

# Teoria dell'Informazione e Inferenza

Alessandro Verri

24 Gennaio 2024

Queste note raccolgono il materiale utilizzato per le lezioni di *Teoria dell'Informazione e Inferenza*. Le lezioni dalla 1 alla 12, raggruppate nel capitolo 1, illustrano i rudimenti della **Teoria della Probabilità**. Molto del materiale di queste lezioni è tratto dal *Ross* [1]. La parte più avanzata, invece, dal *Motwani & Raghavan* [2]. Le lezioni dalla 13 alla 18 formano il capitolo 2 e introducono i principali concetti della **Teoria dell'Informazione** con alcune incursioni nella **Teoria dei Codici**. La principale fonte è il *McKay* [3] con occasionali puntate al *Khinchin* [4] e al *Reza* [5]. Il capitolo 3, che comprende le ultime cinque lezioni, apre all'**Inferenza** attingendo a piene mani da diverse fonti le più importanti delle quali sono il *Duda e Hart* [6], le note di un corso a NYU di *Miranda Holmes-Cefron* [7] e il *Brémaud* [8]. **Alla fine di ogni lezione trovi una lista dei concetti e dei risultati principali che devi aver capito per poter svolgere gli esercizi con costruito.** Le lezioni o le sezioni marcate con un asterisco coprono argomenti avanzati che non fanno parte necessariamente del programma. Il quarto e ultimo capitolo riporta diversi esercizi svolti inerenti la Teoria della Probabilità. Tutti sono benvenuti a inviare ad [alessandro.verri@unige.it](mailto:alessandro.verri@unige.it) rilievi e correzioni.

# Bibliografia

- [1] S. Ross. *A First Course on Probability*. Prentice Hall, 2010.
- [2] R. Motwani & P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [3] D.J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [4] A.I. Khinchin. *Fondamenti Matematici della Teoria dell'Informazione*. Cremonese, s.l., 1978.
- [5] F.M. Reza. *An introduction to Information Theory*. Dover, 1994.
- [6] R.O. Duda e P.E.Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.
- [7] M. Holmes-Cerfon. *Applied Stochastic Analysis: lecture notes* Courant Institute of Mathematical Sciences, 2019.
- [8] P. Brémaud. *Markov Chains* Springer, 1999



# Indice

<b>1</b>	<b>Elementi di Teoria della Probabilità</b>	<b>7</b>
1.1	Impariamo a contare . . . . .	7
1.2	Definizione assiomatica di probabilità . . . . .	10
1.3	Probabilità condizionata . . . . .	13
1.4	Teorema di Bayes . . . . .	16
1.5	Variabili casuali discrete . . . . .	19
1.6	Distribuzioni discrete di probabilità . . . . .	22
1.7	Variabili casuali continue . . . . .	25
1.8	Distribuzioni continue di probabilità . . . . .	28
1.9	Distribuzioni congiunte e indipendenza . . . . .	31
1.10	Proprietà dei valori attesi . . . . .	34
1.11	Risultati asintotici . . . . .	36
1.12	* Problemi di occupazione . . . . .	39
<b>2</b>	<b>Elementi di Teoria dell'Informazione</b>	<b>43</b>
2.14	Informazione ed entropia di Shannon . . . . .	44
2.15	Entropia congiunta e condizionata . . . . .	48
2.16	Lo stretto indispensabile sulla teoria dei codici . . . . .	52
2.17	Codifiche in assenza di rumore . . . . .	56
2.18	Codifica aritmetica . . . . .	60
2.19	Codifiche in presenza di rumore . . . . .	64
<b>3</b>	<b>Elementi di Inferenza</b>	<b>69</b>
3.20	Metodi Monte Carlo . . . . .	70
3.21	Catene di Markov . . . . .	74
3.22	* Catene di Markov <i>Monte Carlo</i> . . . . .	78
3.23	Inferenza Bayesiana . . . . .	82
3.24	Inferenza frequentista . . . . .	86
<b>4</b>	<b>Esercizi risolti di Teoria della Probabilità</b>	<b>91</b>



# Capitolo 1

## Elementi di Teoria della Probabilità

### 1.1 Impariamo a contare

Introduciamo gli elementi principali del calcolo combinatorio, utili per calcolare la probabilità associata a eventi nel caso discreto.

#### Principio base

Nel seguito faremo uso della nozione di *esperimento*, ovvero di un'azione il cui *risultato* è un elemento di un insieme costituito da tutti i *possibili risultati* di quell'azione. Tutto quello che vedremo è ottenibile mediante l'applicazione di un principio molto semplice che enunciamo di seguito.

##### Principio 1.1.1. *Uno alla volta*

Se un esperimento fornisce  $m$  possibili risultati e se per ciascuno di essi un secondo esperimento fornisce  $n$  possibili risultati, allora i due esperimenti forniscono  $m \times n$  possibili risultati.  $\square$

Due le vere difficoltà che si incontrano nell'applicazione di questo principio: *definire* in modo univoco ogni esperimento e *individuare* il numero dei possibili risultati di ogni esperimento definito.

##### Esercizio 1.1.1. *Tutte le targhe*

Se una targa è formata da 4 lettere e 3 cifre, quante sono le targhe possibili?

*Soluzione*

Assumiamo che le lettere possibili siano 26 e le cifre 10. La scelta di ogni elemento della targa può essere visto come un esperimento, l'elemento scelto uno dei possibili risultati.

**Ripetizioni ammesse:** 26 risultati per il primo, 26 per il secondo, 26 per il terzo, 26 per il quarto, 10 per il quinto, 10 per il sesto e 10 per il settimo e ultimo esperimento. In tutto i diversi risultati possibili sono

$$26^4 \times 10^3 = 456,976,000$$

**Ripetizioni non ammesse:** 26 risultati per il primo, 25 per il secondo, 24 per il terzo, 23 per il quarto, 10 per il quinto, 9 per il sesto e 8 per il settimo e ultimo esperimento. In tutto i diversi risultati possibili sono

$$26 \times 25 \times 24 \times 23 \times 10 \times 9 \times 8 = 258,336,000$$

Consideriamo ora tre casi importanti: permutazioni, disposizioni e combinazioni.

## Permutazioni

Una *permutazione* è un particolare ordinamento di  $n$  oggetti. Applicando il principio base per contare le permutazioni possibili, otteniamo  $n!$  poichè abbiamo  $n$  scelte per il primo oggetto,  $n - 1$  per il secondo e così via fino alla scelta obbligata dell' $n$ -esimo oggetto, ultimo rimasto. In una permutazione tutti gli  $n$  oggetti sono distinguibili.

### Esercizio 1.1.2. Tutti gli ordinamenti

In quanti modi è possibile ordinare su uno scaffale 2 libri di chimica, 3 di fisica, 4 di matematica e 5 di informatica in modo che i libri di una stessa materia siano in un unico gruppo di libri consecutivi?

*Soluzione*

**Principio base applicato alle materie:** Se ignoriamo l'ordine con cui sono disposti i libri di una stessa materia, abbiamo 4 scelte per la prima, 3 per la seconda e 2 per la terza materia. In tutto i possibili ordinamenti per materia sono

$$4! = 24$$

**Principio base applicato anche ai libri di ogni materia:** Se invece consideriamo anche l'ordine con cui sono disposti i libri di una stessa materia dobbiamo moltiplicare per tutte le permutazioni possibili dei libri di ogni singola materia. Pertanto, in tutto i possibili ordinamenti in questo caso sono

$$24 \times 2! \times 3! \times 4! \times 5! = 829,440$$

Notiamo che in assenza di qualunque vincolo le possibili permutazioni sono molte di più, ovvero

$$14! = 87,178,291,200$$

## Disposizioni

Una *disposizione* è un particolare ordinamento di  $i$  oggetti scelti da  $n$  oggetti con  $i \leq n$ . Se applichiamo il principio base per contare le disposizioni possibili, otteniamo

$$n(n-1) \dots (n-i+1)$$

perchè abbiamo  $n$  scelte per il primo,  $n - 1$  per il secondo e così via fino alla scelta dell' $i$ -esimo oggetto tra gli  $n - i + 1$  oggetti rimasti.

### Osservazione 1.1.1. Pensando in termini di permutazioni

Dall'identità

$$n(n-1) \dots (n-i+1) = \frac{n!}{(n-i)!}$$

segue che le disposizioni possibili di  $i$  oggetti scelti tra  $n$  possono essere ottenute anche ragionando in modo diverso. Ovvero considerando distinguibili gli  $i$  oggetti scelti e indistinguibili gli  $n-i$  oggetti della cui disposizione non ci curiamo. Delle  $n!$  permutazioni possibili di  $n$  oggetti, consideriamo equivalenti quelle in cui gli  $i$  oggetti scelti si trovano nelle stesse posizioni. Queste permutazioni sono  $(n-i)!$  ovvero tante quante le possibili permutazioni degli  $n-i$  oggetti non scelti che consideriamo indistinguibili.

### Esercizio 1.1.3. Tutti gli anagrammi

Gli anagrammi di *CINEMA* (la maggior parte dei quali non fornisce parole di senso compiuto) sono  $6! = 720$ . Quanti sono, invece, gli anagrammi di *ERRORE*?

*Soluzione*

Dividendo per  $3!$  (per le 3 *R*) e per  $2!$  (per le 2 *E*) i possibili  $6!$  anagrammi di *ERRORE* otteniamo

$$\frac{6!}{3! \times 2!} = 5 \times 4 \times 3 = 60$$



## Combinazioni

Una *combinazione* è una scelta di  $i$  oggetti da  $n$  oggetti con  $i \leq n$ . Se applichiamo il principio base per contare le combinazioni possibili e teniamo presente che l'ordine, questa volta, è irrilevante sia per gli  $i$  oggetti scelti sia per gli  $n - i$  oggetti non scelti, otteniamo

$$\frac{n!}{i!(n-i)!} = \binom{n}{i}$$

### Esercizio 1.1.4. Tutti i comitati

Quanti comitati di tre persone possiamo formare partendo da un gruppo di 20 persone?

*Soluzione*

$$\binom{20}{3} = \frac{20!}{17! \times 3!} = \frac{20 \times 19 \times 18}{6} = 1140$$

Le nozioni di permutazione, disposizione e combinazione sono illustrate in un caso particolare nel diagramma di Venn di Figura 1.1.

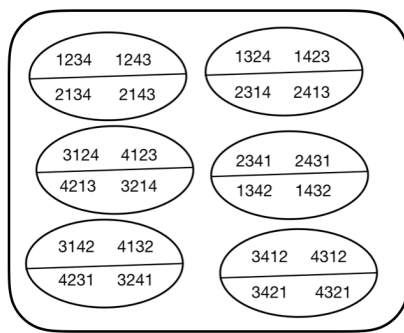


Figura 1.1: Le  $4! = 24$  permutazioni di 1, 2, 3 e 4 sono rappresentate come tutte le sequenze possibili delle quattro cifre. Le  $4 \times 3 = 12$  disposizioni di 1 e 2 sono le 12 coppie all'interno della parte all'alta e della parte bassa delle 6 ellissi. Poiché nelle disposizioni le posizioni degli oggetti scelti contano, in ognuna delle 12 coppie la posizione di 1 e 2 non cambia. Infine le  $\binom{4}{2} = 6$  combinazioni sono rappresentate dalle ellissi. Per le quattro coppie in ogni ellisse sia le posizioni di 1 e 2 sia le posizioni di 3 e 4 possono essere scambiate.

**Cose che devi rimanerti di questa lezione:** L'unico vero concetto è il *principio base*. Sincerati di sapere come applicarlo per risolvere semplici conteggi. Non serve imparare a memoria alcuna formula.

## 1.2 Definizione assiomatica di probabilità

Definiamo ora la probabilità in modo assiomatico discutendo brevemente alcune proprietà nel caso semplice, ma importante, in cui sia possibile individuare eventi equiprobabili. Ci limitiamo al caso discreto in modo da non dover affrontare il tema di quali sono gli eventi misurabili, ovvero cui è possibile associare una probabilità, e quelli che non lo sono.

### Nozioni fondamentali

*Spazio campionario*: l'insieme  $S$  dei possibili risultati di un esperimento (*testa e croce, le sei facce di un dado, oppure il numero di persone in coda a una cassa*).

*Evento*: un qualunque sottoinsieme  $E$  di  $S$  che si *realizza* se il risultato dell'esperimento appartiene a  $E$  (*testa nel lancio di una moneta, faccia dispari nel lancio di un dado, o quattro persone in coda*).

Nel caso discreto, un evento  $E$  è un qualunque sottoinsieme di  $S$ , ovvero un elemento dell'insieme delle parti di  $S$ . Indichiamo con  $E \cup F$  l'unione degli eventi  $E$  e  $F$  e con  $EF$  la loro *intersezione*. Inoltre, due eventi  $E$  e  $F$  tali che  $EF = \emptyset$  sono *mutuamente esclusivi*. L'evento  $E^c$  tale che  $E \cup E^c = S$  è il *complementare* di  $E$ . Il caso continuo è considerabilmente più complicato e non lo cureremo in dettaglio.

### Assiomi

Una probabilità  $P(\cdot)$  risulta ben definita sugli eventi di uno spazio campionario  $S$  se

$$\mathbf{A1} \quad 0 \leq P(E) \leq 1 \quad \forall E \subseteq S$$

$$\mathbf{A2} \quad P(S) = 1$$

$\mathbf{A3}$  Se gli eventi  $E_i$ , con  $i = 1, 2, \dots$  sono mutuamente esclusivi, allora

$$P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$$

Sono conseguenze immediate di **A1**, **A2** e **A3**

$$(i) \quad \forall E, P(E^c) = 1 - P(E)$$

$$(ii) \quad \forall E \text{ e } F, \text{ se } E \subseteq F \text{ allora } P(E) \leq P(F)$$

$$(iii) \quad \forall E \text{ e } F, P(E \cup F) = P(E) + P(F) - P(EF)$$

*Dimostrazione*: la prima segue dal fatto che  $E^c E = \emptyset$  e  $E^c \cup E = S$ , per cui  $P(E) + P(E^c) = 1$ . La seconda si dimostra notando che  $F = E \cup E^c F$  con  $E$  e  $E^c F$  mutuamente esclusivi. Per la terza e ultima conviene scrivere  $E \cup F$  e  $EF$  come unione di insiemi disgiunti, ovvero  $E \cup F = E \cup E^c F$  e  $F = EF \cup E^c F$  e applicare l'assioma **A3**. ■

#### Esercizio 1.2.1. Diagrammi di Venn

L'uso dei concetti di base della teoria degli insiemi consente di trattare in modo preciso, ma allo stesso tempo intuitivo, diverse proprietà delle probabilità. Disegna un diagramma di Venn per  $S$ ,  $E$  ed  $F$  e verifica le conseguenze (i), (ii) e (iii) utilizzando gli assiomi **A1**, **A2** e **A3**.

**Esercizio 1.2.2. Chi canta e chi suona**

Quanti sono i componenti di un complesso in cui 3 cantano, 3 suonano la chitarra e 2 entrambe le cose?

*Soluzione*

Se  $E = \{i \text{ tre cantanti}\}$  ed  $F = \{i \text{ tre chitarristi}\}$  abbiamo che  $EF = \{i \text{ due cantanti chitarristi}\}$ . Dalla proprietà (iii), pertanto, segue che

$$3 + 3 - 2 = 4$$

**Eventi equiprobabili**

Supponiamo che  $S$  sia costituito da un insieme finito di  $N$  risultati che indichiamo con i primi  $N$  numeri naturali, ovvero  $S = \{1, 2, \dots, N\}$ . Se le probabilità  $P(i)$  sono tutte uguali allora  $P(i) = 1/N$ . La probabilità di  $E$ , in questo caso, si calcola come frazione del numero di risultati in  $E$ ,  $\#E$ , sul numero di risultati in  $S$ ,  $\#S$ .

**Esercizio 1.2.3. Lancio di 2 dadi**

Calcola la probabilità  $P$  di ottenere 7 lanciando 2 dadi.

*Soluzione*

Le coppie di risultati ottenibili dal lancio di due dadi sono 36. I casi favorevoli sono sei in tutto: (1,6), (2,5), (3,4), (4,3), (5,2) e (6,1), per cui  $P = 6/36 = 1/6$ .

**Osservazione 1.2.1. Attenzione all'ordinamento!**

Le coppie sono ordinate: il primo elemento è il risultato del lancio del primo dado, il secondo elemento il risultato del lancio del secondo dado.

**Esercizio 1.2.4. Palline bianche e rosse**

Calcola la probabilità  $P$  di estrarre 1 pallina bianca e 2 palline rosse da un'urna con 6 palline bianche e 5 rosse.

*Soluzione con le combinazioni*

Se consideriamo l'insieme delle palline estratte come non ordinato, i casi possibili sono le combinazioni di 3 palline scelte tra 11, i favorevoli quelle di 1 pallina bianca scelta tra 6 e 2 nere scelte tra 5, ovvero

$$\binom{11}{3} = 165 \text{ e } \binom{6}{1} \binom{5}{2} = 60$$

da cui segue  $P = 60/165 = 4/11$ .

*Soluzione con le disposizioni*

Consideriamo ora invece rilevante l'ordine col quale estraiano le palline. I casi possibili sono le disposizioni di 3 palline scelte tra 11, ovvero  $11 \times 10 \times 9 = 990$ . Dividiamo i casi favorevoli in 3 gruppi. Nel primo gruppo la pallina bianca è estratta per prima,  $6 \times 5 \times 4$  casi favorevoli, nel secondo per seconda,  $5 \times 6 \times 4$  casi favorevoli, e nel terzo per terza,  $5 \times 4 \times 6$  casi favorevoli. Otteniamo ancora  $P = 3 \times 120/990 = 4/11$ .

**Esercizio 1.2.5. Paradosso del compleanno (o delle collisioni nell'hashing)**

Calcola la probabilità  $P_n$  che  $n$  individui festeggino il compleanno in  $n$  giorni diversi.

*Soluzione*

Consideriamo l'evento complementare ovvero che nessuno tra  $n$  individui sia nato lo stesso giorno. Perché un secondo individuo non sia nato nel giorno del primo i casi favorevoli sono  $(365-1) = 364$ , per un terzo  $365-2=363$  e per l' $n$ -esimo  $365 - n + 1$ . Applicando il principio base nel caso di eventi equiprobabili, la probabilità che tutti gli  $n$  individui siano nati in giorni diversi è allora

$$P_n = \frac{(365-1)(365-2) \dots (365-n+1)}{365^{n-1}}$$

**Osservazione 1.2.2. Scommettiamo?**

Al crescere di  $n$  la probabilità  $P_n$  diminuisce velocemente. Per  $n = 100$  abbiamo che  $P_{100} < 0.000001$  (esperimento in classe). L'ultimo valore di  $n$  per cui  $P_n < 1/2$  è 23.

**Probabilità soggettiva**

Capita di associare il concetto di probabilità a eventi incerti non ripetibili (probabilità di pioggia a Milano o di vittoria in un incontro di scherma). In questi casi parliamo di *probabilità soggettiva*.

**Esercizio 1.2.6. Piove o non piove?**

La probabilità dell'evento *oggi pioverà* è del 40% e che *domani pioverà* è del 30%. Se la probabilità che *oggi o domani pioverà* è del 60% dimostra che non è possibile che la probabilità che *oggi e domani pioverà* sia del 20%.

*Soluzione*

Se  $E$  è l'evento *oggi pioverà* e  $F$  *domani pioverà* sappiamo che  $P(E \cup F) = P(E) + P(F) - P(EF)$ . Nel nostro caso, invece, abbiamo che

$$40\% + 30\% - 20\% = 50\% \neq 60\%$$

**Caso generale**

Che cosa succede se l'ipotesi di equiprobabilità non vale? A parte il fatto che il calcolo delle probabilità può diventare molto più complicato non cambia molto.

**Esercizio 1.2.7. Dado truccato**

La probabilità delle sei facce di un dado sono

$$P(1) = 0.5, \quad P(2) = 0.3, \quad P(3) = 0.1, \quad P(4) = 0.05, \quad P(5) = 0.03, \quad \text{e} \quad P(6) = 0.02$$

Calcola la probabilità di ottenere un numero pari con un solo lancio e la probabilità di ottenere 6 con due lanci.

*Soluzione*

Nel primo caso  $S$  è costituito dai 6 possibili risultati. Pertanto avremo

$$P(\text{ottenere un numero pari con un lancio}) = P(2) + P(4) + P(6) = 0.3 + 0.05 + 0.02 = 0.37$$

Nel secondo  $S$  è costituito da tutte le 36 coppie dei possibili risultati in cui il primo elemento è il risultato del primo lancio e il secondo il risultato del secondo. La probabilità di ciascuna coppia è data dal prodotto delle probabilità. La probabilità di ottenere 6 in due lanci è allora quella delle coppie (1, 5), (2, 4), (3, 3), (4, 2) e (5, 1). Pertanto avremo

$$P(\text{ottenere 6 con due lanci}) = 2P(1)P(5) + 2P(2)P(4) + P(3)P(3) = 0.07$$

**Cose che devi rimanerti di questa lezione:** Allenati a riconoscere quale sia lo spazio campionario, l'unione e l'intersezione di eventi e l'evento complementare in alcuni semplici esempi. Assicurati di aver capito che la probabilità è una *misura*: disegna un diagramma di Venn nel caso in cui gli eventi sono equiprobabili e nel caso in cui non lo sono e impara a valutare le conseguenze.

### 1.3 Probabilità condizionata

Introduciamo ora il concetto forse più importante della teoria della probabilità, la probabilità condizionata, ovvero *la probabilità di un evento una volta che si venga a conoscenza della realizzazione di un altro evento casuale*.

#### Esercizio 1.3.1. Probabilità che cambia

Lancia una volta un dado onesto. Qual è la probabilità di aver ottenuto 1 se sai che il risultato del lancio è tra 1 e 3 o se invece sai che il risultato è tra 4 e 6?

*Soluzione*

Se  $E$  è l'evento di ottenere 1 e  $F$  l'evento di ottenere  $\{1, 2, 3\}$ , abbiamo

$$P(F) = \frac{3}{6} \text{ e } P(EF) = P(E) + P(F) - P(E \cup F) = \frac{1}{6} + \frac{3}{6} - \frac{3}{6}$$

La probabilità dell'evento  $E$ , condizionata alla realizzazione dell'evento  $F$ , riduce i possibili risultati dell'esperimento da 6 a 3 ed è quindi  $1/3$ , pari al rapporto tra  $P(EF)$  e  $P(F)$ . Ragionando in modo analogo nel caso in cui  $F = \{4, 5, 6\}$  otteniamo una probabilità condizionata pari a 0. In conclusione la probabilità di ottenere 1, inizialmente uguale a  $1/6$ , si modifica in funzione delle informazioni aggiuntive che potrebbero rendersi disponibili.  $\square$

In generale, dati due eventi  $E$  e  $F$ , siamo interessati a calcolare la probabilità di  $E$  quando sappiamo che si è realizzato  $F$ . La probabilità di  $E$  condizionata a  $F$ , indicata con  $P(E|F)$ , è definita come

$$P(E|F) = \frac{P(EF)}{P(F)}$$

#### Osservazione 1.3.1. Riduzione dello spazio campionario

Se si realizza anche  $E$  dopo che si è realizzato  $F$ , significa che il risultato appartiene all'intersezione  $EF$ . Lo spazio campionario per la realizzazione di  $E$  successiva alla realizzazione di  $F$  si è quindi ridotto da  $S$  a  $F$  e misura  $P(F)$ .

#### Esercizio 1.3.2. Sempre una probabilità

Dimostra che  $P(\cdot|F)$  è una probabilità e quindi

**A1**  $\forall E \ 0 \leq P(E|F) \leq 1$

**A2**  $P(S|F) = 1$

**A3** Se gli eventi  $E_i$  con  $i = 1, \dots$  sono mutuamente esclusivi allora

$$P((\cup_i E_i) | F) = \sum_i P(E_i | F)$$

*Dimostrazione:* l'assioma **A1** è soddisfatto in quanto poiché  $\forall E \ EF \subseteq F$  e quindi  $P(EF) \leq P(F)$ , l'assioma **A2** in quanto  $P(SF) = P(F)$ . Per la verifica dell'assioma **A3**, invece, ricordiamo che per via della mutua esclusività degli  $E_i$  abbiamo

$$(\cup_i E_i)F = \cup_i E_i F \text{ e } P(\cup_i E_i F) = \sum_i P(E_i F)$$

da cui otteniamo

$$P((\cup_i E_i) | F) = \frac{P((\cup_i E_i) F)}{P(F)} = \frac{P(\cup_i E_i F)}{P(F)} = \frac{\sum_i P(E_i F)}{P(F)} = \sum_i P(E_i | F) \quad \blacksquare$$

**Esercizio 1.3.3.** *L'importanza di una congiunzione*

Hai due monete,  $A$  e  $B$ , indistinguibili. La moneta  $A$  restituisce *testa* con una probabilità del 70%, la moneta  $B$  con probabilità del 40%. Prendi una moneta a caso e la lanci. Con quale probabilità prendi la moneta  $A$  e ottieni *testa*?

*Soluzione*

Siano  $A$  l'evento *prendo la moneta A* e  $T$  *ottengo testa*. Per la probabilità dell'evento intersezione  $A$  e  $T$  otteniamo

$$P(AT) = P(T|A)P(A) = 0.7 \times 0.5 = 35\%$$

Osserviamo che  $P(AT)$  è molto diversa da  $P(T|A)$ . Quasi più facile interpretare il risultato ottenuto in termini degli eventi complementari: il complementare dell'evento  $AT$  è l'unione dei complementari dei singoli eventi (*prendo la moneta B o ottengo croce*), mentre il complementare dell'evento  $T|A$  è *ottengo croce con la moneta A*.

**Osservazione 1.3.2.** *Regola della moltiplicazione*

Generalizziamo l'identità  $P(EF) = P(F)P(E|F)$  al caso dell'intersezione di  $n$  eventi. Abbiamo che

$$P(E_1 E_2 \dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2) \dots P(E_n|E_1 E_2 \dots E_{n-1})$$

Per dimostrare la validità di questa uguaglianza è sufficiente utilizzare la definizione di probabilità condizionata in modo iterativo su ognuno degli  $n - 1$  fattori. Ad esempio, per tre insiemi  $E, F, G$ , si ha

$$P(EFG) = \frac{P(EFG)}{P(EF)} \frac{P(EF)}{P(E)} P(E) = P(G|EF)P(F|E)P(E)$$

**Esercizio 1.3.4.** *Assi equidistribuiti*

In un mazzo di cinquantadue carte che contiene quattro assi qual è la probabilità che dividendo le carte in quattro pile da tredici ogni pila contenga un asso?

*Soluzione*

Poniamo

$E_1$  : l'asso di picche è in uno qualunque delle quattro pile

$E_2$  : l'asso di picche e l'asso di cuori sono in due pile diverse

$E_3$  : l'asso di picche, l'asso di cuori e l'asso di quadri sono in tre pile diverse

$E_4$  : l'asso di picche, l'asso di cuori, l'asso di quadri e l'asso di fiori sono in quattro pile diverse

Dobbiamo calcolare  $P(E_1 E_2 E_3 E_4)$ . Per la regola della moltiplicazione

$$P(E_1 E_2 E_3 E_4) = P(E_1)P(E_2|E_1)P(E_3|E_1 E_2)P(E_4|E_1 E_2 E_3)$$

Ora,  $P(E_1) = 1$ , mentre  $P(E_2|E_1) = 39/51$ , perché dobbiamo sottrarre a 52 l'asso di picche al denominatore e le tredici carte della pila che lo contiene al numeratore,  $P(E_3|E_1 E_2) = 26/50$  perché dobbiamo sottrarre l'asso di cuori al denominatore e le altre tredici carte della pila che lo contiene al denominatore e  $P(E_4|E_1 E_2 E_3) = 13/49$  perché dobbiamo sottrarre l'asso di quadri al denominatore e le altre tredici carte della pila che lo contiene al numeratore. Mettendo tutto assieme otteniamo

$$P(E_1 E_2 E_3 E_4) = \frac{39}{51} \frac{26}{50} \frac{13}{49} = \frac{13 \times 26 \times 39}{49 \times 50 \times 51} \approx 0.105$$

**Esercizio 1.3.5.** *Prima una e poi l'altra*

Un'urna contiene otto palline rosse e quattro bianche. Se estraiamo due palline qual è la probabilità che entrambe siano rosse?

*Soluzione*

Se  $R_1$  è l'evento che la prima pallina è rossa e  $R_2$  l'evento che la seconda pallina è rossa avremo  $P(R_1) = 2/3$  e  $P(R_2|R_1) = 7/11$ . Pertanto

$$P(R_1 R_2) = P(R_1)P(R_2|R_1) = \frac{2}{3} \frac{7}{11} = \frac{14}{33}$$

Otteniamo lo stesso risultato mediante le combinazioni: i casi favorevoli sono contati come la scelta di due palline rosse tra le otto, contro tutti i casi possibili, due palline tra le dodici, ovvero

$$P(R_1 R_2) = \frac{\binom{8}{2}}{\binom{12}{2}} = \frac{28}{66} = \frac{14}{33}$$

**Esercizio 1.3.6.** *Un risultato sorprendente*

Lancia due volte una moneta onesta e calcola la probabilità di ottenere 2 volte testa,  $T$ , se (i) il risultato del primo lancio è  $T$ , o se (ii) il risultato di uno dei due lanci è  $T$ .

*Soluzione*

Nel caso (i) gli eventi sono  $E = \{T, T\}$  e  $F = \{\{T, C\}, \{T, T\}\}$ , e poichè

$$P(E) = P(\{T, T\}) \text{ e } P(F) = P(\{\{T, C\}, \{T, T\}\})$$

allora

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{1/4}{2/4} = \frac{1}{2}$$

Nel caso (ii) gli eventi sono  $E = \{T, T\}$  e  $F = \{\{T, C\}, \{T, T\}, \{C, T\}\}$  e, poichè

$$P(E) = P(\{T, T\}) \text{ e } P(F) = P(\{\{T, C\}, \{T, T\}, \{C, T\}\})$$

allora

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{1/4}{3/4} = \frac{1}{3}$$

La differenza tra (i) e (ii) è da imputarsi al diverso numero di risultati possibili nei due casi!

**Osservazione 1.3.3.** *Notazione pesante ma necessaria per capire e non cadere in tentazione*

L'uso delle parentesi graffe appesantisce la notazione ma chiarisce che la probabilità di un evento è la misura di un sottoinsieme dello spazio campionario (in questo caso le coppie di possibili risultati ottenibili lanciando due volte una moneta onesta). Lasciarsi guidare dall'intuizione con la probabilità è raramente una buona idea!

**Cose che devi rimanerti di questa lezione:** La nozione di probabilità condizionata è di importanza fondamentale. Familiarizza con la notazione e assicurati di aver capito la differenza tra  $P(E)$  e  $P(E|F)$ . Costruisci esempi in cui  $P(E|F) < P(E)$ ,  $P(E|F) = P(E)$  e  $P(E|F) > P(E)$ . Gli esercizi in questa lezione servono a consolidare il concetto di probabilità condizionata. Imparare a risolverli senza aver capito non serve assolutamente a nulla.

## 1.4 Teorema di Bayes

Appena dietro la nozione di probabilità condizionata arriva il teorema di Bayes, risultato di fondamentale importanza in moltissime applicazioni.

Per ogni coppia di eventi  $E$  e  $F$ , applicando la definizione di probabilità condizionata, possiamo riscrivere  $P(F|E)$  in termini di  $P(E|F)$  (o viceversa). Ovvero

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)} \quad (1.1)$$

o

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

Supponiamo ora che  $E$  sia l'unione dei due (o più) eventi mutuamente esclusivi,  $EF$  e  $EF^c$ . In questo modo, applicando la definizione di probabilità condizionata, la probabilità  $P(E)$  può essere scritta come probabilità totale, ovvero

$$P(E) = P(EF) + P(EF^c) = P(F)P(E|F) + P(F^c)P(E|F^c) \quad (1.2)$$

dove nel primo passaggio abbiamo utilizzato la mutua esclusività mentre nel secondo la definizione di probabilità condizionata.

Sostituendo la formula (1.2) nella (1.1) otteniamo infine la formula nota come *teorema di Bayes*

$$P(F|E) = \frac{P(F)P(E|F)}{P(F)P(E|F) + P(F^c)P(E|F^c)}$$

Vediamone alcune applicazioni.

### Esercizio 1.4.1. Probabilità totale

Hai due monete uguali  $A$  e  $B$ . La probabilità di *testa* per  $A$  è  $1/2$ , per  $B$   $1/10$ . Qual è la probabilità di ottenere *testa* lanciando una moneta a caso?

*Soluzione*

Abbiamo  $P(A) = P(B) = 1/2$ ,  $P(\text{testa}|A) = 1/2$  e  $P(\text{testa}|B) = 1/10$ . Pertanto

$$P(\text{testa}) = P(\text{testa}|A)P(A) + P(\text{testa}|B)P(B) = \frac{1}{2} \times \frac{1}{2} + \frac{1}{10} \times \frac{1}{2} = \frac{3}{10}$$

### Esercizio 1.4.2. Le apparenze possono ingannare

Un tizio si reca dal medico perché sospetta di essere affetto da una certa patologia. Effettua un test che ha una *sensibilità* del 95%, probabilità che una persona affetta dalla patologia risulti positiva, e una *specificità* del 99%, probabilità che una persona sana risulti negativa. Sapendo che la patologia colpisce lo 0.2% della popolazione, con quale probabilità il tizio soffre della patologia se risulta positivo al test?

*Soluzione*

Se  $Pos$  è l'evento di risultare positivo al test e  $Neg$  l'evento di risultare negativo per le verosimiglianze abbiamo

$$\Pr(Pos | malato) = 95\% \quad \text{e} \quad \Pr(Neg | sano) = 99\%$$

Per la probabilità *a priori*  $\Pr(malato)$  abbiamo  $\Pr(malato) = 0.2\%$ , mentre

$$\Pr(Pos | sano) = 1 - \Pr(Neg | sano) = 1\%$$

Dal teorema di Bayes, pertanto, per la probabilità *a posteriori*  $\Pr(malato | Pos)$  otteniamo

$$\begin{aligned} \Pr(malato | Pos) &= \frac{\Pr(malato)\Pr(Pos | malato)}{\Pr(malato)\Pr(Pos | malato) + \Pr(sano)\Pr(Pos | sano)} \\ &= \frac{0.002 \times 0.95}{0.002 \times 0.95 + 0.998 \times 0.01} = \frac{0.00190}{0.00190 + 0.00998} \approx 1.6\% \end{aligned}$$



La probabilità *a posteriori* ottenuta è circa 8 volte più grande di quella *a priori*, ma per la rarità della patologia la probabilità che il paziente sia effettivamente affetto dalla patologia continua a essere piuttosto piccola. Più del 98% dei soggetti positivi al test, inoltre, sono sani!

**Osservazione 1.4.1.** Interpretazione della probabilità di essere sano o malato

In fin dei conti una persona o è sana o è malata. Che senso possiamo attribuire alla probabilità di essere sano o malato? Ci aspettiamo che, ripetendo il test su molti individui, la frazione dei malati tra i soggetti positivi al test si avvicini a  $\Pr(\text{malato} \mid \text{Pos})$ .

**Esercizio 1.4.3.** Paradosso delle tre carte

Ci sono tre carte *A*, *B* e *C*. La carta *A*, è rossa sul lato 1 e sul lato 2, la carta *B* su rossa sull'1 e bianca sul 2, mentre la carta *C* è bianca su entrambi i lati. Ponendo su un tavolo una delle tre carte, scelta a caso, ottengo che il lato visibile è di colore rosso. Qual è la probabilità che anche il lato non visibile sia di colore rosso?

*Prima soluzione*

Calcoliamo i casi favorevoli e tutti i casi possibili. Estruendo una carta e posandola sul tavolo si possono verificare i sei casi equiprobabili in Figura 1.2.













	vis nasc		vis nasc	
<b>A</b>				
<b>B</b>				
<b>C</b>				

Figura 1.2: Vedi testo.

Escludendo gli ultimi tre casi col lato visibile bianco, rimangono tre casi col lato visibile rosso, due dei quali nascondono un lato anch'esso rosso. La probabilità è quindi pari a  $2/3$ .

*Seconda soluzione*

Utilizzando il teorema di Bayes e tenendo conto che la carta *C* non ha lati rossi, otteniamo

$$P(A|\text{vis rosso}) = \frac{P(\text{vis rosso}|A)P(A)}{P(\text{vis rosso})} = \frac{P(\text{vis rosso}|A)P(A)}{P(A)P(\text{vis rosso}|A) + P(B)P(\text{vis rosso}|B)}$$

Ora,  $P(A) = P(B) = 1/3$ . Inoltre, la carta *A* ha due lati rossi, per cui  $P(\text{vis rosso}|A) = 1$ , mentre la carta *B* uno solo per cui  $P(\text{vis rosso}|B) = 1/2$ . Mettendo tutto assieme otteniamo

$$\begin{aligned} P(\text{vis rosso}) &= P(A)P(\text{vis rosso}|A) + P(B)P(\text{vis rosso}|B) \\ &= \frac{1}{3} \times 1 + \frac{1}{3} \times \frac{1}{2} = \frac{1}{2} \end{aligned}$$

per cui

$$P(A|\text{vis rosso}) = \frac{1 \times 1/3}{1/2} = \frac{2}{3}$$

**Esercizio 1.4.4.** Monty Hall

Dietro una delle tre porte *a*, *b* e *c* c'è una macchina, dietro ognuna delle altre due una capra. Scopo del

gioco è indovinare la porta dietro la quale si trova la macchina. Scegli la porta  $a$  e Monty, che vede che cosa si nasconde dietro le tre porte, scopre la porta  $c$  mostrandoti una capra. È meglio mantenere la scelta e chiedere di aprire la porta  $a$  o cambiare scelta e chiedere di aprire la porta  $b$ ?

*Soluzione*

Sia  $R_c$  l'evento *Monty sceglie di aprire la porta  $c$* . Indichiamo con  $X$  l'evento *dietro la porta  $x$  c'è una macchina*. Valutiamo prima di tutto  $P(R_c|A)$ ,  $P(R_c|B)$  e  $P(R_c|C)$ . Se la macchina è dietro la porta  $a$ , Monty può aprire le porte  $b$  e  $c$  con uguale probabilità, per cui  $P(R_c|A) = 1/2$ . Se la macchina è dietro la porta  $b$ , Monty può aprire solo la porta  $c$ , per cui  $P(R_c|B) = 1$ . Se la macchina è dietro la porta  $c$ , Monty non può aprire la porta  $c$ , per cui  $P(R_c|C) = 0$ . Pertanto

$$P(A|R_c) = \frac{P(A)P(R_c|A)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3} \frac{1}{2}}{\frac{1}{3} \frac{1}{2} + \frac{1}{3} 1 + \frac{1}{3} 0} = \frac{1}{3}$$

$$P(B|R_c) = \frac{P(B)P(R_c|B)}{P(A)P(R_c|A) + P(B)P(R_c|B) + P(C)P(R_c|C)} = \frac{\frac{1}{3} 1}{\frac{1}{3} \frac{1}{2} + \frac{1}{3} 1 + \frac{1}{3} 0} = \frac{2}{3}$$

Se questo risultato ci lascia perplessi proviamo a pensare di ripetere il gioco con 1,000,000 porte, una macchina e 999,999 capre e con Monty che, dopo la nostra scelta, apre 999,998 porte mostrandoci altrettante capre...

## Eventi indipendenti

Due eventi sono *indipendenti* se la realizzazione di uno non modifica la probabilità dell'altro. Se  $E$  e  $F$  sono indipendenti allora

$$P(EF) = P(E)P(F)$$

ovvero  $P(E|F) = P(E)$  e  $P(F|E) = P(F)$ .

### Esercizio 1.4.5. Ancora un doppio lancio

Lancia un dado onesto due volte. Verifica che l'evento  $E_1$  *la somma dei due risultati è 7* e l'evento  $F$  *il primo risultato è 4* sono indipendenti, mentre l'evento  $E_2$  *la somma dei due risultati è 6* e  $F$  non lo sono.

*Soluzione*

Chiaramente abbiamo che  $P(F) = 1/6$  e  $P(E_1F) = P(E_2F) = 1/36$ . Per quanto riguarda  $E_1$  abbiamo 6 casi favorevoli su 36 possibili, per cui  $P(E_1) = 1/6$ . Per  $E_2$ , invece, i casi favorevoli sono 5 e, quindi,  $P(E_2) = 5/36$ .

### Osservazione 1.4.2. Strano ma vero

Il motivo per cui le cose cambiano è dovuto al fatto che mentre la probabilità di ottenere 7 con due lanci non dipende dal primo risultato, la probabilità di ottenere 6 richiede che il primo risultato non sia 6.

**Cose che devi rimanerti di questa lezione:** Assieme alla nozione di probabilità condizionata il teorema di Bayes è uno dei fulcri alla base dell'inferenza. Assicurati di aver capito che cosa siano la probabilità a priori, la verosimiglianza, la probabilità a posteriori e la probabilità totale.

## 1.5 Variabili casuali discrete

Con l'introduzione delle variabili casuali nel caso discreto e delle nozioni essenziali di valore atteso e varianza entriamo nel vivo della nostra breve incursione nella teoria della probabilità.

### Variabili casuali

Molto spesso le quantità di interesse in un esperimento non sono i risultati ma una qualche funzione del risultato nota come *variabile casuale*. **Una variabile casuale è una funzione a valori reali definita sullo spazio campionario:** nel caso discreto, per esempio, il numero di teste in  $n$  lanci o il numero di assi in tredici carte. Una variabile casuale  $X$  è completamente definita in termini della probabilità con la quale assume ognuno dei suoi possibili valori.

#### Esercizio 1.5.1. Numero di teste in tre lanci di una moneta

Lancia 3 volte una moneta. Il numero  $X$  di teste ottenute è una variabile casuale i cui possibili valori sono 0, 1, 2 e 3. Fissa lo spazio campionario e determina la variabile casuale come funzione dallo spazio campionario ai reali.

*Soluzione*

Lo spazio campionario è l'insieme delle otto possibili triple

$$S = \{TTT, TTC, TCT, TCC, CTT, CTC, CCT, CCC\}$$

Valutiamo ora la probabilità con la quale  $X$  assume i valori 0, 1, 2 e 3 nell'assunzione che la moneta sia onesta. Poiché tutte le 8 triple sono ugualmente probabili abbiamo

$$\begin{aligned} P(X = 0) &= P(\{CCC\}) = 1/8 & P(X = 1) &= P(\{TCC, CTC, CCT\}) = 3/8 \\ P(X = 2) &= P(\{CTT, TCT, TTC\}) = 3/8 & P(X = 3) &= P(\{TTT\}) = 1/8 \end{aligned}$$

#### Esercizio 1.5.2. Palline numerate

Estrai casualmente 3 palline senza reinserimento tra 20 palline numerate da 1 a 20. Il numero estratto più grande  $X$  è una variabile casuale i cui possibili valori sono 3, 4, ..., 20. I risultati possibili dell'esperimento, ovvero lo spazio campionario, sono le combinazioni di 3 numeri scelti tra 1, 2, ..., 20. Valuta la probabilità con la quale  $X$  assume il valore  $i$  (con  $3 \leq i \leq 20$ ) nell'assunzione di equiprobabilità.

*Soluzione*

Se  $i$  è il numero estratto più grande, le altre due palline sono numerate con una delle possibili coppie di numeri diversi compresi tra 1 e  $i - 1$ . Avremo pertanto

$$P(X = i) = p(i) = \binom{i-1}{2} / \binom{20}{3}, \text{ per } i = 3, 4, \dots, 20$$

che fatti i calcoli fornisce

$i$	$p(i)$	$i$	$p(i)$	$i$	$p(i)$	$i$	$p(i)$	$i$	$p(i)$	$i$	$p(i)$
3	1/1140	4	3/1140	5	6/1140	6	10/1140	7	15/1140	8	21/1140
9	28/1140	10	36/1140	11	45/1140	12	55/1140	13	66/1140	14	78/1140
15	91/1140	16	105/1140	17	120/1140	18	136/1140	19	153/1140	20	171/1140

### Funzione di probabilità di massa

Nel caso di una variabile casuale  $X$  a valori discreti  $x_i$  con  $i = 1, 2, \dots$ , la *funzione di probabilità di massa*  $p(\cdot)$  definita sulla retta reale, o *pmf* o anche solo *funzione di probabilità*, contiene tutta l'informazione necessaria per descrivere completamente  $X$ . Si ha che  $p(x_i) = P(X = x_i) \geq 0$  con  $\sum_i p(x_i) = 1$ . La *pmf* per un dado onesto è illustrata a sinistra nella Figura 1.3.

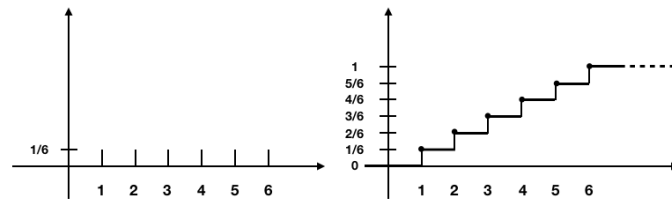


Figura 1.3: Funzione di probabilità di massa e funzione di probabilità cumulata per un dado onesto.

## Funzione di probabilità cumulata

Ordiniamo i valori  $x_i$  in modo tale che  $x_1 < x_2 < \dots < x_i < \dots$  e introduciamo la *funzione di probabilità cumulata*  $F(a)$ , o *cdf*, definita come

$$F(a) = \sum_{x_i \leq a} p(x_i)$$

È facile verificare che la funzione  $F$  è continua da destra e *crescente* da 0 a 1. Una *cdf* di una *pmf* è una funzione a gradini. Se i valori sono in numero finito, la *cdf* vale 0 a sinistra del valore più piccolo e 1 dal valore più grande in poi. L' $i$ -esimo gradino è localizzato nel punto  $x_i$  e il salto corrispondente vale  $p(x_i)$ . La somma di tutti i gradini, ovviamente, è sempre 1. La *cdf* per un dado onesto è illustrata a destra nella Figura 1.3.

## Valore atteso

Introduciamo una delle nozioni centrali dell'intera teoria: il valore atteso di una variabile casuale. Dovremo attendere la legge dei grandi numeri per apprezzarne appieno l'importanza.

Il *valore atteso*  $\mu$  di una variabile casuale  $X$  è indicato con  $\mathbb{E}[X]$  ed è la media pesata dei valori  $x_i$  che può assumere  $X$ . Ogni  $x_i$  è pesato con la sua probabilità  $p(x_i)$  e quindi si ha

$$\mu = \mathbb{E}[X] = \sum_i x_i p(x_i)$$

*Il valore atteso di una variabile casuale non è casuale!*

**Esercizio 1.5.3.** *Valore atteso del numero di teste in tre lanci di una moneta onesta*

Valuta il valore atteso della variabile casuale dell'**Esercizio 1.5.1**.

*Soluzione*

$$\mu = 0 \times 1/8 + 1 \times 3/8 + 2 \times 3/8 + 3 \times 1/8 = 1.5$$

## Valore atteso di una funzione di variabile casuale

Per calcolare il valore atteso di una funzione  $g$  di una variabile casuale discreta  $X$  possiamo determinare la *pmf* della variabile casuale discreta  $g(X)$ , oppure calcolare il valore atteso come media pesata.

**Esercizio 1.5.4.** *Due modi diversi di calcolare il valore atteso*

Sia  $Y = X^2$ . Calcola  $\mathbb{E}[X^2]$  per una variabile casuale  $X$  con

$$P(X = -1) = 0.2, \quad P(X = 0) = 0.5 \quad \text{e} \quad P(X = 1) = 0.3$$

*Soluzione*

Poiché  $P(Y = 1) = P(X = -1) + P(X = 1) = 0.5$  e  $P(Y = 0) = 0.5$  otteniamo

$$\mathbb{E}[X^2] = 1 \times 0.5 + 0 \times 0.5 = 0.5$$

Calcoliamo ora il valore atteso come  $\mathbb{E}[g(X)] = \sum_i g(x_i)p(x_i)$ . In questo caso scriviamo

$$\mathbb{E}[X^2] = 1 \times 0.2 + 1 \times 0.3 = 0.5$$

**Esercizio 1.5.5. Linearità**

Valuta  $\mathbb{E}[aX + b]$  con  $a$  e  $b \in \mathbb{R}$  in funzione di  $\mathbb{E}[X] = \mu = \sum_i x_i p(x_i)$ .

*Soluzione*

Coerentemente con la struttura lineare del valore atteso, abbiamo

$$\mathbb{E}[aX + b] = \sum_i (ax_i + b)p(x_i) = a \sum_i x_i p(x_i) + b = a\mathbb{E}[X] + b = a\mu + b$$

**Varianza**

Una seconda quantità che cattura proprietà importanti di una variabile casuale  $X$  è la *varianza*  $Var(X)$  definita come  $Var(X) = \mathbb{E}[(X - \mu)^2]$ .

**Osservazione 1.5.1. Quadrato è meglio**

Il quadrato nella definizione di varianza è fondamentale per ovviare al fatto che le differenze dal valore atteso hanno valore atteso nullo per qualunque  $X$ . Infatti

$$\mathbb{E}[(X - \mu)] = \mathbb{E}[X] - \mathbb{E}[\mu] = \mu - \mu = 0$$

**Esercizio 1.5.6. Una formula utile**

Dimostra che  $Var(X) = \mathbb{E}[X^2] - \mu^2$

*Dimostrazione*

$$Var(X) = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2 + \mu^2 - 2\mu X] = \mathbb{E}[X^2] + \mu^2 - 2\mu^2 = \mathbb{E}[X^2] - \mu^2$$

**Esercizio 1.5.7. Nonlinearità della varianza**

Dimostra che  $Var(aX + b) = a^2 Var(X)$ .

*Dimostrazione*

$$Var(aX + b) = \mathbb{E}[(aX + b - a\mu - b)^2] = \mathbb{E}[(aX - a\mu)^2] = a^2 Var(X)$$

**Osservazione 1.5.2. Deviazione standard**

Una quantità molto usata è la radice quadrata della varianza, nota come *deviazione standard*, o

$$SD(X) = \sqrt{Var(X)}$$

**Cose che devi rimanerti di questa lezione:** Definizione di variabile casuale discreta, *pmf* e *cdf* e loro proprietà. Saper ricavare la *cdf* di una *pmf* data e viceversa. Valore atteso e varianza di una variabile casuale discreta.

## 1.6 Distribuzioni discrete di probabilità

Prendiamo ora in considerazione importanti distribuzioni di probabilità nel caso discreto.

### Bernoulli

Una variabile casuale di *Bernoulli*  $X$  assume due soli valori, 0 e 1 (talvolta associati al fallimento e al successo di un esperimento), con funzione di probabilità di massa  $P(X = 0) = 1 - p$  e  $P(X = 1) = p$  con  $0 < p < 1$ .

**Esercizio 1.6.1.** Calcoliamo il valore atteso e la varianza di una variabile casuale di Bernoulli.

Da una diretta applicazione delle definizioni di valore atteso e varianza si ha,

$$\mu = \mathbb{E}[X] = p \times 1 + (1 - p) \times 0 = p \quad \text{e} \quad \text{Var}(X) = \mathbb{E}[X^2] - \mu^2 = p - p^2 = p(1 - p)$$

### Binomiale

La variabile casuale *binomiale*  $X$  conta i successi in una sequenza di  $n$  realizzazioni indipendenti di una variabile casuale di Bernoulli con  $p(1) = p$ . La sua funzione di probabilità di massa si scrive come

$$p(i) = \binom{n}{i} p^i (1 - p)^{n-i} \quad i = 0, 1, \dots, n$$

Il coefficiente binomiale  $\binom{n}{i}$  conta in quanti modi diversi si possono realizzare  $i$  successi in una sequenza di  $n$  realizzazioni indipendenti.

**Esercizio 1.6.2.** Somma sempre uguale a 1

Verifica che  $\sum_i p(i) = 1$ .

*Soluzione*

Riconoscendo nella scrittura della somma il binomio di Newton si ha

$$\sum_{i=0}^n \binom{n}{i} p^i (1 - p)^{n-i} = (p + (1 - p))^n = 1$$

**Osservazione 1.6.1.** Valore atteso e varianza della binomiale

Per il valore atteso di una binomiale abbiamo

$$\mathbb{E}[X] = np$$

mentre per la varianza

$$\text{Var}(X) = np(1 - p)$$

La derivazione di questi due risultati richiede passaggi algebrici un po' noiosi. Tra qualche lezione vedremo che entrambi discendono prontamente da proprietà di base dei valori attesi. Il risultato sul valore atteso dalle proprietà di linearità del valore atteso applicato alla somma di  $n$  variabili casuali di Bernoulli, quello sulla varianza dal fatto che le  $n$  variabili sono indipendenti.

### Geometrica

La variabile casuale *geometrica*  $X$  vale  $n$  se si ottiene un successo dopo  $n - 1$  fallimenti in una sequenza di  $n$  realizzazioni indipendenti di una variabile casuale di Bernoulli. La *pmf*, vedi Figura 1.4, è

$$P(X = n) = (1 - p)^{n-1} p \quad \text{per } n = 1, 2, \dots$$

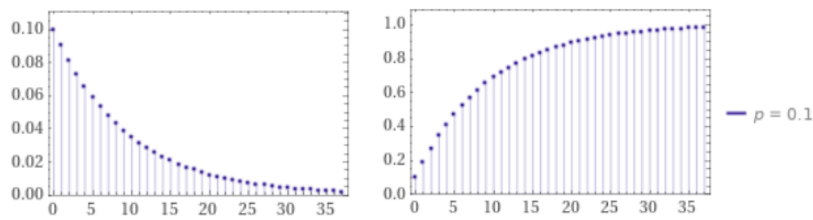


Figura 1.4: *Pmf e cdf per una distribuzione geometrica con  $p = 1/10$ .*

**Esercizio 1.6.3.** *Verifica della somma a 1*

Verifica che  $\sum_{i=1}^{\infty} (1-p)^{i-1} p = 1$ .

*Soluzione*

Per  $0 < p < 1$  abbiamo che

$$\sum_{i=0}^{\infty} p^i = \frac{1}{1-p} \quad \text{per } p \rightarrow 1-p \text{ fornisce } \sum_{i=0}^{\infty} (1-p)^i = \frac{1}{1-(1-p)} = \frac{1}{p}$$

Pertanto abbiamo

$$\sum_{i=1}^{\infty} (1-p)^{i-1} p = \sum_{i=0}^{\infty} (1-p)^i p = \frac{1}{p} \times p = 1$$

**Esercizio 1.6.4.** *Valore atteso di una geometrica*

Decomponiamo la somma in due addendi

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i(1-p)^{i-1} p = \sum_{i=1}^{\infty} (i-1+1)(1-p)^{i-1} p = \sum_{i=1}^{\infty} (i-1)(1-p)^{i-1} p + \sum_{i=1}^{\infty} (1-p)^{i-1} p$$

Ponendo  $j = i - 1$  per la seconda serie otteniamo  $\sum_{i=1}^{\infty} (1-p)^{i-1} p = \sum_{j=0}^{\infty} (1-p)^j p$ . Inoltre, sempre ponendo  $j = i - 1$  si ha

$$\sum_{i=1}^{\infty} (i-1)(1-p)^{i-1} p = \sum_{j=0}^{\infty} j(1-p)^j p = \sum_{j=1}^{\infty} j(1-p)^j p = (1-p) \sum_{j=1}^{\infty} j(1-p)^{j-1} p = (1-p) \mathbb{E}[X]$$

dove abbiamo escluso il termine nullo, raccolto  $1-p$ , e infine ricordato l'espressione del valore atteso. Combinando queste equazioni si ha che

$$\mathbb{E}[X] = (1-p)\mathbb{E}[X] + 1$$

da cui abbiamo che  $\mathbb{E}[X] = 1/p$ . Questo risultato ci riconcilia con l'intuizione che se la probabilità di ottenere *testa* con una moneta truccata in cui la probabilità  $p$  di testa è  $1/10$  ci aspettiamo di ottenere *testa* una volta ogni 10 lanci. Dal grafico della *cdf* di Figura 1.4 notiamo che la probabilità di ottenere almeno una volta testa con 10 lanci è intorno al %70.

**Osservazione 1.6.2.** *Ancora sulla varianza*

In modo simile si ottiene  $\text{Var}(X) = (1-p)/p^2$ .

**Poisson**

Una variabile casuale di *Poisson* è definita dalla *pmf*

$$P(X = i) = \frac{\mu^i}{i!} e^{-\mu} \quad \text{con } i = 0, 1, 2, \dots$$

con

$$\sum_{i=0}^{\infty} \frac{\mu^i}{i!} e^{-\mu} = e^{-\mu} \sum_{i=0}^{\infty} \frac{\mu^i}{i!} = e^{-\mu} e^{\mu} = 1 \quad \text{in quanto} \quad e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \forall x \in \mathbb{R}$$

Il numero di errori di stampa per pagina, il numero di ultracentenari di una comunità, il numero di numeri di telefono sbagliati da un centralino, il numero di pacchi di biscotti venduti in un giorno, il numero di clienti in un ufficio postale sono esempi di variabili casuali di Poisson. La Figura 1.5 mostra tre distribuzioni di Poisson al variare di  $\mu$ .

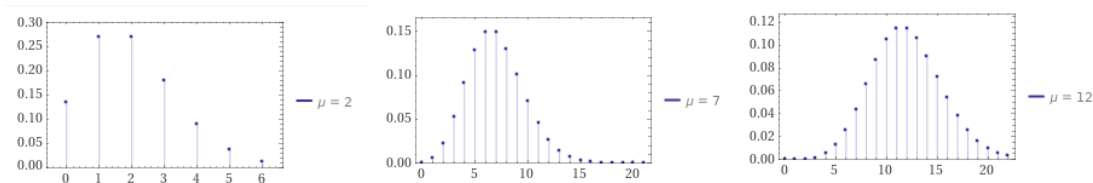


Figura 1.5: Vedi testo.

### Esercizio 1.6.5. Valore atteso di una Poissoniana

Calcola il valore atteso di una variabile casuale di Poisson.

*Soluzione*

Si ha che

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = \sum_{i=1}^{\infty} i e^{-\mu} \frac{\mu^i}{i!} = \mu \sum_{i=1}^{\infty} e^{-\mu} \frac{\mu^{i-1}}{(i-1)!} = \mu \sum_{j=0}^{\infty} e^{-\mu} \frac{\mu^j}{j!} = \mu$$

dove abbiamo escluso il termine nullo, semplificato  $i$ , raccolto  $\mu$  e, infine, rinominato gli indici.

### Osservazione 1.6.3. Varianza di una Poissoniana

La varianza di una variabile casuale di Poisson è ancora  $\mu$ .

### Osservazione 1.6.4. Poisson e binomiale

Per  $n$  grande e  $p$  piccolo la *pmf* della binomiale tende a una Poissoniana con  $\mathbb{E}[X] = \mu$ . Infatti possiamo scrivere

$$\begin{aligned} P(X = i) &= \binom{n}{i} p^i (1-p)^{n-i} \sim \frac{n!}{(n-i)! i!} \left(\frac{\mu}{n}\right)^i \left(1 - \frac{\mu}{n}\right)^{n-i} \\ &= \frac{n(n-1) \dots (n-i+1)}{n^i} \frac{\mu^i}{i!} \left(1 - \frac{\mu}{n}\right)^{-i} \left(1 - \frac{\mu}{n}\right)^n \\ &\sim 1 \times \frac{\mu^i}{i!} \times 1 \times e^{-\mu} = e^{-\mu} \frac{\mu^i}{i!} \end{aligned}$$

**Cose che devi rimanerti di questa lezione:** Valori attesi e varianza per le quattro distribuzioni viste a lezione (Bernoulli, Binomiale, Geometrica e Poisson) che dovresti anche saper riconoscere dal grafico delle *pmf* e *cdf* corrispondenti.



## 1.7 Variabili casuali continue

Estendiamo il nostro studio al caso di variabili casuali che assumono valori nel continuo.

### Funzione densità di probabilità

L'insieme dei valori che può assumere una variabile casuale spesso non è finito o numerabile (pensiamo al tempo di vita di un componente, all'ora d'arrivo di un treno o al tempo di percorrenza di un viaggio in auto). Una variabile casuale  $X$  è continua se esiste una funzione  $f : \mathbb{R} \rightarrow \mathbb{R}^+$  tale che

$$P(X \in B) = \int_B f(x) dx \quad (1.3)$$

su ogni sottoinsieme misurabile  $B \subset \mathbb{R}$  (la misurabilità è una condizione tecnica che, per i nostri scopi, non ha conseguenze rilevanti in quanto tutti i sottoinsiemi di nostro interesse sono misurabili). La funzione  $f$  è la *densità di probabilità*, o *pdf*.

#### Osservazione 1.7.1. Diversamente dal caso discreto

La probabilità che una variabile casuale continua  $X$  assuma un determinato valore  $x$  è sempre 0. Se  $B = [x - \epsilon/2, x + \epsilon/2]$  la probabilità 1.3 diventa

$$P(x - \epsilon/2 \leq X \leq x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} f(t) dt \approx \epsilon \times f(x)$$

Fissato un intervallo di ampiezza  $\epsilon$ , pertanto, la *pdf* esprime la probabilità che il valore assunto da  $X$  sia vicino a  $x$ , ovvero compreso nell'intervallo  $B$ .

#### Esercizio 1.7.1. Una questione di normalizzazione

Se  $f(x) = C(4x - 2x^2)$  per  $0 < x < 2$  e 0 altrimenti, calcola il valore di  $C$  per cui la funzione  $f$  è una densità di probabilità e valuta  $P(X > 1)$ .

*Soluzione*

La costante  $C$  si ottiene imponendo la condizione di normalizzazione  $P(X \in S) = 1$  che, in questo caso, diventa

$$\frac{1}{C} = \int_0^2 (4x - 2x^2) dx = \left( 2x^2 - \frac{2}{3}x^3 \right) \Big|_0^2 = \frac{8}{3}$$

da cui ricaviamo  $C = 3/8$ . Calcolando l'integrale definito tra 1 e 2 della *pdf* normalizzata otteniamo  $P(X > 1) = 1/2$ .  $\square$

Di seguito vediamo come le definizioni introdotte per le variabili casuali discrete si estendono al caso continuo, considerando le *pdf* al posto delle *pmf* e integrali invece che sommatorie.

### Funzione di distribuzione cumulata

La *funzione di distribuzione cumulata*  $F : \mathbb{R} \rightarrow [0, 1]$ , o *cdf*, è definita  $\forall a \in \mathbb{R}$  come

$$F(x) = \int_{-\infty}^x f(t) dt$$

Come nel caso discreto la *cdf* è una funzione crescente non negativa compresa tra 0 e 1. La *pdf* e la *cdf* forniscono due caratterizzazioni equivalenti delle variabili casuali continue.

**Osservazione 1.7.2.** *Per il teorema fondamentale del calcolo integrale*

Applicando la prima parte di questo teorema troviamo che se  $f : (a, b) \rightarrow \mathbb{R}$  e

$$F(x) = \int_{-\infty}^x f(t)dt$$

allora

$$\frac{dF}{dx}(x) = f(x)$$

**Osservazione 1.7.3.** *Che succede con la discontinuità*

Se la densità presenta punti di discontinuità, la *cdf* è sempre continua ma non è più ovunque derivabile. La funzione densità è ottenibile come la derivata della *cdf* escludendo i punti angolosi (vedi Figura 1.6).

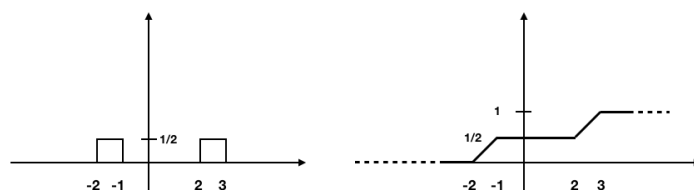


Figura 1.6: Funzione densità con punti di discontinuità e corrispondente *cdf*. La *cdf*, l'integrale definito della funzione densità da  $-\infty$  a  $x$ , è sempre una funzione continua. I punti di discontinuità della funzione densità creano spigoli (punti di non derivabilità) della *cdf*.

**Valore atteso e varianza**

In piena analogia con il caso discreto definiamo il valore atteso  $\mu$  e la varianza  $Var(X)$  di una variabile casuale  $X$  continua come

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx \quad \text{e} \quad Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

**Esercizio 1.7.2.** *Un semplice calcolo*

Mostra che  $\mathbb{E}[X] = 2/3$  per la variabile casuale  $X$  con densità di probabilità

$$f(x) = \begin{cases} 2x & \text{per } 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

*Soluzione*

Per il valore atteso abbiamo

$$\mathbb{E}[X] = \int_0^1 2x^2 dx = \frac{2}{3}$$

**Esercizio 1.7.3.** *Uno un po' più difficile*

Calcola  $\mathbb{E}[e^X]$  per la variabile casuale  $X$  con densità di probabilità

$$f(x) = \begin{cases} 1 & \text{per } 0 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

*Soluzione*

In analogia col caso discreto scriviamo  $\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$ . Quindi,

$$\mathbb{E}[e^X] = \int_{-\infty}^{+\infty} e^x f(x) dx = \int_0^1 e^x dx = e - 1$$

## Distribuzione di una funzione di variabile casuale

Data una variabile casuale  $X$  di distribuzione nota vogliamo trovare la distribuzione di  $g(X)$  per una funzione  $g$  data. Nel caso di  $g$  monotona è sufficiente un po' di attenzione.

### Esempio 1.7.1. *Prima le cose facili*

Sia  $X$  distribuita uniformemente tra 0 e 1. Abbiamo quindi  $f(x) = 1$  e  $F(x) = x$  tra 0 e 1. Se  $Y = X^n$  allora

$$F_Y(y) = P(Y \leq y) = P(X^n \leq y) = P(X \leq y^{1/n}) = F(y^{1/n}) = y^{1/n}$$

Pertanto, per  $0 \leq y \leq 1$ , derivando  $F_Y(y)$  otteniamo

$$f_Y(y) = \frac{y^{(1-n)/n}}{n}$$

### Esempio 1.7.2. *Poi quelle un po' più difficili*

Data  $f(x)$  per  $X$ , troviamo  $f_Y$  per  $Y = X^2$ . Per  $y \geq 0$  abbiamo

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y})$$

la cui derivata per  $y \geq 0$  fornisce

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

### Esercizio 1.7.4. *Una seconda soluzione per l'Esercizio 1.7.3*

Per  $0 \leq x \leq 1$  abbiamo che  $1 \leq e^x \leq e$ . Ora, per  $1 \leq a \leq e$ ,

$$F_Y(a) = P\{Y \leq a\} = P\{e^X \leq a\} = P\{X \leq \ln a\} = \int_0^{\ln a} f(x) dx = \int_0^{\ln a} 1 dx = \ln a$$

Segue che

$$f_Y(a) = \frac{dF_Y(a)}{da} = \frac{1}{a} \quad \text{per } 1 \leq a \leq e$$

e 0 altrimenti. Pertanto,

$$\mathbb{E}[e^X] = \mathbb{E}[Y] = \int_1^e x \left(\frac{1}{x}\right) dx = e - 1$$

**Cose che devi rimanerti di questa lezione:** Definizione di variabile casuale continua, *pdf* e *cdf* e loro proprietà. Valore atteso e varianza di una variabile casuale continua.

## 1.8 Distribuzioni continue di probabilità

Anche nel caso continuo vale la pena soffermarsi su alcune funzioni di distribuzioni importanti.

### Digressione sull'integrazione per parti

Siano  $f$  e  $g$  due funzioni continue e derivabili. La derivata del prodotto delle due funzioni é

$$\frac{d}{dx}(f(x)g(x)) = \frac{df(x)}{dx}g(x) + f(x)\frac{dg(x)}{dx} = f'(x)g(x) + f(x)g'(x)$$

Considerando l'integrale di entrambi i membri, applicando il teorema fondamentale del calcolo integrale e riordinando i termini, otteniamo

$$\int f'(x)g(x)dx = f(x)g(x) - \int f(x)g'(x)dx + C$$

dove  $C$  é una costante arbitraria. La forza di questo metodo risiede nella capacità di individuare, quale tra le funzioni  $f$  e  $g$  sia più facilmente derivabile o integrabile in modo da poter semplificare l'integrale. Nel caso di integrale definito su un intervallo  $[a, b]$  la costante  $C$  scompare e si ottiene

$$\int_a^b f'(x)g(x)dx = f(x)g(x)\Big|_a^b - \int_a^b f(x)g'(x)dx$$

### Distribuzione normale (o Gaussiana)

Una variabile casuale normale  $X = \mathcal{N}(\mu, \sigma^2)$ , con  $\mu$  e  $\sigma^2 > 0$  entrambi parametri reali fissati, ha come pdf

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

#### Osservazione 1.8.1. Normalizzazione garantita

Omettiamo la verifica che l'integrale di  $f$  vale 1 perché troppo laboriosa. Ci limitiamo a osservare che il valore di  $f(0)$  è inversamente proporzionale alla costante  $\sigma$ . Tre distribuzioni normali per diversi valori di  $\mu$  e  $\sigma$  sono mostrate in Figura 1.7.

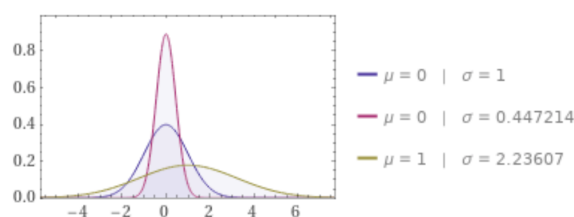


Figura 1.7: Vedi testo.

#### Esercizio 1.8.1. Valore atteso e varianza per la normale standard $Z = \mathcal{N}(0, 1)$

Determina  $\mathbb{E}[Z]$  e  $\text{Var}(Z)$ .

*Soluzione*

Abbiamo

$$\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} xe^{-x^2/2}dx = 0$$

perché la funzione integranda è dispari. Per la varianza otteniamo

$$Var(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} dx = 1$$

Infatti, ponendo

$$f(x) = \frac{x}{\sqrt{2\pi}} \implies f'(x) = \frac{1}{\sqrt{2\pi}}$$

e

$$g(x) = -e^{-\frac{x^2}{2}} \implies g'(x) = xe^{-\frac{x^2}{2}}$$

la formula di integrazione per parti fornisce

$$\int_{-\infty}^{+\infty} \frac{x^2 e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx = -\frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = 0 + 1 = 1$$

**Osservazione 1.8.2.** Una gaussiana è per sempre

Se  $X = \mathcal{N}(\mu, \sigma^2)$  e  $Y = aX + b$  con  $a$  e  $b$  reali qualunque, abbiamo che la variabile casuale  $Y$  è gaussiana con  $Y = \mathcal{N}(a\mu + b, a^2\sigma^2)$ . Infatti,

$$F_Y(x) = P(Y \leq x) = P(aX + b \leq x) = P\left(X \leq \frac{x-b}{a}\right) = F_X\left(\frac{x-b}{a}\right) \quad e$$

$$f_Y(x) = \frac{dF_Y(x)}{dx} = \frac{f_X\left(\frac{x-b}{a}\right)}{a} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-((x-b)/a-\mu)^2/2\sigma^2} = \frac{1}{a\sigma\sqrt{2\pi}} e^{-(x-a\mu-b)^2/2a^2\sigma^2}.$$

In particolare, ponendo  $a = 1/\sigma$  e  $b = -\mu/\sigma$  otteniamo

$$Y = (X - \mu)/\sigma = \mathcal{N}(0, 1) = Z$$

ovvero una normale standard! Invertendo la relazione, poichè  $X = \sigma Z + \mu$ , abbiamo che per una variabile normale  $X = \mathcal{N}(\mu, \sigma^2)$

$$\mathbb{E}[X] = \mu \quad e \quad Var(X) = \sigma^2$$

## Distribuzione esponenziale

La densità di probabilità di una variabile casuale esponenziale è  $f(x) = \lambda e^{-\lambda x}$  per  $x \geq 0$  e 0 altrimenti. Tre distribuzioni esponenziali per diversi valori di  $\lambda$  sono mostrate in Figura 1.8.

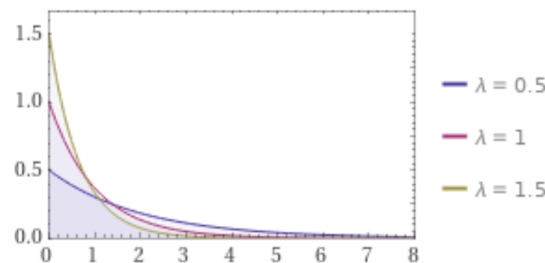


Figura 1.8: Vedi testo.

**Osservazione 1.8.3.** *La solita proprietà di normalizzazione*

Verifica che

$$\int_0^{+\infty} \lambda e^{-\lambda x} dx = 1$$

*Soluzione*

$$\int_0^{+\infty} e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x} \Big|_0^{+\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}$$

**Osservazione 1.8.4.** *Funzione di probabilità cumulata*

$$F(x) = P\{X \leq x\} = -\lambda \frac{1}{\lambda} e^{-\lambda x} \Big|_0^x = 1 - e^{-\lambda x} \quad \text{per } x \geq 0$$

**Osservazione 1.8.5.** *Una distribuzione smemorata*

Una distribuzione per la quale  $P(X > s+t | X > t) = P(X > s)$  per tutti gli  $s, t \geq 0$ , è *senza memoria*. L'esponenziale è senza memoria, infatti,

$$P(X > s+t | X > t) = \frac{P(X > s+t, X > t)}{P(X > t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s)$$

**Esercizio 1.8.2.**  $\mathbb{E}[X] = 1/\lambda$ 

Integrando per parti si ha

$$\mathbb{E}[X] = \int_0^{+\infty} x \lambda e^{-\lambda x} dx = -x e^{-\lambda x} \Big|_0^{+\infty} - \left(-\int_0^{+\infty} e^{-\lambda x} dx\right) = 0 + 1/\lambda$$

**Esercizio 1.8.3.**  $\text{Var}(X) = 1/\lambda^2$ 

Poiché  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$  è sufficiente calcolare  $\mathbb{E}[X^2]$ . Integrando per parti si ha

$$\mathbb{E}[X^2] = \int_0^{+\infty} x^2 \lambda e^{-\lambda x} dx = -x^2 e^{-\lambda x} \Big|_0^{+\infty} - \left(-\int_0^{+\infty} 2x e^{-\lambda x} dx\right) = 2 \int_0^{+\infty} x e^{-\lambda x} dx$$

Moltiplicando e dividendo per  $\lambda$  si ottiene

$$\mathbb{E}[X^2] = \frac{2}{\lambda} \mathbb{E}[X] = \frac{2}{\lambda^2}$$

per cui

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

**Cose che devi rimanerti di questa lezione:** Valori attesi e varianza per le due distribuzioni viste a lezione (normale ed esponenziale) che dovresti anche saper riconoscere dal grafico delle *pdf* e *cdf* corrispondenti.

## 1.9 Distribuzioni congiunte e indipendenza

Estendiamo quanto visto precedentemente al caso di più variabili casuali.

### Caso discreto

La distribuzione di una coppia di variabili congiunte  $(X, Y)$  con valori in  $\{x_1, \dots, x_N\}$  e  $\{y_1, \dots, y_M\}$  è data da una *pmf*

$$P(X = x_i, Y = y_j) = p(x_i, y_j) \quad \text{per } i = 1, \dots, N, \text{ e } j = 1, \dots, M$$

**Distribuzioni marginali** A partire dalla *pmf congiunta* di due variabili casuali è immediato determinare le *pmf* per le singole variabili, note come *marginali*, sommando su tutti i possibili valori assunti dall'altra variabile casuale. Ovvero

$$p_X(x_i) = P(X = x_i) = \sum_{j=1}^M p(x_i, y_j) \quad \text{e} \quad p_Y(y_j) = P(Y = y_j) = \sum_{i=1}^N p(x_i, y_j)$$

#### Esercizio 1.9.1. Palline di tre colori

Estrai 3 palline a caso da un'urna che ne contiene 3 rosse, 4 bianche e 5 blu. Se  $X$  conta le rosse estratte e  $Y$  le bianche, calcola la distribuzione congiunta di probabilità di massa e le probabilità marginali.

*Soluzione*

Abbiamo  $\binom{12}{3} = 220$  casi possibili. Per i casi favorevoli otteniamo

$$(0, 0) \rightarrow \binom{5}{3} = 10 \quad (0, 1) \rightarrow \binom{4}{1} \binom{5}{2} = 40 \quad (0, 2) \rightarrow \binom{4}{2} \binom{5}{1} = 30 \quad (0, 3) \rightarrow \binom{4}{3} = 4$$

$$(1, 0) \rightarrow \binom{3}{1} \binom{5}{2} = 30 \quad (1, 1) \rightarrow \binom{3}{1} \binom{4}{1} \binom{5}{1} = 60 \quad (1, 2) \rightarrow \binom{3}{1} \binom{4}{2} = 18$$

$$(2, 0) \rightarrow \binom{3}{2} \binom{5}{1} = 15 \quad (2, 1) \rightarrow \binom{3}{2} \binom{4}{1} = 12$$

$$(3, 0) \rightarrow \binom{3}{3} = 1$$

Le probabilità congiunte e le marginali, che si chiamano così perché collocate nel margine destro per la  $X$  e nel margine inferiore per la  $Y$ , sono

$p(0, 0) = 10/220$	$p(0, 1) = 40/220$	$p(0, 2) = 30/220$	$p(0, 3) = 4/220$	$p_X(0) = 84/220$
$p(1, 0) = 30/220$	$p(1, 1) = 60/220$	$p(1, 2) = 18/220$		$p_X(1) = 108/220$
$p(2, 0) = 15/220$	$p(2, 1) = 12/220$			$p_X(2) = 27/220$
$p(3, 0) = 1/220$				$p_X(3) = 1/220$

$$p_Y(0) = 56/220 \quad p_Y(1) = 112/220 \quad p_Y(2) = 48/220 \quad p_Y(3) = 4/220$$

È appena il caso di rimarcare che  $p(i, j) \neq p_X(i)p_Y(j)$ .

**Esercizio 1.9.2. Da 0 a 3 figli**

In un paese il 15% delle famiglie non ha figli, il 20% uno, il 35% due e il 30% tre. Se la probabilità di essere maschio ( $M$ ) o femmina ( $F$ ) per un figlio è la stessa, determina la distribuzione di probabilità congiunta per  $P(M = i, F = j)$  e le marginali.  $i, j = 0, \dots, 3$ .

*Soluzione*

Abbiamo che  $P(0, 0) = P(0 \text{ figli}) = 0.15$ . Dal fatto che  $P(0, 1) = P(1 \text{ figlio})P(F|1 \text{ figlio})$  segue che  $P(0, 1) = 0.2 \times 0.5 = 0.1$  e lo stesso per  $P(1, 0)$ . Per  $P(2, 0)$ , e similmente per  $P(0, 2)$ , abbiamo  $P(2 \text{ figli})P(2 F|2 \text{ figli}) = 0.35 \times 0.25 = 0.0875$ . Poiché  $P(2 \text{ figli}) = 0.35$  abbiamo che  $P(1, 1) = 0.175$  (la figlia femmina può essere la prima o la seconda). Ragionando in modo analogo otteniamo le restanti probabilità riassunte nella tabella

---

	$M = 0$	$M = 1$	$M = 2$	$M = 3$	
$F = 0$	0.15	0.10	0.0875	0.0375	0.3750
$F = 1$	0.10	0.175	0.1125	0	0.3875
$F = 2$	0.0875	0.1125	0	0	0.20
$F = 3$	0.0375	0	0	0	0.0375
	0.3750	0.3875	0.20	0.0375	

---

**Funzione di distribuzione cumulata congiunta** È data da

$$F(a, b) = \sum_{i: x_i \leq a} \sum_{j: y_j \leq b} p(x_i, y_j)$$

Come nel caso di singola variabile casuale abbiamo che  $0 \leq F(a, b) \leq 1$ .

**Cumulate marginali** A partire dalla *cdf congiunta*  $F(a, b)$  di due variabili casuali è immediato definire le *cdf* per le singole variabili,  $F_X(a)$  e  $F_Y(b)$ , note come *cdf marginali*

$$F_X(a) = P(X \leq a) = P(X \leq a, Y < +\infty) = F(a, +\infty)$$

$$F_Y(b) = P(Y \leq b) = P(X < +\infty, Y \leq b) = F(+\infty, b)$$

**Esercizio 1.9.3. Spesso assieme**

Dimostra che  $P(X > a, Y > b) = 1 + F(a, b) - F_X(a) - F_Y(b)$ .

$$\begin{aligned} P(X > a, Y > b) &= 1 - P((X > a, Y > b)^c) = 1 - P((X > a)^c \cup (Y > b)^c) \\ &= 1 - P((X \leq a) \cup (Y \leq b)) \\ &= 1 - (P(X \leq a) + P(Y \leq b) - P(X \leq a, Y \leq b)) \\ &= 1 + F(a, b) - F_X(a) - F_Y(b). \end{aligned}$$

**Caso continuo**

Il caso continuo è in completa analogia col discreto, ma richiede l'uso di integrali doppi. Per valutare la probabilità che la coppia  $X, Y$  assuma valori nel dominio  $C$  dobbiamo saper calcolare

$$P((X, Y) \in C) = \int \int_{(x, y) \in C} f(x, y) dx dy = P(X \in A, Y \in B) = \int_B \left( \int_A f(x, y) dx \right) dy$$



con

$$F(a, b) = \int_{-\infty}^b \left( \int_{-\infty}^a f(x, y) dx \right) dy \quad \text{e} \quad f(a, b) = \frac{\partial^2 F(a, b)}{\partial a \partial b}$$

Inoltre, per le probabilità marginali, se

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy \quad \text{e} \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

abbiamo

$$P\{X \in A\} = P\{X \in A, Y < +\infty\} = \int_A \left( \int_{-\infty}^{+\infty} f(x, y) dy \right) dx = \int_A f_X(x) dx$$

e

$$P\{X \in B\} = P\{X < +\infty, Y \in B\} = \int_B \left( \int_{-\infty}^{+\infty} f(x, y) dx \right) dy = \int_B f_Y(y) dy$$

### Variabili casuali indipendenti

Due variabili casuali sono indipendenti se per ogni coppia di insiemi  $A$  e  $B$

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

Una definizione equivalente richiede che per ogni coppia  $a$  e  $b$  di numeri reali

$$P(X \leq a, Y \leq b) = P(X \leq a)P(Y \leq b)$$

Se  $X$  e  $Y$  sono indipendenti, allora

$$F(a, b) = F_X(a)F_Y(b)$$

per le funzioni di distribuzione cumulate e

$$f(x, y) = f_X(x)f_Y(y) \quad \text{e} \quad p(x, y) = p_X(x)p_Y(y)$$

rispettivamente per le *pdf* nel caso continuo e le *pmf* nel caso discreto.

**Cose che devi rimanerti di questa lezione:** Definizione di probabilità congiunta e marginale. Collegamento con la nozione di probabilità condizionata (almeno nel caso discreto).

## 1.10 Proprietà dei valori attesi

La capitale importanza della nozione di valore atteso merita un'attenzione speciale. Partiamo dal ruolo giocato dal valore atteso nel caso di somme di variabili casuali.

### Funzione di variabili casuali

Il valore atteso di  $g(X, Y)$ , nel caso discreto, può essere calcolato come

$$\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y)p(x, y)$$

Se  $g(X, Y) = X + Y$  abbiamo

$$\begin{aligned}\mathbb{E}[X + Y] &= \sum_x \sum_y (x + y)p(x, y) = \sum_x \sum_y xp(x, y) + \sum_y \sum_x yp(x, y) \\ &= \sum_x \sum_y xp(y|x)p_X(x) + \sum_y \sum_x yp(x|y)p_Y(y) \\ &= \sum_x xp_X(x) + \sum_y yp_Y(y) = \mathbb{E}[X] + \mathbb{E}[Y]\end{aligned}$$

**Osservazione 1.10.1.** *Valore atteso di una variabile casuale binomiale*

Poiché  $X = \sum_i X_i$  è la somma di variabili casuali di Bernoulli con  $\mathbb{E}[X_i] = p$  per tutti gli  $i$ , abbiamo

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^n X_i\right] = np$$

### Media e varianza campionaria

Se le  $X_i$  per  $i = 1, \dots, n$  sono variabili casuali identicamente e indipendentemente distribuite con valore atteso  $\mu$  e varianza  $\sigma^2$  definiamo la *media campionaria*  $\langle X_n \rangle$  e la *varianza campionaria*  $S^2$  come

$$\langle X_n \rangle = \frac{\sum_i X_i}{n} \quad \text{e} \quad S_n^2 = \frac{\sum_i (X_i - \langle X_n \rangle)^2}{n - 1}$$

Calcola  $\mathbb{E}[\langle X_n \rangle]$ ,  $Var(\langle X_n \rangle)$  e  $\mathbb{E}[S_n^2]$ .

*Soluzione*

$$\begin{aligned}\mathbb{E}[\langle X_n \rangle] &= \mathbb{E}\left[\frac{\sum_i X_i}{n}\right] = \frac{\mathbb{E}[\sum_i X_i]}{n} = \frac{\sum_i \mathbb{E}[X_i]}{n} = \mu \\ Var(\langle X_n \rangle) &= \frac{1}{n^2} Var\left(\sum_i X_i\right) = \frac{1}{n^2} \sum_i Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \\ \mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{\sum_i (X_i - \langle X_n \rangle)^2}{n - 1}\right] = \frac{\mathbb{E}[\sum_i (X_i - \mu + \mu - \langle X_n \rangle)^2]}{n - 1} \\ &= \frac{\mathbb{E}[\sum_i (X_i - \mu)^2] + \mathbb{E}[\sum_i (\mu - \langle X_n \rangle)^2] - 2\mathbb{E}[\sum_i (\mu - X_i) \sum_i (\mu - \langle X_n \rangle)]}{n - 1} \\ &= \frac{n\sigma^2 + nVar(\langle X_n \rangle) - 2nVar(\langle X_n \rangle)}{n - 1} = \frac{n\sigma^2 - nVar(\langle X_n \rangle)}{n - 1} = \frac{(n - 1)\sigma^2}{n - 1} = \sigma^2\end{aligned}$$

## Covarianza e varianza di somme

Se  $X$  e  $Y$  sono indipendenti,  $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ . Infatti

$$\mathbb{E}[g(X)h(Y)] = \sum_x \sum_y g(x)h(y)p(x, y) = \sum_x g(x)p(x) \sum_y h(y)p(y) = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$

La covarianza di due variabili casuali  $X$  e  $Y$  è definita come  $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

### Esercizio 1.10.1. Come per la varianza

Dimostra che  $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

*Soluzione*

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[(XY - \mathbb{E}[X]Y - \mathbb{E}[Y]X + \mathbb{E}[X]\mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

### Osservazione 1.10.2. Indipendenza e covarianza

Se  $X$  e  $Y$  sono indipendenti, allora  $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$  e, quindi,  $Cov(X, Y) = 0$ .

### Osservazione 1.10.3. Varianza di una variabile casuale binomiale

Poiché  $X = \sum_i X_i$  è la somma di variabili casuali di Bernoulli indipendenti e identicamente distribuite con  $Var(X_i) = p(1 - p)$  per tutti gli  $i$ , abbiamo che

$$Var(X) = Var\left(\sum_{i=1}^n X_i\right) = np(1 - p)$$

### Esercizio 1.10.2. Varianza di una somma

Esprimi la varianza di una somma di due variabili casuali in termini delle varianze delle singole variabili e della loro covarianza.

$$\begin{aligned} Var(X + Y) &= \mathbb{E}[(X + Y - (\mathbb{E}[X + Y]))^2] = \mathbb{E}[(X - \mathbb{E}[X] + Y - \mathbb{E}[Y])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= Var(X) + Var(Y) + 2Cov(X, Y) \end{aligned}$$

### Esercizio 1.10.3. Moltiplicazione per una costante

Dimostra che se  $a$  è una costante, allora  $Cov(aX, Y) = aCov(X, Y)$ .

$$Cov(aX, Y) = \mathbb{E}[(aX - \mathbb{E}[aX])(Y - \mathbb{E}[Y])] = a\mathbb{E}[XY] - a\mathbb{E}[X]\mathbb{E}[Y] = aCov(X, Y)$$

## Correlazione

La correlazione  $\rho(X, Y)$  di due variabili casuali  $X$  e  $Y$  è definita come

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Abbiamo che  $0 \leq \rho(X, Y) \leq 1$ , poiché ponendo  $Var(X) = \sigma^2$  e  $Var(Y) = \tau^2$  otteniamo

$$0 \leq Var\left(\frac{X}{\sigma} + \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} + 2\frac{Cov(X, Y)}{\sigma\tau} = 2(1 + \rho(X, Y))$$

e

$$0 \leq Var\left(\frac{X}{\sigma} - \frac{Y}{\tau}\right) = \frac{Var(X)}{\sigma^2} + \frac{Var(Y)}{\tau^2} - 2\frac{Cov(X, Y)}{\sigma\tau} = 2(1 - \rho(X, Y))$$

**Cose che devi rimanerti di questa lezione:** Valore atteso e varianza di una somma di variabili casuali. Media e varianza campionaria.

## 1.11 Risultati asintotici

Enunciamo ora due risultati fondamentali della teoria della probabilità: la legge dei grandi numeri e il teorema centrale del limite che legano il valore atteso alla media empirica.

### Disuguaglianze fondamentali

**Disuguaglianza di Markov** Sia  $X$  una variabile casuale che assume valori non negativi ed  $f(x)$  la sua densità di probabilità. Allora per ogni  $a > 0$  vale la disuguaglianza

$$P\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a} \quad (1.4)$$

Infatti, dato che  $xf(x) \geq 0$  abbiamo

$$\mathbb{E}[X] = \int_0^{+\infty} xf(x)dx \geq \int_a^{+\infty} xf(x)dx$$

Moltiplicando e dividendo per  $a$  e notando che  $x > a$  e quindi  $x/a > 1$  si ottiene

$$\mathbb{E}[X] \geq a \int_a^{+\infty} \frac{x}{a} f(x)dx \geq a \int_a^{+\infty} f(x)dx$$

La disuguaglianza segue allora dalla definizione di funzione di probabilità cumulata

$$P\{X \geq a\} = \int_a^{+\infty} f(x)dx$$

**Disuguaglianza di Chebyshev** Sia  $X$  una variabile casuale con valore atteso  $\mu$  e varianza  $\sigma^2$  finiti. Allora per tutti gli  $\epsilon > 0$  vale la disuguaglianza

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \quad (1.5)$$

La disuguaglianza di Chebyshev si ottiene dalla disuguaglianza di Markov applicata alla variabile casuale non negativa  $(X - \mu)^2$  con  $a = \epsilon^2$  e notando che

$$\forall \epsilon > 0, P\{|X - \mu| \geq \epsilon\} = P\{|X - \mu|^2 \geq \epsilon^2\}$$

**Osservazione 1.11.1.** *Nessuna ipotesi*

La portata fondamentale delle disuguaglianze di Markov e Chebyshev è dovuta al fatto che valgono per qualunque distribuzione di probabilità.

**Esercizio 1.11.1.** *Il vantaggio di saperne di più*

Sia  $X$  uniformemente distribuita tra 0 e 10. Quanto vale la probabilità che  $X$  si scosti di 4 dal suo valore medio? Confronta il risultato ottenuto utilizzando la disuguaglianza di Chebyshev con quello ottenuto sfruttando il fatto che la distribuzione è uniforme.

*Soluzione*

Abbiamo che  $\mathbb{E}[X] = 5$  e  $\sigma^2 = 25/3$ . Se applichiamo la disuguaglianza di Chebyshev con  $\epsilon = 4$  otteniamo  $P\{|X - 5| \geq 4\} \leq 25/48 \approx 0.52$ . Utilizzando l'informazione sulla forma della distribuzione di  $X$  otteniamo

$$P\{|X - 5| \geq 4\} = \frac{2}{10} \int_0^1 dx = \frac{1}{5} = 0.2$$

## Legge dei Grandi Numeri

Questo risultato chiarisce in che senso la frequenza o la media empirica converge a un valore atteso ed è alla base di tutti i metodi che stimano una quantità ignota a partire da un numero finito di osservazioni.

### **Teorema 1.11.1.** *Legge (debole) dei grandi numeri*

Siano  $X_i$  con  $i = 1, 2, \dots, n$  variabili casuali indipendenti e identicamente distribuite con  $\mathbb{E}[X_i] = \mu$ . Allora

$$\forall \epsilon > 0 \lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_i X_i - \mu \right| \geq \epsilon \right\} = 0$$

Con l'ipotesi aggiuntiva (e non necessaria) che la varianza  $\sigma^2$  sia finita, poichè

$$\mathbb{E} \left[ \frac{1}{n} \sum_i X_i \right] = \mu \quad e \quad Var \left( \frac{1}{n} \sum_i X_i \right) = \frac{\sigma^2}{n}$$

applicando la disuguaglianza di Chebyshev per  $k = \epsilon$  otteniamo infine

$$P \left\{ \left| \frac{1}{n} \sum_i X_i - \mu \right| \geq \epsilon \right\} \leq \frac{\sigma^2}{n\epsilon^2} \quad \blacksquare$$

Per il **Teorema 1.11.1**, quindi, al crescere di  $n$  la probabilità che la media empirica differisca dal valore atteso tende a 0. Differenze significative per  $n$  sufficientemente grande possono essere rilevate, ma non frequentemente.

### **Esercizio 1.11.2.** *Passeggiata dell'ubriaco*

Un ubriaco si muove con passi di lunghezza unitaria, indipendenti e in una direzione  $\Theta$  uniformemente distribuita tra 0 e  $2\pi$ . Dopo  $n$  passi a quale distanza  $D$  si troverà dal punto di partenza?

*Soluzione*

Se  $\theta_i$  è la direzione del passo  $i$ -esimo, dopo  $i$  passi l'ubriaco si troverà nella posizione  $(\sum_i X_i, \sum_i Y_i)$  con  $(X_i, Y_i) = (\cos \theta_i, \sin \theta_i)$ . Pertanto, dopo  $n$  passi avremo

$$\begin{aligned} D^2 &= \left( \sum_i X_i \right)^2 + \left( \sum_i Y_i \right)^2 \\ &= \sum_i (\cos^2 \theta_i + \sin^2 \theta_i) + \left( \sum_i \cos \theta_i \sum_{j \neq i} \cos \theta_j \right) + \left( \sum_i \sin \theta_i \sum_{j \neq i} \sin \theta_j \right) \\ &= n + \left( \sum_i \cos \theta_i \sum_{j \neq i} \cos \theta_j \right) + \left( \sum_i \sin \theta_i \sum_{j \neq i} \sin \theta_j \right) \end{aligned}$$

Poichè  $\mathbb{E}[\cos \Theta] = \mathbb{E}[\sin \Theta] = 0$  per  $n$  grande avremo  $\sum_i \cos \theta_i \approx \sum_i \sin \theta_i \approx 0$  e, pertanto,  $D \approx \sqrt{n}$ .

### **Compito 1.11.1.** *Verifica empirica della legge dei grandi numeri*

Simula un dado truccato con

$$p_1 = 0.4, \quad p_2 = p_3 = 0.2, \quad p_4 = 0.1 \quad e \quad p_5 = p_6 = 0.05$$

campionando  $u$  da una distribuzione uniforme in  $[0, 1]$ .

Calcola  $(\#_n i)/n$ , ovvero la frequenza con la quale ottieni la faccia  $i$  su  $n$  lanci. Poiché ogni faccia  $i$  è una variabile casuale di Bernoulli con  $\mu_i = p_i$  e  $\sigma_i^2 = p_i(1 - p_i)$ , ponendo  $\epsilon = 10^{-2}$  per il

**Teorema 1.11.1** si ha

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \right\} \leq \frac{p_i(1 - p_i)}{n10^{-4}}$$

Ripeti per  $m = 1000$  volte  $n = 10^5$  lanci e verifica che, per ogni faccia  $i$ ,  $f_i \leq p_i(1 - p_i)/10$  approssimando

$$P \left\{ \left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \right\} \quad \text{con} \quad f_i = \frac{1}{m} \#_m \left( \left| \frac{(\#_n i)}{n} - p_i \right| \geq 10^{-2} \text{ è vera} \right)$$

## Teorema Centrale del Limite

Presentiamo ora uno dei risultati più importanti della matematica.

**Teorema 1.11.2.** *Come si distribuiscono le stime di un valore atteso*

Siano le  $X_i$  con  $i = 1, 2, \dots, n$  variabili casuali indipendenti e identicamente distribuite con  $\mathbb{E}[X_i] = \mu$  e  $\text{Var}(X_i) = \sigma^2$ . Allora

$$\frac{\sum_i X_i - \mu}{\sigma/\sqrt{n}} \rightarrow \mathcal{N}(0, 1) \quad \text{per } n \rightarrow \infty \quad \blacksquare$$

Il **Teorema 1.11.2** garantisce che le stime di un valore atteso, al crescere di  $n$ , si distribuiscono come una normale standard centrata sul valore atteso **indipendentemente** dalla distribuzione sottostante.

**Osservazione 1.11.2.** *Aguzzate la vista*

La legge dei grandi numeri ci assicura che la stima di un valore atteso al crescere di  $n$ , con grande probabilità, è vicina a piacere al valore atteso. Tuttavia, non ci dice nulla su quanto debba essere grande  $n$  e non ci consente di stimare la velocità con la quale la stima si avvicina al valore atteso. Il **Teorema 1.11.2**, invece, è un risultato molto più forte perché garantisce che al crescere di  $n$  la distribuzione della stime approssima una distribuzione normale centrata sul valore atteso con varianza  $\sigma^2/n$ .

**Compito 1.11.2.** *Verifica empirica del CLT*

Simula una moneta onesta con  $p = 1/2$ , con 1 il valore attribuito alla realizzazione dell'evento *testa* e 0 a *croce*. Lancia la moneta  $n$  volte e calcola la frequenza  $t_n$  con la quale ottieni *testa* come

$$t_n = \frac{\#_n \text{testa}}{n}$$

Sapendo che per una moneta onesta  $\mu = p = 0.5$  e  $\sigma^2 = p(1 - p) = 1/4$ , ripeti per 1000 volte  $n$  lanci e confronta il grafico della distribuzione normale standard  $\mathcal{N}(0, 1)$  con l'istogramma delle tre distribuzioni empiriche ottenute

$$2\sqrt{n}(t_n - \mu) \quad \text{per } n = 10^2, 10^4 \text{ e } 10^6$$

**Cose che devi rimanerti di questa lezione:** Legame tra frequenza empirica e probabilità dato dalla *Legge dei Grandi Numeri*. Distribuzione delle stime di un valore atteso dato dal *teorema centrale del limite*.

## 1.12 \* Problemi di occupazione

L'analisi del problema del bilanciamento di un carico, ovvero di come distribuire un carico su risorse multiple, è centrale nello studio dell'allocazione dinamica di risorse e nell'*hashing*. Adottiamo come modello il lancio di  $m$  palline indistinguibili in  $n$  contenitori nell'ipotesi di probabilità uniforme e lanci indipendenti e stimiamo valori attesi importanti.

### Formule utili

**Maggiorazione del coefficiente binomiale** Per ogni numero naturale  $n$  e  $k \leq n$  abbiamo che

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n \times (n-1) \dots (n-(k-1))}{k!} \leq \frac{n^k}{k!} \quad (1.6)$$

**Maggiorazione del reciproco del fattoriale** Dallo sviluppo in serie della funzione esponenziale otteniamo che, per qualunque  $k > 0$  intero,

$$\frac{k^k}{k!} < 1 + \frac{k^1}{1!} + \frac{k^2}{2!} + \dots + \frac{k^k}{k!} + \dots = \sum_{i=0}^{+\infty} \frac{k^i}{i!} = e^k$$

da cui segue che

$$\frac{1}{k!} < \left(\frac{e}{k}\right)^k \quad (1.7)$$

**Somma della serie geometrica** Sia  $S = \sum_{i=k}^n a^i$  per  $a > 0$ . Poiché

$$aS = \sum_{i=k}^n a^{i+1}$$

abbiamo che

$$aS - S = a^{n+1} - a^k$$

da cui segue

$$S = \frac{a^k - a^{n+1}}{1 - a} \quad (1.8)$$

**Disuguaglianza di Boole** Dalla definizione di probabilità segue che la probabilità dell'unione di  $n$  eventi arbitrari  $E_i$  con  $i = 1, \dots, n$ , non necessariamente indipendenti, non è più grande della somma delle loro probabilità, ovvero

$$\Pr(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n \Pr(E_i) \quad (1.9)$$

### Concetti e risultati

**Con quale probabilità due palline finiscono in uno stesso contenitore?** Se  $C_{ij}$  è la variabile casuale indicatrice di una *collisione* tra le palline  $i$  e  $j$ , ovvero dell'evento *la pallina  $i$  e la pallina  $j$  finiscono nello stesso contenitore*, e  $R_i^k$  la variabile casuale indicatrice dell'evento *il contenitore  $k$  riceve la pallina  $i$* , avremo

$$\Pr(C_{ij}) = \sum_{k=1}^n \Pr(R_i^k) \Pr(R_j^k | R_i^k) = \sum_{k=1}^n \frac{1}{n} \times \frac{1}{n} = \frac{1}{n}$$

**Quale è il numero atteso di collisioni?** Se indichiamo con  $C$  la variabile casuale che conta le collisioni,  $C = \sum_{i \neq j} C_{ij}$ , avremo

$$E[C] = E \left[ \sum_{i \neq j} C_{ij} \right] = \sum_{i \neq j} E[C_{ij}] = \sum_{i \neq j} \Pr(C_{ij}) = \binom{m}{2} \frac{1}{n}$$

**Problema 1.12.1. Il paradosso del compleanno**

La numerosità minima per avere un compleanno in comune con probabilità maggiore del 50% è 23. Il più piccolo  $m$  per il quale  $\frac{1}{365} \binom{m}{2} \geq 1$  è 28. Come spieghi questa differenza?

**Qual è il numero atteso di contenitori vuoti?** Sia  $V_j$  la variabile casuale indicatrice dell'evento *il contenitore  $j$  è vuoto* e  $V$  la variabile casuale che conta il numero di contenitori vuoti. Dal fatto che la probabilità che una pallina non cada in un particolare contenitore è  $1 - 1/n$  e nell'ipotesi  $m = n$  otteniamo

$$\Pr(V_j) = \prod_{i=1}^n \left(1 - \frac{1}{n}\right) = \left(1 - \frac{1}{n}\right)^n$$

da cui ricaviamo che il valore atteso di  $V$ , per la linearità del valore atteso e per  $n$  grande, è

$$E[V] = E \left[ \sum_{j=1}^n V_j \right] = \sum_{j=1}^n E[V_j] = \sum_{j=1}^n \Pr(V_j) = \sum_{j=1}^n \left(1 - \frac{1}{n}\right)^n \approx \sum_{j=1}^n \frac{1}{e} = \frac{n}{e}$$

**Con quale probabilità un contenitore dato riceve esattamente  $k$  palline?** Sia  $RE_j$  la variabile casuale indicatrice dell'evento *il contenitore  $j$  riceve esattamente  $k$  palline*. Utilizzando le disuguaglianze (1.6) e (1.7), e sempre nell'ipotesi  $m = n$ , otteniamo

$$\Pr(RE_j) = \binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \leq \binom{n}{k} \left(\frac{1}{n}\right)^k \leq \frac{n^k}{k!} \left(\frac{1}{n}\right)^k = \frac{1}{k!} < \left(\frac{e}{k}\right)^k$$

**Con quale probabilità un contenitore riceve almeno  $k$  palline?** Sia  $RA_j$  la variabile casuale indicatrice dell'evento *il contenitore  $j$  riceve almeno  $k$  palline*. Usando la disuguaglianza (1.9) e la formula (1.8), e sempre nell'ipotesi  $m = n$ , abbiamo

$$\Pr(RA_j) \leq \sum_{i=k}^n \left(\frac{e}{i}\right)^i \leq \sum_{i=k}^n \left(\frac{e}{k}\right)^i = \left(\frac{e}{k}\right)^k \frac{1 - (e/k)^{n+1-k}}{1 - e/k} < \left(\frac{e}{k}\right)^k \frac{1}{1 - e/k}$$

Poniamo

$$k^* = \frac{3 \ln n}{\ln(\ln n)}$$

Poiché  $e/k^* < 0.5$  abbiamo

$$\begin{aligned} \left(\frac{e}{k^*}\right)^{k^*} \frac{1}{1 - e/k^*} &= \left(\frac{e \ln(\ln n)}{3 \ln n}\right)^{\left(\frac{3 \ln n}{\ln(\ln n)}\right)} \\ &= \exp \left( \frac{3 \ln n}{\ln(\ln n)} (1 + \ln(\ln(\ln n)) - \ln(\ln(n)) - \ln 3) \right) \\ &< \exp \left( -3 \ln n + \frac{3 \ln n \times \ln(\ln(\ln n))}{\ln(\ln n)} \right) < \exp(-2 \ln n) = \frac{1}{n^2} \end{aligned}$$

dove l'ultima disuguaglianza vale per  $n$  sufficientemente grande.



Usando nuovamente la disuguaglianza di Boole e tenendo presente che il complementare dell'unione di eventi è l'intersezione dei complementari abbiamo

$$\begin{aligned}\Pr(\text{un contenitore qualunque riceve almeno } k^* \text{ palline}) &\leq n \times \frac{1}{n^2} = \frac{1}{n} \\ \Pr(\text{tutti i contenitori ricevono al massimo } k^* \text{ palline}) &\geq 1 - \frac{1}{n}\end{aligned}$$

**Osservazione 1.12.1.** *Caso  $m = n$*

In conclusione abbiamo che quando il valore atteso dell'occupazione è pari a 1 ( $m = n$ ), il numero massimo di palline ricevute da un contenitore è dell'ordine di  $\ln n / \ln(\ln n)$ .

## Grandi deviazioni

Completiamo l'analisi del problema dell'occupazione trattando il caso di  $m = n \ln n$ . Il risultato principale è che il numero massimo di occupazione è dello stesso ordine,  $O(\ln n)$ , del valore atteso. Abbiamo bisogno di una disuguaglianza fondamentale.

### Disuguaglianza di Chernoff

Siano  $X_1, \dots, X_n$  variabili casuali *indipendenti* di Bernoulli con  $\Pr(X_i = 1) = p$  per  $i = 1, \dots, n$ . Se per la variabile casuale somma  $X = \sum_i X_i$  abbiamo

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = np = \mu$$

allora per ogni  $\epsilon > 0$

$$\Pr(X > (1 + \epsilon)\mu) < \left( \frac{e^\epsilon}{(1 + \epsilon)^{1+\epsilon}} \right)^\mu \quad (1.10)$$

*Dimostrazione:* per ogni  $t > 0$  applicando l'esponenziale è sempre vero che

$$P(X > (1 + \epsilon)\mu) = P(e^{tX} > e^{t(1+\epsilon)\mu})$$

La disuguaglianza di Markov (1.4) per la variabile casuale positiva  $e^{tX}$ , tenuto conto che  $X = \sum_i X_i$ , fornisce

$$P(e^{tX} > e^{t(1+\epsilon)\mu}) < \frac{\mathbb{E}[e^{tX}]}{e^{t(1+\epsilon)\mu}} = \frac{\prod_{i=1}^n \mathbb{E}[e^{tX_i}]}{e^{t(1+\epsilon)\mu}} \quad (1.11)$$

Ora

$$\mathbb{E}[e^{tX_i}] = pe^t + (1 - p) = 1 - p(1 - e^t)$$

e siccome  $1 - x < e^x$  ponendo  $x = p(1 - e^t)$  possiamo scrivere

$$\mathbb{E}[e^{tX_i}] < e^{p(1-e^t)}$$

Pertanto, poiché  $np = \mu$ , abbiamo

$$\prod_{i=1}^n \mathbb{E}[e^{tX_i}] < \prod_{i=1}^n e^{p(1-e^t)} = e^{np(1-e^t)} = e^{\mu(1-e^t)}$$

che, sostituito nella disequazione (1.11), fornisce

$$P(e^{tX} > e^{t(1+\epsilon)\mu}) < \frac{e^{\mu(1-e^t)}}{e^{t(1+\epsilon)\mu}} = e^{\mu(1-e^t-t(1+\epsilon))}$$

Poiché questa disuguaglianza è valida per ogni  $t > 0$ , cerchiamo il valore di  $t$  per il quale  $e^{\mu(e^t - t - t\epsilon - 1)}$  è minimo. Annullando la derivata di  $(e^t - t - t\epsilon - 1)$  rispetto a  $t$  troviamo

$$\frac{d}{dt}(e^t - t - t\epsilon - 1) = e^t - 1 - \epsilon = 0 \implies \bar{t} = \ln(1 + \epsilon)$$

In conclusione per  $t = \bar{t}$ , e sbarazzandoci dell'esponenziale nell'argomento della probabilità, abbiamo

$$P(X > (1 + \epsilon)\mu) < e^{\mu(1 + \epsilon - \ln(1 + \epsilon) - \epsilon \ln(1 + \epsilon) - 1)} = e^{\mu(\epsilon - (1 + \epsilon) \ln(1 + \epsilon))} = \left( \frac{e^\epsilon}{(1 + \epsilon)^{(1 + \epsilon)}} \right)^\mu \quad \blacksquare$$

### Ancora sulla probabilità che un contenitore riceva almeno $k$ palline

Guardiamo a che cosa succede nel caso in cui  $m = n \ln n$ . Per la linearità del valore atteso in questo caso ricaviamo che il numero atteso di palline ricevute da ogni contenitore è

$$\mu = \Pr \left( \bigcup_{i=1}^{n \ln n} E_i \right) \leq \sum_{i=1}^{n \ln n} \frac{1}{n} = \ln n$$

Vediamo che cosa ci dicono le tre disuguaglianze se ipotizzassimo l'esistenza di un contenitore con  $10 \ln n$  palline.

**Markov** Dalla disuguaglianza (1.4) con  $\mu = \ln n$  e  $a = 10 \ln n$  ricaviamo

$$\Pr(X > 10 \ln n) = \frac{\ln n}{10 \ln n} = \frac{1}{10}$$

**Chebyshev** La varianza  $\sigma^2$  di una distribuzione binomiale con  $m$  lanci e probabilità  $p = 1/n$  è

$$\sigma^2 = m \frac{1}{n} \left( 1 - \frac{1}{n} \right) \leq \frac{m}{n}$$

Pertanto, la disuguaglianza (1.5) per  $\mu = m/n = \ln n$  e  $k = 9 \ln n$  diventa

$$\Pr\{X \geq \ln n + 9 \ln n\} \leq \frac{\ln n}{81 \ln^2 n} = \frac{1}{81 \ln n}$$

**Chernoff** Con  $\epsilon = 9$  la disuguaglianza (1.10) diventa

$$\Pr(X \geq (1 + 9) \ln n) \leq \left( \frac{e^9}{10^{10}} \right)^{\ln n} < \left( \frac{e^9}{e^{20}} \right)^{\ln n} = \frac{1}{n^{11}}$$

### Osservazione 1.12.2. Non c'è paragone

La disuguaglianza di Chernoff mostra che la probabilità di grandi variazioni dal valore atteso, almeno per il caso di somma di variabili casuali indipendenti, decresce molto più rapidamente di quanto si possa inferire dalle disuguaglianze di Markov e Chebyshev. La probabilità di grandi variazioni dal valore atteso, per  $n$  grande, decresce esponenzialmente.

## Capitolo 2

# Elementi di Teoria dell'Informazione

In questa parte del corso sviluppiamo alcuni dei concetti principali della teoria dell'informazione. Introduciamo **l'informazione e l'entropia di Shannon** e illustriamo **le proprietà e le relazioni fondamentali che intercorrono tra entropia congiunta ed entropia condizionata** di due variabili casuali. Dopo una lezione **dedicata ai minimalia della Teoria dei Codici**, discutiamo **la codifica di Huffman**, codifica per simbolo ottimale anche se spesso lontana dal minimo teorico della compressione. Un'intera lezione è dedicata alla **la codifica aritmetica**, codifica che lavorando sul flusso dei dati in input raggiunge la migliore compressione possibile sotto ipotesi molto blande. La codifica di Huffman e quella aritmetica sono entrambe utilizzabili in assenza di rumore. Chiudiamo questa parte con una lezione sulla **codifica convoluzionale**, codifica in grado di contrastare l'eventuale presenza di rumore nella trasmissione dell'informazione.

## 2.14 Informazione ed entropia di Shannon

Introduciamo la nozione di informazione di Shannon e familiarizziamo con le sue principali proprietà.

### Misura di informazione

Una misura della quantità di informazione che si acquisisce una volta che si sia realizzato un evento casuale di probabilità  $p > 0$  è l'*informazione di Shannon* definita come

$$\log_2 \frac{1}{p}$$

L'unità di misura dell'informazione di Shannon è il *bit*.

#### Esempio 2.14.1. Lancio di una moneta onesta

Il risultato del lancio di una moneta può essere testa,  $T$ , o croce,  $C$ ; nel caso in cui i due eventi sono equiprobabili, ovvero se  $p(T) = p(C) = 1/2$ , l'informazione di Shannon che acquisiamo osservando il risultato è sempre

$$\log_2 2 = 1\text{bit}$$

#### Esempio 2.14.2. Lancio di una moneta truccata

Consideriamo ora una moneta in cui la probabilità che esca testa sia molto più grande di croce, tipo  $p(T) = 7/8$ . In questo caso, l'informazione di Shannon acquisita alla realizzazione dell'evento  $T$  è

$$\log_2 \frac{8}{7} \approx 0.19\text{bit}$$

mentre nel caso dell'evento  $C$ , poiché  $p(C) = 1/8$ , è

$$\log_2 8 = 3\text{bit}$$

Supponiamo ora di lanciare due volte la moneta truccata. L'informazione di Shannon acquisita dalla realizzazione dell'evento  $TT$  è

$$\log_2 \left( \frac{8}{7} \times \frac{8}{7} \right) = \log_2 \frac{8}{7} + \log_2 \frac{8}{7} \approx 0.39\text{bit}$$

mentre nel caso dell'evento  $CC$  è

$$\log_2 (8 \times 8) = \log_2 8 + \log_2 8 = 6\text{bit}$$

Nei casi  $TC$  e  $CT$  si ottiene

$$\log_2 \left( 8 \times \frac{8}{7} \right) = \log_2 \frac{8}{7} + \log_2 8 \approx 3.19\text{bit}$$

La dipendenza logaritmica dalla probabilità garantisce che l'informazione acquisita dalla realizzazione di eventi indipendenti sia pari alla somma delle informazioni di Shannon associate a ogni evento.

### Due giochi

Vediamo ora come l'informazione di Shannon sia legata al numero di bit necessari a rappresentare uno qualunque tra  $N$  numeri attraverso l'analisi di due semplici giochi. In entrambi i casi, assumiamo tacitamente che tutte le scelte del nostro avversario siano ugualmente probabili.

**Esempio 2.14.3. Indovina il numero**

L'avversario sceglie un numero  $n$  compreso tra 0 e 1023; il nostro scopo è indovinare il numero scelto formulando il minimo numero di domande alle quali l'avversario può rispondere solo *si* o *no*. Procediamo per dimezzamenti progressivi dell'intervallo originale. Se pensiamo alla codifica binaria di un qualunque  $n$  compreso tra 0 e 1023 ognuno dei 10 bit necessari può essere uguale a 0 o a 1 con probabilità  $p = 1/2$ , per cui ogni dimezzamento corrisponde all'acquisizione di una quantità di informazione pari a

$$\log_2 \frac{1}{p} = \log_2 2 = 1bit$$

I 10bit di informazione acquisiti dopo 10 domande, alla fine del gioco, ricostruiscono i 10 bit necessari per la codifica di  $n$ . Con questa strategia a ogni passo si ottiene sempre 1bit di informazione. Inoltre, l'informazione complessiva acquisita alla fine del gioco è la stessa che avremmo acquisito se avessimo indovinato al primo tentativo.

**Esempio 2.14.4. Trova il sottomarino**

Scopo del gioco, versione semplificata della battaglia navale, è indovinare in quale delle  $N = 8 \times 8 = 64$  caselle l'avversario abbia posizionato un sottomarino. Nuovamente ipotizziamo che il sottomarino possa essere con la stessa probabilità in qualunque casella. Al primo tentativo abbiamo 1 probabilità su 64 di indovinare e 63 su 64 di fallire. Se falliamo, la quantità di informazione acquisita è

$$\log_2 64 - \log_2 63 \approx 0.02bit$$

Se indoviniamo, invece,

$$\log_2 64 = 6bit$$

In questo secondo caso il gioco finisce e abbiamo acquisito tutta l'informazione di Shannon che era effettivamente disponibile in un solo colpo. Nell'ipotesi di aver fallito, il gioco prosegue. Calcoliamo come aumenta l'informazione di Shannon a ogni passo. Al secondo tentativo abbiamo 1 probabilità su 63 di indovinare e 62 su 63 di fallire. Nel caso in cui falliamo acquisiamo una quantità di informazione pari a

$$\log_2 64 - \log_2 63 + (\log_2 63 - \log_2 62) = \log_2 64 - \log_2 62 \approx 0.05bit$$

La quantità di informazione acquisita è ancora piuttosto scarsa e non troppo diversa da prima (solo 2 tentativi invece che 1 per 64 caselle). Se indoviniamo, invece, si ha

$$S = \log_2 64 - \log_2 63 + \log_2 63 = 6bit$$

È un caso? No: in effetti, in qualunque momento finisca il gioco, l'informazione di Shannon acquisita è sempre uguale a 6bit. Anche per questo gioco, quindi, l'informazione di Shannon complessiva è uguale al numero di bit, 6, richiesti per identificare ciascuna delle 64 caselle.

Che cosa succede se fallissimo per 32 volte? L'informazione accumulata sarebbe

$$S = \log_2 64 - \log_2 32 = 6 - 5 = 1bit$$

Dopo 32 tentativi a vuoto, infatti, abbiamo confinato il sottomarino in metà delle caselle della tabella originale, guadagnando 1bit di informazione!

**Unicità**

Dimostriamo ora che la forma funzionale  $S = S(p)$  dell'informazione di Shannon è univocamente determinata, a meno di un fattore costante arbitrario, da quattro assunzioni.

**A1** Non si acquisisce informazione dalla realizzazione di un evento certo, ovvero  $S(1) = 0$

**A2** Minore la probabilità di un evento, maggiore la quantità di informazione ottenuta dalla sua realizzazione, ovvero se  $p < q$  allora  $S(p) > S(q)$

**A3** A piccoli cambiamenti di  $p$  corrispondono piccoli cambiamenti di  $S(p)$ , ovvero  $S(p)$  è una funzione continua di  $p$ .

**A4** Siano  $E$  e  $F$  due eventi indipendenti con probabilità rispettivamente  $p$  e  $q$ . La quantità di informazione acquisita dalla realizzazione dell'evento  $E$  non è modificata dal fatto di conoscere che l'evento  $F$  si è realizzato, ovvero  $S(pq) = S(p) + S(q)$

**Teorema 2.14.1.** *Unicità della forma funzionale dell'informazione di Shannon*

Se  $S(\cdot)$  soddisfa gli Assiomi **A1**, **A2**, **A3** e **A4** e per  $C$  un numero positivo arbitrario abbiamo

$$S(p) = C \log_2 \frac{1}{p}$$

*Dimostrazione*

$$\begin{array}{ll} \text{da A4} & S(p^2) = S(p) + S(p) = 2S(p) \\ \text{per induzione} & S(p^m) = mS(p) \\ \forall n \in \mathbb{N} & S(p) = S(p^{1/n} p^{1/n} \dots p^{1/n}) = nS(p^{1/n}) \\ \text{conseguentemente} & S(p^{m/n}) = mS(p^{1/n}) = \frac{m}{n} S(p) \end{array}$$

Per **A3** quest'ultima relazione vale non solo per ogni numero razionale positivo  $m/n$ , ma anche per ogni numero reale positivo  $w$ , ovvero

$$S(p^w) = wS(p)$$

Per ogni  $p > 0$ , ponendo  $w = \log_2(1/p)$  si ha  $p = 2^{-w}$  per cui

$$S(p) = S\left(\frac{1}{2^w}\right) = wS\left(\frac{1}{2}\right) = C \log_2 \frac{1}{p}$$

poiché, per **A1** e **A2**,  $C = S(1/2) > S(1) = 0$ . Se  $C = 1$  otteniamo proprio l'informazione di Shannon.

■

Introduciamo ora la nozione chiave dell'intera Teoria dell'Informazione.

## Entropia di Shannon

Sia  $\mathcal{X} = \{x_1, x_2, \dots, x_N\}$  l'insieme dei valori che possono essere assunti da una variabile casuale  $X$  con probabilità  $p_X(x_1), p_X(x_2), \dots, p_X(x_N)$  e  $\sum_i p_X(x_i) = 1$ . Chiaramente  $|\mathcal{X}| = N$ .

**Definizione 2.14.1.** *Entropia di una variabile casuale*

L'entropia di  $X$  è il valore atteso dell'informazione di Shannon

$$H(X) = \sum_{i=1}^N p_X(x_i) \log_2 \frac{1}{p_X(x_i)}$$

dove, se  $p_X(x_i) = 0$ , poniamo  $p_X(x_i) \log_2(1/p_X(x_i)) = 0$ .

Nel caso in cui  $X$  assuma due soli valori, rispettivamente con probabilità  $p$  e  $1-p$ , l'entropia diventa

$$H_2(X) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$

Uguagliando a 0 la derivata di  $H_2(X)$  rispetto a  $p$  otteniamo

$$\log_2 \frac{1}{p} - \log_2 \frac{1}{1-p} = 0$$

da cui segue che il massimo di  $H_2$  si ottiene per  $p = 1/2$  e vale 1.

**Esempio 2.14.5.** *Quante volte testa in due lanci*

Consideriamo il caso di una moneta equa lanciata due volte. Per la variabile casuale che conta il numero di volte in cui esce *testa* abbiamo  $p(0) = 1/4$ ,  $p(1) = 1/2$  e  $p(2) = 1/4$ . Per l'entropia  $H$  abbiamo

$$H = \frac{1}{4} \log_2 4 + \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$$

□

Dalla concavità della funzione logaritmica discende che quanto abbiamo dimostrato per  $H_2$  è un risultato di portata più generale: **l'entropia è massima nel caso di eventi ugualmente probabili**. Infatti, per ogni funzione convessa  $f$  vale la disuguaglianza di Jensen, ovvero

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Nel nostro caso, poichè il logaritmo è concavo dobbiamo invertire la disuguaglianza e abbiamo

$$H(X) = \mathbb{E} \left[ \log_2 \frac{1}{p_X(x_i)} \right] \leq \log_2 \mathbb{E} \left[ \frac{1}{p_X(x_i)} \right] = \log_2 \left( \sum_{i=1}^N \frac{p_X(x_i)}{p_X(x_i)} \right) = \log_2 \sum_{i=1}^N 1 = \log_2 N$$

**Esempio 2.14.6.** *Dado equo a otto facce*

Consideriamo il caso di un dado equo con otto facce. Ogni faccia  $i$  ha probabilità  $p_i = 1/8$  e per l'entropia  $H_e$  abbiamo

$$H_e = \sum_{i=1}^8 p_i \log_2 \frac{1}{p_i} = \sum_{i=1}^8 \frac{1}{8} \log_2 8 = \log_2 8 = 3$$

**Esempio 2.14.7.** *Dado iniquo (sempre a 8 facce)*

Consideriamo il caso di un dado iniquo a otto facce in cui le probabilità sono

$$p(1) = p(2) = p(3) = p(4) = \frac{1}{16}, \quad p(5) = p(6) = \frac{1}{8}, \quad \text{e} \quad p(7) = p(8) = \frac{1}{4}$$

Ci aspettiamo che in questo caso l'entropia del dado iniquo  $H_i$  sia minore di  $H_e$ , infatti

$$H_i = \frac{4}{16} \log_2 16 + \frac{2}{8} \log_2 8 + \frac{2}{4} \log_2 4 = \frac{4}{4} + \frac{3}{4} + \frac{2}{2} = \frac{11}{4} = 2.75 < 3 = H_e$$

**Assicurati di ...**

1. saper rispondere a semplici domande, come negli esempi dei due giochi, su informazione ed entropia di Shannon nel caso di eventi equiprobabili;
2. saper valutare l'entropia di Shannon per una variabile casuale a partire dalla sua distribuzione di probabilità

## 2.15 Entropia congiunta e condizionata

Il concetto di entropia assume maggiore importanza quando si estende al caso di più variabili casuali. Partiamo definendo l'entropia congiunta e l'entropia condizionata, concetti che, non sorprendentemente, si fondano sulla probabilità congiunta e la probabilità condizionata.

### Definizioni

Sia  $\mathcal{Y} = \{y_1, y_2, \dots, y_M\}$  l'insieme dei valori che possono essere assunti da una variabile casuale  $Y$  con probabilità marginale  $p_Y(y_1), p_Y(y_2), \dots, p_Y(y_M)$  e  $\sum_j p_Y(y_j) = 1$ . Sia  $p(x_i, y_j)$  la probabilità congiunta dell'evento  $X = x_i$  e  $Y = y_j \forall i, j$  con  $\sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) = 1$ . Indichiamo con  $p(x_i|y_j)$  la probabilità dell'evento  $X = x_i$  condizionata alla realizzazione dell'evento  $Y = y_j$ ; per ogni  $j$  fissato abbiamo  $\sum_{i=1}^N p(x_i|y_j) = 1$ .

Per l'entropia congiunta  $H(X, Y)$  abbiamo

$$H(X, Y) = \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)}$$

L'entropia  $H(X)$  è modificata dalla realizzazione dell'evento  $Y = y_j$  secondo la formula

$$H(X|Y = y_j) = \sum_{i=1}^N p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)}$$

Ne segue che l'entropia di  $X$  condizionata alla realizzazione di  $Y$  si scrive come

$$H(X|Y) = \sum_{j=1}^M p_Y(y_j) H(X|Y = y_j)$$

### Esempio 2.15.1. Ancora sul dado equo con otto facce

Sia  $X$  la variabile casuale che assume i valori  $i = 1, \dots, 8$  con  $p_X(i) = 1/8$  e  $Y$  la variabile casuale che assume i valori  $j = 0$  se l'esito del lancio è pari e  $j = 1$  se dispari con  $p_Y(j) = 1/2$ . Calcoliamo le entropie  $H(X)$  - quale faccia del dado,  $H(Y)$  - quale parità,  $H(X|Y)$  - quale faccia nota quale parità,  $H(Y|X)$  - quale parità nota quale faccia, e  $H(X, Y)$  - quale faccia e quale parità congiunte. Osserviamo che da

$$p(\text{faccia} = i \mid \text{parità} = j) = \begin{cases} 1/4 & \text{se } i \equiv j \pmod{2} \\ 0 & \text{altrimenti} \end{cases}$$

$$p(\text{parità} = i \mid \text{faccia} = j) = \begin{cases} 1 & \text{se } j \equiv i \pmod{2} \\ 0 & \text{altrimenti} \end{cases}$$

e

$$p(\text{faccia} = i, \text{parità} = j) = \begin{cases} 1/8 & \text{se } j \equiv i \pmod{2} \\ 0 & \text{altrimenti} \end{cases}$$



ricaviamo

$$\begin{aligned}
H(X) &= \sum_{i=1}^8 p_X(i) \log_2 \frac{1}{p_X(i)} = \sum_{i=1}^8 \frac{1}{8} \log_2 8 = \log_2 8 = 3 \\
H(Y) &= \sum_{j=0}^1 p_Y(j) \log_2 \frac{1}{p_Y(j)} = \frac{1}{2} \log_2 2 + \frac{1}{2} \log_2 2 = 1 \\
H(X|Y) &= \sum_{j=0}^1 p_Y(j) H(X|j) = \sum_{j=0}^1 p_Y(j) \sum_{i=1}^8 p(i|j) \log_2 \frac{1}{p(i|j)} \\
&= \frac{1}{2} \left( 4 \times \frac{1}{4} \log_2 4 \right) + \frac{1}{2} \left( 4 \times \frac{1}{4} \log_2 4 \right) = \log_2 4 = 2 \\
H(Y|X) &= \sum_{i=1}^8 p_X(i) H(Y|i) = \sum_{i=1}^8 p_X(i) \sum_{j=0}^1 p(j|i) \log_2 \frac{1}{p(j|i)} = 0 \\
H(X, Y) &= \sum_{i=1}^8 \sum_{j=0}^1 p(x_i, j) \log_2 \frac{1}{p(x_i, j)} = \log_2 8 = 3
\end{aligned}$$

Notiamo che il guadagno atteso dell'informazione di Shannon legato alla conoscenza di quale faccia sia uscita, nota la parità, si riduce perché le facce possibili passano da 8 a 4. Nel caso inverso, invece, il guadagno si azzerava perché nota la faccia una delle due parità è certa e l'altra impossibile.

**Esempio 2.15.2.** *Un dado equo a otto facce e una moneta equa*

Sia  $X$  ancora la variabile casuale che assume i valori da 1 a 8 con probabilità  $p_X(i) = 1/8$ , lancio di un dado equo. Ora  $Y$  è la variabile casuale che vale 0 se l'esito del lancio di una moneta equa è *testa* e 1 se *croce* (in entrambi i casi quindi con probabilità  $p_Y(j) = 1/2$ ). Le due variabili sono chiaramente indipendenti. Pertanto  $p(i|0) = p(i|1) = 1/8$  perché l'esito del lancio della moneta non modifica le probabilità del dado,  $p(0|i) = p(1|i)$  perché l'esito del lancio del dado non modifica le probabilità di testa e croce. Infine,  $p(i, j) = 1/16$  in virtù dell'indipendenza.

Calcoliamo le entropie  $H(X)$  - quale faccia del dado sia uscita,  $H(Y)$  - quale faccia della moneta,  $H(X|Y)$  - quale faccia del dado nota la faccia della moneta,  $H(Y|X)$  - quale faccia della moneta nota la faccia del dado, e  $H(X, Y)$  - quale faccia del dado e quale della moneta congiunte.

$$\begin{aligned}
H(X) &= \sum_{i=1}^8 p_X(i) \log_2 \frac{1}{p_X(i)} = 8 \times \left( \frac{1}{8} \log_2 8 \right) = \log_2 8 = 3 \\
H(Y) &= \sum_{j=0}^1 p_Y(j) \log_2 \frac{1}{p_Y(j)} = 2 \times \left( \frac{1}{2} \log_2 2 \right) = 1 \\
H(X|Y) &= \sum_{j=0}^1 p_Y(j) H(X|j) = \sum_{j=0}^1 \frac{1}{2} \sum_{i=1}^8 p(i|j) \log_2 \frac{1}{p(i|j)} = \log_2 8 = 3 \\
H(Y|X) &= \sum_{i=1}^8 p_X(i) H(Y|i) = \sum_{i=1}^8 \frac{1}{8} \sum_{j=0}^1 p(j|i) \log_2 \frac{1}{p(j|i)} = 8 \times \left( \frac{1}{8} \log_2 2 \right) = 1 \\
H(X, Y) &= \sum_{i=1}^8 \sum_{j=0}^1 p(x_i, j) \log_2 \frac{1}{p(x_i, j)} = \log_2 16 = 4
\end{aligned}$$

Per via dell'equiprobabilità, tutte le entropie non condizionate sono massime.

**Osservazione 2.15.1.** *Le apparenze non sempre ingannano*

In entrambi gli esempi risulta che

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

e che

$$H(X|Y) \leq H(X)$$

$$H(Y|X) \leq H(Y)$$

Dimostriamo ora che queste relazioni, in effetti, sono di portata generale.

### Uguaglianza fondamentale

**Proposizione 2.15.1.** *Entropia congiunta come somma di entropie*

Per ogni coppia di variabili casuali discrete  $X$  e  $Y$  si ha

$$H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$$

*Dimostrazione:* siccome  $p(x_i, y_j) = p_Y(y_j)p(x_i|y_j)$ , abbiamo

$$\begin{aligned} H(X, Y) &= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i, y_j)} \\ &= \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p_Y(y_j)} + \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} \\ &= \sum_{j=1}^M p_Y(y_j) \log_2 \frac{1}{p_Y(y_j)} \sum_{i=1}^N p(x_i|y_j) + \sum_{j=1}^M p_Y(y_j) \sum_{i=1}^N p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} \\ &= \sum_{j=1}^M p_Y(y_j) \log_2 \frac{1}{p_Y(y_j)} \times 1 + H(X|Y) = H(Y) + H(X|Y) \end{aligned}$$

### Disuguaglianza fondamentale

**Proposizione 2.15.2.** *La realizzazione di  $Y$  non può aumentare l'entropia di  $X$*

Se  $X$  e  $Y$  sono variabili casuali, allora

$$H(X|Y) \leq H(X)$$

*Dimostrazione:* ricordando che  $\ln t = \log_2 t / \log_2 e$ , riscriviamo  $\ln t \leq (t - 1)$ , vedi Figura 2.1, come

$$\log_2 t \leq (t - 1) \log_2 e \quad (2.1)$$

Notiamo che il segno di uguaglianza vale solo per  $t = 1$ . Pertanto, usando l'equazione (2.1), otteniamo

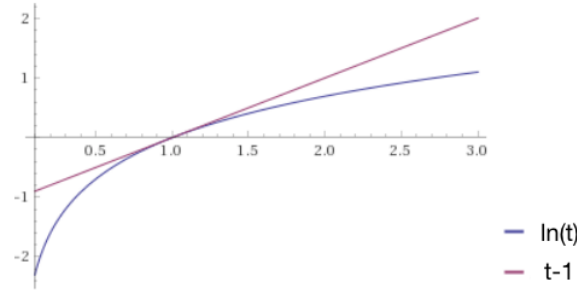


Figura 2.1: Vedi testo.

$$\begin{aligned}
 H(X|Y) - H(X) &= \sum_{i=1}^N \sum_{j=1}^M p_Y(y_j) p(x_i|y_j) \log_2 \frac{1}{p(x_i|y_j)} - \sum_{i=1}^N p_X(x_i) \log_2 \frac{1}{p_X(x_i)} \\
 &= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p(x_i|y_j)} - \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{1}{p_X(x_i)} \\
 &= \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \log_2 \frac{p_X(x_i)}{p(x_i|y_j)} \leq \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \left( \frac{p_X(x_i)}{p(x_i|y_j)} - 1 \right) \log_2 e \\
 &= \left( \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \frac{p_X(x_i)}{p(x_i|y_j)} - \sum_{i=1}^N \sum_{j=1}^M p(x_i, y_j) \right) \log_2 e \\
 &= \left( \sum_{i=1}^N \sum_{j=1}^M p_X(x_i) p_Y(y_j) - 1 \right) \log_2 e = (1 - 1) \log_2 e = 0
 \end{aligned}$$

**Osservazione 2.15.2.** *Entropia di variabili indipendenti*

Combinando le due relazioni otteniamo che se le variabili  $X$  e  $Y$  sono indipendenti

$$H(X, Y) = H(X) + H(Y)$$

ovvero l'entropia congiunta è uguale alla somma delle entropie.

**Compito 2.15.1.** *Dado a 16 facce*

Considera un dado equo con 16 facce. Se  $X$  è la variabile casuale che assume i 16 possibili valori e  $Y$  la variabile casuale che vale 0 se la faccia è pari e 1 se la faccia è dispari, calcola  $H(X)$ ,  $H(Y)$ ,  $H(X, Y)$ ,  $H(X|Y)$ ,  $H(Y|X)$  e  $I(X; Y)$  e commenta i risultati che ottieni.

**Assicurati di ...**

1. saper calcolare entropia congiunta e condizionata per coppie di variabili;
2. saper scrivere l'equazione e la disequazione che legano le varie forme di entropia incontrate e commentarle in termini di dipendenza tra variabili casuali

## 2.16 Lo stretto indispensabile sulla teoria dei codici

Introduciamo ora il tema delle codifiche capaci di ottenere compressione senza perdita di informazione. La nostra attenzione è circoscritta al caso particolare di codifiche binarie.

### Decifrabilità univoca e istantaneità

#### Definizione 2.16.1. Codifica

Sia  $\mathcal{X}$  un insieme finito di simboli che possiamo pensare come i valori possibili di una variabile casuale  $X$ . Una codifica per simbolo  $C$  è una funzione dall'insieme  $\mathcal{X}$  a  $\{0, 1\}^+$ .  $\square$

Indichiamo con  $L_C(x)$  la lunghezza di  $C(x)$  per  $x \in \mathcal{X}$ .

#### Definizione 2.16.2. Codifica estesa

La codifica estesa  $C^+$  è una funzione dall'insieme  $\mathcal{X}^+$  a  $\{0, 1\}^+$  ottenuta concatenando le rappresentazioni senza segni di interpunzione.

#### Esempio 2.16.1. Simboli non equiprobabili

Sia  $\mathcal{X} = \{a, b, c, d\}$  con  $p(a) = 1/2$ ,  $p(b) = 1/4$ ,  $p(c) = 1/8$  e  $p(d) = 1/8$ . Introduciamo una codifica  $C_1$  tale che

$$C_1(a) = 1000 \quad C_1(b) = 0100 \quad C_1(c) = 0010 \quad C_1(d) = 0001$$

Chiaramente  $L_{C_1}(x) = 4$  per tutti gli  $x \in \mathcal{X}$ . Nella codifica estesa  $C_1^+$  la stringa  $acdbac$  è codificata in

$$C_1^+(acdbac) = C_1(a)C_1(c)C_1(d)C_1(b)C_1(a)C_1(c) = 1000 \ 0010 \ 0001 \ 0100 \ 1000 \ 0010$$

dove aggiungiamo lo spazio tra le codifiche dei caratteri per poter leggere più facilmente la codifica estesa. Per quanto riguarda l'entropia per simbolo di  $C_1$  abbiamo

$$H(X) = \frac{1}{2} \log_2 2 + \frac{1}{4} \log_2 4 + \frac{2}{8} \log_2 8 = \frac{1}{2} + \frac{1}{2} + \frac{3}{4} = \frac{7}{4} = 1.75 \quad \square$$

Esaminiamo ora due proprietà importanti che consentono di caratterizzare le codifiche per simbolo senza perdita di informazione: decifrabilità univoca e istantaneità.

#### Definizione 2.16.3. Decifrabilità univoca

Un codifica  $C$  è *univocamente decifrabile* se per la codifica estesa  $C^+$

$$\forall x, y \in \mathcal{X}^+ \quad x \neq y \rightarrow C^+(x) \neq C^+(y) \quad \square$$

È immediato verificare che la codifica  $C_1$  è univocamente decifrabile, poiché le codifiche dei simboli sono diverse tra loro e tutte della stessa lunghezza.

#### Esempio 2.16.2. Codifiche con perdita di informazione

La codifica  $C_2$  per cui

$$C_2(a) = 1 \quad C_2(b) = 0 \quad C_2(c) = 10 \quad C_2(d) = 01$$

non è univocamente decifrabile. Per esempio

$$\begin{aligned} C_2^+(acdbac) &= 1 \ 10 \ 01 \ 0 \ 1 \ 10 \\ C_2^+(aabbadc) &= 1 \ 1 \ 0 \ 0 \ 1 \ 01 \ 10 \end{aligned}$$

#### Osservazione 2.16.1. Codice Morse

La decodifica del codice Morse è possibile grazie agli spazi tra punti e trattini, ovvero alle pause tra colpi secchi e accentuati. I simboli utilizzati dal codice Morse in effetti sono tre: “.”, “—” e “fine-simbolo”. Nella versione a soli due simboli, “.” e “—”, il codice Morse non è univocamente decifrabile.

**Definizione 2.16.4. Istantaneità**

Una codifica  $C$  è *istantanea*, o *istantaneamente decifrabile*, se per ogni  $x \in \mathcal{X}$   $C(x)$  non è un prefisso di qualunque altra rappresentazione.

**Osservazione 2.16.2. Istantaneità vuol dire efficienza**

Una codifica istantanea consente l'identificazione di fine rappresentazione di ogni simbolo senza dover attendere l'arrivo del simbolo successivo e quindi consente implementazioni efficienti.

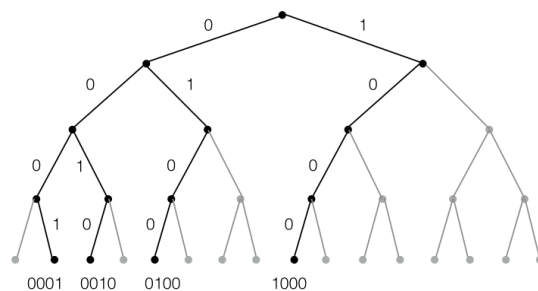


Figura 2.2: Un possibile albero per la codifica istantanea  $C_1$ . Ogni simbolo è una foglia la cui profondità è data dalla lunghezza della rappresentazione che si ottiene concatenando i bit del cammino che parte dalla radice e raggiunge la foglia.

**Osservazione 2.16.3. Istantaneità implica decifrabilità univoca**

È facile verificare che una codifica istantanea è univocamente decifrabile. Il vincolo del prefisso consente di associare all'intera codifica un albero binario (vedi Figura 2.2).

Consideriamo ora la codifica  $C_3$ , sempre per lo stesso insieme di simboli  $a, b, c$  e  $d$ , con

$$C_3(a) = 1 \quad C_3(b) = 10 \quad C_3(c) = 100 \quad C_3(d) = 1000$$

La codifica  $C_3$  è univocamente decifrabile ma, chiaramente, non istantanea.

**Lunghezza attesa e disuguaglianza di Kraft-McMillian**

Una codifica deve anche mirare a ottenere, in media, quanta più compressione possibile. Al riguardo definiamo la lunghezza attesa di un codice e rispondiamo a una domanda fondamentale attraverso il teorema di Kraft-McMillian.

**Definizione 2.16.5. Lunghezza attesa di una codifica**

La *lunghezza attesa*  $L(C, \mathcal{X})$  di una codifica  $C$  è

$$L(C, \mathcal{X}) = \sum_{x \in \mathcal{X}} p(x) L_C(x)$$

Per quanto riguarda gli esempi visti sopra abbiamo

$$\begin{aligned} L(C_1, \mathcal{X}) &= \frac{1}{2} \times 4 + \frac{1}{4} \times 4 + \frac{1}{8} \times 4 + \frac{1}{8} \times 4 = 4 \\ L(C_2, \mathcal{X}) &= \frac{1}{2} \times 1 + \frac{1}{4} \times 1 + \frac{1}{8} \times 2 + \frac{1}{8} \times 2 = 1.25 \\ L(C_3, \mathcal{X}) &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 4 = 1.875 \end{aligned}$$

Dato un insieme  $\mathcal{X}$  di simboli  $x_i$  e di interi  $L_i$  con  $i = 1, \dots, |\mathcal{X}|$ , esiste una codifica univocamente decifrabile  $C$  che abbia gli interi  $L_i$  come lunghezze delle rappresentazioni  $C(x_i)$ ?

**Teorema 2.16.1. Kraft-McMillian**

Le lunghezze  $L_i$  delle rappresentazioni  $C(x_i)$  di una codifica  $C$  univocamente decifrabile soddisfano la disuguaglianza

$$\sum_{i=1, \dots, |\mathcal{X}|} 2^{-L_i} \leq 1$$

*Dimostrazione:* definiamo  $A = \sum_i 2^{-L_i}$  e, per qualche intero  $n$ , consideriamo la quantità

$$A^n = \left( \sum_i 2^{-L_i} \right)^n = \sum_{i_1} \sum_{i_2} \dots \sum_{i_n} 2^{-(L_{i_1} + L_{i_2} + \dots + L_{i_n})}$$

La quantità  $L_{i_1} + L_{i_2} + \dots + L_{i_n}$  è la lunghezza della rappresentazione di  $x = x_{i_1} x_{i_2} \dots x_{i_n}$ . Abbiamo un termine della somma per ogni stringa  $x$  di  $n$  simboli e tutti i termini diversi per via del fatto che  $C$  è univocamente decifrabile. Introduciamo un vettore  $v_L$  che conta quante stringhe  $x$  nella somma hanno una rappresentazione di lunghezza  $L$ . Siano  $L_m$  ed  $L_M$  rispettivamente il valore minimo e massimo delle lunghezze delle rappresentazioni di ogni simbolo. Possiamo scrivere

$$A^n = \sum_{L=nL_m, \dots, nL_M} 2^{-L} v_L$$

Poiché ci sono al più  $2^L$  stringhe binarie di lunghezza  $L$ , abbiamo  $v_L \leq 2^L$  e, quindi,

$$A^n = \sum_{L=nL_m, \dots, nL_M} 2^{-L} v_L \leq \sum_{L=nL_m, \dots, nL_M} 1 < nL_M$$

Ora, se  $A$  fosse maggiore di 1 questa disuguaglianza non potrebbe valere per tutti gli  $n$ , poiché il termine  $A^n$  crescerebbe più velocemente del termine lineare  $nL_M$ . Pertanto  $A \leq 1$ .

■

**Osservazione 2.16.4. Verifica**

Verifichiamo la disuguaglianza di Kraft-McMillian per le codifiche  $C_1$ ,  $C_2$  e  $C_3$ .

$C_1$  è univocamente decifrabile poiché  $L_i = 4$  per ogni  $i$  e

$$2^{-4} + 2^{-4} + 2^{-4} + 2^{-4} = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{1}{4} < 1$$

$C_2$  non è univocamente decifrabile poiché

$$2^{-1} + 2^{-1} + 2^{-2} + 2^{-2} = \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} = \frac{3}{2} > 1$$

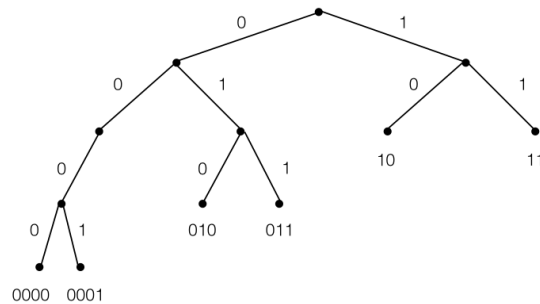
$C_3$  è univocamente decifrabile poiché  $L_i = i + 1$  per  $i = 0, 1, 2$  e  $3$  e

$$2^{-1} + 2^{-2} + 2^{-3} + 2^{-4} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16} < 1$$

**Alberi binari associati a una codifica istantanea e teorema inverso**

Se assumiamo che la codifica  $C$  sia anche istantanea, la disuguaglianza di Kraft-McMillian può essere apprezzata partendo dall'albero binario associato a  $C$ . Assumiamo di rilasciare un'unità di un liquido dalla radice dell'albero. In ogni nodo, ad ogni livello, il liquido si divide in due parti uguali. In una foglia alla profondità  $L_i$  troveremo, pertanto, una frazione dell'unità pari a  $2^{-L_i}$  che associamo al valore  $x_i$ . Se alcuni rami non sono utilizzati parte del liquido andrà perso e il totale del liquido raccolto nelle foglie corrispondenti ai valori rappresentati sarà strettamente minore dell'unità. Se tutti i rami e le foglie sono utilizzati, invece, il totale del liquido raccolto sarà pari a 1 e la codifica è *completa*.

In effetti vale anche l'inverso del teorema **2.16.1**.



**Figura 2.3:** Costruzione di una codifica istantanea da un insieme di interi che soddisfano la disuguaglianza di Kraft-McMillian. Date le lunghezze  $L(x_1) = L(x_2) = 4$ ,  $L(x_3) = L(x_4) = 3$  e  $L(x_5) = L(x_6) = 2$ , otteniamo un albero binario di profondità 4. Partendo dall'alto si assegna il primo nodo disponibile di profondità 2 alla prima rappresentazione cancellando tutti i discendenti. Si ripete l'operazione ottenendo  $C(x_1) = 0000$ ,  $C(x_2) = 0001$ ,  $C(x_3) = 010$ ,  $C(x_4) = 011$ ,  $C(x_5) = 10$  e  $C(x_6) = 11$  e una codifica completa.

### **Teorema 2.16.2.** *Inverso di Kraft-McMillian*

Sia  $x$  una variabile casuale che assume valori nell'insieme  $\mathcal{X}$ . Se un insieme di interi  $L_i$ , con  $i$  che varia da 1 a  $|\mathcal{X}|$ , soddisfa la disuguaglianza di Kraft-McMillian, allora esiste una codifica istantanea  $C$  per la quale  $L_C(x_i) = L_i$ .

*Dimostrazione:* disponiamo ogni simbolo  $x_i$  alla profondità appropriata,  $L_i$ , in un albero binario partendo dalla profondità minima ed eliminando tutti i discendenti (vedi Figura 2.3). Questa operazione è sempre possibile perché le lunghezze soddisfano la disuguaglianza di Kraft-McMillian. La codifica ottenuta è istantanea per costruzione.

■

### **Osservazione 2.16.5.** *Codifiche univocamente decifrabili e istantanee*

Per una codifica univocamente decifrabile  $C$  vale il **Teorema 2.16.1**. Allo stesso tempo abbiamo appena visto che, per il **Teorema 2.16.2**, è sempre possibile costruire una codifica  $C'$  istantanea che usa le stesse lunghezze ottenute con  $C$ . Nella pratica, pertanto, se disponiamo di una codifica univocamente decifrabile possiamo sempre pensare che sia anche istantanea (o comunque modificabile in una codifica istantanea con simboli codificati con le stesse lunghezze).

### **Assicurati di ...**

1. saper usare la disuguaglianza di Kraft-McMillian per determinare se una codifica è univocamente decifrabile
2. aver capito la differenza tra codifiche univocamente decifrabili e istantanee attraverso l'utilizzo di alberi binari
3. saper calcolare la lunghezza attesa di una codifica

## 2.17 Codifiche in assenza di rumore

Enunciamo innanzitutto una proprietà fondamentale della compressione in assenza di rumore e discutiamo il ruolo dell'entropia come limite inferiore di comprimibilità.

### Generalità sulla compressione

Che cosa significa comprimere una data quantità di informazione? La risposta è legata al numero di bit necessari a rappresentare i valori che può assumere una variabile casuale.

#### Definizione 2.17.1. Quantità di informazione grezza

Sia  $\mathcal{X}$  l'insieme dei possibili valori della variabile casuale  $X$ . La quantità di informazione grezza,  $H_0$ , è

$$H_0 = \log_2 |\mathcal{X}|$$

ed è uguale all'entropia associata a  $X$  assumendo che tutti i suoi  $|\mathcal{X}|$  valori siano equiprobabili.

#### Esercizio 2.17.1. Limite invalicabile

Se  $L_C(x)$  è la lunghezza della rappresentazione del valore  $x \in \mathcal{X}$  in una codifica  $C$ . Esiste una codifica binaria  $C^*$  tale che  $\forall x \in \mathcal{X}, L_{C^*}(x) < H_0$ ?

*Soluzione*

Conosciamo già la risposta: abbiamo visto che i bit necessari per rappresentare  $n$  numeri sono  $\log_2 n$ . Sappiamo quindi che per codificare  $|\mathcal{X}|$  valori sono necessari *almeno*  $H_0 = \log_2 |\mathcal{X}|$ . Questo risultato può essere visto come un'applicazione del *pigeon-hole principle*: se dobbiamo sistemare  $n + k$  lettere in  $n$  buche con  $k \geq 1$ , almeno una buca conterrà due lettere.

#### Osservazione 2.17.1. Non esistono compressori perfetti

La legge del buco della piccionaia è alla base di un risultato generale sull'impossibilità di comprimere *tutti* i possibili file di dimensione  $M$  bit in file di dimensione strettamente minore di  $M$ . I possibili file di dimensione  $M$  bit sono  $2^M$ . Tutti i file di dimensione minore di  $M$  bit sono dati dalla somma  $S = 2 + 2^2 + \dots + 2^{M-1}$ . Poiché  $2S = 2^2 + \dots + 2^{M-1} + 2^M$ , abbiamo

$$S = 2S - S = 2^2 + \dots + 2^{M-1} + 2^M - (2 + 2^2 + \dots + 2^{M-1}) = 2^M - 2 < 2^M$$

Non ci sono abbastanza file per comprimere *tutti* i file di dimensione  $M$  bit in file di dimensione al più  $M - 1$ . Pertanto, può capitare che un compressore reale espanda le dimensioni di un particolare file!

### Ruolo dell'entropia nella compressione

Siamo ora in grado di stabilire un limite inferiore per la lunghezza attesa di una codifica univocamente decifrabile.

#### Teorema 2.17.1. Entropia come limite insuperabile

Per una codifica univocamente decifrabile  $C$   $L(C, \mathcal{X}) \geq H(X)$ .

*Dimostrazione:* siano  $z = \sum_i 2^{-L_i} \leq 1$  e  $q_i = 2^{-L_i}/z$  per  $i = 1, \dots, |\mathcal{X}|$  con  $\sum_i q_i = 1$ . Applicando il logaritmo a  $q_i = 2^{-L_i}/z$  si ottiene

$$L_i = \log_2 \frac{1}{q_i} - \log_2 z$$

Ora, se  $p_i = p(x_i) > 0$  per  $i = 1, \dots, |\mathcal{X}|$  e  $\sum_i p_i = 1$  sono le probabilità dei simboli in  $\mathcal{X}$  e  $t_i = q_i/p_i$ , usando la disequazione (2.1) è immediato ottenere che

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{q_i} \geq \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{p_i} = H(X)$$



Infatti

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{p_i} - \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{q_i} = \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 t_i \leq (\log_2 e) \sum_{i=1}^{|\mathcal{X}|} p_i (t_i - 1) = (\log_2 e) \sum_{i=1}^{|\mathcal{X}|} (q_i - p_i) = 0$$

con l'uguaglianza che vale se e solo se  $q_i = p_i$  per tutti gli  $i$ . Pertanto,

$$L(C, \mathcal{X}) = \sum_{i=1}^{|\mathcal{X}|} p_i L_i = \sum_{i=1}^{|\mathcal{X}|} p_i \log_2 \frac{1}{q_i} - \log_2 z \geq H(X) - \log_2 z \geq H(X)$$

dove l'uguaglianza si ha se e solo se  $z = 1$  e le lunghezze delle rappresentazioni soddisfano le uguaglianze  $L_i = \log_2 1/p_i$ .

■

### Osservazione 2.17.2. Controllo di consistenza

Coerentemente, le lunghezze attese  $L(C_1, \mathcal{X})$  e  $L(C_3, \mathcal{X})$  sono entrambe maggiori di  $H(X)$ . La lunghezza attesa di  $C_2$ , invece, viola la disuguaglianza del teorema ma non genera contraddizione perché la codifica non è univocamente decifrabile.

## Codifica di Huffman

Una codifica di Huffman è il risultato di un algoritmo di compressione per simbolo molto semplice. Nel caso in cui la probabilità di ogni simbolo sia nota, una codifica di Huffman è una codifica per simbolo ottimale. Come vedremo in seguito, questo non significa che raggiunga il limite inferiore dato dall'entropia. In molti casi questo limite può essere avvicinato solo ricorrendo a compressori basati su flusso di dati.

Una codifica di Huffman costruisce un albero da un insieme di  $|\mathcal{X}|$  foglie in  $|\mathcal{X}| - 1$  fusioni. Ogni carattere  $x \in \mathcal{X}$  è un oggetto con un attributo dato dalla probabilità  $p$  con cui  $x$  compare nel testo da comprimere. Per identificare i due oggetti con la probabilità più piccola si utilizza una coda con priorità con chiave basata su  $p$ . La probabilità del nuovo oggetto è data dalla somma delle probabilità degli oggetti identificati. In caso di valori uguali la priorità è casuale e la codifica finale può non essere univoca.

Consideriamo l'esempio in Figura 2.4 con  $\mathcal{X} = \{a, b, c, d\}$  e  $p(a) = 1/2$ ,  $p(b) = 1/4$ ,  $p(c) = 1/8$  e  $p(d) = 1/8$ . L'algoritmo parte assegnando a  $Q$ , coda con priorità, i caratteri  $a, b, c$  e  $d$ . Il primo dei tre cicli rimuove dalla coda  $c$  e  $d$ , i nodi con la probabilità più bassa, e inserisce nella coda un nuovo nodo  $z$  che è il risultato dell'identificazione dei nodi  $c$  e  $d$ . Il nodo  $z$  ha come figlio sinistro  $c$  e come figlio destro  $d$ . L'arco sinistro codifica 0 e l'arco destro 1. L'algoritmo restituisce l'unico nodo che rimane nella coda, la radice dell'albero della codifica. Una codifica di ogni carattere si legge concatenando i bit degli archi che uniscono la radice a ogni foglia.

### Algoritmo 2.17.1. HuffmanCoding

*Input:*  $\mathcal{X}$ , insieme di  $|\mathcal{X}|$  simboli con  $p(x)$  probabilità del simbolo  $x$  e  $\sum_{x \in \mathcal{X}} p(x) = 1$ .

*Output:*  $T$ , albero binario in cui ognuna delle  $|\mathcal{X}|$  foglie contiene una codifica binaria di uno degli  $x$ .

---

```

 $Q \leftarrow \mathcal{X}$ 
for  $i = 1 \rightarrow (|\mathcal{X}| - 1)$ 
     $x \leftarrow \text{left}(z) \leftarrow \text{EXTRACT-MIN}(Q)$ 
     $y \leftarrow \text{right}(z) \leftarrow \text{EXTRACT-MIN}(Q)$ 
     $p(z) \leftarrow p(x) + p(y)$ 
    INSERT( $Q, z$ )
return EXTRACT-MIN( $Q$ )

```

---

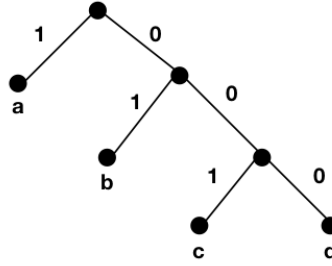


Figura 2.4: Codifica di Huffman:  $C(a) = 1$ ,  $C(b) = 01$ ,  $C(c) = 001$  e  $C(d) = 0001$ .

### Osservazione 2.17.3. Cenni alla correttezza

La dimostrazione di correttezza di *HuffmanCoding*, algoritmo di tipo *greedy*, è piuttosto elaborata e consiste di due passi. Nel primo si dimostra che una codifica ottimale può sempre essere modificata in modo tale che i due simboli con probabilità minore  $a$  e  $b$  siano figli dello stesso nodo e si trovino alla profondità massima dell'albero (e quindi differiscano nella codifica per il solo ultimo bit). Nel secondo, per induzione, si dimostra che l'albero ottenuto sostituendo i simboli a probabilità minima,  $a$  e  $b$ , con il nuovo simbolo  $ab$  di probabilità  $p(ab) = p(a) + p(b)$  è ottimale.

### Osservazione 2.17.4. Ridondanza

Nell'esempio precedente sappiamo che  $H(X) = 1.75$ . Come mostrato in Figura 2.4 l'algoritmo produce le rappresentazioni  $C(a) = 1$ ,  $C(b) = 01$ ,  $C(c) = 001$  e  $C(d) = 0001$ . La lunghezza attesa è

$$L(C, \mathcal{X}) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75 = H(X)$$

Quando, come in questo caso,  $L(C, \mathcal{X}) = H(X)$  la codifica è a ridondanza nulla e non ci sono margini di miglioramento per la compressione ottenibile.

### Osservazione 2.17.5. Punti di forza

Una codifica di Huffman, istantanea per costruzione, è ottimale. Nessun'altra codifica istantanea per simboli può fornire una migliore compressione media. Il suo grande successo è dovuto anche alla semplicità con la quale può essere implementata.

### Osservazione 2.17.6. Punti di debolezza

Alcune rigidità della codifica di Huffman possono presentare un conto salato. Consideriamo il caso di un file binario in cui  $p(0) \approx 1$ . L'entropia è quasi 0 ma la codifica di Huffman è inchiodata alla scelta binaria. La lunghezza media raggiunta resta lontana dal valore minimo dall'entropia anche nei casi in cui le probabilità sono lontane dall'inverso di potenze di 2, come nel caso di 3 soli possibili valori ugualmente probabili.

### Osservazione 2.17.7. Blocchi di simboli

Prima di concludere vediamo, per mezzo di un semplice esempio, come sia possibile migliorare l'efficienza della codifica di Huffman ricorrendo a una codifica per blocchi di simboli anziché per simbolo. Sia  $X$  una variabile casuale binaria con  $p(1) = 3/4$  e  $p(0) = 1/4$ . Valutiamo l'entropia per simbolo.

$$H(X) = \frac{3}{4} \log_2 \frac{4}{3} + \frac{1}{4} \log_2 4 \approx 0.81$$

Ciononostante, trattandosi di un alfabeto binario, la codifica di Huffman non può portare ad alcuna compressione. Consideriamo allora una codifica di Huffman per la rappresentazione di blocchi di  $N$ . Poniamo  $N = 4$ . Una codifica di Huffman  $C$  per i 16 possibili blocchi di 4 simboli è mostrata in tabella.

blocco	probabilità	codifica	blocco	probabilità	codifica	blocco	probabilità	codifica
1111	81/256	01	1010	9/256	00110	0010	3/256	000011
1110	27/256	111	0110	9/256	00101	0100	3/256	000010
1101	27/256	110	1001	9/256	00100	1000	3/256	0000001
1011	27/256	101	0101	9/256	00011	0000	1/256	0000000
0111	27/256	100	0011	9/256	00010			
1100	9/256	00111	0001	3/256	000001			

Per la lunghezza attesa per simbolo della rappresentazione di un blocco di 4 simboli in questa codifica otteniamo

$$\frac{L(C, \mathcal{X}^4)}{4} = \frac{1 \times 2 \times 81 + 4 \times 3 \times 27 + 6 \times 5 \times 9 + 3 \times 6 \times 3 + 1 \times 7 \times 3 + 1 \times 7 \times 1}{256 \times 4} \approx 0.82$$

ora più vicina al minimo valore possibile.

**Osservazione 2.17.8.** *Frequenza dei simboli e comprimibilità*

È istruttivo osservare che cosa succede alla frequenza relativa degli 0 e degli 1 nella codifica di Huffman. Se tutte le probabilità sono l'inverso di potenze di 2, ovvero quando la lunghezza attesa è uguale all'entropia, 0 e 1 compaiono nelle stesse proporzioni. Nella codifica di Huffman in cui  $C(a) = 1$  con  $p(a) = 1/2$ ,  $C(b) = 01$  con  $p(b) = 1/4$ ,  $C(c) = 001$  con  $p(c) = 1/8$  e  $C(d) = 000$  con  $p(d) = 1/8$ , le occorrenze di 0 e 1 - pesate con le rispettive probabilità - sono uguali. Che cosa possiamo dire quando la lunghezza attesa non è uguale all'entropia? Contiamo gli 1 e gli 0 nelle due colonne blocco della tabella di sopra, pesati con la corrispondente probabilità. La proporzione di partenza è 75% contro 25%. Se ripetiamo il conteggio dopo la codifica di Huffman a blocchi di 4 simboli, invece, otteniamo 51% contro 49%. La codifica a blocchi, quindi, produce sequenze difficili da comprimere ulteriormente.

Nella pratica la codifica di Huffman richiede di conoscere o di stimare le probabilità dei simboli o di blocchi di simboli. Ci soffermiamo brevemente su due strategie possibili. Nella prima si stimano le frequenze dei simboli su un *grande numero di file* pubblicando i risultati in un luogo accessibile al compressore e al decompressore. La seconda strategia stima le probabilità con le frequenze dei simboli presenti nel *file* e le include nel *file* compresso.

**Compito 2.17.1.** *Compressione di un testo*

Ogni carattere ASCII occupa in memoria un byte. Fissa un file in input che consiste di almeno  $10^5$  caratteri (spazi bianchi inclusi). Supponiamo che il file contenga  $M$  caratteri diversi. Poni la frequenza empirica del carattere  $x_i$  del file uguale alla probabilità  $p_i$  e calcola l'entropia di Shannon  $H(X)$  associata con  $\mathcal{X} = \{x_1, \dots, x_M\}$ . Implementa una codifica di Huffman  $C$  per l'alfabeto  $\mathcal{X}$  e confronta la lunghezza attesa  $L(C, \mathcal{X})$  con  $H(X)$ . Comprimi il testo usando la codifica e valuta la compressione assumendo che, nella codifica di Huffman, ogni 0 o 1 sia immagazzinato in un bit.

**Assicurati di ...**

1. saper confrontare lunghezza attesa di una codifica con l'entropia ed essere in grado di commentare il risultato ottenuto
2. saper trovare la codifica di Huffman in casi semplici utilizzando la coda con priorità

## 2.18 Codifica aritmetica

I metodi di compressione basati su flusso di dati aggirano molti dei problemi della codifica di Huffman (e di tutte le altre codifiche per simbolo). In particolare possono raggiungere una compressione vicina a quella ottimale. Discutiamo ora in dettaglio forse il più elegante metodo di compressione basato su flusso di dati in grado, con alta probabilità, di ottenere una lunghezza attesa uguale all'entropia per simbolo. Ci restringiamo al caso in cui i simboli sono cifre.

### Algoritmo di codifica

Per ogni  $N$  fissato, la codifica aritmetica associa biunivocamente a ogni stringa  $x = x_1x_2 \dots x_N$  un intervallo  $\Phi_N(x) \subset [0, 1)$  di ampiezza

$$p(x|N) = p(x_1)p(x_2) \dots p(x_N)$$

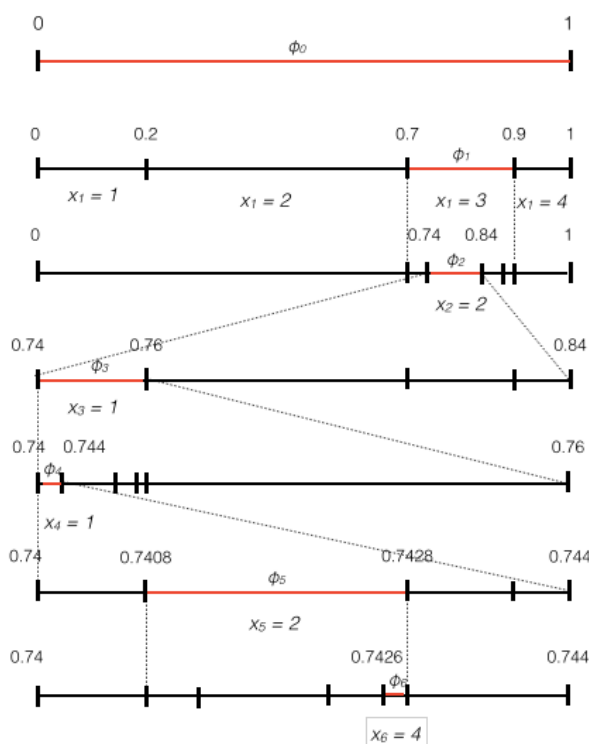


Figura 2.5: Codifica aritmetica di  $x = 321124$  con  $p(1) = 0.2, p(2) = 0.5, p(3) = 0.2$  e  $p(4) = 0.1$ .

**Esempio 2.18.1.** Sia  $\mathcal{X} = \{1, 2, 3, 4\}$  con

$$p(1) = 0.2, \quad p(2) = 0.5, \quad p(3) = 0.2 \text{ e } p(4) = 0.1$$

Per l'entropia otteniamo

$$H(\mathcal{X}) \approx 1.76$$

Poniamo  $N = 6$  e consideriamo  $x = 321124$ .

Costruiamo una sequenza di intervalli innestati  $\Phi_k$  chiusi a sinistra e aperti a destra

$$\Phi_k = [\alpha_k, \beta_k) \text{ per } k = 0, 1, \dots, 6$$

con  $\alpha_k$  e  $\beta_k$  numeri reali tali che  $0 \leq \alpha_k \leq \alpha_{k+1}$  e  $\beta_{k+1} \leq \beta_k \leq 1$ . Come mostrato nella Figura 2.5, nel primo passo dividiamo l'intervallo  $\Phi_0$ , ovvero l'intervallo  $[0, 1)$ , in 4 sotto-intervalli di lunghezza proporzionale a  $p(1)$ ,  $p(2)$ ,  $p(3)$  e  $p(4)$ . Il primo simbolo della stringa,  $x_1 = 3$ , seleziona l'intervallo  $\Phi_1 = [0.7, 0.9)$  di ampiezza  $p(3) = 0.2$ . A questo punto la suddivisione in 4 sotto-intervalli è eseguita nuovamente, rispettando le proporzioni, sul sotto-intervallo  $\Phi_1$ . Il secondo simbolo,  $x_2 = 2$ , seleziona il sotto-intervallo  $\Phi_2 = [0.74, 0.84)$  di ampiezza  $p(3)p(2) = 0.10$ . Sia l'ampiezza di  $\Phi_2$ , sia il suo estremo inferiore, sono determinati da  $x_1$  e  $x_2$ . Per ogni  $k$ , si ripete la suddivisione in 4 sotto-intervalli e si procede alla selezione del sotto-intervallo sulla base del  $k$ -esimo simbolo. Sia la posizione, sia l'ampiezza di ogni  $\Phi_k$ , sono determinati da  $x_k$  e da tutti i simboli precedenti.

Denotiamo con  $\mathcal{L}_k = \beta_k - \alpha_k$  l'ampiezza dell'intervallo  $\Phi_k = [\alpha_k, \beta_k)$ . Facendo uso della funzione di probabilità cumulata

$$cdf(x) = \sum_{i \leq x} p(i)$$

siamo ora in grado di descrivere un algoritmo per la codifica aritmetica.

**Algoritmo 2.18.1.** *ArithmeticCoding(x)*

*Input:*  $x = x_1 \dots x_N$ , stringa di  $N$  cifre  $x_k$  per  $k = 1, \dots, N$

*Output:*  $\Phi_N$ , intervallo  $\in [0, 1)$  chiuso a sinistra e aperto a destra

---

1  $\Phi_0 = [0, 1)$ ,  $\alpha_0 = 0$ ,  $\mathcal{L}_0 = 1$

2 **for**  $k = 1 \rightarrow N$

$$\Phi_k = [\alpha_k, \beta_k) = [\alpha_{k-1} + \mathcal{L}_{k-1}cdf(x_k - 1), \alpha_{k-1} + \mathcal{L}_{k-1}cdf(x_k))$$


---

**Osservazione 2.18.1.** *Correttezza*

Per quanto visto prima e per la ripetizione del passo 2,  $N$  volte, otteniamo che l'ampiezza  $\mathcal{L}_N$  dell'intervallo  $\Phi_N$  è data da

$$\mathcal{L}_N = p(x|N) = p(x_1)p(x_2) \dots p(x_N)$$

Procedendo per bisezioni successive dell'intervallo  $[0, 1)$ , quindi, è possibile associare univocamente all'intervallo  $\Phi_N$ , e alla stringa di lunghezza  $N$  una frazione binaria - esprimibile con  $L$  bit - dove

$$L = \left\lceil \log_2 \frac{1}{p(x|N)} \right\rceil = \left\lceil \log_2 \frac{1}{p(x_1)} + \log_2 \frac{1}{p(x_2)} + \dots + \log_2 \frac{1}{p(x_N)} \right\rceil$$

**Osservazione 2.18.2.** *Efficienza*

Rispetto alla codifica di Huffman, la codifica aritmetica è più efficiente perché arrotonda all'intero più vicino solo la lunghezza finale  $L$  della rappresentazione di  $x$  e non la lunghezza della rappresentazione di ognuna delle  $N$  cifre della stringa. Consideriamo il rapporto  $L/N$  al crescere di  $N$ ; se raggruppiamo la somma di sopra in  $|\mathcal{X}|$  somme parziali e indichiamo con  $f_j = \#j/N$  la frequenza con cui  $j$  compare nella stringa originale otteniamo

$$\frac{L}{N} = \left\lceil f_1 \log_2 \frac{1}{p(1)} + f_2 \log_2 \frac{1}{p(2)} + \dots + f_{|\mathcal{X}|} \log_2 \frac{1}{p(|\mathcal{X}|)} \right\rceil$$

Per la legge dei grandi numeri, al crescere di  $N$ , abbiamo che per tutti i  $j$

$$f_j \rightarrow p(j) \quad \text{e, quindi,} \quad \frac{L}{N} \rightarrow H$$

**Osservazione 2.18.3. Compressioni a confronto**

Nel nostro esempio, la codifica aritmetica ha determinato un intervallo di ampiezza  $A = 0.0002$  per cui la stringa  $x = 321124$  è rappresentata da una frazione dell'unità che in base 2 ha lunghezza

$$L = \left\lceil \log_2 \frac{1}{A} \right\rceil = \lceil \log_2 5000 \rceil = \lceil 12.28 \rceil = 13$$

Poiché

$$H(\mathcal{X}) \approx 6 \times 1.76 \approx 10.56$$

la codificata ottenuta non è ottimale. Questa inefficienza dipende da due motivi distinti. In primo luogo, il valore calcolato non tiene conto del fatto che gli 0 che seguono l'ultima cifra significativa dopo la virgola sono eliminabili. Nel caso specifico, l'unica frazione dell'unità che appartiene all'intervallo  $[0.7426, 0.7428)$  costituita da 13 bit è

$$\theta = 0.74267578125 = 0.1011111000100_2$$

Eliminando gli ultimi due 0, che non sono significativi, la lunghezza della rappresentazione si riduce a 11 bit. La lunghezza media della rappresentazione di tutte le possibili stringhe di 6 simboli si avvicina asintoticamente a  $L - 1$  al crescere di  $N$ , ma manca ancora qualcosa. Il secondo motivo è dovuto alla discrepanza tra le probabilità *a priori* dei simboli e la loro frequenza empirica nella stringa. In altre parole,  $N$  non è abbastanza grande per far scattare le conseguenze della legge dei grandi numeri. Per valori piccoli di  $N$  i vantaggi della codifica aritmetica non sono sempre apprezzabili.

**Algoritmo di decodifica**

Analizziamo la decodifica sullo stesso esempio. Sia  $\theta$  la rappresentazione ottenuta con la codifica aritmetica e  $N$  la lunghezza della stringa originale. La prima cifra della stringa,  $x_1$ , è ricostruita determinando a quale dei 4 sotto-intervalli dell'insieme  $[0, 1)$  appartenga  $\theta$ . Quindi si procede ad aggiornare il valore di  $\theta$ , ovvero nel sottrarre al valore corrente di  $\theta$  la funzione di probabilità cumulata valutata in  $x_1$ ,  $cdf(x_1)$ , e nel dividere il risultato ottenuto per la probabilità di  $x_1$ ,  $p(x_1)$ . La sottrazione determina la posizione di  $\theta$  all'interno del sotto-intervallo rispetto all'estremo inferiore, mentre la divisione per la probabilità aggiusta il fattore di scala. La procedura descritta è ripetuta per il nuovo valore di  $\theta$  e, dopo  $N - 1$  aggiornamenti, termina con la ricostruzione della stringa. La tabella qui sotto illustra i passi della codifica e della decodifica per l'esempio della stringa  $x = 321124$ . Il numero  $\theta$  è rappresentato in base decimale. La tabella mostra anche che cosa succede all'iterazione  $N + 1$ : se mancasse l'informazione sulla lunghezza della stringa originale, infatti, la decodifica non saprebbe quando fermarsi.

$k$	carattere da codificare	estremo inferiore di $\Phi_k$	ampiezza di $\Phi_k$	$\theta_{k-1}$	carattere codificato
0	-	0	1	-	-
1	3	0.7	0.2	0.74267578125	3
2	2	0.74	0.1	0.21337890625	2
3	1	0.74	0.02	0.0267578125	1
4	1	0.74	0.004	0.1337890625	1
5	2	0.7408	0.002	0.6689453125	2
6	4	0.7426	0.0002	0.937890625	4
7				0.37890625	2

Supponiamo inizialmente di conoscere  $p(x)$  per  $x = 1, 2, \dots, |\mathcal{X}|$  e la funzione di probabilità cumulata  $cdf(x)$ .

**Algoritmo 2.18.2.** *ArithmeticDecoding*( $N, \theta$ )*Input:*  $N$  lunghezza della stringa codificata,  $\theta$  codifica aritmetica in notazione decimale*Output:*  $x$ , stringa decodificata

1.  $\theta_0 = \theta$
2. **for**  $k = 1 \rightarrow N$

$$x_k = \{x \in \mathcal{X} : cdf(x-1) < \theta_{k-1} < cdf(x)\}$$

$$\theta_k = \frac{\theta_{k-1} - cdf(x_{k-1})}{p(x_k)}$$

**Osservazione 2.18.4.** *Separazione del modello dalla codifica*

Il fatto che la decodifica proceda nella ricostruzione della stringa originale elaborando la stringa codificata un simbolo alla volta apre a variazioni capaci di sfruttare la separazione del modello probabilistico sottostante dalla codifica. Si può costruire un modello nel quale le probabilità sono aggiornate ricorsivamente sulla base della lettura del flusso di simboli. Questa strategia incorpora nella codifica il modello probabilistico utilizzato e ne consente la ricostruzione in decodifica senza richiedere alcun passaggio esplicito di informazioni. **La codifica aritmetica, pertanto, è in grado di ottenere compressioni efficienti anche quando le probabilità *a priori* dei simboli non siano disponibili o in presenza di dati caratterizzati da frequenze e occorrenze empiriche particolari.**

**Osservazione 2.18.5.** *Dati come flusso*

La codifica aritmetica opera su un flusso di dati in input. Sia la codifica, sia la decodifica, infatti, elaborano esclusivamente l'informazione acquisita fino a quel momento.

**Compito 2.18.1.** *Codifica e decodifica adattive*

Implementa una versione di *ArithmeticCoding* che codifica il  $k$ -esimo bit  $b_k$  per  $k = 1, \dots, N$  con la probabilità  $p_{k-1}(b_k = 1) = 1 - p_{k-1}(b_k = 0)$ . Se inizialmente 0 e 1 sono equiprobabili, poni  $p_0(b_1 = 0) = p_0(b_1 = 1) = 1/2$ . Per i successivi utilizza le ricorsioni

$$p_k(0) = p_{k-1}(0) + \frac{(1 - b_k) - p_{k-1}(0)}{k + 2} \quad \text{e} \quad p_k(1) = p_{k-1}(1) + \frac{b_k - p_{k-1}(1)}{k + 2}$$

Il “+2” al denominatore tiene conto del fatto che tutte le sequenze, inizialmente, sono costituite da una coppia 01 non codificata e responsabile dell'inizializzazione  $p_0(b_1 = 0) = p_0(b_1 = 1) = 1/2$ .

Sempre con  $p_0(0) = p_0(1) = 1/2$ , implementa *ArithmeticDecoding* usando al passo  $k$

$$cdf_k(x) = \sum_{i \leq x} p_k(i)$$

**Assicurati di ...**

1. saper trovare il sotto-intervallo determinato dalla codifica aritmetica in casi semplici come stringhe binarie
2. saper spiegare il motivo per il quale la codifica aritmetica si avvicina asintoticamente alla compressione ottimale

## 2.19 Codifiche in presenza di rumore

Un segnale digitale trasmesso attraverso un canale può essere affetto da rumore. Nel caso dell'archiviazione di dati in memoria, per esempio, la grande quantità di bit da archiviare rende molto alta la probabilità che qualche bit sia invertito. Nel caso delle telecomunicazioni spaziali, invece, il rumore è particolarmente intenso. Invariabilmente il problema è lo sviluppo di una codifica in grado di proteggere il segnale dal rumore. Le soluzioni individuate si basano sempre su codifiche ridondanti.

### Distanza di Hamming

Prima di discutere una codifica in presenza di rumore, introduciamo una nozione di distanza utile per confrontare sequenze di bit.

#### **Definizione 2.19.1.** *Distanza di Hamming*

Per due sequenze di bit  $u = u_1 u_2 \dots u_N$  e  $v = v_1 v_2 \dots v_N$  di lunghezza  $N$  la distanza di Hamming  $d_H(u, v)$  è definita come il numero di posizioni nelle quali i bit delle due sequenze sono diversi.

#### **Proposizione 2.19.1.** *La distanza di Hamming è una metrica.*

*Dimostrazione*

*Simmetria:* per tutte le sequenze di bit  $u$  e  $v$  di lunghezza  $N$  il fatto che  $d_H(u, v) = d_H(v, u)$  segue immediatamente dal fatto che  $d_H(u, v)$  e  $d_H(v, u)$  contano lo stesso numero di posizioni.

*Non negatività:* per tutte le sequenze di bit  $u$  e  $v$  di lunghezza  $N$  il fatto che  $d_H(u, v) \geq 0$  e  $d_H(u, v) = 0$  se e solo se  $u = v$  segue immediatamente dalla definizione

*Disuguaglianza triangolare:* per tutte le sequenze di bit  $u, v$  e  $w$  di lunghezza  $N$  deve valere

$$d_H(u, w) + d_H(w, v) \geq d_H(u, v)$$

Partiamo osservando che per ogni posizione  $i$  in cui  $u_i = v_i$  la disuguaglianza è soddisfatta. Possiamo pertanto limitarci alle sole posizioni nelle quali  $u$  e  $v$  differiscono. Per ogni posizione  $i$  in cui  $u_i \neq v_i$  non può essere che  $u_i = w_i$  e  $v_i \neq w_i$ . Pertanto, quando  $d_H(u, v)$  aumenta di un'unità, anche  $d_H(u, w) + d_H(w, v)$  aumenta di un'unità. ■

Fissiamo un intero  $K > 0$  e supponiamo di voler trasmettere tutte le  $2^K$  sequenze possibili di lunghezza  $K$  attraverso un canale affetto da rumore. Per ogni sequenza  $x$  di lunghezza  $K$  ci sono  $K$  sequenze  $x^i$  a distanza di Hamming uguale a 1 da  $x$ , con  $x^i$  la sequenza che differisce da  $x$  nel bit  $i$ -esimo con  $i = 1, \dots, K$ . Se nella trasmissione di  $x$  si invertisse un solo bit, pertanto, la sequenza ricevuta sarebbe una di queste  $K$  sequenze e non saremmo in grado di rilevare la presenza di un errore. L'idea della codifica in presenza di rumore è allora aggiungere  $M$  bit a ogni blocco in modo tale da ottenere  $2^K$  sequenze di  $K + M$  bit che possano mantenersi distinguibili in presenza di errori nella trasmissione, in quanto a distanza di Hamming più grande fra loro.

### Codifica convoluzionale

La codifica convoluzionale fornisce un'elegante soluzione al problema di trovare codifiche in presenza di rumore. L'idea alla base della codifica convoluzionale è molto semplice. Nella fase di codifica una finestra di lunghezza  $K$  scorre su una sequenza in input di  $N$  bit avanzando di una posizione alla volta. In ogni posizione i  $K$  bit all'interno della finestra sono combinati per calcolare un blocco di  $P$  bit di parità che è poi trasmesso attraverso il canale. La sequenza originale di  $N$  bit è pertanto codificata in una sequenza di  $N$  blocchi di  $P$  bit,  $N \times P$  bit in tutto. È immediato rendersi conto che la velocità di trasmissione della codifica convoluzionale è  $V = 1/P$ .

Indichiamo con  $x[n]$  l' $n$ -esimo bit della sequenza  $x$  in input con  $n = 1, 2, \dots, N$ . Sia  $K$  la lunghezza dei vincoli, ovvero la dimensione della finestra  $W$  che scorre su  $x$  avanzando di una posizione a



ogni passo e  $P$  il numero di sottoinsiemi dei  $K$  bit selezionati per il calcolo dei valori dei  $P$  bit di parità  $C_n = (y_1[n], y_2[n], \dots, y_P[n])$  per  $n = 1, \dots, N$ . Posto  $x[n] = 0$  per i  $K - 1$  valori di  $n$  compresi tra  $-K + 2$  e  $0$ , il blocco di  $P$  bit di parità  $C_n = (y_1[n]y_2[n] \dots y_P[n])$  per ogni  $n = 1, \dots, N$  è ottenuto dalle equazioni

$$\begin{aligned} y_1[n] &\equiv w_1[0]x[n] + w_1[1]x[n-1] + \dots + w_1[K-1]x[n-K+1] \pmod{2} \\ y_2[n] &\equiv w_2[0]x[n] + w_2[1]x[n-1] + \dots + w_2[K-1]x[n-K+1] \pmod{2} \\ &\dots \equiv \dots \\ y_P[n] &\equiv w_P[0]x[n] + w_P[1]x[n-1] + \dots + w_P[K-1]x[n-K+1] \pmod{2} \end{aligned}$$

dove i polinomi generatori  $w_1, w_2, \dots, w_P$  agiscono da selezionatori dei bit della sequenza di input all'interno della finestra  $W$ .

$$f : \mathcal{S} \times \{0, 1\} \rightarrow \mathcal{S} \text{ con}$$

$$f(s_n, x[n]) = s_{n+1} \text{ ovvero}$$

$$f(x[n-2]x[n-1], x[n]) = x[n-1]x[n]$$

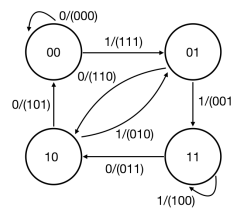


Figura 2.6: Funzione di transizione e FSM per l'esempio 2.19.1.

### Osservazione 2.19.1. Macchina a stati finiti per la codifica

Nel linguaggio delle macchine a stati finiti (FSM), l'input è il bit  $x[n]$ , lo stato corrente  $s_n$  i  $K - 1$  bit  $x[n-1]x[n-2] \dots x[n-K+1]$  e l'output della FSM il blocco di  $P$  bit  $C_n = (y_1[n]y_2[n] \dots y_P[n])$ . Gli stati possibili sono  $2^{K-1}$ . Il comportamento della FSM è completamente determinato dalle equazioni di parità. Il diagramma nella Figura 2.6 si riferisce al caso del prossimo esempio. Lo stato corrente è scritto all'interno di ogni nodo, mentre ogni arco punta al nuovo stato riportando il valore dell'input e il blocco dei  $P = 3$  bit codificati. La funzione di transizione  $f$  che sulla base dello stato  $s_n$  e dell'input  $x[n]$  restituisce lo stato  $s_{n+1}$  è esplicitata nella Figura 2.6.

**Esempio 2.19.1.** Per  $K = 3$  e  $P = 3$ , se  $w_1 = [1 \ 1 \ 1]$ ,  $w_2 = [1 \ 1 \ 0]$  e  $w_3 = [1 \ 0 \ 1]$  per le equazioni di parità abbiamo

$$\begin{aligned} y_1[n] &\equiv x[n] + x[n-1] + x[n-2] \pmod{2} \\ y_2[n] &\equiv x[n] + x[n-1] \pmod{2} \\ y_3[n] &\equiv x[n] + x[n-2] \pmod{2} \end{aligned}$$

Per ogni  $n$  da 1 a  $N + 1$  lo stato  $s_n$  appartiene all'insieme  $\mathcal{S} = \{00, 01, 10, 11\}$ . Utilizzando la macchina a stati finiti in Figura 2.6 per codificare la sequenza  $x = 1101$  otteniamo

$$C(x) = (C_1(x))(C_2(x))(C_3(x))(C_4(x)) = (111)(001)(011)(010)$$

### Algoritmo di Viterbi

Presentiamo ora un algoritmo di decodifica. Descriviamo il suo funzionamento sullo stesso esempio utilizzato per la codifica. Sappiamo che la sequenza  $x = 1101$  è codificata in output nella sequenza di  $N = 4$  blocchi di  $P = 3$  bit,  $C_1, \dots, C_4$ . Se per effetto del rumore il terzo bit di  $C_1^{ric}$  e il primo bit di  $C_3^{ric}$  sono invertiti all'uscita del canale riceveremo

$$C^{ric} = (C_1^{ric})(C_2^{ric})(C_3^{ric})(C_4^{ric}) = (110)(001)(111)(010)$$

Misuriamo l'accordo tra  $C(x)$  e  $C^{ric}$  contando il numero di posizioni  $L(x)$  in cui le due sequenze di 4 blocchi di 3 bit non coincidono. Agendo sulle 4 triplette separatamente, otteniamo

$$L(x) = d_H(C(x), C^{ric}) = \sum_{n=1}^4 d_H(C_n(x), C_n^{ric})$$

Chiaramente,  $L(x) = 0$  se e solo se non ci sono errori di trasmissione, ovvero se e solo se  $C^{ric} = C(x)$ . Pertanto possiamo enunciare un principio generale.

**Principio 2.19.1.** *Piccolo è bello*

La sequenza in input è quella che codifica in output la sequenza di bit di parità a distanza di Hamming minima dalla sequenza ricevuta in uscita dal canale.  $\square$

Poichè la tripletta  $C_n(x)$  è funzione del bit  $x[n]$  da codificare e dello stato  $s_n = x[n-2]x[n-1]$ , possiamo scrivere  $C_n(x) = C_n(s_n, x[n])$  e, quindi,

$$L(s_1, x[1], \dots, x[N]) = \sum_{n=1}^N d_H(C_n(s_n, x[n]), C_n^{ric})$$

L'algoritmo di Viterbi per la decodifica di  $C^{ric}$  si basa sulla ricerca della sequenza  $x^*$  tale che

$$\begin{aligned} x^* = x^*[1] \dots x^*[N] &= \arg \min_{x[1] \dots x[N]} L(s_1, x[1], \dots, x[N]) \\ \text{con } x[n] &= \{0, 1\} \text{ per } n = 1, \dots, N \text{ e} \\ s_1 &= 00 \end{aligned} \quad (2.2)$$

Nel nostro caso, ovviamente, abbiamo  $N = 4$ . Il valore del minimo,  $L^*(s_1)$ , è pari al numero di bit di parità invertiti nella trasmissione.

**Osservazione 2.19.2.** *Una forma distinguibile*

Il **Problema 2.2** ha una struttura particolare: **la parte finale della decodifica ottimale per l'intera sequenza ricevuta è la decodifica ottimale per la corrispondente parte finale della sequenza ricevuta** ed è quindi affrontabile con metodologie di programmazione dinamica (*DP*). In *DP* la sequenza di bit di input  $x[1] \dots x[N]$  è la sequenza di controllo e  $x^*[1] \dots x^*[N]$  la sequenza ottimale.  $\square$

Prima di procedere con la descrizione dell'algoritmo di decodifica, precalcoliamo le distanze di Hamming tra ogni possibile blocco ricevuto al passo  $n$  con il blocco dalla FSM di Figura 2.6 per i due possibili diversi input,  $x[n] = 0$  e  $x[n] = 1$  e per tutti i possibili stati. Otteniamo la seguente tabella

$x[n-2]x[n-1]$ stato: 00	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (000) \quad 1 \rightarrow (111)$	$x[n-2]x[n-1]$ stato: 10	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (101) \quad 1 \rightarrow (010)$
blocco ricevuto	distanza di Hamming	blocco ricevuto	distanza di Hamming
(000)	0      3	(000) (011) (110)	2      1
(001) (010) (100)	1      2	(001) (100) (111)	1      2
(011) (101) (110)	2      1	(010)	3      0
(111)	3      0	(101)	0      3
$x[n-2]x[n-1]$ stato: 01	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (110) \quad 1 \rightarrow (001)$	$x[n-2]x[n-1]$ stato: 11	$x[n] \rightarrow (y_1[n]y_2[n]y_3[n])$ $0 \rightarrow (011) \quad 1 \rightarrow (100)$
blocco ricevuto	distanza di Hamming	blocco ricevuto	distanza di Hamming
(000) (011) (101)	2      1	(000) (101) (110)	2      1
(001)	3      0	(001) (010) (111)	1      2
(010) (100) (111)	1      2	(011)	0      3
(110)	0      3	(100)	3      0

Per la decodifica ricorriamo al traliccio di Figura 2.7. Ogni riga si riferisce a uno dei quattro stati possibili, indicati nella colonna di sinistra. Le quattro triplette ricevute sono indicate nella riga superiore nell'ordine nel quale compaiono nella sequenza ricevuta. Ogni colonna di rettangoli si riferisce allo stato corrente  $n = 1, \dots, 5$  includendo lo stato finale nell'ultima colonna.

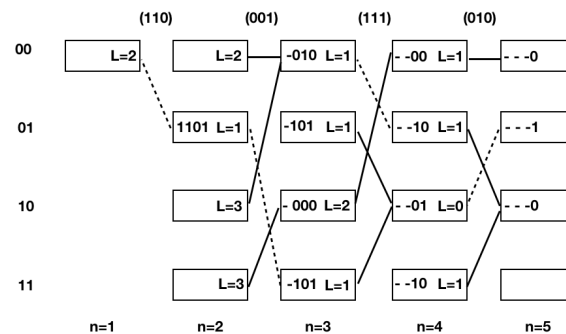


Figura 2.7: Vedi testo.

Sfruttando la programmazione dinamica possiamo riempire il traliccio da destra a sinistra in quattro passi. Partiamo dall'arrivo dell'ultima tripletta,  $C_4^{ric}$ . Dobbiamo determinare, per ogni possibile stato di partenza della colonna  $n = 4$ , quale dei due possibili valori di  $x[4]$  produce la tripletta  $C_4(s_4, x[4])$  più vicina a  $C_4^{ric}$ . In Figura 2.7 delle due possibili linee (piena se  $x[4] = 1$  e tratteggiata se  $x[4] = 0$ ) che partendo da ognuno dei quattro possibili stati della colonna  $n = 4$  terminano in uno stato della colonna  $n = 5$  è indicata sola quella corrispondente alla codifica della tripletta a distanza più piccola da  $C_4^{ric}$ . Il bit decodificato, il quarto, è rappresentato all'interno del rettangolo corrispondente allo stato finale cui punta la linea. Nel rettangolo da cui parte la linea è indicata la distanza di Hamming tra la tripletta ricevuta e la tripletta codificata.

Consideriamo ora la tripletta  $C_3^{ric}$ . Nuovamente, da ognuno dei quattro possibili stati nella colonna  $n = 3$  delle due possibili linee (piena se  $x[3] = 1$  e tratteggiata se  $x[3] = 0$ ) tracciamo solo quella che corrisponde al percorso che *minimizza la somma* delle distanze di Hamming delle triplette decodificate con le triplette ricevute per  $n = 3$  e  $n = 4$ . Se le due somme forniscono lo stesso valore, scegliamo una delle due linee a caso. I bit decodificati, il terzo e il quarto, sono mostrati sulla sinistra del rettangolo corrispondente allo stato di arrivo (colonna  $n = 4$ ), mentre il valore corrente della somma delle distanze di Hamming per le ultime due triplette decodificate ottimali è indicato nel rettangolo corrispondente allo stato di partenza (colonna  $n = 3$ ). Ripetendo per altre due volte la stessa procedura otteniamo una decodifica ottimale, nonché il numero e le posizioni dei bit ricevuti invertiti dal rumore.

### Assicurati di ...

1. sapere come si definisce la distanza di Hamming
2. saper progettare la FSM che implementa una codifica convoluzionale
3. saper descrivere, almeno a grandi linee, l'algoritmo di Viterbi



## Capitolo 3

# Elementi di Inferenza

Muoviamo ora i primi passi nel mondo dell'inferenza. Ci occupiamo dapprima dei metodi Monte Carlo, basati sul campionamento ripetuto da distribuzioni di probabilità note. Una lezione è dedicata a un **metodo Monte Carlo** applicato a **variabili casuali indipendenti e identicamente distribuite**, in un'altra, invece, introduciamo le **catene di Markov** e discutiamo il caso in cui le **variabili casuali sono dipendenti**. La lezione successiva copre un argomento avanzato, le **catene di Markov Monte Carlo** applicate al modello di Ising in due dimensioni. Questa lezione non fa parte del programma, ma può essere scelta come tema per una prova finale. Le rimanenti due lezioni introducono due principi basilari per l'inferenza. Nella prima discutiamo alcuni semplici esempi di inferenza Bayesiana: il **principio di massima probabilità a posteriori** che, appoggiandosi al teorema di Bayes, consente di *prendere decisioni ottimali* nel caso in cui le distribuzioni di probabilità in gioco sono note e **l'apprendimento bayesiano** per *determinare la distribuzione di probabilità dalla quale sono stati campionati i dati disponibili* nel caso in cui sia nota la forma parametrica della distribuzione. Nella seconda, infine, discutiamo una soluzione frequentista a quest'ultimo problema basata sul **principio di massima verosimiglianza**.

## 3.20 Metodi Monte Carlo

I metodi Monte Carlo risolvono un problema deterministico campionando molte volte da una distribuzione di probabilità. Ci limiteremo al caso più semplice, ovvero il calcolo di un integrale definito. Riprendiamo innanzitutto il risultato fondamentale che è alla base di tutti i metodi Monte Carlo.

### Legge dei grandi numeri

Siano  $X_i$  con  $i = 1, 2, \dots, n$  variabili casuali *indipendenti e identicamente distribuite*. Indichiamo con  $\langle X \rangle_n$  la media empirica di una realizzazione delle  $n$  variabili casuali

$$\langle X \rangle_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Se  $\mathbb{E}[X_n] = \mu$  e  $\text{Var}(X_n) = \sigma^2$ , per il valore atteso e la varianza di  $\langle X \rangle_n$  sappiamo che

$$\mathbb{E}[\langle X \rangle_n] = \mu \text{ e } \text{Var}(\langle X \rangle_n) = \frac{\sigma^2}{n}$$

e, dalla legge dei grandi numeri, otteniamo che con alta probabilità per  $n$  grande  $\langle X \rangle_n \approx \mu$ , ovvero

$$\forall \epsilon > 0 \lim_{n \rightarrow \infty} \Pr \{ |\langle X \rangle_n - \mu| \geq \epsilon \} = 0$$

### Area del cerchio

Dobbiamo calcolare l'area  $A$  di un cerchio di raggio  $R$  ma non ricordiamo la formula  $A = \pi R^2$ . Abbiamo trovato da qualche parte che

$$A = 2 \int_{-R}^R \sqrt{R^2 - x^2} dx$$

ma non sappiamo come calcolare l'integrale. Come possiamo uscirne?

Se sappiamo campionare da una distribuzione uniforme tra  $[-R, R]$  possiamo stimare  $A$  valutando  $\langle f \rangle_n$ , ovvero la frazione di  $n$  punti  $(x, y)$  campionati nel quadrato azzurro di lato  $2R$  (vedi Figura 3.1, a sinistra) che cadono all'interno del cerchio giallo inscritto, ovvero per i quali  $x^2 + y^2 \leq R^2$ . L'area del

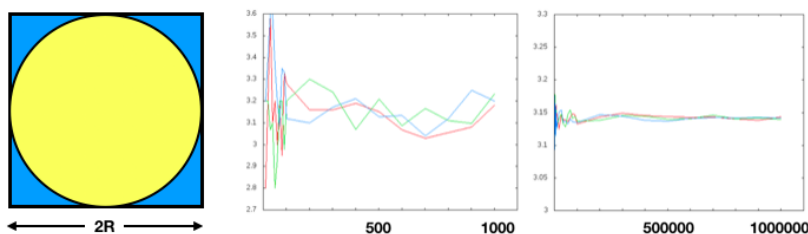


Figura 3.1: Vedi testo.

quadrato è  $4R^2$  per cui se  $f$  è il rapporto tra le aree del cerchio e del quadrato abbiamo  $A = 4fR^2$ . I due grafici in Figura 3.1, al centro e a destra rispettivamente, mostrano stime di  $\pi = 4f$  ottenute come  $4\langle f \rangle_n$ , con  $\langle f \rangle_n$  la stima *Monte Carlo* di  $f$  ottenuta per  $n$  da 1 a  $10^3$  e da  $10^3$  a  $10^6$ . Al crescere di  $n$  le oscillazioni di  $\langle f \rangle_n$  diminuiscono in ampiezza: questo risultato non ci sorprende perché stiamo di fatto stimando un valore atteso,  $f$ , con una media empirica,  $\langle f \rangle_n$ , e la deviazione standard sottostante decresce come  $1/\sqrt{n}$ . Stimato  $\pi$  come  $4\langle f \rangle_n$ , infine, otteniamo  $A = 4\langle f \rangle_n R^2$ .

## Caso generale

Supponiamo di dover calcolare l'integrale

$$I = \int_a^b f(x) dx$$

per una qualche  $f$  continua nell'intervallo  $[a, b]$ . Per il *teorema del valor medio* sappiamo che

$$\int_a^b f(x) dx = (b - a) \times \bar{f}$$

con  $\bar{f}$  un qualche valore assunto dalla funzione  $f$  nell'intervallo. Per stimare  $\bar{f}$  basta campionare  $n$  valori nell'intervallo  $[a, b]$  e valutare la media

$$\langle f \rangle_n = \frac{1}{n} \sum_{i=1}^n f(x_i)$$

Mettendo insieme i pezzi abbiamo infine che, per  $n$  grande, con alta probabilità

$$\int_a^b f(x) dx \approx (b - a) \langle f \rangle$$

### Osservazione 3.20.1. Implementazione efficiente

Nel caso dovessimo ripetere il calcolo della media al variare di  $n$ , è preferibile calcolare  $\langle f \rangle_n$  ricorsivamente. Poniamo  $\langle f \rangle_0 = 0$  e, se  $x_n$  è l' $n$ -esimo valore campionato uniformemente nell'intervallo  $[a, b]$  e  $\langle f \rangle_{n-1}$  la media di  $n - 1$  campioni, aggiorniamo il valor medio con

$$\langle f \rangle_n = \frac{(n - 1) \langle f \rangle_{n-1} + f(x_n)}{n} = \langle f \rangle_{n-1} + \frac{f(x_n) - \langle f \rangle_{n-1}}{n}$$

## Campionare dove serve per ridurre la varianza

Per integrare una funzione che assume valori diversi da 0 su parti piccole del dominio di integrazione, il campionamento da una distribuzione uniforme presenta due inconvenienti seri: la convergenza si otterrebbe per valori di  $n$  spropositati e la varianza del risultato sarebbe inoltre molto grande. La Figura (3.2) a sinistra mostra che nel calcolo dell'integrale

$$\int_0^{10} e^{-2|x-5|} dx$$

il campionamento da una densità uniforme finirebbe con lo scegliere punti nei quali il valore della funzione integranda è molto piccolo. È immediato rendersi conto che in questo caso la convergenza rallenta e, conseguentemente, la varianza della stima aumenta.

Una possibilità per ovviare a questi difetti è fare uso di una densità  $f$  dalla quale si sappia campionare concentrata nelle regioni importanti del dominio di integrazione (regioni nelle quali la funzione integranda assume valori diversi da 0) e moltiplicare la funzione integranda per il reciproco di  $f$ . Nel caso di prima, vedi Figura (3.2) a destra, si potrebbe per esempio utilizzare la densità

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2}$$

e riscrivere l'integrale come

$$\int_0^{10} \frac{e^{-2|x-5|}}{f(x)} f(x) dx = \int_0^{10} e^{-2|x-5|} e^{(x-5)^2/2} \sqrt{2\pi} \frac{1}{\sqrt{2\pi}} e^{-(x-5)^2/2} dx.$$

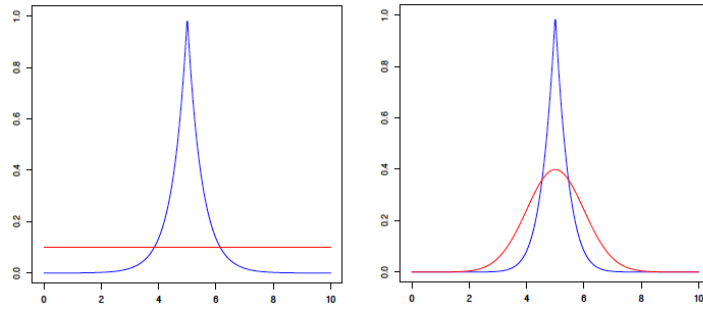


Figura 3.2: Vedi testo.

Il calcolo dell'integrale è del tutto simile al caso precedente con la differenza che i punti sono campionati da  $f(x)$  e la funzione integranda è modificata dividendo la funzione originale per  $f(x)$ .

Non sempre disponiamo di una procedura per campionare da una distribuzione di probabilità. Vediamo brevemente, allora, un metodo generale anche se non sempre efficiente per campionare da una distribuzione arbitraria.

### \* Accettazione e rifiuto

Supponiamo ora di dover campionare una variabile casuale  $X$  da una funzione di densità di probabilità  $f(x)$ . Non sappiamo campionare da  $f$  ma sappiamo campionare da una densità  $g(x)$  tale che, per qualche costante  $M$ ,  $f(x) \leq Mg(x)$ .

#### Algoritmo 3.20.1. *AcceptReject*

- 
1. campiona  $x$  dalla densità  $g$  e  $u$  uniformemente nell'intervallo  $(0, 1)$
  2. if  $u \leq f(x)/(Mg(x))$   
     return  $x$   
   else  
     goto 1.
- 

#### Osservazione 3.20.2. *Correttezza*

Verifichiamo che la funzione di densità da cui campioniamo con *AcceptReject* è proprio  $f(x)$ . Dobbiamo dimostrare che

$$\Pr\{X \leq a\} = \int_{-\infty}^a f(x)dx = F(a)$$

dove  $F(a)$  è la funzione cumulativa di probabilità. Supponiamo che le iterazioni necessarie a ottenere un campionamento accettato siano state  $N$ . Abbiamo

$$\Pr\{X \leq a\} = \Pr\left\{X \leq a \mid U \leq \frac{f(X)}{Mg(X)}\right\} = \frac{\Pr\left\{X \leq a, U \leq \frac{f(X)}{Mg(X)}\right\}}{\Pr\left\{U \leq \frac{f(X)}{Mg(X)}\right\}}.$$

Per quanto riguarda il numeratore l'indipendenza dei due eventi consente di scrivere la funzione di densità congiunta come

$$g(x)\mathbf{1}_{[0,1]}(u)$$



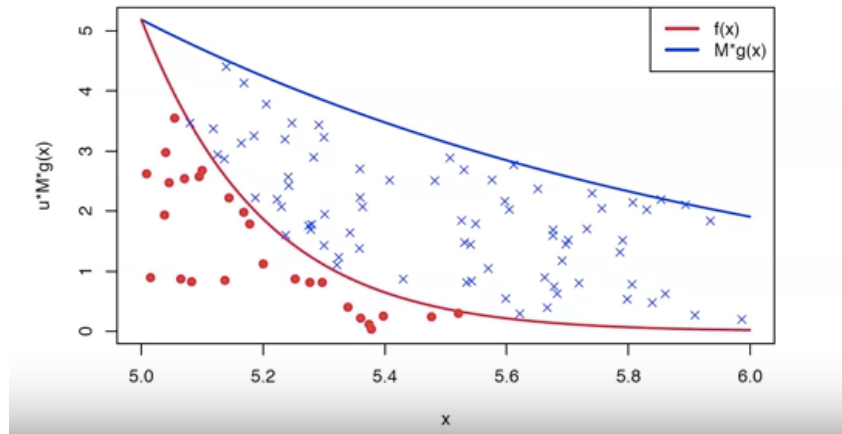


Figura 3.3: I campioni estratti da *AcceptReject* sono accettati solo se cadono sotto il grafico della densità  $f(x)$ . Quelli che cadono sopra, ovviamente sotto il grafico della funzione  $Mg(x)$ , sono invece rifiutati.

per cui abbiamo

$$\begin{aligned} \Pr \left\{ X \leq a, U \leq \frac{f(X)}{Mg(X)} \right\} &= \int_{-\infty}^a \left( \int_0^{f(x)/(Mg(x))} du \right) g(x) dx \\ &= \int_{-\infty}^a \frac{f(x)}{Mg(x)} g(x) dx = \frac{1}{M} \int_{-\infty}^a f(x) dx = \frac{1}{M} F(a). \end{aligned}$$

Pertanto, poiché  $F(a) \rightarrow 1$  per  $a \rightarrow +\infty$ , abbiamo

$$\Pr \left\{ U \leq \frac{f(X)}{Mg(X)} \right\} = \Pr \left\{ X \leq +\infty, U \leq \frac{f(X)}{Mg(X)} \right\} = \frac{1}{M}$$

e infine

$$\Pr\{X \leq a\} = F(a) \quad \blacksquare$$

### Osservazione 3.20.3. Importanza della costante

Il numero di iterazioni  $N$  segue una distribuzione geometrica con valore atteso  $M$ . Conseguentemente, se la costante  $M$  è troppo grande, *AcceptReject* tende a scartare la maggior parte dei campionamenti e perde in efficienza.

### Assicurati di ...

1. aver capito la relazione tra il *teorema del valor medio* e la *legge dei grandi numeri*;
2. aver capito come aumenta al crescere di  $n$  l'accuratezza della stima Monte Carlo di un valore atteso;
3. \* saper valutare l'effetto di una scelta della costante  $M$  troppo grande (anche ispirandosi alla Figura 3.3).

### 3.21 Catene di Markov

Introduciamo le catene di Markov discrete e omogenee nel tempo e con spazio degli stati finito. Anche se le variabili casuali non sono più indipendenti siamo comunque in grado di inferire proprietà di una catena a tempi lunghi sfruttando il campionamento da una distribuzione nota e la legge dei grandi numeri.

#### Concetti fondamentali

Data una sequenza di variabili casuali discrete  $S_t$  con  $t = 0, 1, \dots$ , un *processo di Markov* (MP) è costituito da una coppia  $(\mathcal{S}, \mathbf{P})$  dove

$\mathcal{S} = \{s_1, s_2, \dots, s_N\}$  è lo *spazio degli stati*, insieme dei possibili  $N$  valori che possono essere assunti dalle variabili casuali  $S_t$

$\mathbf{P} \in \mathbb{R}^{N \times N}$  è la *matrice di transizione* (TM) che esprime la probabilità di transizione allo stato  $s_j$  **al tempo**  $t + 1$  condizionata al fatto che il processo si trova **al tempo**  $t$  nello stato  $s_i$  con  $i, j = 1, \dots, N$ . Pertanto per l'elemento di riga  $i$  e colonna  $j$  scriviamo

$$(\mathbf{P})_{ij} = P_{ij} = \Pr(S_{t+1} = s_j | S_t = s_i) \quad (3.1)$$

con  $0 \leq P_{ij} \leq 1$  per ogni  $i$  e  $j$  e  $\sum_j P_{ij} = 1$  per ogni riga  $i$ .  $\square$

Restringiamoci al caso **omogeneo nel tempo** in cui gli elementi  $P_{ij}$  descrivono probabilità che **non dipendono dal tempo in cui avviene la transizione**. Possiamo riscrivere la (3.1) semplicemente come

$$P_{ij} = \Pr(s_j | s_i)$$

#### Definizione 3.21.1. Catena di Markov (MC)

Partendo da uno stato iniziale  $\bar{s} \in \mathcal{S}$  al tempo  $t = 0$ , una **catena di Markov** (MC) è una realizzazione del MP sottostante, ovvero una sequenza di stati per la quale se lo stato al tempo  $t$  è  $s_i$ , lo stato della catena al tempo  $t + 1$  è campionato dalla distribuzione di probabilità data dalla  $i$ -esima riga di  $\mathbf{P}$ .  $\square$

Omogenea o no nel tempo, una MC è senza memoria:  $P_{ij}$ , la probabilità di transizione allo stato  $s_j$  da  $s_i$ , non dipende dagli stati eventualmente occupati dalla MC nei tempi precedenti.

#### Esempio 3.21.1. Sole o pioggia?

Assumiamo che lo stato del tempo in una città sia descritto da *sole*,  $s_1$ , o *pioggia*,  $s_2$ , e che la TM sia

$$\mathbf{P} = \frac{1}{5} \begin{pmatrix} 3 & 2 \\ 1 & 4 \end{pmatrix}$$

Qual è la frazione di giorni di sole che possiamo aspettarci per  $t$  grande?

#### Esempio 3.21.2. Gioco d'azzardo

A ogni mano di un gioco  $G$  Alice può trovarsi in uno di cinque stati  $s_1, \dots, s_5$ . Nello stato  $s_i$  Alice ha un patrimonio di  $(i - 1)\$$  e da  $s_i$  con  $i = 2, 3$  e  $4$  o passa a  $s_{i+1}$  con probabilità  $p$  e vince  $1\$$ , o passa allo stato  $s_{i-1}$  e perde  $1\$$ . Una volta che si trova in  $s_1$  o  $s_5$ , invece, non succede più nulla: in  $s_1$  Alice è rovinata, mentre in  $s_5$  ha vinto definitivamente e ha un patrimonio di  $4\$$ . La TM  $\mathbf{P}$  è data da

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1-p & 0 & p & 0 & 0 \\ 0 & 1-p & 0 & p & 0 \\ 0 & 0 & 1-p & 0 & p \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Che cosa possiamo dire sul patrimonio di Alice per  $t$  grande?

**Esempio 3.21.3. Visita a caso in un grafo diretto e pesato**

In un cammino casuale in un grafo diretto pesato il vertice da visitare tra i vertici raggiunti dagli archi uscenti da un vertice origine è campionato con probabilità proporzionale al peso degli archi uscenti (come nel caso del grafo e della TM di Figura 3.4). *Google's Page Rank* è basato su un modello simile: internet è il grafo, i nodi le pagine e un arco connette  $A$  a  $B$  se  $A$  contiene un collegamento a  $B$ . Se il navigatore segue collegamenti a caso quale frazione di tempo trascorre su ogni pagina per  $t$  grande?

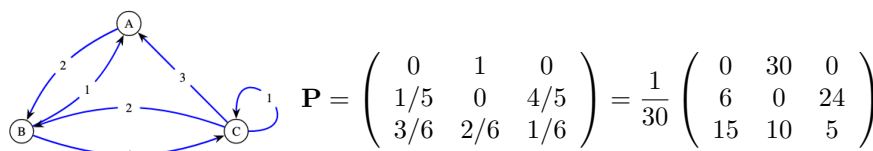


Figura 3.4: Vedi testo.

**Esempio 3.21.4. Saltatore ostinato**

Consideriamo infine un esempio utile per capire una proprietà importante di alcune MC: se la catena è nello stato  $s_1$  al tempo  $t$  sarà sempre nello stato  $s_2$  al tempo  $t + 1$  e viceversa, o

$$\mathbf{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

**Evoluzione di una catena di Markov**

Troviamo la distribuzione di probabilità dell'evoluzione nel tempo di una MC in termini di  $\mathbf{P}$ .

**Due passi:** per calcolare la probabilità che l'elemento  $t + 2$  di una MC sia  $s_j$  se l'elemento  $t$  è  $s_i$  dobbiamo sommare su tutti i possibili stati  $s_k$  in cui la MC può trovarsi al tempo  $t + 1$ , ovvero

$$\Pr(s_i \rightarrow s_j \text{ in 2 passi}) = \sum_k \Pr(s_j | s_k, s_i) \Pr(s_k | s_i) = \sum_k \Pr(s_j | s_k) \Pr(s_k | s_i) = \sum_k P_{kj} P_{ik} = (\mathbf{P}^2)_{ij}$$

dove  $\Pr(s_j | s_k, s_i) = \Pr(s_j | s_k)$  perché si tratta di una MC.

**Caso generale:** applicando ripetutamente i passaggi del punto precedente per  $t$  arbitrario, otteniamo l'equazione di Chapman-Kolmogorov, ovvero

$$\Pr(s_i \rightarrow s_j \text{ in } t \text{ passi}) = (\mathbf{P}^t)_{ij}$$

**Esercizio 3.21.1. Sempre una matrice di transizione**

Dimostra che  $\mathbf{P}^t$  è una TM per tutti i  $t \in \mathbb{N}$ .

*Dimostrazione:* se  $\mathbf{1} = (1 \ 1 \dots 1)^\top$ , abbiamo  $\mathbf{P}^t \mathbf{1} = \mathbf{P}^{t-1} \mathbf{P} \mathbf{1} = \mathbf{P}^{t-1} \mathbf{1} = \dots = \mathbf{P} \mathbf{1} = \mathbf{1}$ .  $\square$

Introduciamo ora due concetti utili per descrivere l'importante proprietà di raggiungibilità.

**Definizione 3.21.2. Irriducibilità**

Una TM  $\mathbf{P}$  è *irriducibile* se  $\forall s_i$  ed  $s_j$  con  $i, j = 1, \dots, N \exists t(i, j)$  tale che  $(\mathbf{P}^{t(i, j)})_{ij} > 0$

**Definizione 3.21.3. Regolarità**

Una TM  $\mathbf{P}$  è *regolare* se  $\forall s_i$  ed  $s_j$  con  $i, j = 1, \dots, N \exists \bar{t}$  tale che per  $\forall t > \bar{t}$  si ha  $(\mathbf{P}^t)_{ij} > 0$   $\square$

Intuitivamente, in una MC irriducibile per ogni coppia  $s_i$  ed  $s_j$ , se la MC si trova in  $s_i$ , prima o poi si troverà in  $s_j$ . Il tempo minimo di attesa potrebbe essere più lungo per alcune coppie ma sempre finito. Se la MC è regolare, esiste un tempo minimo di attesa uguale per tutte le coppie. Chiaramente una MC regolare è irriducibile, poiché è sufficiente porre  $\tau(i, j) = \bar{\tau}$  per tutti le coppie  $i$  e  $j$ .

**Esercizio 3.21.2.** Una MC regolare è irriducibile ma non viceversa

Determina se le MC degli esempi 3.21.1, 3.21.2, 3.21.3 e 3.21.4 sono regolari o irriducibili.

*Soluzione*

**3.21.1:** banalmente regolare, e quindi irriducibile, poiché  $\bar{\tau} = \tau(i, j) = 1$

**3.21.2:** non irriducibile poiché gli ultimi quattro elementi della prima riga di  $\mathbf{P}^t$  sono sempre nulli

**3.21.3:** regolare con  $\bar{\tau} = 3$  poiché

$$\mathbf{P}^2 = \frac{1}{180} \begin{pmatrix} 36 & 0 & 144 \\ 72 & 84 & 24 \\ 27 & 100 & 53 \end{pmatrix} \quad \text{e} \quad \mathbf{P}^3 = \frac{1}{5400} \begin{pmatrix} 2160 & 2520 & 720 \\ 864 & 2400 & 2136 \\ 1395 & 1340 & 2665 \end{pmatrix}$$

**3.21.4:** irriducibile poiché  $\tau(i, j) = 2$  if  $i = j$  and  $\tau(i, j) = 1$  if  $i \neq j$  ma non regolare poiché per le potenze dispari e pari di  $\mathbf{P}$  si ha

$$\mathbf{P}^{2t-1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{e} \quad \mathbf{P}^{2t} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

## Distribuzioni limite e stazionarie

**Esercizio 3.21.3.** Da una distribuzione di probabilità a un'altra

Un vettore  $\mathbf{p}$  in  $N$  dimensioni con componenti non negativi che sommano a 1 è una distribuzione di probabilità per gli stati  $s_1, \dots, s_N$  ed è noto come vettore *stocastico*. Dimostra che se  $\mathbf{p}$  è un vettore stocastico e  $\mathbf{P}$  una TM il vettore  $\mathbf{q}$  ottenuto come  $\mathbf{q}^\top = \mathbf{p}^\top \mathbf{P}$  è stocastico.

*Soluzione*

Per i vincoli di non negatività degli elementi di  $\mathbf{p}$ , per  $j = 1, \dots, N$ , abbiamo  $q_j = \sum_i p_i P_{ij} \geq 0$ . Inoltre, poiché  $\sum_j p_j = 1$  e  $\sum_j P_{ij} = 1$  per  $i = 1, \dots, N$ , otteniamo

$$\sum_j q_j = \sum_j \sum_i p_i P_{ij} = \sum_i p_i \left( \sum_j P_{ij} \right) = \sum_i p_i = 1$$

**Definizione 3.21.4.** Distribuzione stazionaria

Una distribuzione di probabilità  $\pi$  con  $\pi_i \geq 0$  per ogni  $s_i$  con  $i = 1, \dots, N$  e  $\sum_i \pi_i = 1$  è *stazionaria* per una MC con TM  $\mathbf{P}$  se

$$\pi^\top = \pi^\top \mathbf{P}, \quad \text{ovvero} \quad \pi_j = \sum_i \pi_i P_{ij} \quad j = 1, \dots, N \quad \square$$

Poiché  $\mathbf{P}\mathbf{1} = \mathbf{1}$ , 1 è sempre un autovalore di  $\mathbf{P}$ . È facile dimostrare che è anche il più grande. Se  $\lambda > 1$  fosse un autovalore e  $\mathbf{x}$  il corrispondente autovettore, infatti, avremmo  $\mathbf{P}^n \mathbf{x} = \lambda^n \mathbf{x}$ . Ma per valori crescenti di  $n$  questo porterebbe a una TM  $\mathbf{P}^n$  con elementi maggiori di 1!

Enunciamo ora un importante teorema basato sul teorema del punto fisso di Brouwer.

**Teorema 3.21.1.** *Perron Frobenius*

Se una TM  $\mathbf{P}$  è regolare, l'autovettore sinistro  $\pi$  corrispondente all'autovalore 1 è unico e con tutti gli elementi strettamente positivi.  $\square$

Una distribuzione stazionaria  $\pi$  è quindi una soluzione del sistema lineare

$$\pi^\top \mathbf{P} = \pi^\top$$

**Esercizio 3.21.4. Ancora i quattro esempi**

Trova le distribuzioni stazionarie per gli esempi 3.21.1, 3.21.2, 3.21.3 e 3.21.4.

*Soluzione*

$$\pi_{sr} = \left( \frac{1}{3} \quad \frac{2}{3} \right)^\top, \quad \pi_l = (1 \ 0 \ 0 \ 0)^\top \text{ e } \pi_w = (0 \ 0 \ 0 \ 1)^\top, \quad \pi_{rw} = \left( \frac{17}{66} \quad \frac{25}{66} \quad \frac{24}{66} \right)^\top, \quad \pi_{sj} = \left( \frac{1}{2} \quad \frac{1}{2} \right)^\top$$

**Definizione 3.21.5. Distribuzione limite**

Un vettore  $\lambda$  con  $\lambda_j \geq 0$  per  $j = 1, \dots, N$  e  $\sum_j \lambda_j = 1$  è una *distribuzione limite* per  $P$  se

$$\lim_{t \rightarrow \infty} (P^t)_{ij} = \lambda_j \quad \text{per } i = 1, \dots, N$$

Se una MC ammette una distribuzione limite ed è inizializzata da  $\lambda$ , a ogni passo l'evoluzione della MC sarà sempre determinata da  $\lambda$ .  $\square$

Quando una distribuzione limite esiste,  $P^t$  converge alla matrice con tutte le righe uguali a  $\lambda$ , ovvero

$$P^t \xrightarrow{t \rightarrow \infty} \begin{pmatrix} \lambda_1 & \lambda_2 & \dots & \lambda_N \\ \lambda_1 & \lambda_2 & \dots & \lambda_N \\ \lambda_1 & \lambda_2 & \dots & \lambda_N \end{pmatrix}$$

In questo caso, la probabilità di transizione allo stato  $s_j$  for  $j = 1, \dots, N$  al tempo  $t + 1$ , per  $t$  sufficientemente grande, non dipende dallo stato della MC al tempo  $t = 0$ .

Per una MC regolare l'esistenza della distribuzione limite è assicurata da un teorema - basato sulla legge dei grandi numeri - che enunciamo ma non dimostriamo.

**Teorema 3.21.2. Senza via di fuga**

Una MC regolare ha un'unica distribuzione limite  $\lambda$  with  $\lambda_j > 0$  for  $j = 1, \dots, N$ . Inoltre, poiché

$$\sum_i \lambda_i P_{ij} = \sum_i \lim_{t \rightarrow \infty} (P^t)_{ki} P_{ij} = \lim_{t \rightarrow \infty} \sum_i (P^t)_{ki} P_{ij} = \lim_{t \rightarrow \infty} (P^{t+1})_{kj} = \lambda_j$$

$\lambda$  è anche l'unica distribuzione stazionaria per  $P$ .

**Esercizio 3.21.5. Un'ultima volta i quattro esempi**

Discuti l'esistenza della distribuzione limite per i nostri quattro esempi.

*Soluzione*

**3.21.1:** la MC è regolare e  $\pi_{sr}$ , quindi, è anche la distribuzione limite. Per  $t$  grandi, e indipendentemente dallo stato iniziale, possiamo attenderci 1/3 di giorni di sole e 2/3 di pioggia.

**3.21.2:** non può avere una distribuzione limite poiché  $\pi_l$  e  $\pi_w$ , che corrispondono a due stati *assorbenti*, sono stazionari. Per  $t$  grandi, la MC si troverà in uno dei due stati assorbenti: Alice, quindi, o sarà in rovina o avrà sbancato. Se il banco fosse infinitamente ricco, Alice finirebbe sempre in rovina e otteniamo un caso particolare del *teorema della rovina del giocatore*.

**3.21.3:** la MC è regolare e  $\pi_{rw}$  è anche la distribuzione limite che fornisce una misura del tempo atteso trascorso su ogni pagina.

**3.21.4:** nessuna distribuzione limite. La MC oscilla indefinitamente tra i due stati.

**Assicurati di ...**

1. saper enunciare la definizione di processo di Markov
2. sapere come calcolare la probabilità di transizione dallo stato  $s_i$  allo stato  $s_j$  in  $n$  passi
3. conoscere sotto quali ipotesi esiste una distribuzione limite e come trovare una distribuzione stazionaria

### 3.22 \* Catene di Markov *Monte Carlo*

Le catene di Markov *Monte Carlo* sono metodi utilizzati per simulare l'estrazione di una sequenza di campioni da una distribuzione di probabilità arbitraria. La sequenza ottenuta costituisce uno strumento computazionale molto utile sia per approssimare una distribuzione di probabilità sia per stimare valori attesi in alte dimensioni.

#### Bilancio dettagliato

Consideriamo un'ulteriore proprietà di cui può godere una MC.

**Definizione 3.22.1.** *Bilancio dettagliato*

Una MC con TM  $\mathbf{P}$  è reversibile se esiste una distribuzione di probabilità  $\pi$  tale che

$$\pi_j P_{ji} = \pi_i P_{ij} \quad \forall i, j \quad (3.2)$$

**Osservazione 3.22.1.** *Stazionarietà garantita*

La distribuzione  $\pi$  dell'equazione (3.2) è stazionaria per  $\mathbf{P}$ . Infatti sommando su  $j$  entrambi i membri e tenendo conto che  $\sum_j P_{ij} = 1$ , si ottiene

$$\sum_j \pi_j P_{ji} = \sum_j \pi_i P_{ij} = \pi_i \sum_j P_{ij} = \pi_i$$

#### Metropolis-Hastings

La reversibilità espressa dall'equazione (3.2), ovvero il fatto che la probabilità di passare allo stato  $i$  dallo stato  $j$  è la stessa di passare dallo stato  $j$  allo stato  $i$ , è alla base dell'algoritmo di *Metropolis-Hastings*.

**Algoritmo 3.22.1.** *Metropolis-Hastings*

*Input:*  $\pi$  distribuzione di probabilità discreta per  $N$  stati,  $\mathbf{H}$  matrice di transizione regolare di dimensione  $N \times N$  utilizzabile per proporre un nuovo stato,  $\bar{s}$  stato iniziale della catena arbitrario e  $T \in \mathbb{N}$  grande  
*Output:* Catena di Markov  $X_t$  con matrice di transizione  $\mathbf{P}$  e distribuzione limite  $\pi$

---

for  $t = 0$  to  $T$

1. proponi la transizione allo stato  $s_j$  campionato con probabilità  $\Pr(X_{t+1} = s_j | X_t = s_i) = H_{ij}$

2. calcola

$$a_{ij} = \min \left\{ 1, \frac{\pi_j H_{ji}}{\pi_i H_{ij}} \right\} \quad (3.3)$$

3. campiona un numero  $U$  in  $[0, 1]$  con probabilità uniforme

4. if  $U \leq a_{ij}$

    then  $X_{t+1} = s_j$

    else  $X_{t+1} = s_i$

La matrice  $\mathbf{P}$

$$P_{ij} = \begin{cases} H_{ij} a_{ij} & (i \neq j) \\ 1 - \sum_{i \neq j} H_{ij} a_{ij} & (i = j) \end{cases} \quad (3.4)$$

è la matrice di transizione della catena  $X_t$  e ha per distribuzione limite  $\pi$ .

---

**Correttezza** Poiché  $\mathbf{H}$  è regolare anche la matrice  $\mathbf{P}$  è regolare. Sostituendo inoltre l'equazione (3.3) nell'equazione (3.4), se  $\pi_j H_{ji} < \pi_i H_{ij}$  otteniamo

$$P_{ij} = H_{ij} \min \left\{ 1, \frac{\pi_j H_{ji}}{\pi_i H_{ij}} \right\} = \frac{\pi_j H_{ji}}{\pi_i} \quad \text{e} \quad P_{ji} = H_{ji} \min \left\{ 1, \frac{\pi_i H_{ij}}{\pi_j H_{ji}} \right\} = H_{ji}$$

da cui ricaviamo

$$\pi_i P_{ij} = \pi_i \frac{\pi_j}{\pi_i} H_{ji} = \pi_j H_{ji} = \pi_j P_{ji}$$

È facile verificare che per  $\pi_j H_{ji} > \pi_i H_{ij}$  si ottiene nuovamente l'equazione (3.2). In entrambi i casi, quindi, la matrice regolare  $\mathbf{P}$  soddisfa l'equazione del bilancio dettagliato e la distribuzione di probabilità  $\pi$  è la distribuzione limite e stazionaria di  $\mathbf{P}$ .

### **Osservazione 3.22.2.** *Matrice $H$ e probabilità di accettazione della proposta*

La scelta della matrice  $H$  deve ricadere su una distribuzione di probabilità dalla quale sia possibile estrarre campioni a caso. Se  $H$  è simmetrica,  $a_{ij}$  - nota come *probabilità di accettazione della proposta del nuovo stato* - dipende solo dal rapporto  $\pi_j/\pi_i$ .

### **Osservazione 3.22.3.** *Conoscenza parziale della distribuzione $\pi$*

La dipendenza di  $a_{ij}$  dal rapporto  $\pi_j/\pi_i$  rende possibile utilizzare il *Metropolis-Hastings* anche nel caso in cui la distribuzione di probabilità  $\pi$  sia nota a meno di una costante di proporzionalità. Questa proprietà ha importanti ripercussioni nell'ambito della statistica bayesiana.

### **Osservazione 3.22.4.** *Convergenza*

Per raggiungere il regime asintotico è necessario scartare i primi elementi della catena, operazione nota come *burn-in*, in modo tale da ottenere una sequenza che non dipende dalla scelta dello stato iniziale.

### **Osservazione 3.22.5.** *Dipendenza*

Un limite particolarmente pronunciato del *Metropolis-Hastings*, e che affligge le catene di Markov *Monte Carlo* ottenute anche con altri algoritmi, è quello di produrre sequenze di campioni correlati. Al fine di ottenere campioni indipendenti spesso si sottocampiona la sequenza.

## **Modello di Ising in 2D**

Consideriamo un esempio dal mondo fisico. Un modello di Ising in 2D è un sistema di  $A$  atomi disposti su un reticolo quadrato regolare che useremo come semplice modello di un materiale ferromagnetico.

### **Descrizione del modello**

Ogni atomo  $a$  con  $a = 1, \dots, A$  può trovarsi con il proprio spin  $\sigma_a$  *up* ( $\sigma_a = 1$ ) o *down* ( $\sigma_a = -1$ ). Per ogni configurazione degli  $A$  spin  $\sigma = (\sigma_1, \dots, \sigma_A)$  o *stato* del sistema, l'energia è

$$E(\sigma) = E(\sigma_1, \dots, \sigma_A) = -\frac{1}{2} \sum_{a=1}^A \sum_{t \in I(a)} \sigma_a \sigma_t \quad (3.5)$$

dove  $I(a)$  è l'insieme dei 4 atomi primi vicini di  $a$  sul reticolo. Poiché se  $t \in I(a)$  allora  $a \in I(t)$ , il fattore  $1/2$  evita che il contributo di una coppia di spin vicini sia contata due volte. Fissata la temperatura, **all'equilibrio il sistema si troverà in uno stato a energia minima**. Dalla fisica sappiamo che se  $\beta = 1/T$  è un parametro inversamente proporzionale alla temperatura ed  $\mathcal{S}$  lo spazio dei possibili  $2^A$  stati  $\sigma$ , la distribuzione stazionaria di probabilità degli stati è

$$\pi(\sigma) = \frac{e^{-\beta E(\sigma)}}{Z} \quad \text{con} \quad Z = \sum_{\sigma \in \mathcal{S}} e^{-\beta E(\sigma)} \quad (3.6)$$

A temperatura fissata, pertanto, stati con più spin adiacenti allineati hanno energia minore e sono più probabili. Al crescere della temperatura le differenze in energia si assottigliano e spin adiacenti possono più facilmente assumere valori opposti.

La magnetizzazione del reticolo,

$$M(\sigma) = \frac{1}{A} \sum_{a=1}^A \sigma_a$$

misura la differenza tra la frazione di spin *up* e spin *down*. Dall'esperienza sappiamo che a basse temperature un materiale ferromagnetico può mostrare una magnetizzazione permanente importante. A temperature più alte, tuttavia, lo stesso materiale perde questa proprietà: una calamita dimenticata vicino a un calorifero si smagnetizza. La fisica ci dice che questa transizione avviene in modo brusco a una temperatura ben precisa, nota come *temperatura critica*  $T^*$ . Nel caso di reticolo bidimensionale per  $A \rightarrow \infty$  è possibile ottenere che la transizione avviene per un valore della temperatura

$$T^* = \frac{2}{\ln(1 + \sqrt{2})} \approx 2.26$$

Per verificare in che misura il modello di Ising è in accordo con l'esperienza vogliamo stimare

$$\mathbb{E}[|M|] = \frac{1}{Z} \sum_{\sigma \in \mathcal{S}} e^{-\beta E(\sigma)} |M(\sigma)| = \frac{1}{AZ} \sum_{\sigma \in \mathcal{S}} e^{-\beta E(\sigma)} \left| \sum_{a=1}^A \sigma_a \right|$$

al variare del parametro  $\beta$  **all'equilibrio**.

#### **Osservazione 3.22.6.** *Dimensioni sproporzionate*

Anche limitandoci a un sistema costituito da soli  $32 \times 32 = 1024$  atomi lo spazio degli stati conta  $N = 2^{1024}$  elementi. La matrice di transizione  $\mathbf{P} \in \mathbb{R}^{2^{1024}} \times \mathbb{R}^{2^{1024}}$  e la distribuzione  $\pi \in \mathbb{R}^{2^{1024}}$  appartengono entrambe a spazi dimensionalmente e computazionalmente intrattabili.

#### **Stima del valore atteso della magnetizzazione**

Supponiamo che il sistema si trovi in un stato  $\sigma$  e riscriviamo la funzione energia dell'equazione (3.5) isolando i termini che dipendono dallo spin di un particolare atomo  $s$ . Avremo

$$E(\sigma_1, \dots, \sigma_s, \dots, \sigma_A) = -\sigma_s \sum_{t \in I(s)} \sigma_t - \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t \quad (3.7)$$

dove nella prima somma il fattore  $1/2$  scompare per via del fatto che ogni coppia di primi vicini di  $s$  compare due volte. Per la differenza in energia  $\Delta E$  tra due stati che differiscono tra loro solo per la variazione dello spin di  $s$  da  $\sigma_s$  a  $-\sigma_s$  avremo, utilizzando l'equazione (3.7) due volte,

$$\begin{aligned} \Delta E &= E(\sigma_1, \dots, -\sigma_s, \dots, \sigma_A) - E(\sigma_1, \dots, \sigma_s, \dots, \sigma_A) \\ &= -(-\sigma_s) \sum_{t \in I(s)} \sigma_t - \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t + \sigma_s \sum_{t \in I(s)} \sigma_t + \frac{1}{2} \sum_{a \neq s} \sum_{t \in I(a) \setminus s} \sigma_a \sigma_t \\ &= 2\sigma_s \sum_{t \in I(s)} \sigma_t \end{aligned}$$

Per l'additività della funzione energia, quindi,  $\Delta E$  dipende solo dall'orientamento dello spin dell'atomo  $s$  e da quello dei suoi primi vicini. Per la forma esponenziale della distribuzione di probabilità  $\pi$ , inoltre, anche il rapporto delle probabilità dipende solo da  $\Delta E$  ovvero

$$\frac{\pi(\sigma_1, \dots, -\sigma_s, \dots, \sigma_A)}{\pi(\sigma_1, \dots, \sigma_s, \dots, \sigma_A)} = e^{-\beta \Delta E} \quad (3.8)$$



h

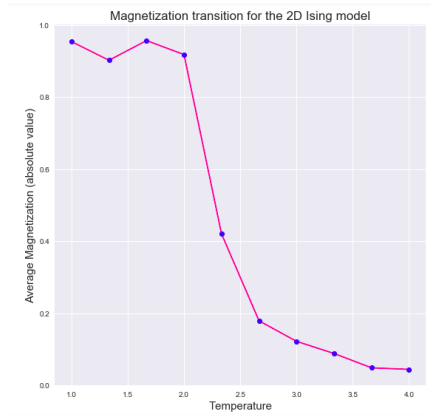


Figura 3.5: Stima di  $\mathbb{E}[|M|]$  ottenuta su un reticolo  $32 \times 32$  con *burn-in* di  $10^7$  e media empirica calcolata su 10 catene di  $10^5$  passi. Nello stato iniziale ogni spin è *up* o *down* con probabilità  $1/2$ .

Abbiamo ora tutti gli elementi per stimare  $\mathbb{E}[|M|]$  in funzione di  $\beta$ . La distribuzione di probabilità  $\pi$  per gli  $N = 2^A$  stati è data dall'equazione (3.6).

**Osservazione 3.22.7.** *Matrice regolare per la proposta di un nuovo stato*

Per ogni stato  $i$ ,  $\mathbf{H}$  propone la transizione a uno degli  $A$  stati che differiscono da  $i$  per la variazione di uno solo degli  $A$  spin con probabilità uniforme e uguale a  $1/A$ .

**Osservazione 3.22.8.** *Regolarità di  $\mathbf{H}$*

La drastica riduzione del numero degli stati verso i quali si può muovere la catena, da  $2^A$  ad  $A$ , è sufficiente a rendere la simulazione computazionalmente possibile. Pur consentendo transizioni verso una frazione molto piccola degli stati possibili la regolarità di  $\mathbf{H}$  non è in discussione: modificando uno spin alla volta, infatti, è evidente che è possibile raggiungere qualunque stato  $j$  indipendentemente dallo stato iniziale  $i$  con un numero di passi uguale al numero di spin il cui orientamento nello stato  $j$  è diverso rispetto allo stato  $i$ . Il calcolo di  $a_{ij}$  nell'equazione (3.3), infine, è basato sull'equazione (3.8).

**Esercizio 3.22.1.** *Verifica dell'esistenza di un punto critico*

Prova a verificare la capacità del modello di Ising di catturare la brusca variazione nel valore atteso della magnetizzazione al variare di  $\beta$ . Dovresti ottenere un grafico simile a quello di Figura 3.5.

### 3.23 Inferenza Bayesiana

Consideriamo ora un semplice esercizio che chiarisce come l'approccio bayesiano si fondi sull'accumulo di evidenze basate sui dati. Partiamo assumendo di conoscere le distribuzioni di probabilità *a priori* e le verosimiglianze.

#### Ipotesi a confronto

##### Esercizio 3.23.1. Quale tipo di moneta?

Sono dati tre tipi di monete. Le monete di tipo *A* per le quali la probabilità di ottenere *testa*, *T*, con un lancio è  $1/2$ , quelle di tipo *B* per le quali è  $3/5$  e quelle di tipo *C* per le quali è  $9/10$ . Un cassetto contiene due monete di tipo *A*, una di tipo *B* e una di tipo *C*. Pesco una moneta a caso. Qual è la probabilità che la moneta sia di tipo *A*, *B* o *C*? Se lanciando la moneta una prima volta ottengo *testa*, come cambiano le probabilità?

Per ottenere la risposta alla prima domanda basta applicare le probabilità *a priori*. Per la seconda, invece, il teorema di Bayes. Se indichiamo con  $T_1$  il risultato *testa* al primo lancio il teorema di Bayes ci dice che la probabilità *a posteriori* che la moneta sia di tipo *A* si ottiene come

$$\Pr(A|T_1) = \frac{\Pr(A)\Pr(T_1|A)}{\Pr(A)\Pr(T_1|A) + \Pr(B)\Pr(T_1|B) + \Pr(C)\Pr(T_1|C)}$$

e, similmente, se di tipo *B* o *C*. La tabella qui sotto riporta le ipotesi, le probabilità *a priori*, le verosimiglianze e le probabilità *a posteriori* non normalizzate e normalizzate.

<i>H</i>	$\Pr(H)$	$\Pr(T_1 H)$	$\Pr(T_1 H)\Pr(H)$	$\Pr(H T_1)$
<i>A</i>	0.5	0.5	0.25	0.4
<i>B</i>	0.25	0.6	0.15	0.24
<i>C</i>	0.25	0.9	0.225	0.36
totale	1		0.625	1

La forza dell'inferenza bayesiana risiede nel fatto che lo schema appena visto può essere iterato nel caso si accumuli ulteriore evidenza. Per esempio: come si modificano le probabilità se lanciando una seconda volta la moneta ottenessimo nuovamente *testa*? Per rispondere a questa domanda ripetiamo l'aggiornamento delle probabilità *a posteriori* non normalizzate moltiplicandole per le verosimiglianze (che non sono cambiate) e imponiamo la normalizzazione all'ultimo passo. I risultati sono mostrati nella tabella qui sotto.

<i>H</i>	$\Pr(T_1 H)\Pr(H)$	$\Pr(T_1 H)\Pr(T_2 H)\Pr(H)$	$\Pr(H T_1T_2)$
<i>A</i>	0.25	0.125	0.299
<i>B</i>	0.15	0.09	0.216
<i>C</i>	0.225	0.2025	0.485
totale	0.625	0.41750	1

#### Aggiornamento delle probabilità predittive

Riconsideriamo lo schema descritto nell'**Esercizio 3.23.1** e poniamoci il problema di stimare le probabilità di ottenere *testa* nelle tre condizioni, ovvero prima di effettuare lanci, dopo aver ottenuto *testa* con un primo lancio,  $T_1$ , e dopo aver ottenuto *testa* anche con un secondo lancio,  $T_2$ .

Prima di effettuare lanci, la probabilità di ottenere *testa* con un primo lancio è

$$\begin{aligned}\Pr(T_1) &= \Pr(T_1|A)\Pr(A) + \Pr(T_1|B)\Pr(B) + \Pr(T_1|C)\Pr(C) \\ &= 0.5 \times 0.5 + 0.25 \times 0.6 + 0.25 \times 0.9 = 0.625\end{aligned}$$

Osserviamo che nessuna delle monete restituisce *testa* con questa probabilità! Se l'esito del primo lancio è *testa*, invece, la probabilità di ottenere *testa* con un secondo lancio è

$$\begin{aligned}\Pr(T_2|T_1) &= \Pr(T_2|A)\Pr(A|T_1) + \Pr(T_2|B)\Pr(B|T_1) + \Pr(T_2|C)\Pr(C|T_1) \\ &= 0.5 \times 0.4 + 0.6 \times 0.24 + 0.9 \times 0.36 = 0.668\end{aligned}$$

**Osservazione 3.23.1.** *Probabilità predittive*

Sia  $\Pr(T_1)$  sia  $\Pr(T_2|T_1)$  esprimono una predizione. Nel primo caso,  $\Pr(T_1)$  è una probabilità *predittiva a priori*. Nel secondo,  $\Pr(T_2|T_1)$  è una probabilità *predittiva a posteriori*. Notiamo che il risultato del primo lancio aumenta la probabilità di ottenere *testa* in un secondo lancio. In entrambi i casi riscriviamo  $\Pr(T_1)$  e  $\Pr(T_2|T_1)$  come valore atteso delle probabilità condizionate alla scelta del tipo di moneta di ottenere *testa* al primo e al secondo lancio,  $\Pr(T_1|\cdot)$  e  $\Pr(T_2|\cdot)$ , rispetto alla probabilità di aver scelto quel tipo di moneta,  $\Pr(\cdot)$  e  $\Pr(\cdot|T_1)$ .  $\square$

Nell'approccio bayesiano l'acquisizione di osservazioni modifica la distribuzione delle probabilità *a priori* sia delle ipotesi sia dei risultati, trasformandole da probabilità *a priori* a probabilità *a posteriori*. Dopo ogni acquisizione le probabilità *a posteriori* diventano le nuove probabilità *a priori*.

## Parametri di una distribuzione di probabilità come variabili casuali

Consideriamo ora il caso nel quale la distribuzione di probabilità è nota nella forma parametrica ma non nei valori dei parametri. Vogliamo inferire la migliore stima dei parametri avendo a disposizione dati campionati dalla distribuzione. Nell'inferenza *bayesiana* sia i dati sia i parametri che definiscono la distribuzione di probabilità sono variabili casuali. Si ipotizza una distribuzione di probabilità *a priori* per i parametri, nota **prima** di aver acquisito i dati, e si modifica in una *a posteriori*, **dopo** aver acquisito i dati. Dal teorema di Bayes sappiamo che la distribuzione di probabilità *a posteriori* è proporzionale al prodotto della probabilità *a priori* con la verosimiglianza.

Ci restringiamo al caso in cui la distribuzione di probabilità è normale. Abbiamo quindi a disposizione  $n$  realizzazioni delle variabili  $X_i$  con  $i = 1, \dots, n$  identicamente e indipendentemente distribuite secondo la distribuzione  $\mathcal{N}(\mu_0, \sigma_0^2)$ . Per semplicità ci restringiamo al caso in cui  $\mu_0$  è ignoto, mentre  $\sigma_0^2$  è noto.

Supponiamo infine di sapere che il valore ignoto  $\mu_0$  sia la realizzazione di una **una variabile casuale**  $\mu$  distribuita secondo una distribuzione normale  $\mathcal{N}(\mu_{pr}, \sigma_{pr}^2)$ , ovvero

$$\Pr(\mu) = \frac{1}{\sqrt{2\pi\sigma_{pr}^2}} e^{-(\mu_{pr}-\mu)^2/2\sigma_{pr}^2}$$

Dimostriamo ora che, sotto queste condizioni, anche la distribuzione di probabilità *a posteriori* del parametro  $\mu$  è normale. Per la verosimiglianza, ovvero per la probabilità con la quale ci aspettiamo di ottenere i dati che abbiamo effettivamente osservato in funzione del valore di  $\mu$ , abbiamo

$$L(\mathbf{x}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i-\mu)^2/2\sigma_0^2}$$

con  $\mathbf{x} = (x_1 \dots x_n)$ . Da Bayes sappiamo che la probabilità *a posteriori* è proporzionale alla verosimiglianza moltiplicata per la probabilità *a priori*.

Applicando il logaritmo alla distribuzione *a posteriori*

$$\Pr(\mu|\mathbf{x}) = \frac{\Pr(\mu)L(\mathbf{x}|\mu)}{\Pr(\mathbf{x})}$$

si ottiene

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_{pr}^2} (\mu_{pr} - \mu)^2 + a$$

dove la costante  $a$  raggruppa tutti i termini che non dipendono da  $\mu$ , ovvero

$$a = -n \ln \sqrt{2\pi} - \ln \sqrt{2\pi} - \ln \Pr(\mathbf{x})$$

Sviluppando i quadrati e ponendo nuovamente  $\hat{\mu} = \sum_{i=1}^n x_i/n$ , si ha

$$\begin{aligned} \ln \Pr(\mu|\mathbf{x}) &= -\frac{1}{2} \left( \frac{\sum_{i=1}^n x_i^2 + n\mu^2 - 2n\hat{\mu}\mu}{\sigma_0^2} + \frac{\mu_{pr}^2 + \mu^2 - 2\mu_{pr}\mu}{\sigma_{pr}^2} \right) + a \\ &= -\frac{1}{2} \left( \frac{(\sigma_0^2 + n\sigma_{pr}^2)\mu^2 - 2(\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu})\mu}{\sigma_0^2\sigma_{pr}^2} \right) + b \end{aligned} \quad (3.9)$$

dove  $b$  riassume tutti i termini che non dipendono da  $\mu$ , ovvero

$$b = a - \frac{1}{2} \frac{\sum_{i=1}^n x_i^2}{\sigma_0^2} - \frac{1}{2} \frac{\mu_{pr}^2}{\sigma_{pr}^2}$$

L'equazione (3.9) può essere riscritta come

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{\sigma_0^2 + n\sigma_{pr}^2}{2\sigma_0^2\sigma_{pr}^2} \left( \mu^2 - 2\frac{\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu}}{\sigma_0^2 + n\sigma_{pr}^2} \mu \right) + b$$

Ponendo

$$\mu_{post} = \frac{\sigma_0^2\mu_{pr} + \sigma_{pr}^2 n\hat{\mu}}{\sigma_0^2 + n\sigma_{pr}^2} \quad \text{e} \quad \sigma_{post}^2 = \frac{\sigma_0^2\sigma_{pr}^2}{\sigma_0^2 + n\sigma_{pr}^2}$$

e completando i quadrati otteniamo allora

$$\ln \Pr(\mu|\mathbf{x}) = -\frac{1}{2\sigma_{post}^2} (\mu^2 - 2\mu_{post}\mu + \mu_{post}^2) + \frac{1}{2} \frac{\mu_{post}^2}{\sigma_{post}^2} + b$$

Raggruppando nuovamente tutti i termini che non dipendono da  $\mu$  in una sola costante

$$c = \frac{1}{2} \frac{\mu_{post}^2}{\sigma_{post}^2} + b$$

e ripristinando l'esponenziale, otteniamo

$$\Pr(\mu|\mathbf{x}) = d e^{-\frac{(\mu - \mu_{post})^2}{2\sigma_{post}^2}} = \mathcal{N}(\mu_{post}, \sigma_{post}^2)$$

con  $d = e^c$  che non dipende da  $\mu$ .

Basandoci sull'acquisizione delle  $n$  osservazioni  $\mathbf{x} = (x_1 \dots x_n)$  modifichiamo la distribuzione di probabilità *a priori* in quella *a posteriori*

$$\mathcal{N}(\mu_{pr}, \sigma_{pr}^2) \rightarrow \mathcal{N}(\mu_{post}, \sigma_{post}^2)$$

Come ci aspettiamo dalla legge dei grandi numeri, per  $n$  grande  $\mu_{post} \approx \hat{\mu} \approx \mu_0$  con  $\sigma_{post}^2 \approx \sigma_0^2/n$ . Ovvero, al crescere di  $n$ , la distribuzione *a posteriori* si stringe attorno al valore  $\mu_0$ .

### Compito 3.23.1. Confronto

Fissa  $\sigma_0^2 = 1$  e campiona  $\mu_0$  dalla distribuzione normale  $\mathcal{N}(\mu_{pr}, \sigma_{pr}^2)$ . Campiona  $n$  punti  $x_i$  (con  $n = 2, 10$  e  $20$ ) dalla distribuzione normale  $\mathcal{N}(\mu_0, 1)$  e confronta  $\hat{\mu}$  e  $\mu_{post}$  con  $\mu_0$  al variare di  $n$  ripetendo l'esperimento 100 volte.

### **Assicurati di ...**

1. saper scrivere il teorema di Bayes e individuare in un problema le verosimiglianze, le probabilità *a priori*, la probabilità totale e le probabilità *a posteriori*;
2. aver capito come nell'inferenza Bayesiana le probabilità *a posteriori* diventino le nuove probabilità *a priori* in presenza di nuovi dati.

## 3.24 Inferenza frequentista

Il secondo principio di inferenza che incontriamo è il *principio di massima verosimiglianza*.

### Principio di massima verosimiglianza

Disponiamo di  $n$  dati  $\mathbf{x} = (x_1 \ x_2 \ \dots \ x_n)$  realizzazioni di  $n$  variabili casuali  $X_i$  *indipendenti e identicamente distribuite, iid*, con  $i = 1, 2, \dots, n$ . Supponiamo nuovamente di sapere che la funzione di distribuzione sottostante  $f$ , funzione di *probabilità di massa* nel caso discreto e di *densità* nel caso continuo, dipende da un parametro  $\theta$  **che non conosciamo**. Scriviamo pertanto  $f$  come  $f_\theta$ . Vogliamo *inferire* dagli  $n$  dati il valore di  $\theta$  che determina la distribuzione che ha effettivamente generato i dati. Indichiamo questo valore con  $\theta_0$  e con  $f_{\theta_0}$  la distribuzione corrispondente. Tipicamente il parametro  $\theta$  è reale e appartiene all'insieme  $\Theta \subseteq \mathbb{R}$ , ma più in generale  $\theta$  potrebbe essere un vettore di parametri.

Una possibile strategia è calcolare la verosimiglianza

$$L(\mathbf{x}|\theta) = \prod_{i=1}^n f_\theta(x_i)$$

ovvero la probabilità con la quale ci aspettiamo di ottenere i dati che abbiamo effettivamente osservato in funzione del parametro  $\theta$ .

#### Osservazione 3.24.1. Verosimiglianza o probabilità

La verosimiglianza è una *probabilità* per ogni valore di  $\theta$  ma *non è una probabilità* come funzione di  $\theta$ . Quando scriviamo  $L(\mathbf{x}|\theta)$ , pertanto, non stiamo studiando probabilità di eventi diversi ma quanto sia verosimile il verificarsi di un evento fissato al variare della distribuzione sottostante.  $\square$

Ci aspettiamo che per valori di  $\theta$  vicini  $\theta_0$  la verosimiglianza assuma valori più grandi che per valori lontani da  $\theta_0$ . Stimiamo  $\theta_0$ , pertanto, come il **massimo della verosimiglianza**, ovvero

$$\hat{\theta} = \arg \max_{\theta \in \Theta} \prod_{i=1}^n f_\theta(x_i)$$

Vediamo con qualche esempio in che senso  $\hat{\theta}$  è una *buona* stima di  $\theta_0$ .

**Distribuzione di Bernoulli** Supponiamo di voler inferire il parametro  $p_0$  di una distribuzione di Bernoulli. Indicando il parametro  $\theta$  con  $p$ , abbiamo  $\Theta = (0, 1)$  e  $\theta_0 = p_0$ . I dati  $x_i \in \{0, 1\}$  per  $i = 1, \dots, n$  corrispondono, per esempio, all'esito di  $n$  lanci di una stessa moneta (1 e 0 i valori assunti dalla variabile casuale  $X$  con probabilità  $p_0$  e  $1 - p_0$ ). Per  $p$  fissato, la funzione di probabilità di massa può essere scritta come

$$f_p(X) = p^X (1 - p)^{(1-X)} \quad (3.10)$$

L'equazione (3.10) può essere studiata analiticamente al variare del parametro  $p$  (notiamo che, correttamente,  $f_p(1) = p$  e  $f_p(0) = 1 - p$ ). Per la verosimiglianza abbiamo

$$L(\mathbf{x}|p) = \prod_{i=1}^n p^{x_i} (1 - p)^{(1-x_i)} \quad (3.11)$$

Applichiamo il logaritmo (che in quanto funzione monotona non modifica il massimo)

$$\ln L(\mathbf{x}|p) = \ln p \sum_{i=1}^n x_i + \ln(1 - p) \sum_{i=1}^n (1 - x_i)$$

e annulliamo la derivata rispetto a  $p$  ottenendo

$$\frac{d \ln L(\mathbf{x}|p)}{dp} = 0 = \frac{1}{p} \sum_{i=1}^n x_i - \frac{n}{1-p} + \frac{1}{1-p} \sum_{i=1}^n x_i$$

la cui soluzione è

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$$

Quindi  $\hat{p}$ , la stima di massima verosimiglianza di  $p_0$ , è **la media empirica dei valori campionati**. Non sorprendentemente,  $\hat{p}$  stima  $p_0$  come frazione delle *teste* ottenute su  $n$  lanci.

**Distribuzione uniforme** Indicando con  $S$  il parametro  $\theta$ , consideriamo ora il caso di una distribuzione uniforme  $f_S$

$$f_S(x) = \begin{cases} 1/S & x \in [0, S] \\ 0 & \text{altrove} \end{cases}$$

dove  $S > 0$  è il supporto di  $f_S$ . In questo caso abbiamo  $\Theta = (0, +\infty)$  e  $\theta_0 = S_0$ . I dati  $x_i$  per  $i = 1, \dots, n$  sono valori campionati uniformemente nell'intervallo  $[0, S_0]$  con  $S_0$  fissato ma incognito. Un esempio nel caso discreto è dato da una città in cui tutti i taxi sono registrati con un numero da 0 a  $S_0$ . Vogliamo stimare  $S_0$  osservando  $n$  numeri  $x_1, \dots, x_n$ . Sia  $x_{max}$  il massimo valore tra gli  $x_i$ . La verosimiglianza vale il prodotto di  $n$  fattori uguali a  $1/S$  se  $S \geq x_{max}$  e 0 altrimenti, ovvero

$$L(\mathbf{x}|S) = \begin{cases} 1/S^n & \text{se } S \geq x_{max} \\ 0 & \text{se } S < x_{max} \end{cases} \quad (3.12)$$

In questo caso non ci serve calcolare la derivata per determinare il massimo. La verosimiglianza è massima per il più piccolo valore di  $S$ , e quindi per

$$\hat{S} = x_{max}$$

Pertanto  $\hat{S}$ , la stima di massima verosimiglianza di  $S_0$ , è il **massimo dei valori campionati** (ovvero, nell'esempio dei taxi, il numero di registrazione più grande tra quelli osservati).

Discutiamo ora un caso in cui il parametro  $\theta$  è bi-dimensionale.

**Distribuzione normale** Supponiamo ora che i dati  $x_1, \dots, x_n$  provengano dalla distribuzione normale  $\mathcal{N}(\mu_0, \sigma_0^2)$  con  $\mu_0$  e  $\sigma_0^2$  non noti. In questo caso  $\theta$  corrisponde alla coppia  $\theta = (\mu, \sigma^2)$ . Per la verosimiglianza abbiamo

$$L(\mathbf{x}|\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \mu)^2 / 2\sigma^2} \quad (3.13)$$

Applicando il logaritmo otteniamo

$$\ln L(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Uguagliando a zero la derivata rispetto a  $\mu$  otteniamo

$$\frac{\partial}{\partial \mu} \ln L(\mathbf{x}|\mu, \sigma^2) = \frac{\partial}{\partial \mu} \left( -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

che è risolta da

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

Anche in questo caso  $\hat{\mu}$ , la stima di massima verosimiglianza di  $\mu_0$ , coincide con la **media empirica dei valori campionati**.

Ponendo  $\mu = \hat{\mu}$  e uguagliando a zero la derivata rispetto a  $\sigma^2$  otteniamo

$$\begin{aligned} \frac{\partial}{\partial \sigma^2} \ln L(\mathbf{x}|\mu, \sigma^2) &= \frac{\partial}{\partial \sigma^2} \left( -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu})^2 \right) \\ &= -\frac{n}{2\sigma^2} + \left( \frac{1}{2} \sum_{i=1}^n (x_i - \hat{\mu})^2 \right) \frac{1}{\sigma^4} = 0 \end{aligned}$$

che è risolta da

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2 \quad (3.14)$$

### Corretto o distorto

**Distribuzione binomiale** Sia  $p_0$  il valore del parametro della distribuzione di Bernoulli che ha effettivamente generato i dati osservati. Per  $\hat{p}$  abbiamo

$$\mathbb{E}[\hat{p}] = \mathbb{E} \left[ \frac{1}{n} \sum_i x_i \right] = \frac{1}{n} \mathbb{E} \left[ \sum_i x_i \right] = \frac{1}{n} \sum_i \mathbb{E}[x_i] = \frac{np_0}{n} = p_0$$

Poiché  $\mathbb{E}[\hat{p}] = p_0$ , lo stimatore di massima verosimiglianza  $\hat{p}$  è *corretto*.

**Distribuzione uniforme** Poiché  $x_{max} \leq S_0$ , lo stimatore di massima verosimiglianza  $\hat{S} = x_{max}$  è intrinsecamente *distorto* per difetto. Ripetendo la stima tante volte, in questo caso, i risultati fluttuano sempre dalla stessa parte non essendo possibile osservare un numero di registrazione più grande di  $S_0$ .

**Distribuzione normale** Sia  $\mathcal{N}(\mu_0, \sigma_0^2)$  la distribuzione normale che ha effettivamente generato i dati. Per  $\hat{\mu}$  abbiamo

$$\mathbb{E}[\hat{\mu}] = \mathbb{E} \left[ \frac{1}{n} \sum_i x_i \right] = \frac{1}{n} \mathbb{E} \left[ \sum_i x_i \right] = \frac{1}{n} \sum_i \mathbb{E}[x_i] = \frac{1}{n} \sum_i \mu_0 = \frac{n\mu_0}{n} = \mu_0$$

Lo stimatore  $\hat{\mu}$ , quindi, è *corretto*. Tenuto conto che per la varianza di  $\hat{\mu}$  abbiamo

$$Var(\hat{\mu}) = \mathbb{E}[(\hat{\mu} - \mu_0)^2] = Var \left( \frac{1}{n} \sum_i x_i \right) = \frac{1}{n^2} \sum_i Var(x_i) = \frac{n\sigma_0^2}{n^2} = \frac{\sigma_0^2}{n}$$

per il valore atteso di  $\hat{\sigma}^2$  otteniamo invece

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2] &= \mathbb{E} \left[ \frac{1}{n} \sum_i (x_i - \hat{\mu})^2 \right] = \frac{1}{n} \sum_i \mathbb{E}[(x_i - \hat{\mu})^2] = \frac{1}{n} \sum_i \mathbb{E}[(x_i - \mu_0 + \mu_0 - \hat{\mu})^2] \\ &= \frac{1}{n} \sum_i \mathbb{E}[(x_i - \mu_0)^2] + \frac{1}{n} \sum_i \mathbb{E}[(\mu_0 - \hat{\mu})^2] - 2\mathbb{E} \left[ (\mu_0 - \hat{\mu}) \frac{1}{n} \sum_i (\mu_0 - x_i) \right] \\ &= \frac{1}{n} n\sigma_0^2 + \frac{1}{n} \frac{n\sigma_0^2}{n} - 2\mathbb{E}[(\mu_0 - \hat{\mu})^2] = \sigma_0^2 + \frac{\sigma_0^2}{n} - 2\frac{\sigma_0^2}{n} = \frac{n-1}{n} \sigma_0^2 \end{aligned}$$

Lo stimatore  $\hat{\sigma}^2$ , pertanto, è *distorto*. Tuttavia, poiché  $(n-1)/n \rightarrow 1$  per  $n \rightarrow \infty$ , è *asintoticamente corretto*. Si può *correggere* lo stimatore dividendo per  $n-1$  anziché per  $n$  nella formula (3.14).



## Consistenza

Una seconda importante proprietà è la *consistenza*, ovvero la convergenza in probabilità della stima ottenuta rispetto al valore vero nel senso della legge dei grandi numeri.

Per le *distribuzioni binomiale e normale* la consistenza è garantita dalla convergenza della media empirica al valore atteso.

Per la *distribuzione uniforme*, invece la convergenza è garantita dal fatto che

$$\text{per } n \rightarrow \infty, \forall \epsilon \Pr(S - x_{max} \geq \epsilon) \rightarrow 0$$

### Osservazione 3.24.2. Più osservazioni

I due diversi metodi di inferenza che abbiamo visto possono portare a risultati diversi. In generale, non possiamo concludere che uno sia meglio dell'altro, ma solo osservare che **l'inferenza dipende dal principio che decidiamo di adottare**. Tipicamente, le differenze tra i metodi frequentisti e i metodi bayesiani tendono a ridursi all'aumentare del numero di osservazioni.

### Compito 3.24.1. La forma della verosimiglianza

1. Fissa  $0 < p_0 < 1$  e campiona  $n$  punti  $x_i$  con  $i = 1, \dots, n$  dalla distribuzione uniforme in  $[0, 1]$  e poni  $x_i = 1$  se  $x_i < p_0$  e 0 altrimenti. Produci il grafico della verosimiglianza dell'equazione (3.11), vedi Figura 3.6 a sinistra, per  $0 < p < 1$ . Confronta  $\hat{p}$  con  $p_0$  ripetendo l'esperimento 10 volte con  $n = 10$  ed  $n = 100$  e per diversi valori di  $p_0$ .
2. Fissa  $S_0 > 0$  e campiona  $n$  punti da una distribuzione uniforme in  $[0, S_0]$ . Produci il grafico della verosimiglianza dell'equazione (3.12), vedi Figura 3.6 al centro, per  $0 < S < 2S_0$ . Confronta  $\hat{S}$  con  $S_0$  ripetendo l'esperimento 10 volte con  $n = 10$  ed  $n = 100$ .
3. Fissa  $\mu_0$  e  $\sigma_0^2$  e campiona  $n$  punti da una distribuzione normale  $\mathcal{N}(\mu_0, \sigma_0^2)$ . Produci il grafico della verosimiglianza dell'equazione (3.13), vedi Figura 3.6 a destra, per  $0 < \mu < 2\mu_0$  e  $0 < \sigma^2 < 2\sigma_0^2$ . Confronta  $\hat{\mu}$  con  $\mu_0$  e  $\hat{\sigma}^2$  con  $\sigma_0^2$  ripetendo l'esperimento 10 volte con  $n = 10$  ed  $n = 100$ .

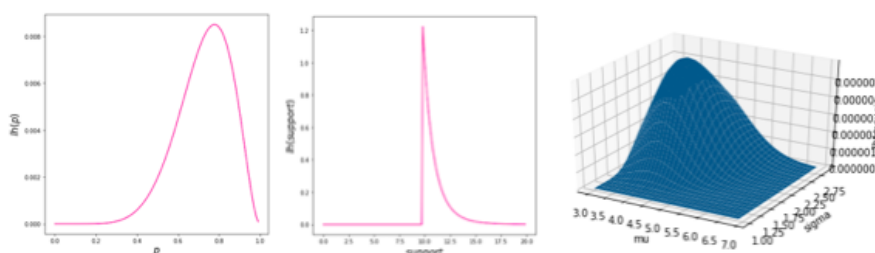


Figura 3.6: Vedi testo

## Assicurati di ...

1. saper spiegare che cosa sia la verosimiglianza, almeno nei tre esempi discussi in dettaglio;
2. saper applicare il principio di massima verosimiglianza nei tre esempi, ovvero annullando la derivata, nei casi continui, o confrontando i valori disponibili, nel caso discreto.



## Capitolo 4

# Esercizi risolti di Teoria della Probabilità

Questi esercizi non sono necessariamente simili a quello che incontrerai nella prova d'esame. Il loro scopo principale è consentirti di valutare se hai capito i concetti fondamentali.

### Principio base

- 1.1** Se una stringa è formata da  $n$  caratteri quante sono le possibili sotto-stringhe?

*In ogni sotto-stringa ognuno degli  $n$  caratteri può esserci e non esserci: 2 scelte per ognuno degli  $n$  caratteri, in tutto  $2^n$ , che è uguale alla cardinalità dell'insieme delle parti di un insieme di  $n$  elementi.*

- 1.2** Hai 3 maglioni, 4 camicie e 2 paia di pantaloni. In quanti ordinamenti diversi puoi trovarli senza mescolare i tipi di indumento?

*I 3 tipi di indumento sono ordinabili in  $3!$  modi. Per ognuno di questi 6 ordinamenti i 3 maglioni sono ordinabili in  $3!$  modi, le 4 camicie in  $4!$  e le 2 paia di pantaloni in  $2!$ , per cui*

$$3! \times (3! \times 4! \times 2!) = 6 \times 6 \times 24 \times 2 = 1728$$

- 1.3** Includendo anche le parole senza senso quanti sono gli anagrammi di RAMARRO?

*Nelle  $7!$  permutazioni delle lettere che compongono RAMARRO, per ogni posizione fissata delle 3 R, ci sono  $3!$  permutazioni indistinguibili (quelle che si ottengono permutando le R tra loro) e, per ogni posizione fissata delle 2 A, ci sono  $2!$  permutazioni indistinguibili (quelle che si ottengono permutando le A tra loro) per cui otteniamo*

$$\frac{7!}{3! \times 2!} = 7 \times 5 \times 4 \times 3 = 420$$

- 1.4** Un campionato è composto da 20 squadre. Calcola quanti mini-campionati diversi si potrebbero disputare composti da 5 squadre.

*La risposta è data dal coefficiente binomiale  $\binom{20}{5}$ . Alternativamente, nelle  $20!$  permutazioni delle squadre, le permutazioni che hanno le stesse 5 squadre che compongono un mini-campionato nelle stesse 5 posizioni sono  $5!$  e le permutazioni che hanno le stesse 15 squadre escluse nelle rimanenti posizioni sono  $15!$ , per cui*

$$\frac{20!}{5! \times 15!} = \frac{20 \times 19 \times 18 \times 17 \times 16}{5 \times 4 \times 3 \times 2} = 19 \times 3 \times 17 \times 16 = 15,584$$

## Eventi equiprobabili e non

- 2.1** Dopo aver definito lo spazio campionario, calcola la probabilità  $P$  di pescare 1 pallina rossa e 2 bianche da un'urna che contiene 2 palline rosse, 2 bianche e 3 azzurre.

*Lo spazio campionario è definito dall'insieme di tutte le possibili estrazioni di tre palline da 2 palline rosse, 2 bianche e 3 azzurre, ovvero*

$$\binom{7}{3} = 7 \times 5 \times 4 \times 3 \times 2 = 840$$

*I casi favorevoli sono invece*

$$\binom{2}{1} \times \binom{2}{2} = 2$$

*per cui  $P = 1/420$ .*

- 2.2** Peschi due carte da un mazzo di 52. Dopo aver definito lo spazio campionario, calcola la probabilità  $P_{AA}$  che siano due Assi,  $P_{AR}$  un Asso e un Re e  $P_{scart}$  due carte dal 10 al 2 compresi.

*Lo spazio campionario è costituito da tutte le possibili coppie di 52 carte, ovvero*

$$\binom{52}{2} = 26 \times 51 = 1326$$

*Per i casi favorevoli abbiamo*

$$\#AA = \binom{4}{2} = 6, \quad \#AR = \binom{4}{1} \times \binom{4}{1} = 16 \quad \text{e} \quad \#scart = \binom{36}{2} = 18 \times 35 = 630$$

*per cui*

$$P_{AA} = 6/1326 \approx 0.004, \quad P_{AR} = 16/1326 \approx 0.12 \quad \text{e} \quad P_{scart} = 630/1326 \approx 0.475$$

- 2.3** Lanci un dado onesto tre volte. Dopo aver definito lo spazio campionario, a quale somma dei risultati corrisponde la probabilità  $P$  più alta?

*Lo spazio campionario consiste delle triplette di tutti i possibili risultati, ovvero  $6^3 = 216$ . I possibili valori della somma sono da 3 a 18 compresi. I valori centrali, 10 e 11 sono quelli che si possono ottenere in più casi, come mostrato in tabella*

somma 10	27 casi	somma 11	27 casi
6 3 1	6 modi	6 4 1	6 modi
6 2 2	3 modi	6 3 2	6 modi
5 4 1	6 modi	5 5 1	3 modi
5 3 2	6 modi	5 4 2	6 modi
4 4 2	3 modi	5 3 3	3 modi
4 3 3	3 modi	4 4 3	3 modi

*da cui otteniamo  $P = 27/216 = 1/8 = 0.125$ .*

- 2.4** Se lanci una moneta truccata con  $P(T) = 0.1$  quattro volte, calcola le probabilità di tutti gli eventi possibili.

*Gli eventi possibili sono  $2^4 = 16$ . Di questi eventi, uno consiste di 0 volte testa con probabilità pari a  $9^4/10000 = 0.6561$ , quattro di 1 volta testa (TCCC, CTCC, CCTC e CCCT) ciascuno con probabilità pari a  $9^3/10000 = 0.0729$ , sei di 2 volte testa (TTCC, ..., CCTT) ciascuno con probabilità pari a  $9^2/10000 = 0.0081$ , quattro di 3 volte testa (TTTC, TTCT, TCTT e CTTT) ciascuno con probabilità pari a  $9/10000 = 0.0009$  e uno di 4 volte testa con probabilità pari a  $1/10000 = 0.0001$ . Ovviamente abbiamo che*

$$0.6561 + 4 \times 0.0729 + 6 \times 0.0081 + 4 \times 0.0009 + 0.0001 = 1$$

## Probabilità condizionata

- 3.1** Dato un mazzo di 52 carte, qual è la probabilità che la prima carta sia un Asso? E quale se l'ultima è un Asso?

*Ragionando in termini di risultati possibili e risultati favorevoli otteniamo*

$$\Pr(\text{prima } A) = 4/52 = 1/13 \approx 0.077 \quad \text{e} \quad \Pr(\text{prima } A \mid \text{ultima } A) = 3/51 \approx 0.059$$

*Poiché dei  $52 \times 51/2$  casi possibili ci sono  $\binom{4}{2} = 6$  casi favorevoli per un  $A$  nella prima e un  $A$  nell'ultima carta, possiamo usare la definizione di probabilità condizionata e, siccome  $\Pr(\text{ultima } A) = 1/13$ , ottenere ugualmente*

$$\Pr(\text{prima } A \mid \text{ultima } A) = \frac{\Pr(\text{prima } A, \text{ultima } A)}{\Pr(\text{ultima } A)} = \frac{6}{26 \times 51} \times 13 = \frac{3}{51} \approx 0.059$$

- 3.2** Raimondo ritiene che la probabilità di prendere 30 nell'esame di Biologia sia pari all'80%, mentre nell'esame di Chimica sia del 70%. Se Raimondo decide quale esame sostenere, Biologia o Chimica, sulla base del lancio di una moneta onesta, con quale probabilità sosterrà Biologia e prenderà 30?

*Per la probabilità dell'intersezione degli eventi se  $T$  è l'evento prendere 30 e  $B$  l'evento sostenere l'esame di Biologia, otteniamo*

$$\Pr(TB) = \Pr(B)\Pr(T|B) = \frac{1}{2} \times 80\% = 40\%$$

*Questo risultato a prima vista sembra sorprendente. Il complementare dell'evento intersezione prendere 30 e sostenere l'esame di Biologia, tuttavia, è l'unione dei complementari, ovvero non prendere 30 o sostenere l'esame di Chimica invece che quello di Biologia!*

- 3.3** Lancia un dado onesto due volte. Qual è la probabilità che uno dei risultati sia 3 se la somma è 8?

*Poiché se la somma è 8 si ottiene un 3 quando l'altro risultato è un 5, dobbiamo calcolare  $\Pr(3 \text{ al secondo} \mid \text{il primo è } 5)$  e  $\Pr(5 \text{ al secondo} \mid \text{il primo è } 3)$ . Ma in entrambi i casi gli eventi sono indipendenti per cui*

$$\begin{aligned} \Pr(3 \text{ al secondo} \mid \text{il primo è } 5) &= \Pr(3 \text{ al secondo}) = \frac{1}{6} \\ \Pr(5 \text{ al secondo} \mid \text{il primo è } 3) &= \Pr(5 \text{ al secondo}) = \frac{1}{6} \end{aligned}$$

*e, quindi,  $P = 1/6 + 1/6 = 1/3 \approx 0.333$ .*

## Teorema di Bayes

- 4.1** Sono date tre urne uguali,  $A$ ,  $B$  e  $C$ . L'urna  $A$  contiene 3 palline rosse e 1 bianca, l'urna  $B$  3 palline bianche e 1 rossa e l'urna  $C$  4 palline rosse. Se peschi una pallina rossa, calcola la probabilità che provenga da  $A$ ,  $B$  o  $C$ .

*Se indichiamo rossa con  $r$ , dobbiamo calcolare  $P(A|r)$ ,  $P(B|r)$  e  $P(C|r)$ . Poiché le tre urne sono uguali abbiamo che  $P(A) = P(B) = P(C) = 1/3$ . Inoltre  $P(r|A) = 3/4$ ,  $P(r|B) = 1/4$  e  $P(r|C) = 1$ . Per la probabilità totale  $P(r)$  abbiamo quindi*

$$P(r) = P(r|A)P(A) + P(r|B)P(B) + P(r|C)P(C) = \left(\frac{3}{4} + \frac{1}{4} + 1\right) \times \frac{1}{3} = \frac{2}{3}$$

Applicando il teorema di Bayes otteniamo

$$\begin{aligned}P(A|r) &= \frac{P(r|A)P(A)}{P(r)} = \frac{3}{4} \times \frac{1}{3} \times \frac{3}{2} = \frac{3}{8} \\P(B|r) &= \frac{P(r|B)P(B)}{P(r)} = \frac{1}{4} \times \frac{1}{3} \times \frac{3}{2} = \frac{1}{8} \\P(C|r) &= \frac{P(r|C)P(C)}{P(r)} = \frac{1}{3} \times \frac{3}{2} = \frac{1}{2}\end{aligned}$$

- 4.2** Una compagnia di assicurazione crede che gli individui possano dividersi in due categorie: gli individui della categoria  $A$  sono predisposti a essere coinvolti in incidenti mentre gli individui della categoria  $A^c$  non lo sono. Se  $A_1$  è l'evento di essere coinvolto in un incidente entro un anno dalla stipula della polizza, la compagnia sa che  $P(A_1|A) = 40\%$  mentre  $P(A_1|A^c) = 20\%$ . Se la compagnia sa che il 30% degli assicurati appartiene alla categoria  $A$  con che probabilità un nuovo assicurato riporterà un incidente nel primo anno? E con quale probabilità, in quel caso, appartiene alla categoria  $A$ ?

Per la probabilità totale abbiamo

$$P(A_1) = P(A_1|A)P(A) + P(A_1|A^c)P(A^c) = 0.4 \times 0.3 + 0.2 \times 0.7 = 26\%$$

mentre per  $P(A|A_1)$ , usando il teorema di Bayes, otteniamo

$$P(A|A_1) = \frac{P(A)P(A_1|A)}{P(A_1)} = \frac{0.3 \times 0.4}{0.26} = \frac{6}{13} \approx 46\%$$

- 4.3** In un esame a risposta multipla Manuela o conosce la risposta corretta o tira a indovinare. Sia  $p$  la probabilità che Manuela conosca la risposta corretta (e quindi  $1 - p$  che tiri a indovinare). Se le possibili risposte sono  $m$  Manuela indovinerà la risposta corretta con probabilità  $1/m$ . Con quale probabilità Manuela conosce la risposta corretta se ha fornito la risposta corretta?

Sia  $F$  l'evento Manuela ha fornito la risposta corretta e  $C$  l'evento Manuela conosce la risposta corretta. Dobbiamo trovare  $P(C|F)$ . Usando Bayes otteniamo

$$\begin{aligned}P(C|F) &= \frac{P(CF)}{P(F)} = \frac{P(C)P(F|C)}{P(C)P(F|C) + P(C^c)P(F|C^c)} \\&= \frac{p \times 1}{p \times 1 + (1 - p)/m} = \frac{mp}{1 + (m - 1)p}\end{aligned}$$

Se  $p = 1/2$  e  $m = 4$ , per esempio, abbiamo  $P(C|F) = 4/5$ .

## Variabili casuali discrete

- 5.1** Calcola  $\mathbb{E}[X]$ ,  $Var(X)$  e  $\mathbb{E}[X^2]$  della variabile casuale  $X$  con

$$p(1) = \frac{1}{2}, \quad p(2) = \frac{1}{4}, \quad p(3) = \frac{1}{8} \quad \text{e} \quad p(4) = \frac{1}{8}$$

Dalla definizione di valore atteso si ha

$$\mathbb{E}[X] = 1 \times \frac{1}{2} + 2 \times \frac{1}{4} + 3 \times \frac{1}{8} + 4 \times \frac{1}{8} = \frac{7}{4}$$

e

$$\mathbb{E}[X^2] = 1 \times \frac{1}{2} + 4 \times \frac{1}{4} + 9 \times \frac{1}{8} + 16 \times \frac{1}{8} = \frac{37}{8}$$

per cui la varianza è data da

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{37}{8} - \frac{49}{16} = \frac{25}{16}$$

## 5.2 Calcola la cdf della variabile casuale $X$ con

$$p(1) = \frac{1}{8}, \quad p(2) = \frac{1}{8}, \quad p(3) = \frac{1}{4} \quad \text{e} \quad p(4) = \frac{1}{2}$$

e  $P(X < 2)$ ,  $P(X \leq 2)$  e  $P(X > 2)$ .

I valori possibili di  $X$  sono i quattro interi 1, 2, 3 e 4. La cdf è definita allora come

$$F(x) = \sum_{x \leq i} p(i)$$

e, quindi,

$$F(x) = \begin{cases} 0 & x < 1 \\ 1/8 & 1 \leq x < 2 \\ 1/4 & 2 \leq x < 3 \\ 1/2 & 3 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

Per le tre probabilità, quindi, abbiamo

$$P(X < 2) = p(1) = \frac{1}{8}, \quad P(X \leq 2) = p(1) + p(2) = \frac{1}{4} \quad \text{e} \quad P(X > 2) = p(3) + p(4) = \frac{3}{4}$$

## 5.3 Data la cdf

$$F(x) = \begin{cases} 0 & x < 1 \\ 2/3 & 1 \leq x < 2 \\ 7/9 & 2 \leq x < 3 \\ 1 & x \geq 3 \end{cases}$$

deriva la pmf e calcola valore atteso e varianza della variabile casuale sottostante. Per ottenere la pmf dalla cdf dobbiamo considerare la differenza tra il limite da destra e il limite da sinistra in ogni punto. La pmf è diversa da 0 solo nei punti di discontinuità di  $F$ . Abbiamo, infatti,

$$p(1) = F(1^+) - F(1^-) = \frac{2}{3}, \quad p(2) = F(2^+) - F(2^-) = \frac{1}{9} \quad \text{e} \quad p(3) = F(3^+) - F(3^-) = \frac{2}{9}$$

Pertanto

$$\mathbb{E}[X] = 1 \times p(1) + 2 \times p(2) + 3 \times p(3) = \frac{2}{3} + \frac{2}{9} + \frac{6}{9} = \frac{14}{9}$$

E poiché

$$\mathbb{E}[X^2] = 1 \times p(1) + 4 \times p(2) + 9 \times p(3) = \frac{2}{3} + \frac{4}{9} + 2 = \frac{28}{9}$$

per la varianza otteniamo

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{28}{9} - \frac{196}{81} = \frac{56}{81}$$

## Distribuzioni discrete di probabilità

**6.1** Sia  $X$  la variabile casuale che conta il numero di 6 ottenuti lanciando un dado onesto tre volte. Determina la pmf e la cdf di  $X$ . Con che probabilità ottieni almeno un 6? E con che probabilità un numero pari di 6? Calcola il valore atteso e la varianza di  $X$ .

I possibili valori assunti da  $X$  sono 0, 1, 2 e 3. Poiché il dado è onesto si tratta di una distribuzione binomiale con  $p = 1/6$  ed  $n = 3$ . Pertanto per la pmf abbiamo

$$p(0) = \frac{5^3}{6^3} = \frac{125}{216}, \quad p(1) = \frac{3 \times 5^2}{216} = \frac{75}{216}, \quad p(2) = \frac{3 \times 5}{216} = \frac{15}{216} \quad \text{e} \quad p(3) = \frac{1}{216}$$

La cdf è una scala con il primo gradino a 0 e altezza  $p(0)$ , il secondo a 1 con altezza  $p(1)$ , il terzo a 2 con altezza  $p(2)$  e il quarto a 3 con altezza  $p(3)$ . La probabilità di ottenere almeno un 6 è  $1 - p(0) = 91/216$ , un numero pari di 6 è  $p(0) + p(2) = 140/216$ . Per il valore atteso otteniamo

$$\mathbb{E}[X] = 1 \times \frac{75}{216} + 2 \times \frac{15}{216} + 3 \times \frac{1}{216} = \frac{108}{216} = \frac{1}{2}$$

Poiché

$$\mathbb{E}[X^2] = 1 \times \frac{75}{216} + 4 \times \frac{15}{216} + 9 \times \frac{1}{216} = \frac{144}{216} = \frac{2}{3}$$

per la varianza otteniamo

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{3} - \frac{1}{4} = \frac{5}{12}$$

**6.2** Se un algoritmo restituisce la risposta corretta il 50% delle volte quante volte devi lanciarlo per ottenere il risultato corretto con probabilità superiore al 99,9%?

Si tratta di una distribuzione geometrica con  $p = 1/2$ . Pertanto

$$\left(1 - \frac{1}{2^n}\right) \geq 0.999 \rightarrow \frac{1}{2^n} \leq 0.001 \rightarrow n \geq \log_2 1000 \approx 10$$

**6.3** Un libro contiene in media un errore di stampa ogni pagina. Con che probabilità contiene due errori in una stessa pagina?

Si tratta di una distribuzione di Poisson con  $\mu = 1$  per cui

$$P(X = 2) = \frac{1}{2}e^{-1} \approx 18\%$$

## Variabili casuali continue

**7.1** Se la pdf di una variabile casuale  $X$  è definita come

$$f(x) = \begin{cases} 1 & \text{se } 0 \leq x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

calcola la cdf,  $\mathbb{E}[X]$  e  $\text{Var}(X)$ . La funzione  $f(\cdot)$  è una densità di probabilità perché il suo integrale definito è uguale a 1. Infatti

$$\int_{-\infty}^{+\infty} f(x)dx = \int_0^1 dx = x \Big|_0^1 = 1$$

Per ogni  $x$  in  $(-\infty, +\infty)$  la cdf  $F$  è definita come

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

Quindi, per  $x \leq 0$   $F$  è identicamente nulla, mentre per  $x \geq 1$  è identicamente uguale a 1. Per  $x \in [0, 1]$  abbiamo

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt = \int_0^x dt = x$$

In conclusione

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$



Poiché la pdf è continua a tratti, la cdf nei punti di continuità per  $f$  è derivabile e  $F'(x) = f(x)$ . Si verifica inoltre facilmente che  $F'(0^-) = f(0^-) = 0$ ,  $F'(0^+) = f(0^+) = 1$ ,  $F'(1^-) = f(1^-) = 1$  e  $F'(1^+) = f(1^+) = 0$ . Per quanto riguarda il valore atteso abbiamo

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx = \int_0^1 x dx = \frac{x^2}{2} \Big|_0^1 = \frac{1}{2}$$

Per  $\text{Var}(X)$  conviene sfruttare la formula  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ . Dal fatto che

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \int_0^1 x^2 dx = \frac{x^3}{3} \Big|_0^1 = \frac{1}{3}$$

otteniamo

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

**7.2** Se  $Y = X^n$ , dove  $X$  è la variabile casuale dell'esercizio precedente, calcola  $\mathbb{E}[Y]$  e commenta il risultato che ottieni per  $n$  grande.

$$\mathbb{E}[X^n] = \int_{-\infty}^{+\infty} x^n f(x) dx = \int_0^1 x^n dx = \frac{x^{n+1}}{n+1} \Big|_0^1 = \frac{1}{n+1}$$

Per  $n$  grande il valore atteso tende a 0!

**7.3** Per quale valore della costante  $C$  la funzione

$$f(x) = \begin{cases} Cx & \text{se } 0 \leq x < 1 \\ C & \text{se } 1 \leq x \leq 2 \\ 0 & \text{altrimenti} \end{cases}$$

per  $x$  nell'intervallo  $[0, 2]$  è una pdf? Calcola la cdf per  $X$ ,  $\mathbb{E}[X]$  e  $\text{Var}(X)$ .

La funzione  $f(\cdot)$  è una densità di probabilità se il suo integrale definito nell'intervallo  $[0, 2]$  è uguale a 1. Poiché

$$1 = \int_0^2 f(x) dx = C \int_0^1 x dx + C \int_1^2 dx = C \left( \frac{x^2}{2} \Big|_0^1 + x \Big|_1^2 \right) = \frac{3C}{2}$$

otteniamo  $C = 2/3$ . Per ogni  $x$  in  $(-\infty, +\infty)$  la cdf  $F$  è definita come

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Quindi, per  $x \leq 0$   $F$  è identicamente nulla, mentre per  $x \geq 2$  è identicamente uguale a 1. Per  $x \in [0, 1]$  abbiamo

$$F(x) = P(X \leq x) = \frac{2}{3} \int_{-\infty}^x t dt = \frac{2}{3} \int_0^x t dt = \frac{2}{3} \frac{x^2}{2} = \frac{x^2}{3}$$

Per  $x \in [1, 2]$  abbiamo

$$\begin{aligned} F(x) &= P(X \leq x) = \frac{2}{3} \int_{-\infty}^x f(t) dt = \frac{2}{3} \int_0^1 t dt + \frac{2}{3} \int_1^x dt \\ &= \frac{1}{3} + \frac{2}{3} \int_1^x dt = \frac{1}{3} + \frac{2}{3}(x-1) = \frac{2x-1}{3} \end{aligned}$$

In conclusione

$$F(x) = \begin{cases} 0 & x < 0 \\ x^2/3 & 0 \leq x \leq 1 \\ (2x-1)/3 & 1 \leq x \leq 2 \\ 1 & 2 \geq x \end{cases}$$

Poiché la pdf è continua ovunque tranne che per  $x = 2$ , la cdf è derivabile e  $F'(x) = f(x)$  per tutti gli  $x \neq 0$ . Si verifica facilmente in particolare che  $F'(0^-) = F'(0^+) = f(0) = 0$  e  $F'(1^-) = F'(1^+) = f(1) = 2/3$ . Per  $x = 2$ , come prima, abbiamo  $F'(2^-) = f(2^-) = 2/3$  e  $F'(2^+) = f(2^+) = 0$ . Per quanto riguarda il valore atteso abbiamo

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx = \frac{2}{3} \left( \int_0^1 x^2 dx + \int_1^2 x dx \right) = \frac{2}{3} \left( \frac{x^3}{3} \Big|_0^1 + \frac{x^2}{2} \Big|_1^2 \right) = \frac{2}{9} + 1 = \frac{11}{9}$$

Per  $\text{Var}(X)$  conviene sfruttare la formula  $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ . Dal fatto che

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \frac{2}{3} \left( \int_0^1 x^3 dx + \int_1^2 x^2 dx \right) = \frac{2}{3} \left( \frac{x^4}{4} \Big|_0^1 + \frac{x^3}{3} \Big|_1^2 \right) = \frac{1}{6} + \frac{14}{9} = \frac{31}{18}$$

otteniamo

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{31}{18} - \frac{121}{81} = \frac{279 - 242}{162} = \frac{37}{162}$$

## Distribuzioni continue di probabilità

**8.1** Calcola la cdf, il valore atteso e la varianza per la variabile casuale continua  $X$  descritta dalla funzione densità uniforme

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$

Per ogni  $x$  in  $(-\infty, +\infty)$  la cdf  $F$  è definita come

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt$$

Quindi, per  $x \leq a$   $F$  è identicamente nulla, mentre per  $x \geq b$  è identicamente uguale a 1. Per  $x \in [a, b]$  abbiamo

$$F(x) = \frac{1}{b-a} \int_a^x dt = \frac{x-a}{b-a}$$

Per tutti gli  $x \in (a, b)$ ,  $F'(x) = 1/(b-a) = f(x)$ . Per il valore atteso abbiamo

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{b-a} \int_a^b x dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

e poiché

$$\mathbb{E}[X^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + b^2 + ab}{3}$$

per la varianza otteniamo

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{a^2 + b^2 + ab}{3} - \frac{a^2 + b^2 + 2ab}{4} = \frac{a^2 + b^2 - 2ab}{12} = \frac{(b-a)^2}{12}$$

## 8.2 Determina il valore della costante $C$ per la quale

$$f(x) = \begin{cases} C(1-x^2) & -1 \leq x \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

è una funzione densità. Quindi calcola valore atteso, varianza e  $P(0 \leq X \leq 1/2)$  per la variabile casuale  $X$  sottostante.

Da

$$\int_{-\infty}^{+\infty} f(x)dx = C \int_{-1}^1 (1-x^2)dx = C \left( x - \frac{x^3}{3} \right) \Big|_{-1}^1 = C \left( 2 - \frac{2}{3} \right) = C \frac{4}{3}$$

otteniamo  $C = 3/4$ . Poiché  $f$  è pari il valore atteso è nullo. Per la varianza abbiamo

$$\frac{3}{4} \int_{-1}^1 x^2(1-x^2)dx = \frac{3}{4} \left( \frac{x^3}{3} - \frac{x^5}{5} \right) \Big|_{-1}^1 = \frac{3}{4} \left( \frac{2}{3} - \frac{2}{5} \right) = \frac{1}{5}$$

Per ogni  $x$  in  $(-\infty, +\infty)$  la cdf  $F$  è definita come

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$$

Quindi, per  $x \leq -1$   $F$  è identicamente nulla, mentre per  $x \geq 1$  è identicamente uguale a 1. Per  $x \in [-1, 1]$  abbiamo

$$F(x) = \frac{3}{4} \int_{-1}^x (1-t^2)dt = \frac{3}{4} \left( x + 1 - \frac{x^3}{3} - \frac{1}{3} \right) = \frac{3}{4} \left( \frac{2}{3} + x - \frac{x^3}{3} \right)$$

Pertanto,

$$P(0 \leq X \leq 1/2) = F\left(\frac{1}{2}\right) - F(0) = \frac{3}{4} \left( \frac{2}{3} + \frac{1}{2} - \frac{1}{24} \right) - \frac{1}{2} = \frac{81}{96} - \frac{1}{2} = \frac{11}{32}$$

## Distribuzioni congiunte e indipendenza

### 9.1 Determina la distribuzione congiunta delle variabili casuali $X$ , faccia di un dado onesto, e $Y$ , parità del risultato nel lancio.

Per  $X$  abbiamo

$$p_X(x = i) = \frac{1}{6} \quad \forall i = 1, \dots, 6$$

mentre per  $Y$

$$p_Y(y = 0) = \frac{1}{2} \quad \text{per } i = 2, 4 \text{ e } 6 \quad \text{e} \quad p_Y(y = 1) = \frac{1}{2} \quad \text{per } i = 1, 3 \text{ e } 5$$

Per le probabilità condizionate del tipo  $p(x|y)$  abbiamo

$$\begin{aligned} p(x = 2|y = 0) &= p(x = 4|y = 0) = p(x = 6|y = 0) = \frac{1}{3} \\ p(x = 1|y = 1) &= p(x = 3|y = 1) = p(x = 5|y = 1) = \frac{1}{3} \end{aligned}$$

da cui, tenuto conto che  $p(x, y) = p(x|y)p_Y(y)$ , ricaviamo la tabella

---

	$Y = 0$	$Y = 1$	
$X = 1$	0	1/6	1/6
$X = 2$	1/6	0	1/6
$X = 3$	0	1/6	1/6
$X = 4$	1/6	0	1/6
$X = 5$	0	1/6	1/6
$X = 6$	1/6	0	1/6
	1/2	1/2	

---

Otterremmo la stessa tabella usando le probabilità condizionate del tipo  $p(y|x)$ , ovvero

$$\begin{aligned} p(y=0|x=2) &= p(y=0|x=4) = p(y=0|x=6) = 1 \\ p(y=1|x=1) &= p(y=1|x=3) = p(y=1|x=5) = 1 \end{aligned}$$

e tenendo conto che  $p(x, y) = p_X(x)p(y|x)$ .

**9.2** Determina le probabilità condizionate e le probabilità marginali data la distribuzione di probabilità congiunta in tabella.

---

	$Y = 1$	$Y = 2$	$Y = 4$
$X = -1$	1/16	3/16	0
$X = 0$	5/16	1/16	3/16
$X = 1$	1/16	1/16	1/16

---

Osserviamo preliminarmente che sommando su tutti i valori possibili di  $X$  e  $Y$  otteniamo 1, infatti

$$\sum_x \sum_y p(x, y) = \frac{1}{16} + \frac{3}{16} + \frac{5}{16} + \frac{1}{16} + \frac{3}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = 1$$

Per le probabilità marginali sommando sulle colonne per ogni riga otteniamo

$$\begin{aligned} p_X(-1) &= p(-1, 1) + p(-1, 2) + p(-1, 4) = \frac{1}{16} + \frac{3}{16} = \frac{4}{16} \\ p_X(0) &= p(0, 1) + p(0, 2) + p(0, 4) = \frac{5}{16} + \frac{1}{16} + \frac{3}{16} = \frac{9}{16} \\ p_X(1) &= p(1, 1) + p(1, 2) + p(1, 4) = \frac{1}{16} + \frac{1}{16} + \frac{1}{16} = \frac{3}{16} \end{aligned}$$

e sommando sulle righe per ogni colonna

$$\begin{aligned} p_Y(1) &= p(-1, 1) + p(0, 1) + p(1, 1) = \frac{1}{16} + \frac{5}{16} + \frac{1}{16} = \frac{7}{16} \\ p_Y(2) &= p(-1, 2) + p(0, 2) + p(1, 2) = \frac{3}{16} + \frac{1}{16} + \frac{1}{16} = \frac{5}{16} \\ p_Y(4) &= p(-1, 4) + p(0, 4) + p(1, 4) = \frac{0}{16} + \frac{3}{16} + \frac{1}{16} = \frac{4}{16} \end{aligned}$$

Pertanto per la probabilità condizionata  $p(X|Y)$  otteniamo

$$\begin{aligned}
p(X = -1|Y = 1) &= \frac{p(X = -1, Y = 1)}{p_Y(Y = 1)} = \frac{1}{16} \times \frac{16}{7} = \frac{1}{7} \\
p(X = 0|Y = 1) &= \frac{p(X = 0, Y = 1)}{p_Y(Y = 1)} = \frac{5}{16} \times \frac{16}{7} = \frac{5}{7} \\
p(X = 1|Y = 1) &= \frac{p(X = 1, Y = 1)}{p_Y(Y = 1)} = \frac{1}{16} \times \frac{16}{7} = \frac{1}{7} \\
p(X = -1|Y = 2) &= \frac{p(X = -1, Y = 2)}{p_Y(Y = 2)} = \frac{3}{16} \times \frac{16}{5} = \frac{3}{5} \\
p(X = 0|Y = 2) &= \frac{p(X = 0, Y = 2)}{p_Y(Y = 2)} = \frac{1}{16} \times \frac{16}{5} = \frac{1}{5} \\
p(X = 1|Y = 2) &= \frac{p(X = 1, Y = 2)}{p_Y(Y = 2)} = \frac{1}{16} \times \frac{16}{5} = \frac{1}{5} \\
p(X = -1|Y = 4) &= \frac{p(X = -1, Y = 4)}{p_Y(Y = 4)} = 0 \times \frac{16}{4} = 0 \\
p(X = 0|Y = 4) &= \frac{p(X = 0, Y = 4)}{p_Y(Y = 4)} = \frac{3}{16} \times \frac{16}{4} = \frac{3}{4} \\
p(X = 1|Y = 4) &= \frac{p(X = 1, Y = 4)}{p_Y(Y = 4)} = \frac{1}{16} \times \frac{16}{4} = \frac{1}{4}
\end{aligned}$$

Per la probabilità condizionata  $p(Y|X)$  invece otteniamo

$$\begin{aligned}
p(Y = 1|X = -1) &= \frac{p(X = -1, Y = 1)}{p_X(X = -1)} = \frac{1}{16} \times \frac{16}{4} = \frac{1}{4} \\
p(Y = 2|X = -1) &= \frac{p(X = -1, Y = 2)}{p_X(X = -1)} = \frac{3}{16} \times \frac{16}{4} = \frac{3}{4} \\
p(Y = 4|X = -1) &= \frac{p(X = -1, Y = 4)}{p_X(X = -1)} = 0 \times \frac{16}{4} = 0 \\
p(Y = 1|X = 0) &= \frac{p(X = 0, Y = 1)}{p_X(X = 0)} = \frac{5}{16} \times \frac{16}{9} = \frac{5}{9} \\
p(Y = 2|X = 0) &= \frac{p(X = 0, Y = 2)}{p_X(X = 0)} = \frac{1}{16} \times \frac{16}{9} = \frac{1}{9} \\
p(Y = 4|X = 0) &= \frac{p(X = 0, Y = 4)}{p_X(X = 0)} = \frac{3}{16} \times \frac{16}{9} = \frac{1}{3} \\
p(Y = 1|X = 1) &= \frac{p(X = 1, Y = 1)}{p_X(X = 1)} = \frac{1}{16} \times \frac{16}{3} = \frac{1}{3} \\
p(Y = 2|X = 1) &= \frac{p(X = 1, Y = 2)}{p_X(X = 1)} = \frac{1}{16} \times \frac{16}{3} = \frac{1}{3} \\
p(Y = 4|X = 1) &= \frac{p(X = 1, Y = 4)}{p_X(X = 1)} = \frac{1}{16} \times \frac{16}{3} = \frac{1}{3}
\end{aligned}$$

**10.1** Calcola  $\mathbb{E}[X + 2]$ ,  $\mathbb{E}[Y^2]$  e  $\mathbb{E}[(X + 2)Y^2]$  per le variabili casuali  $X$  e  $Y$  dell'Esercizio 9.2.

$$\sum_x (x + 2)p_X(x) = \frac{4}{16} + \frac{18}{16} + \frac{9}{16} = \frac{31}{16}$$

$$\sum_y y^2 p_Y(y) = \frac{7}{16} + \frac{20}{16} + \frac{64}{16} = \frac{91}{16}$$

$$\sum_x \sum_y (x+2)y^2 p(x,y) = \frac{1}{16} + \frac{12}{16} + \frac{10}{16} + \frac{8}{16} + \frac{96}{16} + \frac{3}{16} + \frac{12}{16} + \frac{48}{16} = \frac{95}{8}$$

*Le due variabili casuali non sono indipendenti come si evince dal fatto che*

$$\frac{95}{8} = 11.875 \neq \frac{31 \cdot 91}{256} \approx 11.02$$

**10.2** Verifica che per la distribuzione di probabilità dell'**Esercizio 9.2**.

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

*Non scrivendo le somme con probabilità nulla e le somme nulle, abbiamo*

$$\begin{aligned} \mathbb{E}[X + Y] &= \sum_x \sum_y (x + y) p(x, y) \\ &= \frac{3}{16} + \frac{5}{16} + \frac{2}{16} + \frac{12}{16} + \frac{2}{16} + \frac{3}{16} + \frac{5}{16} = 2 \end{aligned}$$

*mentre*

$$\mathbb{E}[X] = \sum_x x p_X(x) = -\frac{4}{16} + \frac{3}{16} = -\frac{1}{16} \quad \text{e} \quad \mathbb{E}[Y] = \sum_y y p_Y(y) = \frac{7}{16} + \frac{10}{16} + \frac{16}{16} = \frac{33}{16}$$