alcuni nodi vengono eliminati ed altri si riaggiungono: l'algoritmo si adatta di conseguenza restituendo sempre ottimi risultati anche in situazioni di grafo dinamico.

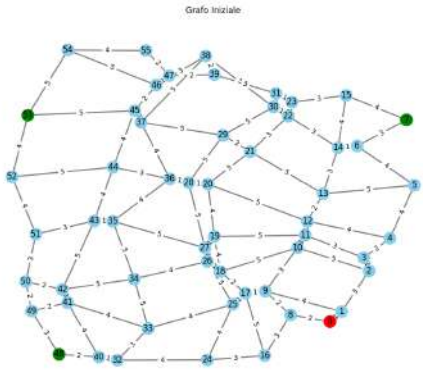


Figura 135: Grafo iniziale con pesi sugli archi. Il nodo rosso rappresenta il nido, i nodi verdi le fonti di cibo

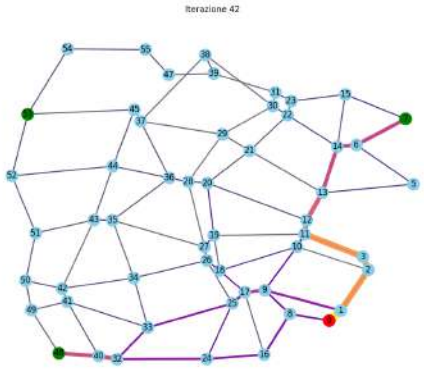


Figura 136: Distribuzione del feromone sugli archi all'iterazione 42. I colori più spessi e accesi indicano la presenza di feromone negli archi.

Più nel dettaglio, l'algoritmo ACO è stato implementato con vari parametri per poter testare la sua efficacia in diverse situazioni. Di seguito, riportiamo una breve descrizione dei parametri più importanti. A questi, vanno aggiunti creazioni di grafi casuali e configurazioni che regolano la dinamicità del grafico.

- `NUM-FORMICHE`: Numero di formiche utilizzate nell'algoritmo. A ogni iterazione, partono dal nido un numero di formiche pari al valore indicato. Valori alti aumentano l'esplorazione in quanto la scelta dei percorsi, anche se vincolata alla probabilità, risulta più ampia. Questo va a scapito dell'esecuzione; valori bassi, invece, riducono la diversità delle soluzioni.

Figura 137: Andamento del percorso di una formica nel corso delle iterazioni del ciclo.

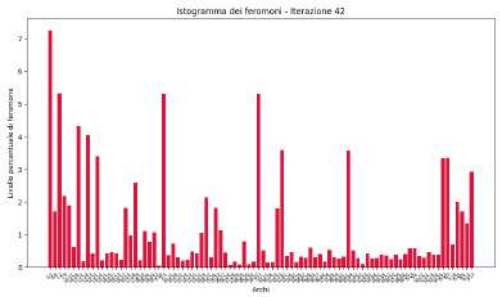


Figura 138: Distribuzione (percentuale) dei feromoni nei vari archi.

- `NUM-ITERAZIONI`: Numero totale di iterazioni dell'algoritmo. Valori alti migliorano la qualità della soluzione ma aumentano il tempo di calcolo; valori bassi possono portare a soluzioni subottimali. Nelle applicazioni reali, le iterazioni vengono lanciate senza un valore massimo fissato, In questo modo, se il grafo cambia dinamicamente, i cammini migliori vengono ricalcolati in tempo reale.
- `ALFA` α: Peso attribuito al feromone nella scelta del cammino. Valori alti favoriscono lo sfruttamento dei cammini già esplorati, rendendo il comportamento delle formiche più conservativo e focalizzato sui percorsi che hanno già dimostrato di essere buoni. Questo può accelerare la convergenza verso una soluzione, ma rischia di bloccare il sistema in minimi locali. Valori bassi, invece, aumentano l'esplorazione di nuovi cammini, favorendo una maggiore diversità nelle soluzioni, ma rallentando la convergenza.
- `BETA` β: Peso attribuito all'euristica (es. distanza) nella scelta del cammino. Valori alti danno maggiore importanza all'euristica, spingendo le formiche a scegliere cammini che sembrano promettenti in base a informazioni statiche come la distanza o il costo. Questo può essere utile in grafi con pesi ben definiti e affidabili. Valori bassi bilanciano l'influenza tra feromone ed euristica, permettendo alle formiche di considerare sia l'esperienza accumulata (feromone) sia le caratteristiche intrinseche del grafo (euristica).
- `EVAPORAZIONE` ρ: Tasso di evaporazione del feromone (valore tra 0 e 1). Questo parametro controlla la velocità con cui il feromone depositato sugli archi diminuisce nel tempo. Valori alti (vicini a 1) fanno sì che le tracce di feromone scompaiano rapidamente, favorendo l'esplorazione di nuovi cammini e riducendo l'influenza delle soluzioni precedenti. Valori bassi (vicini a 0) mantengono più a lungo le tracce di feromone, rafforzando i cammini già esplorati e favorendo lo sfruttamento delle soluzioni esistenti. La scelta del valore dipende dal bilanciamento desiderato tra esplorazione e sfruttamento.
- `FEROMONE-INIZIALE`: Quantità iniziale di feromone su ogni arco. Valori alti favoriscono una scelta più uniforme inizialmente; valori bassi aumentano la casualità.
- `FEROMONE-DEPOSITO` Q: Quantità di feromone depositata dalle formiche. Valori alti rafforzano rapidamente i cammini migliori; valori bassi favoriscono una convergenza più lenta.
- `NODO-PARTENZA`, `NODI-DESTINAZIONE`: Nodi che identificano il nido e le fonti di cibo

36.6.4 Applicazioni dell'algoritmo ACO

L'algoritmo ACO è stato applicato con successo a numerosi problemi di ottimizzazione combinatoria, tra cui:

- il **problema del commesso viaggiatore** (TSP): trovare il percorso più breve che permette a un venditore di visitare un insieme di città una sola volta e tornare al punto di partenza. Questo problema è fondamentale in logistica e trasporti, dove l'ottimizzazione dei percorsi può ridurre significativamente i costi operativi e i tempi di consegna;
- il **routing in reti** di telecomunicazione: ottimizzare il flusso di dati attraverso una rete per minimizzare ritardi e congestioni. Questo è cruciale per garantire la qualità del servizio in applicazioni come streaming video, chiamate VoIP e trasferimenti di file su larga scala;
- la **pianificazione** di operazioni industriali: organizzare sequenze di operazioni per migliorare l'efficienza produttiva e ridurre i tempi morti. Ad esempio, nelle catene di montaggio, una pianificazione ottimale può aumentare la produttività e ridurre i costi di manutenzione;
- il **pathfinding** in robotica e videogiochi: determinare il percorso ottimale per un robot o un personaggio virtuale per raggiungere un obiettivo evitando ostacoli. Questo è essenziale per applicazioni come la navigazione autonoma di robot in ambienti complessi o l'intelligenza artificiale nei giochi per offrire esperienze realistiche e coinvolgenti.

Il grande pregio dell'ACO è la sua capacità di trovare buone soluzioni anche in contesti in cui non esistono strategie deterministiche efficienti, sfruttando la ridondanza, la parallelizzazione e la selezione delle soluzioni migliori.

36.7 ACO vs algoritmi deterministici: Obiettivi e Differenze

36.7.1 L'obiettivo dell'ACO non è la convergenza rigida

A differenza di molti algoritmi classici, il comportamento delle formiche non punta a convergere rapidamente a una soluzione unica e definitiva. L'obiettivo dell'ACO è piuttosto mantenere un bilanciamento tra:

- **Esplorazione** di nuovi cammini potenzialmente migliori;
- **Sfruttamento** dei percorsi che si sono rivelati buoni.

Questa tensione tra curiosità e fiducia è controllata dai parametri α, β, e ρ (rispettivamente peso del feromone, dell'euristica e della velocità di evaporazione).

In questo senso, una convergenza rapida e rigida su un cammino specifico può essere controproducente. Infatti, se le formiche si concentrano troppo presto su un solo percorso, il sistema rischia di bloccarsi in un minimo locale, perdendo la possibilità di esplorare soluzioni migliori.

36.7.2 Quando scegliere ACO rispetto a un algoritmo che trova il cammino in modo deterministico?

In generale, l'algoritmo delle formiche è preferibile quando:

- il grafo cambia nel tempo o contiene incertezze (es. traffico, costi variabili);

- si cercano molte buone soluzioni, non una sola perfetta;
- si vuole una soluzione distribuita e scalabile;
- si affrontano problemi combinatori complessi (es. TSP, vehicle routing, scheduling).

In ambienti controllati e statici, un algoritmo deterministico rimane insuperabile in termini di efficienza e precisione. Ma in molti contesti reali, la natura dinamica e flessibile dell'ACO lo rende una scelta più realistica e potente.