Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

CATENE DI MARKOV E TEORIA DELLE GRANDI DEVIAZIONI

Relatore: Prof. PAOLO DAI PRA
Dipartimento di Matematica

Laureando:
WALTER GENCHI
Matricola N° 1100882

Anno Accademico 2016/2017

A mia nonna,
che per me è da sempre come una mamma.

Indice

Introduzione

Il presente lavoro di tesi vuole essere una descrizione quanto più semplice e accurata possibile di un argomento poco ricorrente nei corsi di una laurea triennale in Statistica: la Teoria delle Grandi Deviazioni, una branca della disciplina statistica che si occupa di calcolare le probabilità di eventi rari.

Vista la grande mole di teoremi e risultati preliminari indispensabili per le applicazioni nel calcolo delle probabilità di tali eventi rari, nel Capitolo 1 si è voluta inizialmente sviluppare una discussione sulle catene di Markov, particolari processi stocastici che hanno numerose applicazioni ai più diversi campi della scienze, fra cui la statistica. Seguendo l'impostazione generale della Teoria delle Grandi Deviazioni per una qualunque successione di variabili aleatorie, nel Capitolo 2 si è entrati specificatamente nel caso dei processi markoviani fino a giungere alla dimostrazione del Principio di Grandi Deviazioni per due particolari distribuzioni empiriche caratterizzanti una catena di Markov.

Si è scelto di trattare le catene di Markov essenzialmente per due ragioni: da un lato questi particolari processi stocastici forniscono applicazioni meno scontate rispetto al caso di variabili indipendenti e identicamente distribuite, che costituisce la parte più corposa e nota della Teoria delle Grandi Deviazioni; dall'altro lato, grazie ad alcuni risultati combinatori noti come *metodo dei tipi*, si è riuscito a dimostrare agilmente il Principio di Grandi Deviazioni restringendosi a questi particolari processi stocastici.

Si tratta di un lavoro che è stato svolto consultando i testi di riferimento indicati per la parte teorica e, all'occorrenza, approfondendo autonomamente le applicazioni più interessanti e talora sorprendenti nei contesti più disparati, di cui è stato fatto cenno alla fine del lavoro di tesi sia per contestualizzare i risultati teorici ottenuti nei capitoli precedenti, sia per stimolare la curiosità del lettore.

Durante la redazione del lavoro di tesi sono stati ripresi principi che in sé attingono originariamente ad altre discipline scientifiche: per esempio, è stata ampiamente trattata la nozione di entropia, un concetto proprio della fisica e della teoria dell'informazione, che tuttavia riveste un ruolo di grande importanza nella

definizione della *rate function* per una successione di variabili casuali. L'entropia relativa, una variante dell'entropia inizialmente definita, risulterà particolarmente utile nel Capitolo 3, dove si affronteranno le applicazioni della Teoria delle Grandi Deviazioni nella teoria dei test in Statistica.

1 Catene di Markov a stati finiti

In questo capitolo, dopo un semplice esempio e la definizione di un processo markoviano a stati finiti (tempo-discreto), si introdurrano concetti generali legati alle catene di Markov, quali le proprietà di irriducibilità e aperiodicità che, come si vedrà, sono entrambe condizioni necessarie per il Teorema di Convergenza. Si passerà a questo punto a stabilire l'esistenza e l'unicità della distribuzione stazionaria dopo averla opportunamente definita, enunciando altresì le equazioni di bilancio dettagliato. Infine si enunceranno il Teorema di Convergenza e quello Ergodico.
Si noti che due ben note varianti dei processi di Markov di seguito descritti sono le catene di Markov a tempo continuo e quelle con spazio degli stati infinito, ma numerabile. Tali varianti non veranno qui prese in esame.
Definizioni, enunciati e dimostrazioni sono stati ripresi da Levin *e altri* (2009).

1.1 Esempio

Supponiamo la situazione in cui un viaggiatore debba percorrere una piccola città formata da quattro strade e da quattro angoli, così come mostrato in Figura 1.1. Al tempo $t = 0$, l'uomo si trova all'angolo v_1 e deve a questo punto lanciare una moneta equilibrata: se esce testa si sposta all'angolo v_2, mentre se esce croce andrà all'angolo v_4. Supponiamo che esca testa e che quindi l'uomo si trovi all'angolo v_2 al tempo $t = 1$. A questo punto l'uomo lancia nuovamente la moneta per decidere in quale dei due angoli adiacenti spostarsi, seguendo la logica precedente: se esce testa si sposta in senso orario (in questo caso andrà all'angolo v_3), mentre se esce croce si sposta in senso antiorario (in questo caso tornerà all'angolo v_1). Questa procedura si ripete al tempo $t = 2, 3, \ldots$.

Sia X_t l'indice dell'angolo in cui l'uomo si trova al tempo t e sia quindi $(X_0, X_1, \ldots)$ un processo stocastico che assume valori nell'insieme $\{1, 2, 3, 4\}$. Per le ipotesi fatte precedentemente si ha che $\mathbf{P}(X_0 = 1) = 1$, mentre $\mathbf{P}(X_1 = 2) = \frac{1}{2}$ e $\mathbf{P}(X_1 = 4) = \frac{1}{2}$.

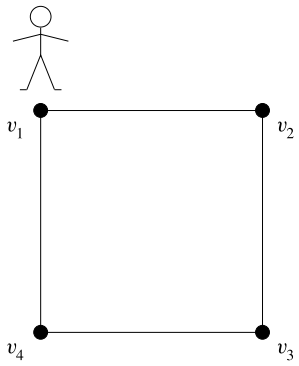

Figura 1.1: Passeggiata aleatoria in una piccola città

Per calcolare la distribuzione di X_t per $t \geq 2$ è utile considerare le probabilità condizionate. Se supponiamo che al tempo t l'uomo si trovi all'angolo v_3, si ha che $\mathbf{P}(X_{t+1} = 4 \mid X_t = 3) = \frac{1}{2}$ e $\mathbf{P}(X_{t+1} = 2 \mid X_t = 3) = \frac{1}{2}$, cioè il lancio della moneta al tempo $t+1$ è indipendente da tutti i lanci precedenti e la distribuzione condizionale di X_{t+1} dipende solo da X_t.

1.2 Definizione

Una catena di Markov a stati finiti è un processo stocastico per cui se x e y appartengono all'insieme finito Ω, il passaggio dallo stato x allo stato y dipende solo dallo stato x e non dai precedenti.
Più precisamente, sia $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità, dove Ω è lo spazio finito degli stati, $\mathcal{F}$ una σ-algebra su Ω e P una misura di probabilità. Una sequenza di variabili casuali $(X_0, X_1, \ldots)$ è una catena di Markov con **spazio degli stati** Ω e **matrice di transizione** P se per tutti gli $x, y \in \Omega$, $t \geq 1$ e tutti gli eventi $H_{t-1} = \cap_{s=0}^{t-1} \{X_s = x_s\}$ tali per cui $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, si ha

$$\mathbf{P}\{X_{t+1} = y \mid H_{t-1} \cap \{X_t = x\}\} = \mathbf{P}\{X_{t+1} = y \mid X_t = x\} = P(x, y). \quad (1.1)$$

La passeggiata aleatoria dell'esempio precedente è una catena di Markov con spazio degli stati $\Omega = \{1, 2, 3, 4\}$ e matrice di transizione

$$P = \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

La 1.1 è spesso chiamata **proprietà di Markov**, che stabilisce che la probabilità condizionata di passare dallo stato x allo stato y non dipende dalla sequenza $x_0, x_1, \ldots, x_{t-1}$ di stati che precede lo stato corrente x. Si noti che il termine destro della 1.1 non dipende dal tempo t. In questo caso la catena di Markov è detta **omogenea**. Ciò significa che la matrice di transizione è la stessa ad ogni passo t.

Si consideri l'x-esima riga di P, che corrisponde alla distribuzione di probabilità $P(x, \cdot)$, dove $P(x, y)$ è la probabilità di passare dallo stato x allo stato y per ogni $x, y \in \Omega$. La matrice di transizione P è **stocastica** essendo che

$$P(x, y) > 0 \ \forall \ x, y \in \Omega \quad \text{e} \quad \sum_{y \in \Omega} P(x, y) = 1 \ \forall \ x \in \Omega.$$

Sia μ_0 la **distribuzione iniziale** del processo $(X_0, X_1, \ldots)$, ossia il vettore riga dato da

$$\mu_0(x) = \mathbf{P}\{X_0 = x\} \ \text{ per tutti gli } x \in \Omega.$$

Nell'esempio della passeggiata aleatoria la distribuzione iniziale è $\mu_0 = (1, 0, 0, 0)$, essendo che $\mathbf{P}(X_0 = 1) = 1$ e $\mathbf{P}(X_0 = 2) = \mathbf{P}(X_0 = 3) = \mathbf{P}(X_0 = 4) = 0$.

Si può mostrare per induzione su $t \geq 0$ che il vettore riga μ_t, ossia la distribuzione di X_t, può essere scritto come

$$\mu_t = \mu_0 P^t \text{ per t} \geq 0. \tag{1.2}$$

Ciò significa che la probabilità di muoversi in t passi da x a y è data dall'elemento di posizione (x, y) della matrice P^t.

1.3 Irriducibilità e Aperiodicità

Per ottenere i risultati più interessanti nella teoria delle catene di Markov, è utile fissare alcune ipotesi sulle catene di Markov che prenderemo in esame: irriducibilità e aperiodicità.
Una catena markoviana si dice **irriducibile** se, presi comunque due stati $x, y \in \Omega$ esiste un intero t tale che $P^t(x, y) > 0$. La condizione di irriducibilità garantisce quindi che è possibile raggiungere qualunque stato da qualunque stato iniziale con una probabilità positiva.

Si consideri per esempio una catena di Markov $(X_0, X_1, \ldots)$ con spazio degli

stati $\Omega = \{1, 2, 3, 4\}$ e matrice di transizione

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.3 & 0.7 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \\ 0 & 0 & 0.8 & 0.2 \end{bmatrix}.$$

Osservando la matrice di transizione P, si nota che se la catena inizia nello stato 1 o 2, allora non potrà mai visitare lo stato 3 o 4. Viceversa, se inizia nello stato 3 o 4, non potrà mai lasciare il sottoinsieme $\{3, 4\}$. Ne consegue che la catena non è irriducibile.
Si noti che, se la catena inzia nello stato 1 o 2, il suo comportamento è esattamente uguale a quello di una catena di Markov con spazio degli stati $\{1, 2\}$ e matrice di transizione

$$\begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \end{bmatrix}.$$

Se invece la catena ha il suo stato inziale in 3 o 4, questa si comporta esattamente come un catena di Markov con spazio degli stati $\{3, 4\}$ e matrice di transizione

$$\begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix}.$$

Questo breve esempio mostra come la condizione di riducibilità di una catena di Markov permetta di studiare il suo comportamento nel lungo periodo considerando una catena con spazio degli stati più semplice.

La definizione di aperiodicità richiede prima che si definisca il **periodo** dello stato x, cioè il massimo comun divisore di $\mathcal{T}(x) := \{t \geq 1 : P^t(x, x) > 0\}$, l'insieme che contiene il numero di passi in cui la catena può ritornare al punto iniziale x. Se la catena gode della proprietà di irriducibilità, il periodo viene definito come il periodo comune a tutti gli stati. In definitiva, una catena è **aperiodica** se tutti gli stati hanno periodo 1, mentre è **periodica** se non è aperiodica.

Si consideri l'esempio della passeggiata aleatoria della Figura 1.1, dove il viaggiatore si trova nell'angolo v_1 al tempo $t = 0$. Affinché egli ritorni all'angolo v_1, è necessario fare un numero pari di passi. Ciò significa che il massimo comun divisore di $\mathcal{T}(v_1) = \{t \geq 1 \colon P^t(v_1, v_1) > 0\} = \{2, 4, 6, \ldots\}$ è 2 e quindi la catena in questione è periodica.

1.4 Stazionarietà

La definizione di una distribuzione stazionaria parte da alcune considerazioni sul comportamento della distribuzione μ_t, così come definita nella 1.2, nel lungo periodo. Si intende cioè studiare il comportamento di μ_t per $t \to \infty$.
Una distribuzione di probabilità π che soddisfa

$$\pi = \pi P \tag{1.3}$$

è detta **distribuzione stazionaria** della catena di Markov con matrice di transizione P. Poiché la definizione di stazionarietà coinvolge solo la matrice di transizione P, spesso si dice che la distribuzione π che soddisfa la 1.3 è stazionaria per P.
In prima battuta si può verificare che π **esiste**, cioé che il numero di passi richiesto affinché μ_t converga a π è finito. In seconda battuta si può altresì mostrare che esiste un'**unica** distribuzione stazionaria π che soddisfa la 1.3. Infine viene affrontata la questione della **convergenza** di π dal Teorema di Convergenza.
Spesso il modo più semplice per mostrare che una particolare distribuzione è stazionaria consiste nel verificare se questa soddisfa le **equazioni di bilancio dettagliato**, particolarmente usate nella meccanica statistica. Una distribuzione π è stazionaria per P se soddisfa l'equazione

$$\pi(x) P(x,y) = \pi(y) P(y,x) \text{ per tutti gli } x, y \in \Omega. \tag{1.4}$$

Si noti che la condizione di bilancio dettagliato è sufficiente ma non necessaria alla stazionarietà: esistono infatti catene di Markov che hanno distribuzioni stazionaria, ma non soddisfano le equazioni di bilancio dettagliato. Si consideri per esempio la catena di Markov $(X_0, X_1, \ldots)$ con spazio degli stati $\Omega = \{1, 2, 3\}$ e matrice di transizione

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Si può provare che la distribuzione di probabilità $\pi = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ soddisfa la 1.3 ed è l'unica distribuzione stazionaria della catena di Markov in questione. Tuttavia, non viene soddifatta la condizione di bilancio dettagliato per $x = 1$ e $y = 2$:

$$\pi(1) P(1,2) = \frac{1}{3} \cdot 1 \neq \frac{1}{3} \cdot 0 = \pi(2) P(2,1).$$

1.5 Teorema di Convergenza

Gran parte dell'attenzione nello studio della convergenza di una catena di Markov irriducibile e aperiodica è rivolta alla stima della velocità di tale convergenza. Il Teorema di Convergenza in questo senso fornisce un limite superiore (dipendente dal tempo t) della "distanza" tra P^t e π, che deve avvicinarsi il più possibile allo zero.

Si esprimano questi in concetti in termini matematici definendo la **distanza in variazione totale** tra due distribuzioni di probabilità μ e ν, entrambe definite nello spazio Ω:

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} \mid \mu(A) - \nu(A) \mid = \frac{1}{2} \sum_{x \in \Omega} \mid \mu(x) - \nu(x) \mid . \tag{1.5}$$

L'intepretazione di tale distanza è puramente probabilistica: si misura infatti la massima distanza tra le probabilità assegnate dalle due distrubuzioni al singolo evento A.

Sia P la matrice di transizione di una catena di Markov irriducibile e aperiodica, con spazio degli stati Ω e una distribuzione stazionaria π. Sotto queste ipotesi, il **Teorema di Convergenza** afferma che esistono due costanti $\alpha \in (0,1)$ e $C > 0$ tali per cui

$$\max_{x \subset \Omega} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t. \tag{1.6}$$

1.6 Teorema Ergodico

La teoria ergodica si occupa principalmente dello studio matematico del comportamento medio, a lungo termine, di sistemi dinamici. L'ipotesi ergodica è che la media "temporale" di ogni osservabile del sistema debba coincidere con la media "spaziale" di tale osservabile rispetto a un'opportuna distribuzione di probabilità sullo spazio delle configurazioni del sistema. Il **Teorema Ergodico** estende un'idea analoga alle catene di Markov.

Sia f una funzione a valore reali definita su Ω. Se (X_t) è una catena di Markov irriducibile, allora per ogni distribuzione μ,

$$\mathbf{P}_\mu \left\{ \lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} f(X_s) = E_\pi(f) \right\} = 1. \tag{1.7}$$

2 La Teoria delle Grandi Deviazioni

L'obbiettivo di questo capitolo è di intodurre la Teoria delle Grandi Deviazioni, una branca della teoria della probabilità che si occupa di definire e analizzare gli eventi rari. Le origini di questa disciplina si fanno risalire agli anni 30' del secolo scorso, in particolare agli studi in ambito attuariale di H. Cramér, matematico e statistico svedese, che affronta il caso di variabili aleatorie a valori reali indipendenti. In seguito altri studiosi hanno esteso l'applicazione dei risultati ottenuti da Cramér a contesti più generali, come per esempio al caso dei processi markoviani.

Inizialmente si enuncierà il prinicipio delle grandi deviazioni (LDP) in termini di *rate function* per una generica successione di variabili aleatorie, fornendo un'interpretazione dei risultati teorici ottenuti attraverso alcuni esempi. In seguito verrano trattate più approfonditamente le applicazioni della Teoria delle Grandi Deviazioni alle catene di Markov, argomento apparentemente ben più complesso rispetto al caso di variabili indipendenti e identicamente distribuite, per le quali vale il teorema di Sanov (1957). In realtà, si riuscirà agilmente a dimostrare il Principio delle Grandi Deviazioni per processi markoviani sfruttando alcuni risultati combinatori ottenuti grazie al *metodo dei tipi*, uno strumento chiave introdotto da Claude Shannon nella teoria dell'informazione.

I risultati generali della Teoria delle Grandi Deviazioni sono stati ripresi da Dembo e Zeitouni (2010), mentre l'approfondimento sulle applicazioni dei risultati generali alle catene di Markov sono stati derivati da Vidyasagar (2009).

2.1 Esempio particolare

Sia $X_1, \ldots, X_n$ una successione di variabili casuali indipendenti e identicamente distribuite. Per semplicità assumiamo che $X_i \sim N(0,1)$ per ogni $i = 1, \ldots n$. Si consideri la media campionaria $\hat{S}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, anche questa distribuita normalmente con media nulla e varianza pari a $1/n$. Per la legge (debole) dei grandi numeri, si ha che per ogni $\delta > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\mid \hat{S}_n \mid \geq \delta\right) = 0,$$

ossia la media campionaria converge in probabilità a 0, la media comune delle X_i.

Si ha inoltre che, per ogni intervallo A,

$$\lim_{n\to\infty} \mathbf{P}\left(\sqrt{n}\hat{S}_n \in A\right) = \frac{1}{\sqrt{2\pi}} \int_A e^{-x^2/2} dx, \tag{2.1}$$

poiché la variabile casuale $\sqrt{n}\hat{S}_n \sim N(0,1)$ e più in generale per il Teorema del Limite Centrale.

Si noti ora che

$$\mathbf{P}\left(\mid \hat{S}_n \mid \geq \delta\right) = \mathbf{P}\left(\mid \sqrt{n}\hat{S}_n \mid \geq \delta\sqrt{n}\right) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{+\delta\sqrt{n}} e^{-x^2/2} dx. \tag{2.2}$$

La 2.1 è la probabilità di deviazioni moderate di $\hat{S}_n$ dalla sua media, mentre la 2.2 è la probabilità di *grandi* deviazioni di $\mid \hat{S}_n \mid$ dalla sua media.

Per calcolare il comportamento asintotico dell'integrale della 2.2 si fa ricorso al *Principio di Laplace*, un teorema basilare nella Teoria delle Grandi Deviazioni e particolarmente usato nella meccanica statistica per determinare il comportamento di un sistema quando la temperatura tende allo zero assoluto.
Sia A un sottoinsieme Lebesgue-misurabile dello spazio euclideo d-dimensionale $\mathbb{R}^d$ e sia $\varphi : \mathbb{R}^d \to \mathbb{R}$ una funzione misurabile con $\int_A e^{-\varphi(x)} dx < \infty$. Si ha allora che

$$\lim_{n\to\infty} \frac{1}{n} \log \int_A e^{-n\varphi(x)} dx = -\inf_{x\in A} \varphi(x). \tag{2.3}$$

In questo caso, siano $A = [-\delta, +\delta]$, $A_n = [-\delta\sqrt{n}, +\delta\sqrt{n}]$ e A^c e A_n^c i loro complementari. Riscriviamo la 2.2 effettuando un cambio di variabili ($x = \sqrt{n}y$) in modo da ottenere un'espressione simile all'integrale della 2.3:

$$1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{+\delta\sqrt{n}} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{A_n^c} e^{-x^2/2} dx = \sqrt{\frac{n}{2\pi}} \int_{A^c} e^{-ny^2/2} dx.$$

Si ottiene quindi, coerentemente con la 2.3:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\mid \hat{S}_n \mid \geq \delta\right) &= \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(\mid \hat{S}_n \mid \in A^c\right) \\ &= -\inf_{y\in A^c} \left\{\frac{y^2}{2}\right\} = -\frac{\delta^2}{2}, \end{aligned} \tag{2.4}$$

dove il termine $\frac{1}{n} \log \sqrt{\frac{n}{2\pi}} \underset{n\to\infty}{\longrightarrow} 0$.

Osservando la 2.4 si coglie il significato dell'oggetto di studio della Teoria delle Grandi Deviazioni, cioè gli eventi rari: all'aumentare della numerosità campionaria, mentre $\hat{S}_n$ assume in media valori dell'ordine $1/\sqrt{n}$, al contrario $\mid \hat{S}_n \mid$

assume valori relativamente grandi con probabilità basse (dell'ordine $e^{-n\delta^2/2}$).

Questo esempio tratta variabili normali indipendenti e identicamente distribuite, ma i risultati trovati possono essere estesi sia ad altre distribuzioni, per le quali valgono i risultati di Cramér, sia al caso di leggera dipendenza tra le variabili coinvolte, assumendo per esempio che le X_i siano realizzazioni di un processo di Markov.

2.2 Principio delle Grandi Deviazioni

Sia $X_1, \ldots, X_n$ una successione di variabili casuali definite su uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ che assume valori in $\mathcal{X}$, uno spazio polacco, cioè separabile e completamente metrizzabile.

La Teoria delle Grandi Deviazioni si concentra sulle variabili casuali $X_1, \ldots X_n$ per cui $\mathbf{P}(X_i \in A)$ converge esponenzialmente a zero per una classe A di σ-algebre di Borel. La velocità di convergenza di queste probabilità è espressa in termini di una funzione I, detta *rate function*, se $I \colon \mathcal{X} \to [0, \infty)$ e se per ogni $M < \infty$ l'insieme di livello $\{x \in \mathcal{X} : I(x) \leq M\}$ è un sottoinsieme chiuso di $\mathcal{X}$. Una *rate function* che soddisfa la seconda proprietà è semicontinua inferiormente.

La successione $X_1, \ldots, X_n$ con una *rate function* I soddisfa il Principio delle Grandi Deviazioni se valgono le due seguenti disuguaglianze:

1. Per ogni insieme aperto $\Gamma \subseteq \mathcal{X}$, si ha

$$-\inf_{x \in \Gamma} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}(X_n \in \Gamma) \tag{2.5}$$

2. Per ogni insieme chiuso $\Gamma \subseteq \mathcal{X}$, si ha

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}(X_n \in \Gamma) \leq -\inf_{x \in \Gamma} I(x) \tag{2.6}$$

Si può mostrare che esiste un'unica *rate function* che soddisfa 2.5 e 2.6.
Sia Γ^o la parte interna di Γ e sia $\bar{\Gamma}$ la chiusura di Γ. La 2.5 e la 2.6 posso essere combinate:

$$\begin{aligned} -\inf_{x \in \Gamma^o} I(x) &\leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}(X_n \in \Gamma) \\ &\leq \limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}(X_n \in \Gamma) \\ &\leq -\inf_{x \in \bar{\Gamma}} I(x) \end{aligned} \tag{2.7}$$

Se I è continua (non solo semicontinua inferiormente) e Γ non ha punti isolati, cioé $\Gamma \subseteq \bar{\Gamma^o}$, allora i due estremi inferiori coincidono e quindi la 2.7 diventa

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left(X_n \in \Gamma\right) = -\inf_{x\in\bar{\Gamma}} I\left(x\right), \tag{2.8}$$

che è per esempio il caso del paragrafo 2.1, dove la *rate function* $I\left(x\right) = \frac{x^2}{2}$ e $\Gamma = (-\infty, \delta) \cup (\delta, +\infty)$. Per i processi markoviani è possibile determinare soltanto il limite inferiore e superiore, non un'uguaglianza come nella 2.8.

2.3 Applicazioni alle catene di Markov

Il Principio delle Grandi Deviazioni è valido anche per variabili aleatorie che assumono valori in insieme finito e che presentano una forma di dipendenza, quali le catene di Markov. È bene a questo proposito fissare la notazione, esplicitando alcune considerazioni sui processi markoviani a stati finiti che erano state omesse per semplicità di lettura nel primo capitolo.

Notazione

Sia $X_1, \ldots, X_n$ un processo markoviano a stati finiti, cioè che assume valori in un insieme finito $\mathbb{A} = \{a_1, a_2, \ldots a_n\}$; per semplicità scriveremo che $\mathbb{A} = \{1, 2, \ldots n\}$. Nella Teoria delle Grandi Deviazioni $\mathbb{A}$ è detto *alfabeto* e $\mid \mathbb{A} \mid = n$ è la cardinalità dell'insieme. $\mathcal{M}\left(\mathbb{A}\right)$ denota lo spazio di tutte le misure di probabilità sull'alfabeto $\mathbb{A}$ ed è identificato con il simplesso standard $\mathbb{S}_n := \left\{ \mathbf{v} \in [0,1]^n : \sum_{i=1}^{n} v_i = 1 \right\}$, cioè l'insieme di tutti i vettori n-dimensionali a valori in $[0,1]$ con componenti la cui somma è pari a 1.
Definiamo $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A}^2)$ come il "vettore delle frequenze doppie" con generico elemento

$$\mu_{ij} = \mathbf{P}\left(X_t X_{t+1} = ij\right), \forall i, j \in \mathbb{A} \tag{2.9}$$

e di conseguenza la distribuzione stazionaria delle catena di Markov è data dal vettore $\bar{\boldsymbol{\mu}} \in \mathbb{S}_n$ con generico elemento

$$\bar{\mu}_i = \sum_{j\in\mathbb{A}} \mu_{ij} = \sum_{j\in\mathbb{A}} \mu_{ji}, \tag{2.10}$$

mentre la matrice di transizione A ha generico elemento

$$a_{ij} = \frac{\mu_{ij}}{\bar{\mu}_i}.$$

Assumiamo qui e nel seguito che $\boldsymbol{\mu}$ sia la distribuzione di equilibrio della catena di Markov (altrimenti $\boldsymbol{\mu}$ dipenderebbe da t) e che contenga tutta l'informazione rilevante sulla catena di Markov, utilizzando la definizione di distribuzione stazionaria su $\mathbb{A}^k$, che corrisponde esattamente alla 2.10 quando $k = 2$. Si dirà quindi che $\boldsymbol{\mu} \in \mathcal{M}_s\left(\mathbb{A}^k\right)$, cioè all'insieme di tutte le distribuzione stazionarie su $\mathbb{A}^k$.
Le distribuzioni di probabilità verrano indicate con le lettere greche in grassetto.

Entropia

La nozione di informazione associata ad una distribuzione di probabilità discreta finita è stata elaborata da Claude Shannon in un celebre articolo del 1948. Nella teoria dell'informazione l'entropia misura la quantità di incertezza di una variabile casuale. Nella Teoria delle Grandi Deviazioni il concetto di entropia è cruciale per calcolare la *rate function* che, per le catene di Markov, è esattamente uguale all'entropia relativa condizionale.

Data una distribuzione $\boldsymbol{\nu} \in \mathbb{S}_n$, la sua entropia è definita come

$$H(\boldsymbol{\nu}) := \sum_{i=1}^{n} \nu_i \log(1/\nu_i) = -\sum_{i=1}^{n} \nu_i \log \nu_i,$$

dove $0 \log 0 = 0$ per convenzione, giustificato dal fatto che $x \log x \to 0$ per $x \to 0$.

Siano $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{S}_n$, allora la loro entropia relativa è definita come

$$D(\boldsymbol{\nu} \parallel \boldsymbol{\mu}) = J(\boldsymbol{\nu}, \boldsymbol{\mu}) - H(\boldsymbol{\nu}), \tag{2.11}$$

dove $J(\boldsymbol{\nu}, \boldsymbol{\mu})$ è una "funzione di perdita" definita da

$$J(\boldsymbol{\nu}, \boldsymbol{\mu}) := \sum_{i=1}^{n} \nu_i \log(1/\mu_i). \tag{2.12}$$

È noto che $D(\cdot \parallel \cdot)$ è una funzione strettamente convessa per entrambi gli argomenti, cioè per ogni fissato $\boldsymbol{\mu}$, $D(\boldsymbol{\nu} \parallel \boldsymbol{\mu})$ è una funzione strettamente convessa di $\boldsymbol{\nu}$. L'entropia relativa corrisponde di fatto alla misura della "distanza" tra due distribuzioni di probabilità, anche se non è simmetrica e non soddisfa la disuguaglianza triangolare. Da un punto di vista statistico, l'entropia relativa $D(\boldsymbol{\nu} \parallel \boldsymbol{\mu})$ misura l'inefficienza di assumere che la distribuzione sia $\boldsymbol{\mu}$, quando la vera distribuzione è $\boldsymbol{\nu}$.

Infine si introduce l'entropia condizionale di X (a valori in $\mathbb{A} = \{1, \dots n\}$) dato Y (a valori in $\mathbb{B} = \{1, \dots m\}$), cioè l'entropia di X condizionata a Y, definita

come

$$H(X \mid Y) := \sum_{j=1}^{m} \mathbf{P}(Y = j) \cdot H(X \mid Y = j). \tag{2.13}$$

Si possono mostrare alcune proprietà dell'entropia condizionale, che risulteranno utili in seguito per alcune dimostrazioni. In particolare, si può ottenere la scomposizione

$$H(X \mid Y) = H(X,Y) - H(Y). \tag{2.14}$$

Si può estendere il concetto di entropia ai processi stocastici stazionari $\{X_t\}_{t\geq 0}$, dove ogni X_t assume valori nell'insieme finito $\mathbb{A}$. Si può calcolare per esempio $H\left(X_t \mid X_0^{t-1}\right)$, ossia l'entropia condizionale della variabile "corrente" X_t date le variabili "passate" $X_0^{t-1} := X_0 X_1 \dots X_{t-1}$. Essendo il processo stazionario, l'entropia condizionale non dipende da t, ovvero

$$H\left(X_{t+\tau} \mid X_t^{t+\tau-1}\right) = H\left(X_t \mid X_0^{t-1}\right). \tag{2.15}$$

Inoltre, $H\left(X_t \mid X_0^{t-1}\right)$ è una funzione monotona decrescente di t, cioè l'entropia non aumenta se si condiziona rispetto a più variabili "passate". Essendo $H\left(X_t \mid X_0^{t-1}\right)$ una distribuzione di probabilità, essa è limitata inferiormente dallo zero, quindi esiste una costante $c > 0$ tale che

$$H\left(X_t \mid X_0^{t-1}\right) \to c \text{ per } t \to \infty.$$

La costante c viene chiamata entropia del processo.
Appare evidente che se le variabili del processo $X_1, X_2, \dots$ sono indipendenti e identicamente distribuite, si ha che

$$H\left(X_t \mid X_0^{t-1}\right) = H(X_t) = H(X_1), \forall t \geq 1.$$

Proposizione: Se $\{X_t\}$ è una catena di Markov così come definita in 2.3, la sua entropia sarà data da

$$c = H(\boldsymbol{\mu}) - H(\bar{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \bar{\mu}_i H\left(\mathbf{a}^i\right), \tag{2.16}$$

dove $\mathbf{a}^i$ denota l'i-esima riga della matrice di transizione A.

Dimostrazione: Essendo $\{X_t\}$ una catena di Markov, per la proprietà di Markov della 1.1, per la proprietà dell'entropia condizionale di un processo stazionario

espressa nella 2.15 e per quanto visto nella 2.14, si ha che

$$\begin{aligned} H\left(X_t \mid X_0^{t-1}\right) &= H\left(X_t \mid X_{t-1}\right) = H\left(X_1 \mid X_0\right) \\ &= H\left(X_1, X_0\right) - H\left(X_0\right). \end{aligned} \tag{2.17}$$

Si noti che il primo e il secondo termine della 2.17 corrispondono rispettivamente a $H(\boldsymbol{\mu})$ e $H(\bar{\boldsymbol{\mu}})$ per come sono stati definiti precedentemente, ovvero $\boldsymbol{\mu}$ come distribuzione di probabilità di (X_0, X_1) e $\bar{\boldsymbol{\mu}}$ come distribuzione di probabilità di X_0. Dalla 2.13 si ottiene la seconda uguaglianza:

$$\begin{aligned} H\left(X_1 \mid X_0\right) &= \sum_{i}^{n} \mathbf{P}\left(X_0 = i\right) \cdot H\left(\mathbf{P}\left(X_1 \mid X_0 = i\right)\right) \\ &= \sum_{i=1}^{n} \mu_i \bar{H}\left(\mathrm{a}^i\right). \end{aligned}$$

■

Introduciamo adesso due quantità che saranno fondamentali più avanti per la dimostrazione della *rate function* di un processo markoviamo. La prima ricorda la definizione di entropia condizionale per un variabile aleatoria data nella 2.14, mentre la seconda quantità riprende il concetto di entropia relativo per due distribuzioni di probabilità.
L'entropia condizionale del "vettore delle frequenze doppie" $\boldsymbol{\mu} \in \mathcal{M}_s\left(\mathbb{A}^2\right)$ è definita come

$$H_c\left(\boldsymbol{\mu}\right) := H\left(\boldsymbol{\mu}\right) - H\left(\bar{\boldsymbol{\mu}}\right), \tag{2.18}$$

mentre la quantità

$$D_c\left(\boldsymbol{\nu}\right) := D\left(\boldsymbol{\nu} \parallel \boldsymbol{\mu}\right) - D\left(\bar{\boldsymbol{\nu}} \parallel \bar{\boldsymbol{\mu}}\right) \tag{2.19}$$

è detta entropia relativa condizionale tra $\boldsymbol{\nu} \in \mathcal{M}_s\left(\mathbb{A}^2\right)$ e $\boldsymbol{\mu}$.

Proprietà di equipartizione asintotica

Verrà qui approfondito un teorema molto famoso nell'ambito della teoria dell'informazione e strettamente connesso con le applicazioni della Teoria delle Grandi Deviazioni ai processi stocastici e, nel caso specifico, alle catene di Markov. La proprietà di equipartizione asintotica (AEP) gioca un ruolo fondamentale nella definizione di *insieme tipico* e in generale nella teoria della compressione, una tecnica di elaborazione di dati che, grazie ad alcuni specifici algoritmi, permette la riduzione della quantità di bit necessari alla rappresentazione in forma digi-

tale di un'informazione. Si rinvia a Cover e Thomas (2012) per un'analisi più approfondita sul teorema di Shannon-McMillan-Breiman (AEP generale).

L'esempio più semplice per comprendere la logica della AEP è il caso di una successione $X_1, \ldots, X_n$ di variabili casuali indipendenti e identicamente distribuite, dove ciascuna variabile assume soltante due valori, cioè $X_i \in \{0, 1\} \, \forall i = 1, \ldots, n$. Si ha inoltre che $\mathbf{P}(X_i = 1) = p$ e $\mathbf{P}(X_i = 0) = q \; \forall i = 1, \ldots, n$.
Supponiamo che $n = 6$ e di avere osservato $(1, 0, 1, 1, 0, 1)$. La probabilità di osservare $(1, 0, 1, 1, 0, 1)$ è pari a

$$\prod_{i=1}^{6} \mathbf{P}(X_i = x_i) = p^{\sum_{i=1}^{6} X_i} q^{6 - \sum_{i=1}^{6} X_i} = p^4 q^2.$$

Essendo $p \neq q$ le 2^6 possibili osservazioni di lunghezza 6 non hanno tutte la stessa probabilità di essere osservate. Ci si può tuttavia chiedere se la successione osservata proviene con una "alta probabilità" da un sottoinsieme di successioni (*insieme tipico* nella teoria dell'informazione) che hanno tutte la stessa probabilità di essere osservate. Si tratta di fatto di una conseguenza della Legge dei grandi numeri e della Teoria Ergodica. La proprietà di equipartizione asintotica afferma in questo caso (variabili indipendenti e identicamente distribuite) che

$$-\frac{1}{n} \log \mathbf{P}(X_1 = x_1, \ldots X_n = x_n) \to H(X) \text{ in probabilità}, \tag{2.20}$$

dove X è la generica distribuzione del processo. Quindi, affinché la 2.20 sia soddisfatta, $\mathbf{P}(X_1 = x_1, \ldots X_n = x_n) \approx e^{-nH(X)}$ con "alta probabilità". Si ritroverà un'espressione simile quando si introdurrà il *metodo dei tipi*, una procedura ancora più forte, dove si considerano solo le successioni che hanno la stessa distribuzione empirica di quella osservata.

Rate function per misure di coppia

In questa sezione deriveremo la *rate function* per le misure di coppia inizialmente e in seguito estenderemo il risultato per la *rate function* della distribuzione stazionaria grazie al principio di contrazione.

Supponiamo di avere osservato i valori $x_1^l = x_1 \ldots x_l$. L'obiettivo in questo caso è stimare una distribuzione empirica del "vettore delle frequenze doppie" $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A}^2)$, così come è stata definita nella 2.9. Esistono almeno due distribuzioni empiriche per stimare $\boldsymbol{\mu}$.

Una possibilità è quella di costruire un vettore $\boldsymbol{\theta} \in \mathbb{S}_{n^2}$ con generico elemento

$$\theta_{ij}\left(x_1^l\right) := \frac{1}{l-1}\sum_{t=1}^{l-1} I_{\{X_t X_{t+1}=ij\}}, \tag{2.21}$$

$$\text{dove } I_{\{X_t X_{t+1}=ij\}} = \begin{cases} 1 & \text{se } X_t X_{t+1} = ij \\ 0 & \text{altrimenti} \end{cases}.$$

Il problema dell'uso di $\boldsymbol{\theta}$ per stimare $\boldsymbol{\mu}$ è che in generale $\boldsymbol{\theta}$ non è stazionaria. Si può mostrare infatti che in generale non vale la seconda uguaglianza della 2.10 per $\boldsymbol{\theta}$ e $\bar{\boldsymbol{\theta}}$, dove quest'ultima è definita come

$$\bar{\theta}_i := \frac{1}{l-1}\sum_{t=1}^{l-1} I_{\{X_t=i\}}.$$

Intuitivamente, il problema in questo caso è che l'osservazione x_1^l è composta da l numeri, ma possono essere fatti soltanto $l-1$ confronti a coppie. Sembra quindi sensato modificare il campione x_1^l aggiungendo una transizione "fantasma" da x_l a x_1, così da creare un ciclo. Si ottiene così il vettore $\boldsymbol{\nu}\left(x_1^l\right)$ con generiche componenti

$$\nu_{ij}\left(x_1^l\right) := \frac{1}{l}\sum_{t=1}^{l} I_{\{x_t x_{t+1}=ij\}}, \forall i,j \in \mathbb{A} \tag{2.22}$$

dove $x_{t+1} = x_1$ per quanto detto precedentemente. In questo caso si dimostra che il vettore $\boldsymbol{\nu}$ soddisfa la 2.10 per $\boldsymbol{\nu}$ e $\boldsymbol{\phi} \in \mathbb{S}_n$, dove quest'ultima è una misura empirica della distribuzione stazionaria del processo $\{X_t\}$ che ha generico elemento

$$\phi_j\left(x_1^l\right) := \frac{1}{l}\sum_{t=1}^{l} I_{\{x_t=j\}}, \forall j \in \mathbb{A}. \tag{2.23}$$

In altre parole $\boldsymbol{\phi}\left(x_1^l\right)$ non è altro che un vettore contenente alla posizione j-esima la frequenza del j-esimo valore dell'insieme $\mathbb{A}$.

Esistono le forme esplicite delle *rate function* sia per $\boldsymbol{\theta}\left(x_1^l\right)$ che per $\boldsymbol{\nu}\left(x_1^l\right)$, ma per la prima si distingue il caso in cui $\boldsymbol{\theta}\left(x_1^l\right) \in \mathcal{M}_s\left(\mathbb{A}^2\right)$ da quello in cui $\boldsymbol{\theta}\left(x_1^l\right) \notin \mathcal{M}_s\left(\mathbb{A}^2\right)$. Di seguito verrà enunciata e dimostrata la forma esplicita della *rate function* di $\boldsymbol{\nu}\left(x_1^l\right)$.

Sia $\{X_t\}$ un processo di Markov, così come è stato definito nella sezione 2.3, e siano $\boldsymbol{\mu} \in \mathcal{M}_s(\mathbb{A}^2)$ (con $\mu_{ij} > 0$, cioè con generico elemento della matrice di transizione $a_{ij} > 0\ \forall i,j \in \mathbb{A}$) e $\bar{\boldsymbol{\mu}} \in \mathbb{S}_n$ il "vettore delle frequenze doppie" e la distribuzione stazionaria del processo, così come definite nella 2.9 e nella 2.10

rispettivamente. Allora il processo $\left\{\boldsymbol{\nu}\left(x_1^l\right)\right\}$ soddisfa il Principio delle Grandi Deviazioni con *rate function* pari a

$$I\left(\boldsymbol{\nu}\right) = \sum_{i\in\mathbb{A}} \bar{\nu}_i \sum_{j\in\mathbb{A}} c_{ij} \log\left(\frac{c_{ij}}{a_{ij}}\right), \tag{2.24}$$

dove $\bar{\boldsymbol{\nu}} \in \mathbb{S}_n$ rispetto a $\boldsymbol{\nu}$ è l'analogo di $\bar{\boldsymbol{\mu}} \in \mathbb{S}_n$ rispetto a $\boldsymbol{\mu}$, precedentemente definiti. Inoltre si ha che

$$c_{ij} = \frac{\nu_{ij}}{\bar{\nu}_i}.$$

Si può dimostrare che la *rate function* appena trovata può essere scritta come la entropia relativa condizionale tra $\boldsymbol{\nu}$ e $\boldsymbol{\mu}$, così come è stata definita nella 2.19:

$$\begin{aligned}
D_c\left(\boldsymbol{\nu} \parallel \boldsymbol{\mu}\right) &= D\left(\boldsymbol{\nu} \parallel \boldsymbol{\mu}\right) - D\left(\bar{\boldsymbol{\nu}} \parallel \bar{\boldsymbol{\mu}}\right) \\
&= \sum_{i\in\mathbb{A}}\sum_{j\in\mathbb{A}} \nu_{ij} \log\frac{\nu_{ij}}{\mu_{ij}} - \sum_{i\in\mathbb{A}} \bar{\nu}_i \log\frac{\bar{\nu}_i}{\bar{\mu}_i} \\
&= \sum_{i\in\mathbb{A}}\sum_{j\in\mathbb{A}} \nu_{ij} \log\frac{\nu_{ij}}{\mu_{ij}} - \sum_{i\in\mathbb{A}} \left[\sum_{j\in\mathbb{A}} \nu_{ij}\right] \log\frac{\bar{\nu}_i}{\bar{\mu}_i} \\
&= \sum_{i\in\mathbb{A}}\sum_{j\in\mathbb{A}} \nu_{ij} \log\frac{\nu_{ij}/\bar{\nu}_i}{\mu_{ij}/\bar{\mu}_i} \\
&= \sum_{i\in\mathbb{A}} \bar{\nu}_i \sum_{j\in\mathbb{A}} c_{ij} \log\left(\frac{c_{ij}}{a_{ij}}\right) = I\left(\boldsymbol{\nu}\right),
\end{aligned}$$

dove c_{ij} e a_{ij} sono esattamente le quantità definite in precedenza.
La dimostrazione verrà data nella sezione 2.3, dopo che sarà stato introdotto il *metodo dei tipi*.

Rate function per la distribuzione stazionaria

Vogliamo ora estendere le considerazioni fatte sulla *rate function* del processo $\left\{\boldsymbol{\nu}\left(x_1^l\right)\right\}$, stima empirica del "vettore delle frequenze doppie" $\boldsymbol{\mu} \in \mathcal{M}(\mathbb{A}^2)$, al processo che stima empiricamente la distribuzione stazionaria. Supponiamo di osservare i valori $x_1^l = x_1 \ldots x_l$. Possiamo costruire una misura empirica $\boldsymbol{\phi}\left(x_1^l\right) \in \mathbb{S}_n$ come

$$\phi_j\left(x_1^l\right) := \frac{1}{l}\sum_{t=1}^{l} I_{\{x_t=j\}}, \forall j \in \mathbb{A}, \tag{2.25}$$

dove $I_{\{x_t=j\}}$ è la funzione indicatrice di x_t, per $t = 1, \ldots, l$. Abbiamo già introdotto la quantità $\boldsymbol{\phi}\left(x_1^l\right) \in \mathbb{S}_n$ nella sezione precedente, osservando che essa non

è altro che un vettore contenente alla posizione j-esima la frequenza del j-esimo valore dell'insieme $\mathbb{A}$.

Sia $\{X_t\}$ un processo di Markov, così come è stato definito nella sezione 2.3, aggiungendo la condizione che la matrice di transizione A sia irriducibile, e siano $\boldsymbol{\mu} \in \mathcal{M}_s(\mathbb{A}^2)$ e $\bar{\boldsymbol{\mu}} \in \mathbb{S}_n$ il "vettore delle frequenze doppie" e cioè la distribuzione stazionaria del processo, così come definite nella 2.9 e nella 2.10 rispettivamente. Allora il processo $\left\{\boldsymbol{\phi}\left(x_1^l\right)\right\}$ soddisfa il Principio delle Grandi Deviazioni con *rate function* pari a

$$I(\boldsymbol{\phi}) = \sup_{\mathbf{u}>0} \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(\mathbf{u}A)_i} = \sup_{\mathbf{u}>0} \sum_{i=1}^{n} \phi_i \log \frac{u_i}{(A\mathbf{u})_i}, \tag{2.26}$$

dove la notazione $\mathbf{u} > 0$ vuole indicare che $u_{ij} > 0 \; \forall i \in \mathbb{A}$.

La dimostrazione della 2.26 è lineare e utilizza un teorema noto come *contraction principle*. La dimostrazione non verrà approfondita ulteriormente, in quanto si basa su un'estensione del risultato ottenuto nella 2.24.

Metodo dei tipi

I risultati teorici della Teoria delle Grandi Deviazioni sono stati notevolmente semplificati alla fine degli anni 70' da Csiszar e Körner, che per primi introdussero il *metodo dei tipi*. Da un punto di vista teorico, il *metodo dei tipi* prende in considerazione l'insieme delle successioni che hanno tutte la medesima distribuzione empirica e calcola la probabilità di ciascuna di esse. Nella pratica, se osserviamo due successioni $x_1^l, y_1^l \in \mathbb{A}^l$, diremo che queste sono equivalenti se hanno la stessa distribuzione empirica, cioè $\boldsymbol{\nu}\left(x_1^l\right) = \boldsymbol{\nu}\left(y_1^l\right)$, per la notazione introdotta nella sezione 2.3.
È bene introdurre in questa fase preliminare alcune notazioni e definizioni che ci saranno particolarmente utili nella dimostrazione della 2.24.

Sia $\varepsilon(l, n)$ l'*insieme dei tipi*, cioè l'insieme di tutte le possibili distribuzioni empiriche $\boldsymbol{\nu}\left(x_1^l\right)$ calcolate, a partire dei dati osservati $x_1^l \in \mathbb{A}^l$, dalla formula 2.22. Si noti la notazione con cui viene identificato l'insieme dei tipi: quest'ultimo infatti dipende sia dalla dimensione del dato osservato che dalla cardinalità dello spazio degli stati $\mathbb{A}$.
Supponiamo per esempio che $\mathbb{A} = \{1, 2\}$ e di avere osservato x_1^l. Ciascun elemento di $\boldsymbol{\nu}\left(x_1^l\right) \in \mathcal{M}_s\left(\mathbb{A}^2\right)$ ha la forma $\nu_{ij} = l_{ij}/l$ e indica il numero di volte in cui si è passati dallo stato i allo stato j divisa per gli l confronti di coppia. In questo caso abbiamo soltanto quattro possibili coppie: $(1,1)$, $(1,2)$, $(2,1)$ e $(2,2)$. Il

corrispondente insieme dei tipi sarà quindi

$$\varepsilon(l,2) = \left\{ \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{21} & \nu_{22} \end{pmatrix} : \begin{pmatrix} 0 & 0 \\ 0 & \frac{l}{l} \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ \frac{1}{l} & \frac{l-1}{l} \end{pmatrix}, \ldots, \begin{pmatrix} \frac{l}{l} & 0 \\ 0 & 0 \end{pmatrix} \right\}.$$

Per ogni $\boldsymbol{\zeta} \in \varepsilon(l,n)$, si definisce $T(\boldsymbol{\zeta}, l)$ l'insieme di tutte le successioni di lunghezza l che generano la distribuzione empirica ζ. Matematicamente

$$T(\boldsymbol{\zeta}, l) := \left\{ x_1^l \in \mathbb{A}^l : \boldsymbol{\nu}\left(x_1^l\right) = \boldsymbol{\zeta} \right\}$$

è detto *type class* di $\boldsymbol{\zeta}$.
Tornando all'esempio precedente in cui $\mathbb{A} = \{1,2\}$, supponiamo per semplicità che la lunghezza della successione osservata sia 3. In generale si hanno 2^3 possibili successioni osservabili, ma in questo caso ci si chiede quante e quali di queste successioni generano la medesima distribuzione empirica $\boldsymbol{\zeta}$, dove

$$\boldsymbol{\zeta} = \begin{pmatrix} \zeta_{11} & \zeta_{12} \\ \zeta_{21} & \zeta_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 \end{pmatrix}.$$

Inizialmente scriviamo in forma estesa l'insieme dei tipi $\varepsilon(3,2)$, derivato direttamente dalle 2^3 possibili successioni osservabili:

$$\varepsilon(3,2) = \left\{ \begin{pmatrix} \nu_{11} & \nu_{12} \\ \nu_{21} & \nu_{22} \end{pmatrix} : \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} \end{pmatrix}, \begin{pmatrix} \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & 0 \end{pmatrix} \right\}.$$

A questo punto $T(\boldsymbol{\zeta}, 3)$ risulta essere uguale a

$$T(\boldsymbol{\zeta}, 3) = \{(121), (211), (112)\}.$$

Si noti che $|\varepsilon(3,2)| = 4$, cioè che tra le 2^3 successioni alcune di queste hanno la stessa distribuzione empirica. In particolare, esistono $|T(\boldsymbol{\zeta}, 3)| = 3$ successioni che hanno distribuzione empirica pari a $\boldsymbol{\zeta}$. Queste due cardinalità sono al centro della dimostrazione della 2.24, così come il logaritmo della probabilità di ciascuna successione nell'insieme $T(\boldsymbol{\zeta}, l)$ e la sua relazione con $\boldsymbol{\zeta}$.

In particolare, uno dei principali risultati del *metodo dei tipi* riguarda la stima della cardinalità della type class $|T(\boldsymbol{\zeta}, l)|$.

Teorema: Se $\boldsymbol{\zeta} \in \varepsilon(l,n)$ allora si ha che

$$(2l)^{-n^2} e^{lH_c(\boldsymbol{\zeta})} \leq |T(\boldsymbol{\zeta}, l)| \leq l e^{lH_c(\boldsymbol{\zeta})}, \tag{2.27}$$

dove $H_c(\boldsymbol{\zeta})$ è l'entropia condizionale, così come è stata definita nella 2.18. La dimostrazione della 2.27 si basa essenzialmente su alcune proprietà combinatorie del *metodo dei tipi* e si trova nell'appendice di Vidyasagar (2009) e nella referenza 6 dell'articolo stesso.

Dimostrazione

A questo punto si può finalmente procedere alla dimostrazione dell'espressione della *rate function* per il processo $\left\{\boldsymbol{\nu}\left(x_1^l\right)\right\}$, definito nella 2.22. Sotto le ipotesi formulate nella sezione 2.3, la tesi è

$$I(\boldsymbol{\nu}) = D_c(\boldsymbol{\nu} \parallel \boldsymbol{\mu}).$$

- Analizziamo prima la probabilità che una qualunque successione delle variabili casuali X_1^l sia proprio uguale alla successione osservata x_1^l. Trattandosi di un processo markoviano, per la proprietà espressa nella 1.1:

$$\begin{aligned}
\mathbf{P}\left(X_1^l = x_1^l\right) &= \mathbf{P}(X_1 \dots X_l = x_1 \dots x_l) \\
&= \mathbf{P}(X_2 \dots X_l = x_2 \dots x_l \mid X_1 = x_1) \cdot \mathbf{P}(X_1 = x_1) \\
&= \left[\prod_{t=1}^{l-1} \mathbf{P}(X_{t+1} = x_{t+1} \mid X_t = x_t)\right] \cdot \mathbf{P}(X_1 = x_1) \\
&= \left[\prod_{t=1}^{l-1} \frac{\mathbf{P}(X_t X_{t+1} = x_t x_{t+1})}{\mathbf{P}(X_t = x_t)}\right] \cdot \mathbf{P}(X_1 = x_1) \\
&= \left[\prod_{t=1}^{l-1} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)}\right] \cdot \bar{\mu}(x_1) \\
&= \left[\frac{\bar{\mu}(x_l)}{\mu(x_l x_1)} \cdot \prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)}\right] \cdot \bar{\mu}(x_1) \\
&= \left[\prod_{t=1}^{l} \frac{\mu(x_t x_{t+1})}{\bar{\mu}(x_t)}\right] \cdot \frac{\bar{\mu}(x_1)\,\bar{\mu}(x_l)}{\mu(x_l x_1)}, \qquad (2.28)
\end{aligned}$$

  dove per esigenze grafiche non utilizziamo più i pedici per indicare gli elementi di $\boldsymbol{\mu}$ e $\bar{\boldsymbol{\mu}}$, come nella definizione 2.9 e 2.10 rispettivamente.
  Essendo $\mu_{ij} > 0 \ \forall i, j$, possiamo affermare che esistano delle costanti $\underline{a}$, $\overline{a}$, $\underline{b}$ e $\overline{b}$ tali che $0 \leq \underline{a} \leq \bar{\mu}_i \leq \overline{a} \ \forall i \in \mathbb{A}$ e $0 \leq \underline{b} \leq \mu_{ij} \leq \overline{b} \ \forall i, j \in \mathbb{A}$. Se ora definiamo due costanti $\underline{c}$ e $\overline{c}$ come

$$\underline{c} = \log \frac{\underline{a}^2}{\overline{b}} \text{ e } \underline{c} = \log \frac{\overline{a}^2}{\underline{b}},$$

allora si può mostrare che

$$\begin{aligned}
\underline{a} \leq \quad & \bar{\mu}_i & \leq \overline{a} \\
\underline{a}^2 \leq \quad & \bar{\mu}_i \bar{\mu}_j & \leq \overline{a}^2 \\
\frac{\underline{a}^2}{\overline{b}} \leq \quad & \frac{\bar{\mu}_i \bar{\mu}_j}{\mu_{ij}} & \leq \frac{\overline{a}^2}{\underline{b}} \\
\log \frac{\underline{a}^2}{\overline{b}} \leq \quad & \log \frac{\bar{\mu}_i \bar{\mu}_j}{\mu_{ij}} & \leq \log \frac{\overline{a}^2}{\underline{b}} \\
\underline{c} \leq \quad & \log \frac{\bar{\mu}_i \bar{\mu}_j}{\mu_{ij}} & \leq \overline{c}, \ \forall i, j \in \mathbb{A}.
\end{aligned}$$

Dalla 2.28 segue immediatamente che

$$\log \left[\prod_{t=1}^{l} \frac{\mu\left(x_t x_{t+1}\right)}{\bar{\mu}\left(x_t\right)} \right] + \underline{c} \leq \log \mathbf{P}\left(X_1^l = x_1^l\right) \leq \log \left[\prod_{t=1}^{l} \frac{\mu\left(x_t x_{t+1}\right)}{\bar{\mu}\left(x_t\right)} \right] + \overline{c}. \tag{2.29}$$

Si vuole esprimere il logaritmo della produttoria precedente non in termini di t, ma in termini di i e j. Possiamo osservare che l'evento $x_t x_{t+1} = ij$ si verifica esattamente $l_{ij} = l \cdot \nu_{ij}$ volte, mentre l'evento $x_t = i$ si verifica esattamente $\bar{l}_i = \sum_{j=1}^{n} l_{ij} = \sum_{j=1}^{n} \nu_{ij} \cdot l = l \cdot \bar{\nu}_i$ volte. Si ottiene quindi, sfruttando le proprietà dei logaritmi, che

$$\begin{aligned}
\log \left[\prod_{t=1}^{l} \frac{\mu\left(x_t x_{t+1}\right)}{\bar{\mu}\left(x_t\right)} \right] &= \sum_{t=1}^{l} \left[\log \mu\left(x_t x_{t+1}\right) - \log \bar{\mu}\left(x_t\right) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} l \cdot \nu_{ij} \log \mu_{ij} - \sum_{i=1}^{n} l \cdot \bar{\nu}_i \log \bar{\mu}_i \\
&= l \cdot \left[- \sum_{i=1}^{n} \sum_{j=1}^{n} \nu_{ij} \log\left(1/\mu_{ij}\right) + \sum_{i=1}^{n} \bar{\nu}_i \log\left(1/\bar{\mu}_i\right) \right] \\
&= l \cdot \left[-J\left(\boldsymbol{\nu}, \boldsymbol{\mu}\right) + J\left(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}\right) \right],
\end{aligned}$$

dove $J\left(\cdot, \cdot\right)$ è la "funzione di perdita" definita nella 2.12.
A questo punto si può riscrivere la 2.29 come

$$\log \mathbf{P}\left(X_1^l = x_1^l\right) \geq l \cdot \left[-J\left(\boldsymbol{\nu}, \boldsymbol{\mu}\right) + J\left(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}\right) \right] + \underline{c} \tag{2.30}$$

$$\log \mathbf{P}\left(X_1^l = x_1^l\right) \leq l \cdot \left[-J\left(\boldsymbol{\nu}, \boldsymbol{\mu}\right) + J\left(\bar{\boldsymbol{\nu}}, \bar{\boldsymbol{\mu}}\right) \right] + \overline{c} \tag{2.31}$$

- Estendiamo adesso i due risultati appena ottenuti alla Teoria delle Grandi Deviazioni. L'interesse in questo caso è calcolare la probabilità che alcune successioni (tra le possibili $|\mathbb{A}|^l$) generino tutte una particolare distribuzione

empirica $\boldsymbol{\zeta}$. In realtà si calcola il logaritmo di tale probabilità e in seguito si opera una normalizzazione per la lunghezza l dell'osservazione.
Per ogni $\boldsymbol{\zeta} \in \varepsilon(l,n)$ definiamo la quantità

$$\delta(l,\boldsymbol{\zeta}) := \frac{1}{l}\log \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) = \boldsymbol{\zeta}\right)$$

e studiamone il comportamento per $l \to \infty$.
Dalla 2.27 si può studiare il comportamento di una trasformazione di $|T(\boldsymbol{\zeta},l)|$ per $l \to \infty$:

$$\begin{aligned}
|T(\boldsymbol{\zeta},l)| &\leq l e^{lH_c(\boldsymbol{\zeta})} \\
\log|T(\boldsymbol{\zeta},l)| &\leq \log l + lH_c(\boldsymbol{\zeta}) \\
\frac{1}{l}\log|T(\boldsymbol{\zeta},l)| &\leq \frac{\log l}{l} + H_c(\boldsymbol{\zeta}) \\
\frac{1}{l}\log|T(\boldsymbol{\zeta},l)| &\leq H_c(\boldsymbol{\zeta}) + o(1/l).
\end{aligned}$$

Combinando il risultato appena ottenuto con la 2.30 e la 2.31 e ricordando le definizioni collegate all'entropia date nella sezione 2.3, si ottiene che

$$\begin{aligned}
\delta(l,\boldsymbol{\zeta}) &\leq H_c(\boldsymbol{\zeta}) - J(\boldsymbol{\zeta},\boldsymbol{\mu}) + J\left(\bar{\boldsymbol{\zeta}},\bar{\boldsymbol{\mu}}\right) + o(1/l) \\
&= H(\boldsymbol{\zeta}) - H\left(\bar{\boldsymbol{\zeta}}\right) - J(\boldsymbol{\zeta},\boldsymbol{\mu}) + J\left(\bar{\boldsymbol{\zeta}},\bar{\boldsymbol{\mu}}\right) + o(1/l) \\
&= -D(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}) + D\left(\bar{\boldsymbol{\zeta}} \parallel \bar{\boldsymbol{\mu}}\right) + o(1/l) \\
&= -D_c(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}) + o(1/l) \qquad (2.32)
\end{aligned}$$

ed analogamente per l'estremo inferiore

$$\delta(l,\boldsymbol{\zeta}) \geq -D_c(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}) + o(1/l). \qquad (2.33)$$

Affinché l'entropia relativa condizionale tra $\boldsymbol{\zeta}$ e $\boldsymbol{\mu}$ sia la *rate function* del processo, $D_c(\boldsymbol{\zeta} \parallel \boldsymbol{\mu})$ deve soddisfare il Principio delle Grandi Deviazioni introdotto nella sezione 2.2 e in particolare le disuguaglianze 2.5 e 2.6.

- Verifichiamo che la *rate function* $D_c(\boldsymbol{\zeta} \parallel \boldsymbol{\mu})$ soddisfi la 2.6. Sia $\Gamma \subseteq \mathcal{M}_s(\mathbb{A}^2)$ un qualunque sottoinsieme (chiuso o aperto) di distribuzioni stazionarie in $\mathbb{A}^2$. Si può mostrare che

$$\begin{aligned}
\mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) \in \Gamma\right) &\leq \sum_{\boldsymbol{\zeta} \in \varepsilon(l,n) \cap \Gamma} \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) = \boldsymbol{\zeta}\right) \\
&= |\varepsilon(l,n)| \sup_{\boldsymbol{\zeta} \in \Gamma} \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) = \boldsymbol{\zeta}\right)
\end{aligned}$$

e pertanto la quantità d'interesse è

$$\frac{1}{l}\log \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right)\in\Gamma\right)\leq\frac{1}{l}\log\left|\varepsilon\left(l,n\right)\right|+\sup_{\boldsymbol{\zeta}\in\Gamma}\delta\left(l,\boldsymbol{\zeta}\right). \tag{2.34}$$

Si può dimostrare che $|\varepsilon(l,n)|\leq(l+1)^2$ e quindi

$$\lim_{l\to\infty}\frac{1}{l}\log\left|\varepsilon\left(l,n\right)\right|=0.$$

Infine calcoliamo il limite superiore del secondo addendo della 2.34, che può essere facilmente ricavato dalla 2.32:

$$\begin{aligned}\limsup_{l\to\infty}\sup_{\boldsymbol{\zeta}\in\Gamma}\delta\left(l,\boldsymbol{\zeta}\right) &\leq \sup_{\boldsymbol{\zeta}\in\Gamma}-D_c\left(\boldsymbol{\zeta}\parallel\boldsymbol{\mu}\right)\\ &= -\inf_{\boldsymbol{\zeta}\in\Gamma}D_c\left(\boldsymbol{\zeta}\parallel\boldsymbol{\mu}\right).\end{aligned} \tag{2.35}$$

Si ottiene quindi un risultato analogo alla 2.6:

$$\limsup_{l\to\infty}\frac{1}{l}\log\mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right)\in\Gamma\right)\leq-\inf_{\boldsymbol{\zeta}\in\Gamma}D_c\left(\boldsymbol{\zeta}\parallel\boldsymbol{\mu}\right).$$

Si noti che la disuguaglianza appena dimostrata è una condizione più forte di quella in realtà richiesta dal Principio delle Grandi Deviazioni, dove è sufficiente che valga la disuguaglianza per $\boldsymbol{\zeta}\in\Gamma^o$, cioè per i punti interni di Γ. Per dimostrare la 2.5 si supporrà che $\boldsymbol{\zeta}\in\Gamma^o$.

- Per dimostrare che l'entropia relativa condizionale verifica la 2.5, supponiamo che $\boldsymbol{\zeta}$ sia un punto interno dell'insieme $\Gamma\subseteq\mathcal{M}_s\left(\mathbb{A}^2\right)$. Per alcuni risultati della Teoria delle Grandi Deviazioni in cui non entreremo nel dettaglio in questa sede, si consideri una successione $\{\boldsymbol{\zeta}_l\}$, dove $\boldsymbol{\zeta}_l\in\varepsilon(l,n)$, tale per cui $\boldsymbol{\zeta}_l\to\boldsymbol{\zeta}$ per l sufficientemente grande. Essendo $\boldsymbol{\zeta}$ un punto interno di Γ, è ragionevole ipotizzare che anche $\boldsymbol{\zeta}_l\in\Gamma$ per $l\to\infty$. Si può a questo punto stabilire la seguente disuguaglianza per l sufficientemente grande:

$$\begin{aligned}\mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right)\in\Gamma\right) &\geq \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right)=\boldsymbol{\zeta}_l\right)\\ \frac{1}{l}\log\mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right)\in\Gamma\right) &\geq \delta\left(l,\boldsymbol{\zeta}_l\right)\\ &\geq -D_c\left(\boldsymbol{\zeta}_l\parallel\boldsymbol{\mu}\right)+o\left(1/l\right),\end{aligned}$$

dove il termine a destra dell'ultima disuguaglianza converge a $-D_c\left(\boldsymbol{\zeta}\parallel\boldsymbol{\mu}\right)$ per $l\to\infty$, poichè $\boldsymbol{\zeta}_l\to\boldsymbol{\zeta}$ e $D_c\left(\cdot\parallel\cdot\right)$ è una funzione continua. Vale pertanto

la disuguaglianza

$$\liminf_{l\to\infty} \frac{1}{l} \log \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) \in \Gamma\right) \geq -D_c\left(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}\right)$$

per ogni $\boldsymbol{\zeta} \in \Gamma^o$. Si può quindi concludere \

$$\begin{aligned} \liminf_{l\to\infty} \frac{1}{l} \log \mathbf{P}\left(\boldsymbol{\nu}\left(x_1^l\right) \in \Gamma\right) &\geq \sup_{\boldsymbol{\zeta}\in\Gamma^o} -D_c\left(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}\right) \\ &= -\inf_{\boldsymbol{\zeta}\in\Gamma^o} D_c\left(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}\right). \end{aligned} \tag{2.36}$$

- La 2.35 e la 2.36 stabiliscono quindi che $I\left(\boldsymbol{\zeta}\right) = D_c\left(\boldsymbol{\zeta} \parallel \boldsymbol{\mu}\right)$ è la *rate function* del processo, essendo questa una funzione continua.

3 Applicazioni nella statistica

La Teoria delle Grandi Deviazioni assume un ruolo centrale in statistica non soltanto nella stima delle probabilità di eventi rari, ma anche in uno degli strumenti ancora oggi più utilizzati nell'ambito della statistica inferenziale: il test di verifica d'ipotesi. Grazie ad alcuni risultati della Teoria delle Grandi Deviazioni si riescono infatti a formulare considerazioni di natura asintotica sulla probabilità di errore del test, che fa parte di uno degli argomenti più dibattuti ancora oggi nel mondo accademico, ovvero l'uso dei p-value nella ricerca scientifica.

Inizialmente si definirà da un punto di vista statistico parametrico il test di verifica d'ipotesi, l'errore di primo tipo e l'errore di secondo tipo, la potenza del test e il lemma di Neyman-Pearson. Quest'ultimo risultato verrà ripreso per definire il comportamento asintotico di α_n e β_n, rispettivamente la probabilità di errore di primo e di secondo tipo, dove il pedice indica la dipendenza degli errori dalla numerosità campionaria, al fine di enunciare il lemma di Stein. Infine si adotterà un approccio bayesiano per minimizzare la probabilità totale di errore di un test introducendo l'informazione di Chernoff.

Da un punto di vista generale la verifica d'ipotesi viene ripresa da Pace e Salvan (1996), mentre per le applicazioni specifiche alla Teoria delle Grandi Deviazioni è stato preso come riferimento Dembo e Zeitouni (2010).

3.1 Il test statistico

Uno dei problemi più ricorrenti in statistica è quello di decidere tra due spiegazioni alternative dei dati: in ambito medico, per esempio, si è soliti verificare l'efficacia di un nuovo farmaco dividendo i pazienti in due gruppi, uno a cui si somministra il nuovo farmaco e un altro gruppo che riceve un placebo; a questo punto si valutano i livelli della patologia per entrambi e ci si chiede se questi siano significamente diversi nei due gruppi. Appare evidente che per potere rispondere a simili domande occore uno strumento statistico che misuri il livello di accordo o disaccordo dei dati con l'ipotesi iniziale.

Un modello statistico parametrico mette a punto una procedura per verificare la conformità dei dati a un sottomodello $\mathcal{F}_\mathbf{0}$ di $\mathcal{F}$, dove $\mathcal{F}$ è stato correttamente specificato e il parametro $\theta \in \Theta$ è identificabile. Poiché vi è una corrispondenza biunivoca tra Θ e $\mathcal{F}$, il sottomodello $\mathcal{F}_\mathbf{0}$ è espresso dall'ipotesi nulla $H_0 : \theta \in \Theta_0$, dove $\Theta_0 \subset \Theta$. L'ipotesi nulla viene anche detta "ipotesi conservativa", mentre l'ipotesi alternativa $H_1 : \theta \in \Theta \setminus \Theta_0$ viene detta "ipotesi innovativa".

Lo strumento matematico che permette di stabilire se i dati mostrano maggiore evidenza verso l'ipotesi nulla oppure verso l'ipotesi alternativa è detto statistica test, una funzione dei dati $t : \mathcal{Y} \to \mathbb{R}$, dove $\mathcal{Y}$ è lo spazio campionario. L'applicazione t divide quindi lo spazio campionario in due sottoinsiemi disgiunti

$$R = \{y \in \mathcal{Y} : t(y) \text{ suggerisce di rifiutare } H_0\} \text{ e}$$

$$A = \{y \in \mathcal{Y} : t(y) \text{ suggerisce di accettare } H_0\},$$

dette rispettivamente regione di rifiuto e regione di accettazione.

Si noti che *accettare* e *rifiutare* vanno intesi non in modo netto, bensì come posizioni conoscitive provvisoriamente prese, alla luce dei dati osservati. Ne è una prova la possibilità di commettere un errore nel discriminare H_0 rispetto ad H_1 o viceversa. Tutti i possibili casi, inclusi quelli di corretta discriminazione, sono riassumibili nella seguente tabella:

		Decisione	
		Accetto H_0	Rifiuto H_0
Verità	H_0	Corretto	Errore di Tipo I
	H_1	Errore di Tipo II	Corretto

Tabella 3.1: Due tipi di errore nel test statistico

In generale, la funzione di potenza di un test con regione di rifiuto R è

$$\pi(\theta) = \mathbf{P}_\theta(Y \in R), \ \theta \in \Theta.$$

A partire dalla funzione di potenza si possono derivare $\alpha(\cdot)$ e $\beta(\cdot)$, rispettivamente la probabilità di errore di primo tipo e di secondo tipo. Si ottiene infatti che

$$\alpha(\theta) = \pi(\theta) \text{ se } \theta \in \Theta_0 \text{ e}$$

$$\beta(\theta) = 1 - \pi(\theta) \text{ se } \theta \in \Theta \setminus \Theta_0.$$

Appare evidente che il test ideale ha $\alpha(\theta) = 0$ per $\theta \in \Theta$ e $\beta(\theta) = 1$ per $\theta \in \Theta \setminus \Theta_0$, cioè ha probabilità sia dell'errore di primo tipo che dell'errore di

secondo tipo pari a zero. Si tratta tuttavia di un test ideale che nella pratica non è possibile ottenere, salvo che in casi banali. Nonostante ciò la teoria della verifica delle ipotesi ha evidenziato alcune proprietà dei test, che in alcuni modelli statistici con una struttura molto specifica possono essere sfruttate per costruire dei test ottimi.

Lemma di Neyman-Pearson

Il risultato principale sui test ottimi è il lemma di Neyman-Pearson, dove si considera un modello statistico costituito da due soli modelli probabilistici, cioè $\Theta = \{\theta_0, \theta_1\}$ e quindi $\mathcal{F} = \{p_{\theta_0}(y), p_{\theta_1}(y)\}$. Si consideri il sistema d'ipotesi

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

e il test con regione di rifiuto

$$R^* = \{y \in \mathcal{Y} : p_{\theta_1}(y) > k p_{\theta_0}(y)\}$$

e regione di accettazione

$$A^* = \{y \in \mathcal{Y} : p_{\theta_1}(y) < k p_{\theta_0}(y)\},$$

per una costante assegnata $k \geq 0$. Sia altresì

$$\alpha = \alpha(k) = \sup_{\theta \in \Theta_0} \mathbf{P}_\theta (Y \in R^*)$$

il livello di significatività del test. Allora tale test è il più potente (UMP) tra tutti i test con livello di significatività non superiore ad α ed è quello che minimizza la probabilità di errore di secondo tipo, fissata la soglia massima α per la probabilità di commettere errore di primo tipo.

Il lemma di Neyman-Pearson verrà in seguito applicato ad un test con regione di rifiuto basata sul log-rapporto di verosimiglianza.

3.2 Comportamento asintotico di α_n e β_n

Sia $Y_1, \ldots, Y_n$ una successione di variabili casuali. Il problema di verifica di ipotesi consiste nel decidere se, alla luce dell'osservazione $(Y_1 = y_1, \ldots Y_n = y_n)$, la legge

di probabilità che ha generato i dati è $\mathbf{P}_{\mu_0}$ (ipotesi nulla) oppure $\mathbf{P}_{\mu_1}$ (ipotesi alternativa). D'ora in avanti assumeremo che le variabili μ_0 e μ_1 siano note a priori e, per semplicità, che siano state ottenute come prodotto delle singole $\mu_i \in \mathcal{M}(\mathbb{A})$, seguendo la notazione introdotta nella sezione 2.3.

Sia $\mathcal{S}$ una successione di funzioni misurabili $\mathcal{S}^n : \mathbb{A}^n \to \{0,1\}$, che ha la seguente interpretazione: se $\mathcal{S}^n(y_1, \dots y_n) = 0$ si accetta H_0, mentre se $\mathcal{S}^n(y_1, \dots y_n) = 1$ si rifiuta H_0. Possiamo quindi scrivere le quantità

$$\alpha_n \doteq \mathbf{P}_{\mu_0}(\mathcal{S}^n \text{ rifiuta } H_0) \text{ e}$$

$$\beta_n \doteq \mathbf{P}_{\mu_0}(\mathcal{S}^n \text{ rifiuta } H_1).$$

Analogamente a quanto visto precedentemente, l'obiettivo è quello di minimizzare β_n vincolando α_n.

Definiamo il rapporto di verosimiglianza in termini di

$$L_{0\|1}(y) = \frac{\mu_0(y)}{\mu_1(y)} \text{ e } L_{1\|0}(y) = \frac{\mu_1(y)}{\mu_0(y)},$$

dove la prima quantità è quella che viene usata nell'inferenza di verosimiglianza, mentre la seconda è l'inversa della prima. Si ottiene così la trasformazione del log-rapporto di verosimiglianza

$$X_j \doteq \log L_{1\|0}(Y_j) = -\log L_{0\|1}(Y_j).$$

Da questa derivano due importanti quantità:

$$\bar{x}_0 \doteq E_{\mu_0}[X_1] = E_{\mu_1}\left[X_1 e^{-X_1}\right] \text{ e} \tag{3.1}$$

$$\bar{x}_1 \doteq E_{\mu_1}[X_1] = E_{\mu_0}\left[X_1 e^{X_1}\right] > E_{\mu_0}[X_1] = \bar{x}_0. \tag{3.2}$$

La 3.1 e la 3.2 possono essere scritte in termini di entropia relativa:

$$\begin{aligned}
\bar{x}_0 &= E_{\mu_0}[X_1] \\
&= \sum_{j \in \mathbb{A}} x_{1j} \cdot \mu_0(x_{1j}) \\
&= \sum_{j \in \mathbb{A}} \log\left(\frac{\mu_1(x_{1j})}{\mu_0(x_{1j})}\right) \cdot \mu_0(x_{1j}) \\
&= -\sum_{j \in \mathbb{A}} \log\left(\frac{\mu_0(x_{1j})}{\mu_1(x_{1j})}\right) \cdot \mu_0(x_{1j}) = -D(\mu_0 \parallel \mu_1),
\end{aligned} \tag{3.3}$$

$$\begin{aligned}\bar{x}_1 &= E_{\mu_1}[X_1] \\ &= \sum_{j\in\mathbb{A}} x_{1j}\cdot\mu_1(x_{1j}) \\ &= \sum_{j\in\mathbb{A}} \log\left(\frac{\mu_1(x_{1j})}{\mu_0(x_{1j})}\right)\cdot\mu_1(x_{1j}) = D(\mu_1\parallel\mu_0) \end{aligned} \tag{3.4}$$

Per il teorema di Neyman-Pearson, esiste un test per cui la log-verosimiglianza normalizzata

$$\hat{S}_n \doteq \frac{1}{n}\sum_{j=1}^{n} X_j$$

può essere confrontata con una soglia γ_n, per la quale si accetta H_0 se $\hat{S}_n \leq \gamma_n$ e si rifiuta H_0 se $\hat{S}_n > \gamma_n$. Nel caso specifico è di particolare interesse confrontare il tasso esponenziale di α_n e β_n al variare di $\gamma \in (\bar{x}_0, \bar{x}_1)$, cioè studiare le grandi deviazioni di $\hat{S}_n$. Si ha che il test di Neyman-Pearson soddisfa

$$\lim_{n\to\infty}\frac{1}{n}\log\alpha_n = -\Lambda_0^*(\gamma) < 0 \tag{3.5}$$

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n = \gamma - \Lambda_0^*(\gamma) < 0, \tag{3.6}$$

dove $\Lambda_0^*(\gamma) = \sup_{\lambda\in\mathbb{R}}\{\lambda\gamma - \Lambda_0(\lambda)\}$ è la trasformata di Fenchel-Legendre, mentre $\Lambda_0(\lambda) = \log E_{\mu_0}\left[e^{\lambda X_1}\right]$ è la funzione generatrice dei cumulanti. Analogamente a quanto visto per il lemma di Neyman-Pearson nel caso generale, anche qui si nota che il comportamento di β_n dipende dal comportamento di α_n. A questo proposito, il lemma di Stein determina il miglior tasso esponenziale di β_n con α_n vincolato ad essere minore di 1.

Sia β_n^ϵ l'estremo inferiore di β_n tra tutti i test che hanno $\alpha_n < \epsilon$. Si ottiene che per $\epsilon < 1$,

$$\lim_{n\to\infty}\frac{1}{n}\log\beta_n^\epsilon = \bar{x}_0 = -D(\mu_0\parallel\mu_1). \tag{3.7}$$

Tale risultato è ottenuto riconducendosi ai test di Neyman-Pearson, dove β_n può essere scritto come un valore atteso, ossia

$$\beta_n = \mathbf{P}_{\mu_1}\left(\hat{S}_n \leq \gamma_n\right) = E_{\mu_1}\left[1_{\hat{S}_n\leq\gamma_n}\right] = E_{\mu_0}\left[1_{\hat{S}_n\leq\gamma_n}\cdot e^{n\hat{S}_n}\right]$$

per la definizione delle X_j come log-rapporti di verosimiglianza. A questo punto si identifica l'estremo superiore

$$\frac{1}{n}\log\beta_n = \frac{1}{n}\log E_{\mu_0}\left[1_{\hat{S}_n\leq\gamma_n}\cdot e^{n\hat{S}_n}\right] \leq \gamma_n.$$

In seguito si calcola il limite superiore e il limite inferiore della quantita $\frac{1}{n}\log\beta_n^\epsilon$, distinguendo i due casi $\bar{x}_0 = -\infty$ e $\bar{x}_0 > -\infty$ e ottenendo infine che, per tutti gli $\epsilon, \eta > 0$

$$\liminf_{n\to\infty} \frac{1}{n}\log\beta_n^\epsilon \geq \bar{x}_0 - \eta \text{ e}$$

$$\limsup_{n\to\infty} \frac{1}{n}\log\beta_n^\epsilon \leq \bar{x}_0 + \eta.$$

È proprio nella 3.7 che si coglie il collegamento con i capitoli precedenti e in particolare con la nozione di entropia nella Teoria delle Grandi Deviazioni utilizzata per trovare un test ottimo, che soddisfa quindi le ipotesi del lemma di Neyman-Pearson, al crescere della numerosità campionaria.
Si noti che che il test appena considerato, anche se asintoticamente ottimo, non è ottimo per qualunque problema fissato di verifica di ipotesi. Il test ottimo che minimizza le probabilità di errore è quello dato dal lemma di Neyman-Pearson.

3.3 Un approccio bayesiano

Fino a questo punto abbiamo considerato il problema di verifica d'ipotesi nel contesto usuale, dove trattiamo le due probabilità di errore α_n e β_n separatamente. Grazie al lemma di Stein abbiamo ottenuto che al crescere di n, $\beta_n \cong e^{-nD(\mu_0\|\mu_1)}$ per $\alpha_n < \epsilon < 1$. Tuttavia questo approccio risulta essere poco simmetrico.

Una scelta migliore è quella di assegnare delle probabilità a priori ad entrambe le ipotesi. In questo scenario l'obiettivo non è più quello di minimizzare l'errore di secondo tipo, bensì l'errore totale, dato dalla somma pesata delle probabilità di errore di primo e di secondo tipo. I pesi sono proprio le probabilità a priori che assegnamo all'ipotesi nulla e all'ipotesi alternativa. La quantità di errore totale è espressa quindi da

$$\mathbf{P}_n^{(e)} \doteq \alpha_n\pi_1 + \beta_n\pi_2,$$

dove π_1 e π_2 sono le probabilità a priori dell'ipotesi nulla e dell'ipotesi alternativa rispettivamente. Se $0 < \pi_1 < 1$, si ha

$$\inf_{\mathcal{S}} \liminf_{n\to\infty} \left\{ \frac{1}{n}\log\mathbf{P}_n^{(e)} \right\} = -\Lambda_0^*(0),$$

dove si calcola l'estremo inferiore tra tutte le funzioni misurabili $\mathcal{S}_n$.

Si ottiene pertanto che la probabilità di errore totale è minimizzata da un test di Neyman-Pearson con soglia $\gamma = 0$. Tale risultato può essere compreso intuitivamente se si osservano la 3.5 e la 3.6, che descrivono il comportamente

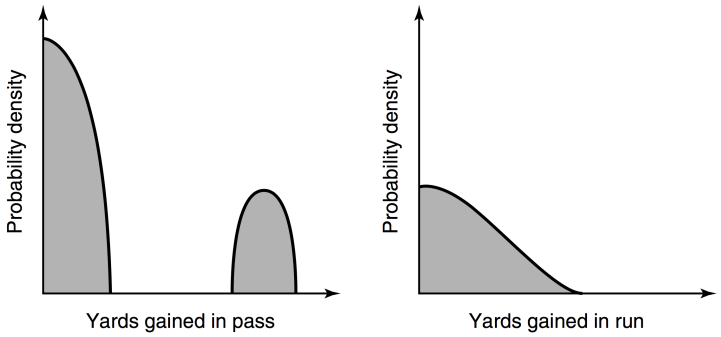

Figura 3.1: Distribuzione delle yards conquistate mediamente con passaggio e con corsa

asintotico di α_n e β_n per $\gamma \in (\bar{x}_0, \bar{x}_1)$. Si noti che per la disuguaglianza di Jensen $\bar{x}_0 < \log E_{\mu_0}\left[e^{X_1}\right] = 0$ e $\bar{x}_1 > -\log E_{\mu_1}\left[e^{-X_1}\right]$, quindi la soglia γ può assumere sia valori positivi che valori negativi. Nella formula della probabilità di errore totale l'obiettivo è quello di minimizzare la somma di α_n e β_n pesata con π_1 e π_2 rispettivamente: il minimo viene raggiunto ponendo asintoticamente $\alpha_n = \beta_n$ e ciò può essere fatto soltanto per $\gamma = 0$, come è evidente nella 3.5 e nella 3.6.

Applicazione nel football americano

Consideriamo come esempio di applicazione della Teoria delle Grandi Deviazioni al mondo reale una versione semplificata del football americano. L'esempio è tratto da Cover e Thomas (2012).

Supponiamo che il punteggio di ogni squadra sia direttamente collegato con il numero di yards conquistate durante un gioco. Per semplicità assumiamo che esistano soltanto due strategie di vittoria: passaggio o corsa, ciascuna con una propria distribuzione di probabilità in funzione del numero di yards conquistate in un gioco.

Si osservi a titolo puramente esemplificativo la figura 3.1. Si noti che la strategia del passaggio risulta più efficace in termini di yards conquistate: si ha un "guadagno" molto alto di yards in un gioco ma con una probabilità molto bassa. Al contrario, la corsa permette di raggiungere poche yards con una probabilità molto alta, mentre si dimostra poco efficace se si vuole conquistare un numero di yards elevato in un gioco.

Assumiamo che ci si trovi nella parte finale della partita ed il risultato sia nettamente a favore della squadra di casa. La squadra ospite può sperare di portare a casa il risultato soltanto se è molto fortunata. Si tratta quindi di un esempio di eventi rari che vengono studiati con grande attenzione non soltanto nello sport, ma anche in altri contesti reali. La Teoria delle Grandi Deviazioni porta un contributo importante almeno per l'impostazione iniziale della risoluzione del problema.

Supponiamo che siano rimasti n giochi prima del termine dell'incontro e che la squadra ospite debba conquistare l yards per vincere la partita. Visto lo svantaggio netto della squadra ospite sui padroni di casa, assumiamo che $l \gg n \cdot q$, dove q è il numero medio di yards conquistate in un gioco. Siano Z_i variabili casuali indipendenti e identicamente distribuite per $i = 1, \ldots, n$, dove ogni Z_i rappresenta il numero di yards conquistate nel i-esimo gioco e assume la distribuzione corrispondente alla strategia scelta per quel gioco.

Il nostro interesse è calcolare la probabilità che $\sum_{i=1}^{n} Z_i \geq n \cdot q$. A tale scopo si possono utilizzare i risultati della Teoria delle Grandi Deviazioni e in particolare il teorema di Sanov, che vale per processi indipendenti e identicamente distribuiti.

Siano $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ le distribuzioni di probabilità corrispondenti rispettivamente alla strategia di passaggio e a quella di corsa, per tutti gli n giochi rimasti. Allora per il teorema di Sanov la probabilità di vincere la partita seguendo la strategia di $\boldsymbol{\mu}_1$ converge a $e^{-nD(\boldsymbol{\nu}_1 \| \boldsymbol{\mu}_1)}$, dove $\left\{\boldsymbol{\phi}\left(z_1^l\right)\right\}$, introdotto nella 2.23, è il processo che soddisfa il Principio delle Grandi Deviazioni con *rate function* $I(\boldsymbol{\nu}) = D(\boldsymbol{\nu} \parallel \boldsymbol{\mu}_1)$. Un risultato analogo si ottiene per la probabilità di vincere l'incontro quando si segue la strategia di $\boldsymbol{\mu}_2$.

Coerentemente con quanto visto nel paragrafo 3.3 ci si potrebbe domandare se sia più ragionevole attuare una strategia "mista", ossia utilizzare sia il passaggio che la corsa negli n giochi rimasti. Si può dimostrare che utilizzare $\boldsymbol{\mu}_\lambda = \lambda \boldsymbol{\mu}_1 + (1 - \lambda)\boldsymbol{\mu}_2$, dove $\lambda \in (0, 1)$, è preferibile rispetto a $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ singolarmente.

4 Conclusioni

Nella relazione appena esposta sono stati introdotti i principi fondamentali di un argomento complesso e articolato come la Teoria delle Grandi Deviazioni, che meriterebbe certamente una trattazione più approfondita, soprattutto nelle applicazioni alle scienze tra loro più diverse: si potrebbe cominciare dalla meccanica statistica in fisica, dove il concetto di entropia si unisce alla seconda legge della termodinamica, per poi passare ad alcuni interessanti parallelismi tra la teoria dell'informazione e la teoria degli investimenti in economia, dove si analizza il comportamento duale tra l'entropia del mercato azionario e la crescita del sistema economico.

In questo elaborato si è scelto tuttavia di soffermarsi sui fondamenti teorici della Teoria delle Grandi Deviazioni introducendo nel Capitolo 2 le quantità matematiche d'interesse, come per esempio l'entropia, l'entropia relativa e l'entropia relativa condizionale tra due distribuzioni di probabilità. Quest'ultima definizione ha creato un forte collegamento con le catene di Markov, particolari processi stocastici esposti nel Capitolo 1, e con il Principio delle Grandi Deviazioni, una base di partenza per potere definire rigorosamente la *rate function* per un qualunque processo stocastico e quindi anche per i processi markoviani. Alla fine del Capitolo 2 si è voluta fornire la dimostrazione dell'espressione della *rate function* per i processi markoviani, operazione che è stata resa possibile soltanto grazie ad alcuni risultati combinatori, noti come *metodo dei tipi*, e che altrimenti avrebbe richiesto sia conoscenze matematiche avanzate, sia una trattazione tanto teorica da allontanarsi dagli obiettivi inizialmente stabiliti.

Nel Capitolo 3 è stato approfondita una parte consistente dei risvolti della Teoria delle Grandi Deviazioni nella disciplina statistica non soltanto in un'ottica di calcolo delle probabilità di eventi rari, ma anche come strumento utile nella teoria dei test per migliorarne la precisione statistica. Dopo avere introdotto l'approccio classico per minimizzare i due tipici errori di un test statistico, si è analizzato il loro comportamento asintotico al crescere della numerosità campionaria esprimendolo in termini di entropia. Infine si è scelto di proporre un approccio alternativo per la minimizzazione dei due errori, prima giustificando-

ne la maggiore precisione rispetto all'approccio classico e, infine, fornendo un esempio reale in cui l'impostazione bayesiana appare evidentemente da preferire.

Bibliografia

Cover T. M.; Thomas J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dembo A.; Zeitouni O. (2010). Large deviations techniques and applications, volume 38 of stochastic modelling and applied probability.

Levin D. A.; Peres Y.; Wilmer E. L. (2009). Markov chains and mixing times.

Pace L.; Salvan A. (1996). Introduzione alla statistica: Inferenza, verosimiglianza, modelli.-2001.-xvi, 422 p.

Vidyasagar M. (2009). An elementary derivation of the large deviation rate function for finite state markov chains. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pp. 1599–1606. IEEE.