Dipartimento di Matematica
Corso di Laurea Magistrale in Matematica Applicata

Rielaborazione degli appunti del corso di

# Elementi di Probabilità e statistica per Data Science

Simmaco Di Lillo
dsimmaco@gmail.com

a.a. 22-23

## Introduzione

Queste note contengono i miei appunti personali presi durante il corso di "Elementi di Probabilità e statistica per Data Science" della Prof.ssa A. Faggionato e dunque non sono dispense ufficiali del corso.
Nelle note potrebbero essere presenti typo o errori, per qualsiasi segnalazione scrivetemi un'e-mail a dsimmaco@gmail.com, la versione aggiornata verrà caricata sul mio sito

# Indice

# 1 Notazione e richiami

Andiamo a fissare alcune notazioni che verranno usate in seguito.

Se $(K, d)$ è metrico, la palla di centro $x$ e raggio $\varepsilon$ è

$$B(x, \varepsilon) = B_\varepsilon(x) = \{y \in K \ : \ d(x, y) \leq \varepsilon\}$$

Se $x \in \mathbb{R}^d$ denotiamo con $||x||$ la norma euclidea, inoltre, se non diversamente specificato, considereremo sempre $\mathbb{R}^d$ con la metrica euclidea.

$$S^{n-1} = \left\{x \in \mathbb{R}^d \ : \ ||x|| = 1\right\}$$

ovvero la sfera unitaria di $\mathbb{R}^n$

**Definizione 1.1.** Un insieme $(X, \leq)$ è un insieme parzialmente ordinato se $\leq$ è una relazione binaria con le seguenti caratteristiche

- riflessiva: $\forall x \in X$ vale $x \leq x$
- transitiva: $\forall x, y, z \in X$ vale $x \leq y$ e $y \leq z \quad \Rightarrow \quad x \leq z$
- antisimmetrica: $\forall x, y \in X$ vale $x \leq y$ e $y \leq x \quad \Rightarrow \quad x = y$

Se inoltre $\forall x, y \in X$ vale $x \leq y$ o $y \leq x$ diremo che l'insieme è **totalmente ordinato**

**Definizione 1.2.** Un sottoinsieme $C$ di un insieme ordinato si dice **catena** se $(C, \leq)$ è totalmente ordinato.
Un elemento di $X$ che è maggiore di tutti gli elementi della catena si dice **maggiorante** per la catena.

**Definizione 1.3.** $y \in X$ è **massimale** se per ogni $x \in X$ con $y \leq x$ si ha $y = x$.

**Lemma 1.1** (di Zorn)**.** *Sia $(X, \leq)$ un insieme non vuoto con un ordine parziale. Se ogni sua catena ha un maggiorante in $X$ allora $X$ contiene almeno un elemento massimale.*

*Osservazione* 1. Il lemma di Zorn è equivalente all'assioma di scelta

## 1.1 Richiami probabilistici

**Definizione 1.4.** Un **campione aleatorio** è una collezione di variabili aleatorie a valori in $\chi$: $X_1, \dots, X_n$ i.i.d.

**Definizione 1.5.** Se $X$ è una variabile aleatoria a valori reali, la sua **funzione di ripartizione** è

$$F: \mathbb{R} \to [0,1] \qquad F(t) = \mathbb{P}(X \leq t)$$

**Proposizione 1.2.** *Le funzioni di ripartizioni sono continue a destra con limite a sinistra, cioè*

$$\exists f(t^-) = \lim_{t \to t^-} f(t)$$

*e*

$$f(t) = \lim_{t \to t^+} f(t)$$

**Definizione 1.6.** Se $Y$ è vettore aleatorio in $\mathbb{R}^n$, la **funzione generatrice dei momenti** è

$$M_Y(t) = \mathbb{E}\left[e^{\langle t, Y \rangle}\right] \quad \forall t \in \mathbb{R}^n$$

**Teorema 1.3.** *Siano $X, Y$ vettori aleatori in $\mathbb{R}^n$ allora*

$$M_X = M_Y \quad \Rightarrow \quad X \sim Y$$

**Teorema 1.4** (Legge forte dei grandi numeri). *Sia $(X_n)_{n \in \mathbb{N}}$ una successione di variabili aleatorie definite sullo stesso spazio i.i.d con media $\mu$. Per ogni $n$ definiamo la media campionaria come*

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

*Allora con probabilità 1 si ha*

$$\lim_{n \to +\infty} \overline{X}_n = \mu$$

**Lemma 1.5** (di Borel Cantelli - prima parte). *Sia $(A_n)$ una successione di eventi.*

$$\sum_{n \in \mathbb{N}} \mathbb{P}(A_n) < +\infty \quad \Rightarrow \quad P\left(A_n \; i.o\right) = 0$$

**Teorema 1.6** (Disuguaglianza di Jensen). *Sia $X$ una v.a. reale con momento primo finito. Sia $\varphi: \mathbb{R} \to \mathbb{R}$ convessa allora vale*

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

## 2 Numeri di ricoprimento e impacchettamento

**Definizione 2.1.** Sia $(K, d)$ uno spazio metrico, dato $\varepsilon > 0$. Diremo che $A \subseteq K$ è un **$\varepsilon$-net** di $K$ se

$$K \subseteq \bigcup_{x \in A} B(x, \varepsilon)$$

In altre parole,

$$\forall x \in K \quad \exists x_0 \in A \quad d(x, x_0) \leq \varepsilon$$

*Osservazione* 2. Per qualsiasi $\varepsilon$ esiste un $\epsilon$-net: tutto lo spazio è un $\varepsilon$-net.

**Definizione 2.2.** Il **numero di ricoprimento** (covering number) di $K$ (di parametro $\varepsilon$) è definito come

$$N(K, d, \varepsilon) = \inf \{|A| \ : \ A \subseteq K \ \varepsilon\text{-net}\}$$

Dalla definizione di spazio compatto, segue la seguente

**Proposizione 2.1.** *Il numero di ricoprimento di uno spazio compatto è finito*

*Dimostrazione.* Gli insiemi $\{B^\circ(x, \varepsilon)\}_{x \in K}$ sono un ricoprimento aperto di $K$. Per compattezza $\exists x_1, \ldots, x_n \in K$ tale che

$$K \subseteq \bigcup_{i=1}^{n} B^\circ(x_i, \varepsilon) \subseteq \bigcup_{i=1}^{n} B^\circ(x_i, \varepsilon)$$

dunque $\{x_1, \ldots, x_n\}$ formano un $\varepsilon$-net di $K$. $N(k, d, \varepsilon) \leq n$.

$\square$

**Definizione 2.3.** Sia $(K, d)$ uno spazio metrico, dato $\varepsilon > 0$. Diremo che $A \subseteq K$ è un **$\varepsilon$-separato** da $K$ se

$$\forall x, y \in A \quad x \neq y \quad \Rightarrow \quad d(x, y) > \varepsilon$$

*Osservazione* 3. Per ogni $\varepsilon$, esiste un insieme $\varepsilon$-separato di $K$. Il vuoto è $\varepsilon$-serpatato.

**Definizione 2.4.** Sia $(K, d)$ uno spazio metrico, dato $\varepsilon > 0$. Diremo che $A \subseteq K$ è un **$\varepsilon$-net** di $K$ se

$$K \subseteq \bigcup_{x \in A} B(x, \varepsilon)$$

**Definizione 2.5.** Il **numero d'impacchettamento** di $K$ (di parametro $\varepsilon$) è definito come

$$P(K, d, \varepsilon) = \sup \{|A| \ : \ A \subseteq K \ \varepsilon\text{-separato}\}$$

## 2.1 Equivalenza tra i due numeri

**Lemma 2.2.** *Sia $(K, d)$ metrico non vuoto. Per ogni $\varepsilon > 0$ esiste un insieme $\varepsilon$-separato massimale (rispetto all'inclusione).*

*Dimostrazione.* Se $K$ è finito la tesi segue banalmente (tra tutti gli $\varepsilon$-separati si prende quello di cardinalità massima).
Se $K$ è infinito consideriamo

$$X = \{A \subseteq X \ : \ A \ \ \varepsilon - \text{separato}\}$$

allora $(X, \subseteq)$ è un insieme parzialmente ordinato.
Se mostriamo che ogni catena ha un maggiorante possiamo concludere per Zorn.
Sia $Y \subseteq X$ una catena, proviamo che $\bigcup Y \in X$ e dunque è un maggiorante per la catena.
Siano $x, z \in \bigcup Y$ allora $x \in A_1, \ z \in A_2$. Essendo $Y$ una catena, possiamo assumere senza perdità di generalità $A_1 \subseteq A_2$ e dunque $x, z \in A_2$. Ma $A_2$ è $\varepsilon$-separato dunnque $d(x, z) > \varepsilon$. □

**Lemma 2.3.** *Un $\varepsilon$-separato massimale è un $\varepsilon$-net*

*Dimostrazione.* Sia $A$ un $\varepsilon$-separato massimale allora dobbiamo provare che

$$\forall x \in K \quad \exists x' \in A \quad d(x, x') \leq \varepsilon$$

- Se $x \in A$ allora basta prendere $x' = x$
- Se $x \notin A$ allora per massimalità $A \cup \{x\}$ non può essere $\varepsilon$-separato. Esistono dunque $y, z \in A \cup \{x\}$ con $d(y, z) \leq \varepsilon$.
  Poichè $A$ è $\varepsilon$-separato, uno tra $y$ e $z$ deve essere $x$. Supponiamo $y = x$, allora basta porre $x' = z$.

**Teorema 2.4** (Equivalenza tra covering e packing number)**.** *Per ogni spazio metrico $(K, d)$ e per ogni $\varepsilon > 0$ vale*

$$P(k, d, 2\varepsilon) \leq N(k, d, \varepsilon) \leq P(k, d, \varepsilon)$$

*Dimostrazione.* Sia $A_\star \subseteq K$ un $\varepsilon$-separato massimale; per il lemma precedente $A_\star$ è anche un $\epsilon$-net.
Dunque vale

$$P(k, d, \varepsilon) = \sup \{|A| \ : \ A \subseteq K \ \varepsilon\text{-separato}\} \geq |A_\star| \geq \inf \{|A| \ : \ A \subseteq K \ \varepsilon\text{-net}\}$$

Proviamo l'altra disuguaglianza. Sia $P \subseteq K$ un $2\varepsilon$-separato e $N \subseteq K$ un $\varepsilon$-net.
Sia $x \in P$ allora poichè $N$ è $\varepsilon$-net, $\exists x_0 \in N$ tale che $d(x, x_0)$. Pongo $f(x) = x_0$.
Ho dunque definito una funzione $f : P \to N$ (facendo ricorso all'assioma di scelta se $N$ non è finito).
Proviamo che $f$ è iniettiva. Siano $x, y \in P$ con $f(x) = f(y) = x_0$ allora dalla definizione di $f$ segue che

$$d(x, x_0), d(y, y_0) \leq \varepsilon \quad \Rightarrow \quad d(x, y) \leq d(x, x_0) + d(x_0, y) \leq 2\varepsilon$$

Poichè $x, y \in P$ che è $2\varepsilon - \mathit{separato}$, deve accadere $x = y$.
Abbiamo dunque provato che $f$ è iniettiva e dunque $|P| \leq |N|$, si conclude per l'arbitrarietà di $P$ e $N$ □

## 2.2 Stime per sottoinsiemi di $\mathbb{R}^n$

**Definizione 2.6.** Dati $A, B \subseteq \mathbb{R}^n$ definiamo la loro **somma di Minkowski** come

$$A + B = \{a + b \; : \; a \in A, \, b \in B\}$$

**Proposizione 2.5.** *Sia $K \subseteq \mathbb{R}^n$ Lebesgue misurabile allora*

$$\frac{l(K)}{l(B(0,\varepsilon))} \leq N(K,\varepsilon) \leq P(K,\varepsilon) \leq \frac{l\left(K + B\left(0, \frac{\varepsilon}{2}\right)\right)}{l\left(B\left(0, \frac{\varepsilon}{2}\right)\right)}$$

*dove $l(A)$ indica la misura n-dimensionale di Lebesgue di $A$*

*Dimostrazione.*

- Proviamo la prima disuguaglianza. Sia $A \subseteq K$ un $\varepsilon$-net

  $$K \subseteq \bigcup_{x \in A} B(x,\varepsilon) \quad \Rightarrow \quad l(K) \leq \sum_{x \in A} l(B(x,\varepsilon)) = |A| \, l(B(0,\varepsilon))$$

  ovvero abbiamo provato che

  $$|A| \geq \frac{l(K)}{l(B(0,\varepsilon))}$$

  passando all'estremo inferiore in $A$ la tesi.

- La seconda disuguaglianza segue dal teorema di equivalenza.

- Proviamo la terza disuguaglianza. Sia $A \subseteq K$ un $\varepsilon$-separato Notiamo che

  $$\bigcup_{x \in A} B\left(x, \frac{\varepsilon}{2}\right) = A + B\left(0, \frac{\varepsilon}{2}\right) \subseteq K + B\left(0, \frac{\varepsilon}{2}\right)$$

  Dunque usando la monotonia della misura di Lebesgue si ha

  $$l\left(\bigcup_{x \in A} B\left(x, \frac{\varepsilon}{2}\right)\right) \leq l\left(K + B\left(0, \frac{\varepsilon}{2}\right)\right)$$

  Ma la prima unione è disgiunta infatti essendo $A$ un $\epsilon$-separato e $\mathbb{R}^d$ normato vale

  $$\Big(\forall x, y \in A \text{ distinti} \quad d(x,y) \geq \varepsilon\Big) \quad \Rightarrow \quad \left(\forall x, y \in A \text{ distinti} \quad B\left(x, \frac{\varepsilon}{2}\right) \cap B\left(y, \frac{\varepsilon}{2}\right) = \emptyset\right)$$

  quindi abbiamo

  $$|A| \, l\left(B\left(0, \frac{\varepsilon}{2}\right)\right) = \sum_{x \in A} l\left(B\left(x, \frac{\varepsilon}{2}\right)\right) \leq l\left(K + B\left(0, \frac{\varepsilon}{2}\right)\right)$$

  Abbiamo dunque

  $$|A| \leq \frac{l\left(K + B\left(0, \frac{\varepsilon}{2}\right)\right)}{l\left(B\left(0, \frac{\varepsilon}{2}\right)\right)}$$

  e passando al sup in $A$ la tesi

□

**Corollario 2.6.** *Se $l\left(K + B\left(0, \frac{\varepsilon}{2}\right)\right) < \infty$ allora i numeri di impacchettamento e di ricoprimento sono finiti. In particolare se $K$ è limitato lo sono*

**Corollario 2.7.** *Sia $B(0,1)$ la palla unitaria in $\mathbb{R}^n$ allora vale*

$$\left(\frac{1}{\varepsilon}\right)^n \leq N(B(0,1),\varepsilon) \leq P(B(0,1),\varepsilon) \leq \left(\frac{2}{\varepsilon}+1\right)^n$$

*Dimostrazione.* Ricordiamo che se denotiamo con $\omega_n = l(B(0,1))$ allora $l(B(0,r)) = r^n\omega_n$. Dalla proposizione precedente vale

$$\frac{l(B(0,1))}{l(B(0,\varepsilon))} \leq N(B(0,1),\varepsilon) \leq P(B(0,1),\varepsilon) \leq \frac{l\left(B(0,1)+B\left(0,\frac{\varepsilon}{2}\right)\right)}{l\left(B\left(0,\frac{\varepsilon}{2}\right)\right)}$$

Ora

$$\frac{l(B(0,1))}{l(B(0,\varepsilon))} = \frac{\omega_n}{\varepsilon^n\omega_n}$$

Inoltre

$$B(0,1)+B\left(0,\frac{\varepsilon}{2}\right) = B\left(0,1+\frac{\varepsilon}{2}\right)$$

e dunque la tesi

□

**Corollario 2.8.** *Sia $S^{n-1}$ la sfera unitaria di $\mathbb{R}^n$ allora vale*

$$N\left(S^{n-1},\varepsilon\right) \leq P\left(S^{n-1},\varepsilon\right) \leq \left(\frac{2}{\varepsilon}+1\right)^n$$

*Dimostrazione.* Dalla proposizione precedente si ha

$$N\left(S^{n-1},\varepsilon\right) \leq P\left(S^{n-1},\varepsilon\right) \leq \frac{l\left(S^{n-1}+B\left(0,\frac{\varepsilon}{2}\right)\right)}{l\left(B\left(0,\frac{\varepsilon}{2}\right)\right)}$$

Ma

$$S^{n-1} \subseteq B(0,1) \quad \Rightarrow \quad S^{n-1}+B\left(0,\frac{\varepsilon}{2}\right) \subseteq B(0,1)+B\left(0,\frac{\varepsilon}{2}\right) = B\left(0,1+\frac{\varepsilon}{2}\right)$$

Usando la monotonia della misura, la tesi.

□

*Osservazione* 4. Nelle applicazioni $\varepsilon$ piccolo. Se $\varepsilon \in (0,1)$ allora $\frac{2}{\varepsilon}+1 \leq \frac{3}{\varepsilon}$ e dunque vale

$$\left(\frac{1}{\varepsilon}\right)^n \leq l\left(B(0,1)\right) \leq \left(\frac{3}{\varepsilon}\right)^n$$

*Osservazione* 5. Gli $\varepsilon$-net sono importanti per fare approssimazione.
Supponiamo ad esempio di avere una funzione $f : K \to \mathbb{R}$ L-lipschitz e siamo interessati a calcolare $\sup f$.
Sia $A$ un $\varepsilon$-net finito di $K$, allora $\forall x \in K$ esiste $x_0 \in A$ con $x \in B(x_0,\varepsilon)$.
Poichè

$$|f(x)-f(x_0)| \leq Ld(x,x_0) \leq L\varepsilon$$

Otteniamo

$$\sup_{x\in K} f(x) \leq \max_{x\in A} f(x) + L\varepsilon$$

# 3 Teorema di Glivenko-Cantelli

**Definizione 3.1.** La **misura empirica** del campione aleatorio $X_1, \dots, X_n$ è data da

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

*Osservazione* 6. $\mathbb{P}_n$ è una misura di probabilità su $\chi$ che è aleatoria: dipende dalla realizzazione del campione. In generale, $\mathbb{P}_n$ è una misura atomica su $\chi$ che associa ad ogni atomo la frequenza relativa con cui appare in $X_1, \dots, X_n$

**Definizione 3.2.** Se $\chi = \mathbb{R}$ possiamo definire la **funzione di ripartizione empirica** come

$$F_n \ : \ \mathbb{R} \to [0,1] \qquad F_n(t) = \mathbb{P}_n[(-\infty, t]) = \frac{1}{n}\sum_{i=1}^{n} 1_{[-\infty,t]}(X_i)$$

**Lemma 3.1.** *Sia $F$ una funzione di ripartizione. Allora $\forall \varepsilon \in (0,1)$ esiste una partizione*

$$-\infty = t_0 < t_1 < \dots < t_{k-1} < t_k = \infty \qquad t.c \qquad F(t_i^-) - F(t_{i-1}^-) \leq \varepsilon \quad \forall i = 1, \dots, k$$

*dove intendiamo che $F(t_k^-) = 1$ e $F(t_0) = 0$.*

*Dimostrazione.* Poniamo

$$t_1 = \sup\left\{ t \in \mathbb{R} \mid F(t) \leq \frac{\varepsilon}{2} \right\}$$

notiamo che tale estremo superiore esiste finito, infatti per le propietà delle funzioni di ripartizione

$$\lim_{x\to-\infty} F(x) = 0 \qquad \lim_{x\to+\infty} F(x) = 1$$

$t_1$ soddisfa la disuguaglianza:

$$F(t_1^-) - F(t_0) = F(t_1^-) = \lim_{t\to t_1^-} F(t) \leq \frac{\varepsilon}{2} < \varepsilon$$

ed inoltre $F(t_1) \geq \frac{\varepsilon}{2}$ infatti se per assurdo $F(t_1) < \frac{\varepsilon}{2}$ allora sfruttando la continuità a destra della $F$, esisterebbe un $s > t_1$ con $F(s) < \frac{\varepsilon}{2}$ contro la definizione di $t_1$.
Definiamo

$$t_2 = \sup\left\{ t \in \mathbb{R} \mid F(t) \leq F(t_1) + \frac{\varepsilon}{2} \right\}$$

Se $t_2 = +\infty$ pongo $k = 2$ e concludo infatti se $t_2 = +\infty$ allora

$$\forall t \in \mathbb{R} \qquad F(t) \leq F(t_1) + \frac{\varepsilon}{2}$$

passando al limite per $t \to +\infty$ ottengo

$$1 \leq F(t_1) + \frac{\varepsilon}{2} \quad \Leftrightarrow \quad 1 - F(t_1) \leq \frac{\varepsilon}{2} < \varepsilon$$

Se $t_2 < \infty$ allora sia

$$t_3 = \sup\left\{ t \in \mathbb{R} \mid F(t) \leq F(t_2) + \frac{\varepsilon}{2} \right\}$$

posso iterare la procedura definendo $t_4, \dots, t_k$ fino a quando si avrà $t_k = +\infty$.
Il procedimento termina in un numero finito di passi infatti come abbiamo provato che $F(t_1) \geq \varepsilon 2$ si prova che

$$t_i < \infty \quad \Rightarrow \quad F(t_{i+1}) \geq F(t_i) + \frac{\varepsilon}{2}$$

e dunque $F(t_i) \geq \frac{i\varepsilon}{2}$ si conclude ricordando che $F$ è limitata.

**Teorema 3.2** (di Glivenko-Cantelli). *Sia $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie definite sullo stesso spazio i.i.d. Se $F$ è la funzione di ripartizione di $X_1$ allora quasi certamente si ha*

$$\lim_{n\to+\infty} ||F_n - F||_\infty = 0$$

*Dimostrazione.* Per la legge forte dei grandi numeri, fissato $t \in \mathbb{R}$ si ha

$$\lim_{n\to+\infty} F_n(t) = F(t) \text{ q.o}$$

infatti

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(X_i)$$

ovvero $F_n$ è la media campionaria di $n$ variabili i.i.d. con media $P(X_1 \leq t) = F(t)$.

Ricordando che le funzioni di ripartizione sono continue a destra con limite a sinistra si ottiene che

$$F_n(t^-) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t)} \to \mathbb{E}\left[1_{(-\infty,t)}(X_1)\right] = F(t^-) \text{ q.o}$$

Sia $t_0 < \cdots < t_k$ una partizione come nel Lemma 3.1 e sia $t \in \mathbb{R}$. Poniamo $i$ come l'unico indice che verifica $t_{i-1} \leq t < t_i$.
Usando la monotonia di $F_n$ e di $F$ ottengo

$$F_n(t) - F(t) \leq F_n(t_i^-) - F(t_i^-) + \varepsilon$$

$$F_n(t) - F(t) \geq F_n(t_{i-1}) - F(t_{i-1}) - \varepsilon$$

ovvero mettendo insieme le due disuguaglianze

$$|F_n(t) - F(t)| \leq \max\left\{\max_i \left|F_n(t_i^-) - F(t_i^-) + \varepsilon\right|, \max_i |F_n(t_{i-1}) - F(t_{i-1}) - \varepsilon|\right\}$$

Per $n \to +\infty$ il membro di destra tende a $\varepsilon$ dunque si ha

$$\forall\varepsilon \qquad \limsup_{n\to+\infty} ||F_n - F||_\infty \leq \varepsilon$$

Per l'arbitrarietà di $\varepsilon$ si ha la tesi.

□

## 3.1 Disuguaglianza DWK

**Teorema 3.3** (Disguaglianza DKW). *Dato un campione aleatorio $\{X_i\}_{i\in\mathbb{N}}$, sia $\mathbb{F}_n$ la funzione di ripartizione empirica e $F$ la funzione di ripartizione di $X_1$.*
*Vale*

$$\mathbb{P}\left(||\mathbb{F}_n - F||_\infty \geq \frac{x}{\sqrt{n}}\right) \leq 2e^{-2x^2}$$

*Osservazione* 7. La disuguaglianza prende il nome dai 3 matematici che l'hanno dimostrata: Dvoretxky, Kiefer e Wolfowitz. Ad onore del vero, venne dimostrata con un valore della costante diverso da 2. La versione con il 2 si deve Massart.

Dalla disuguaglianza DWK segue il Teorema di Glivenko-Cantelli:

**Corollario 3.4.** *$\forall n \in \mathbb{N}$ con probabilità 1 vale*

$$||F_n - F|| \leq \sqrt{\frac{\ln n}{n}}$$

*e dunque in particolare vale il teorema di Givenko-Cantelli.*

*Dimostrazione.* Consideriamo la successione di eventi

$$A_n = \left\{ ||F_n - F|| > \sqrt{\frac{\ln n}{n}} \right\}$$

Allora dalla disuguaglianza DKW si ha

$$\mathbb{P}(A_n) \leq 2e^{-2\ln n} = \frac{2}{n^2}$$

e dunque

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) \leq 2 \sum_{i=1}^{\infty} \frac{1}{n^2} < \infty$$

Si conclude per Borel-Cantelli.

□

## 3.2 Classi di Glivenko-Cantelli

**Definizione 3.3.** Date $\mu, \nu$ due misure su $\chi$ e supponiamo che ogni $f \in \mathfrak{F}$ sia integrabile (rispetto alle due misure) allora

$$||\mu - \nu||_{\mathfrak{F}} = \sup_{f \in \mathfrak{F}} |\mu(f) - \nu(f)|$$

dove

$$\mu(f) = \int_{\chi} f(x)\, \mathrm{d}\mu(x)$$

**Definizione 3.4.** 0 La famiglia $\mathfrak{F}$ si dice di **Glivenko-Cantelli in senso forte** se

$$||\mathbb{P}_n - \mathbb{P}_X||_{\mathfrak{F}} \to 0 \text{ per } n \to +\infty \text{ q.c.}$$

La famiglia $\mathfrak{F}$ si dice di **Glivenko-Cantelli in senso debole** se

$$||\mathbb{P}_n - \mathbb{P}_X||_{\mathfrak{F}} \to 0 \text{ per } n \to +\infty \text{ in probabilità}$$

*Osservazione* 8. Se $\mathcal{F}$ è di Glivenko-Cantelli allora

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)] \right| \to 0 \text{ q.c} \tag{1}$$

ho dunque una famiglia di leggi forti dei grandi numeri: $\forall f \in \mathcal{F}$ ho la legge forte dei grandi numeri per la successione i.i.d $f(X_1), f(X_2), \dots$.
Siccome in (1) ho l'estremo superiore fatto su $\mathcal{F}$, la convergenza a 0 è uniforme: si parla in questo caso di legge uniforme dei grandi numeri.

## 3.3 Applicazioni di Glivenko-Cantelli

**Proposizione 3.5** (Principio plug-in)**.** *Sia*

$$\mathcal{A} = \{G \ : \mathbb{R} \to [0,1] \mid G \text{ funzione di ripartizione}\}$$

*dotato della norma uniforme. Sia $\mathcal{A}_0 \subseteq \mathcal{A}$ e $\gamma : \ \mathcal{A}_0 \to \mathbb{R}$ una funzione continua, $(X_n)_{n\in\mathbb{N}}$ variabili aleatorie i.i.d. definite sullo stesso spazio e supponiamo $\mathbb{F}_n, F \in \mathcal{A}_0$ allora quasi certamente vale*

$$\gamma(\mathbb{F}_n) \to \gamma(F) \text{ per } n \to +\infty$$

*Dimostrazione.* Per il teorema di Glivenko-Cantelli $||\mathbb{F}_n - F||_\infty \to 0$ quasi certamente. Si conclude sfruttando la continuità di $\gamma$.

$\square$

*Osservazione* 9. La proposizione ci dice che uno stimatore ragionevole di $\gamma(F)$ (che non conosciamo, perchè non conosciamo $F$) è $\gamma(\mathbb{F}_n)$ .
Il nome del principio deriva dal fatto che per stimare la quantità: tolgo $F$ e metto dentro $\mathbb{F}_n$

**Definizione 3.5.** Il funzionale di Cramer-Von Mises è

$$\gamma : \mathcal{A} \to \mathbb{R} \qquad \gamma(G) = \int_R (G(t) - F_0(t))^2 \, \mathrm{d}F_0(t)$$

dove $F_0(t)$ è una fissata funzione di ripartizione

*Osservazione* 10. $\gamma(F)$ misura quanto $F$ si discosta da $F_0$

**Lemma 3.6.** *Il funzionale di Cramer-Von Mises è continuo*

*Dimostrazione.* Siano $G, \tilde{G} \in \mathcal{A}$

$$\begin{aligned}\left|\gamma(G) - \gamma\left(\tilde{G}\right)\right| &= \left|\int_{\mathbb{R}} (G(t) - F_0(t))^2 - \left(\tilde{G}(t) - F_0(t)\right)^2 \mathrm{d}F_0(t)\right| = \\ &= \left|\int_{\mathbb{R}} \left(G(t) - \tilde{G}(t)\right)\left(G(t) - F_0(t) + \tilde{G}(t) - F_0(t)\right) \mathrm{d}F_0(t)\right| \leq \\ &\leq \int_{\mathbb{R}} \left|G(t) - \tilde{G}(t)\right| \left(|G(t) - F_0(t)| + \left|\tilde{G}(t) - F_0(t)\right|\right) \mathrm{d}F_0(t)\end{aligned}$$

Ora l'argomento del secondo e terzo modulo è compreso in $[-1,1]$ e dunque

$$\left|\gamma(G) - \gamma\left(\tilde{G}\right)\right| \leq 2\left|\left|G - \tilde{G}\right|\right|_\infty$$

infatti ricordiamo che stiamo integrando rispetto a misure di probabilità.

$\square$

**Esempio 3.7** (Goodness of fit test)**.** *Vogliamo verificate l'ipotesi che la distribuzione della popolazione su $\mathbb{R}$ sia una data $\mathbb{P}_0$ o quasi (equivalentemente che la funzione di ripartizione effettiva sia una data $F_0$ o quasi). Per fare questo posso usare il principio plug-in e stimare $\gamma(F)$ con $\gamma(\mathbb{F}_n)$*

# 4 Complessità di Rademacher

**Definizione 4.1.** Una variabile aleatoria si dice di **Rademacher** se ha distribuzione

$$f(x) = \begin{cases} \frac{1}{2} & \text{se } x = \pm 1 \\ 0 & \text{altrimenti} \end{cases}$$

*Osservazione* 11. A meno di allargare lo spazio di probabilità possiamo supporre che su $\Omega$ sia definito

- $\{X_i\}_{i=1}^n$ ovvero il campione aleatorio
- $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. di Rademacher

ed inoltre con i vettori $(X_i)_{i=1}^n$ e $(\varepsilon_i)_{i=1}^n$ indipendenti.

**Definizione 4.2.** La **complessità di Rademacher** della classe $\mathfrak{F}$ riferita al campione $\{X_i\}_{i=1}^n$ è

$$R_n(\mathfrak{F}) = \mathbb{E}_{X,\varepsilon}\left[\sup_{f \in \mathfrak{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]$$

**Proposizione 4.1.** *Valgono le seguenti propietà per la complessità di Rademacher*

*(i) Date $\mathcal{F}$ e $\mathcal{G}$ classi di funzioni si ha*

$$\mathcal{R}_n(\mathcal{F} + \mathcal{G}) \leq \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$$

*(ii) Data $\mathcal{F}$ classe di funzioni e $g : \chi \to \mathbb{R}$ uniformemente limitata vale*

$$|\mathcal{R}_n(\mathcal{F} + g) - \mathcal{R}_n(\mathcal{F})| \leq \frac{||g||_\infty}{\sqrt{n}}$$

*dove*

$$\mathcal{F} + g = \{f + g \mid f \in \mathcal{F}\}$$

*(iii) Data $\mathcal{F}$ classe di funzioni e $X$ v.a. a valori in $\chi$, definiamo la classe*

$$\overline{\mathcal{F}} = \{f - \mathbb{E}[f(X)] \mid f \in \mathcal{F}\}$$

*Allora si ha*

$$\mathcal{R}_n(\mathcal{F}) - \mathcal{R}_n(\overline{\mathcal{F}}) \leq \sup_{f \in \mathcal{F}} \frac{|f(X)|}{\sqrt{n}}$$

*Dimostrazione.*

(i)

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F} + \mathcal{G}) = \mathbb{E}_{X,\varepsilon}&\left[\sup_{\substack{f \in \mathcal{F} \\ g \in \mathcal{G}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(f(X_i) - g(X_i)\right) \right| \right] \leq_1 \\
\leq_1 \mathbb{E}_{X,\varepsilon}&\left[\sup_{\substack{f \in \mathcal{F} \\ g \in \mathcal{G}}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| + \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) \right| \right] \leq \\
\leq \mathbb{E}_{X,\varepsilon}&\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right| + \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(X_i) \right| \right] = \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})
\end{aligned}$$

dove $\leq_1$ è la disuguaglianza triangolare

(ii) Proviamo prima che

$$\mathcal{R}_n(\mathcal{F}+g) \leq \mathcal{R}_n(\mathcal{F}) + \frac{||g||_\infty}{\sqrt{n}} \tag{2}$$

Se poniamo $\mathcal{G} = \{g\}$ allora vale

$$\mathcal{R}_n(\mathcal{F}+g) = \mathcal{R}_n(\mathcal{F}+\mathcal{G})$$

e dunque dal punto $(i)$ si ha

$$\mathcal{R}_n(\mathcal{F}+\mathcal{G}) \leq \mathcal{R}_n(\mathcal{F}) + \mathcal{R}_n(\mathcal{G})$$

Ora

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right]$$

Dalla disuguaglianza di Cauchy-Swartz si ha

$$\mathbb{E}_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right] \leq \mathbb{E}_{X,\varepsilon}\left[\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right)^2\right]^{\frac{1}{2}} =$$

$$= \frac{1}{n}\left(\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\varepsilon_i\varepsilon_j g(X_i)g(X_j)]\right)^{\frac{1}{2}} = \frac{1}{n}\left(\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\varepsilon_i\varepsilon_j]\mathbb{E}[g(X_i)g(X_j)]\right)^{\frac{1}{2}}$$

dove l'ultima disuguaglianza segue dall'indipendenza dei vettori $(\varepsilon_1, \dots, \varepsilon_n)$ e $(g(X_1), \dots, g(X_n))$. Poichè

$$\mathbb{E}[\varepsilon_i\varepsilon_j] = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}$$

otteniamo

$$E_{X,\varepsilon}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\right|\right] \leq \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}\left[g(X_i)^2\right]\right)^{\frac{1}{2}} \leq \frac{||g||_\infty}{\sqrt{n}}$$

Per l'altra disuguaglianza, usiamo (2) rimpiazzando $\mathcal{F}$ con $\mathcal{F}+g$ e $g$ con $-g$ ottenendo

$$\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n((\mathcal{F}+g)-g) \leq \mathcal{R}_n(\mathcal{F}+g) + \frac{||-g||_\infty}{\sqrt{n}}$$

(iii) Definiamo una nuova classe di funzioni costanti:

$$\mathbb{E}[\mathcal{F}] = \{\mathbb{E}[f(X)] \mid f \in \mathcal{F}\}$$

Notiamo che

$$\mathcal{F} \subseteq \overline{\mathcal{F}} + \mathbb{E}[\mathcal{F}]$$

infatti

$$\forall f \in \mathcal{F} \qquad f = (f - \mathbb{E}[f(X)]) + \mathbb{E}[f(X)]$$

Dal punto $(i)$ si ha

$$\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\overline{\mathcal{F}}) + \mathcal{R}_n(\mathbb{E}[\mathcal{F}])$$

Ma

$$\mathcal{R}_n(\mathbb{E}[\mathcal{F}]) = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{E}[f(X_1)]\right|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] =$$

$$= \sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|\,\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] = \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right]\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|$$

Poichè da Cauchy-Swartz si ha

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right|\right] \leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)^2\right]^{\frac{1}{2}} = \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\varepsilon_i\varepsilon_j\right] = \frac{1}{\sqrt{n}}$$

**Lemma 4.2** (di simmetrizzazione)**.** *Siano $X_1, \dots, X_n, Y_1, \dots, Y_n$ v.a. indipendenti e distribuiti come $X$. Siano $\varepsilon_1, \dots, \varepsilon_n$ variabili di Rademacher indipendenti tra loro e dalle $X_i, Y_j$. Se $f : \chi \to \mathbb{R}$ allora i vettori*

$$(f(X_1) - f(Y_1), \dots, f(X_n) - f(Y_n)) \;\; e \;\; (\varepsilon_1 (f(X_1) - f(Y_1)), \dots, \varepsilon_n (f(X_n) - f(Y_n)))$$

*hanno la stessa legge.*

*Dimostrazione.* Siano $V, W$ i due suddetti vettori aleatori. Poichè le entrate dei due vettori sono indipendenti tra loro, basta provare che $\forall A$ boreliano vale

$$\mathbb{P}(f(X_i) - f(Y_i) \in A) = \mathbb{P}\left(\varepsilon_i(f(X_i) - f(Y_i)) \in A\right)$$

Ora

$$\begin{aligned} &\mathbb{P}\left(\varepsilon_i(f(X_i) - f(Y_i)) \in A\right) = \\ &= \frac{1}{2}\mathbb{P}\left(\varepsilon_i(f(X_i) - f(Y_i)) \in A \mid \varepsilon_i = 1\right) + \frac{1}{2}\mathbb{P}\left(\varepsilon_i(f(X_i) - f(Y_i)) \in A \mid \varepsilon_i = -1\right) = \\ &= \frac{1}{2}\left[\mathbb{P}\left(f(X_i) - f(Y_i) \in A\right) + \mathbb{P}\left(f(Y_i) - f(X_i) \in A\right)\right] \end{aligned}$$

infatti gli eventi $\{f(X_i) - f(Y_i) \in A\}$ e $\{\varepsilon_i = 1\}$ sono indipendenti (così come gli altri due ). Ora poichè $(X_i, Y_i)$ e $(Y_i, X_i)$ hanno entrate indipendenti con medesima legge dunque

$$\mathbb{P}\left(f(X_i) - f(Y_i) \in A\right) = \mathbb{P}\left(f(Y_i) - f(X_i) \in A\right)$$

da cui la tesi.

□

**Teorema 4.3.** *Sia $\mathcal{F} = \{f : \chi \to \mathbb{R}\}$ una classe di funzioni. Allora*

$$\frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}}) \leq \mathbb{E}\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right] \leq 2\mathcal{R}_n(\mathcal{F})$$

*Dimostrazione.* Proviamo la seconda disuguaglianza. Siano $(Y_1, \dots, Y_n)$ un vettore casuale distribuito come $(X_1, \dots, X_n)$ e indipendente. Sia $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ un vettore di Rademacher indipendente dalle $X_i$ e $Y_j$.

$$\mathbb{E}\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right] = \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|\right] =$$

$$= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_Y[f(Y_i)]\right|\right] = \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right]\right|\right] \leq$$

$$\leq \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\mathbb{E}_Y\left[\left|\frac{1}{n}\sum_{i=1}^n (f(X_i) - f(Y_i))\right|\right]\right] \leq_1 \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right] =_2$$

$$=_2 \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\right|\right] \leq_3 \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left(\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right)\right] \leq$$

$$\leq \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right] =$$

$$= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(Y_i)\right|\right] = 2\mathcal{R}_n(\mathcal{F})$$

dove

- $\leq_1$ è il Lemma 15.1,
- $=_2$ è il lemma di simmetrizzazione e
- $\leq_3$ è la disuguaglianza triangolare

Proviamo la prima disuguaglianza

$$\begin{aligned}
\mathcal{R}_n(\overline{\mathcal{F}}) =& \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\overline{\mathcal{F}}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] = E_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - \mathbb{E}[f(X)]\right|\right] = \\
=& E_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - \mathbb{E}[f(Y_i)])\right|\right] == E_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_Y\left[\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\right|\right]\right] \\
\leq_1& \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}n\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i)\right|\right] =_2 \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right] \leq_3 \\
\leq_3& \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right|\right\} + \left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}[f(Y_i)]\right|\right] \leq \\
\leq& \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X_i)]\right|\right] + \mathbb{E}_{X,Y}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}[f(Y_i)]\right| =
\end{aligned}$$

$$= 2\mathbb{E}_X\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right]$$

dove

- $\leq_1$ è è il Lemma 15.1,
- $=_2$ è il lemma di simmetrizzazione
- $\leq_3$ si ottiene sommando e sottraendo $\mathbb{E}[f(X_i)] = \mathbb{E}[f(Y_i)]$ e applicando la disuguaglianza triangolare

$\square$

*Osservazione* 12. Usando il punto (iii) della proposizione della scorsa lezione, otteniamo

$$\frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \sup_{f\in\mathcal{F}} \frac{|\mathbb{E}[f(X)]|}{2\sqrt{n}} \leq \mathbb{E}\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right] \leq 2\mathcal{R}_n(\mathcal{F})$$

Inoltre, nelle applicazioni la classe $\mathcal{F}$ è formata da funzioni limitate uniformemente da una costante $b$ e dunque si ha

$$\frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{b}{2\sqrt{n}} \leq \mathbb{E}\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right] \leq 2\mathcal{R}_n(\mathcal{F})$$

**Definizione 4.3.** Una funzione $g : \chi^n \to \mathbb{R}$ si dice che soddisfa la **proprietà delle differenze limitate** con parametri $(l_1, \dots, l_n)$ se

$$\forall k \in [n] \quad \forall x, x' \in \chi^n \quad (x_i = x_i' \;\forall i \neq k \quad \Rightarrow \quad |g(x) - g(x')| \leq l_k)$$

**Teorema 4.4.** *Sia $g : \chi^n \to \mathbb{R}$ una funzione misurabile che soddisfa la proprietà delle differenza finite con parametro $(l_1, \dots, l_n)$. Se $(X_1, \dots, X_n)$ è un vettore aleatorio su $\chi^n$ con coordinate indipendenti vale*

$$\mathbb{P}\left[|g(X_1, \dots, X_n) - \mathbb{E}[g(X_1, \dots, X_n)]| > t\right] \leq \exp\left(-\frac{2t^2}{\sum l_k^2}\right)$$

**Teorema 4.5.** *Sia $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} \mid \; ||f||_\infty \leq b\}$ e $\delta > 0$.*
*Con probabilità di $1 - 2\exp\left(-\frac{\delta^2 n}{2b^2}\right)$ vale*

$$\frac{1}{2}\mathcal{R}_n(\overline{\mathcal{F}}) - \delta \leq ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta$$

*Dimostrazione.* Data $f \in \mathcal{F}$, pongo $\overline{f} = f - \mathbb{E}[f(X)]$. Definiamo la funzione

$$g : \chi^n \to \mathbb{R} \qquad g(x_1, \dots, x_n) = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_j \overline{f}(x_j)\right|$$

Proviamo che $g$ soddisfa la proprietà delle differenza finite con parametro $\frac{2b}{n}$.
Poichè $g$ è invariante per permutazione delle coordinate, basta provare che

$$|g(x_1, \dots, x_n) - g(y_1, \dots, y_n)| \leq \frac{2b}{n} \quad \text{se } x_j = y_j \quad \forall j \neq 1$$

Fissata $f \in \mathcal{F}$ ho

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n \overline{f}(x_i)\right| - \sup_{h\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \overline{h}(y_i)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^n \overline{f}(x_i)\right| - \left|\frac{1}{n}\sum_{i=1}^n \overline{f}(y_i)\right| \leq \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \overline{f}(x_i) - \frac{1}{n}\sum_{i=1}^n \overline{f}(y_i)\right| = \left|\frac{1}{n}\left(\overline{f}(x_1) - \overline{f}(y_1)\right)\right| = \\
&= \left|\frac{1}{n}\left(f(x_1) - f(y_1)\right)\right| \leq \frac{1}{n}\left(|f(x_1)| + |f(y_1)|\right) \leq \frac{2b}{n}
\end{aligned}$$

Dal Teorema 4.4 otteniamo $\forall \delta >$ vale

$$\mathbb{P}\left(|\, ||\mathbb{P}_n - P_X||_{\mathcal{F}} - \mathbb{E}\left[||\mathbb{P}_n - P_X||_{\mathcal{F}}\right]| \geq \delta\right) \leq 2\exp\left(-\frac{n\delta^2}{2b^2}\right)$$

dunque con probabilità maggiore di $1 - 2\exp\left(-\frac{n\delta^2}{2b^2}\right)$ ho

$$\mathbb{E}[||\mathbb{P}_n - P_X||_{\mathcal{F}} - \delta \leq ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq \mathbb{E}[||\mathbb{P}_n - P_X||_{\mathcal{F}}] + \delta$$

La tesi segue applicando le stime date dal Teorema 4.3 □

*Osservazione* 13. Il teorema ci dice che la stima ottenuta in media vale con una probabilità alta (a meno di $\delta$) puntualmente.

**Corollario 4.6.** *Sia $\mathcal{F}$ una classe di funzioni uniformemente limitate da $b$*

$$\mathcal{F} \text{ è di Glivenko-Cantelli in senso forte rispetto a } P_X \quad \Leftrightarrow \quad \lim_{n\to+\infty} \mathcal{R}_n(\mathcal{F}) = 0$$

*Dimostrazione.* $\Leftarrow$ Supponiamo che $\mathcal{R}_n(\mathcal{F}) \to 0$, fissato $\delta > 0$ consideriamo la successione di eventi

$$A_n = \{||\mathbb{P}_n - P_X||_{\mathcal{F}} > 2\mathcal{R}_n(\mathcal{F}) + \delta\}$$

Allora dal Teorema 4.5 si ha $\mathbb{P}(A_n) \leq 2\exp\left(-\frac{n\delta^2}{2b^2}\right)$ e dunque $\sum \mathbb{P}(A_n) < \infty$. Per il lemma di Borel-Cantelli otteniamo $\mathbb{P}(A_n \text{ i.o }) = 0$ quindi

$$\forall\delta > 0 \text{ q.c } ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta \text{ definitivamente in } n$$

equivalentemente

$$\forall k \in \mathbb{N} \text{ q.c } ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \frac{1}{k} \text{ definitivamente in } n$$

poichè l'unione numerabile di eventi con probabilità 1 ha probabilità 1 si ha

$$\text{q.c } \forall k \in \mathbb{N} \quad ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \frac{1}{k} \text{ definitivamente in } n$$

e dunque

$$\text{q.c } \forall k \in \mathbb{N} \quad \limsup_{n\to+\infty} ||\mathbb{P}_n - P_X||_{\mathcal{F}} \leq \frac{1}{k}$$

Per l'aritrarietà di $k$, la tesi.

$\Rightarrow$ Supponiamo per assurdo che $\mathcal{R}_n(\mathcal{F}) \not\to 0$ e dunque $\limsup \mathcal{R}_n(\mathcal{F}) > 0$.
Dato $\delta > 0$ definiamo la successione di eventi

$$B_n = \left\{\frac{1}{n}\mathcal{R}_n(\mathcal{F}) - \frac{b}{2\sqrt{n}} > ||\mathbb{P}_n - P_X||_{\mathcal{F}}\right\}$$

Per il Teorema 4.5, $\mathbb{P}(B_n) \leq 2\exp\left(-\frac{n\delta^2}{2b^2}\right)$ e dunque $\sum \mathbb{P}(B_n) < \infty$.
Per il lemma di Borel-Cantelli otteniamo $\mathbb{P}(B_n \text{ i.o}) = 0$ quindi

$$\forall\delta \text{ q.c } \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{b}{2\sqrt{n}} \leq ||\mathbb{P}_n - P_X||_{\mathcal{F}} \text{ definitivamente in } n$$

Ragionando come fatto nella dimostrazione dell'altra implicazione otteniamo

$$\text{q.c } \limsup_{n\to+\infty} \mathcal{R}_n(\mathcal{F}) - \frac{1}{k} \leq \limsup_{n\to+\infty} ||\mathbb{P}_n - P_X||_{\mathcal{F}}$$

Per l'arbitrarietà di $k$ otteniamo

$$\text{q.c } \limsup_{n\to+\infty} \mathcal{R}_n(\mathcal{F}) \leq \limsup_{n\to+\infty} ||\mathbb{P}_n - P_X||_{\mathcal{F}}$$

Poichè il termine di sinistra è strettamente positivo, lo è anche quello di destra e contro l'ipotesi che $\mathcal{F}$ sia di Glivenko-Cantelli □

## 4.1 Discriminante polinomiale

D'ora in avanti useremo le due seguenti notazioni:

- La $n$-upla $(x_1, \dots, x_n)$ con $x_1^n$.
- Se $\mathcal{F}$ è una classe di funzioni su $\chi$ definiamo

$$\mathcal{F}(x_1^n) = \{(f(x_1), \dots, f(x_n)) \mid f \in \mathcal{F}\}$$

**Definizione 4.4.** Una classe di funzioni $\mathcal{F}$ su $\chi$ ha **discriminante polinomiale** di ordine $\nu$ se

$$\forall n \geq 1 \ \ \forall x_1^n \in \chi^n \quad |\mathcal{F}(x_1^n)| \leq (n+1)^\nu$$

**Lemma 4.7.** *Sia $\mathcal{F}$ una classe di funzioni a discriminante polinomiale di ordine $\nu \geq 1$. Allora $\forall n$ e $\forall x_1^n \in \chi^n$ vale*

$$\mathbb{E}_\varepsilon \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq 2D(x_1^n) \sqrt{\frac{\nu \ln(n+1)}{n}}$$

*dove*

$$D(x_1^n) = \sup_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n \frac{f(x_i)^2}{n}}$$

*Dimostrazione.* Fissati $x_1, \dots, x_n$ e sia $\lambda > 0$. Applicando la disuguaglianza di Jensen con $\varphi(t) = e^{\lambda t}$ otteniamo

$$\begin{aligned}
\exp\left( \lambda \mathbb{E}_\varepsilon \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \right) &\leq \mathbb{E}_\varepsilon \left[ \exp\left( \lambda \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right) \right] = \\
= \mathbb{E}_\varepsilon \left[ \exp\left( \lambda \sup_{a \in \mathcal{F}(x_1^n)} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \right) \right] &\leq E_\varepsilon \left[ \sup_{a \in \mathcal{F}(x_1^n)} \exp\left( \lambda \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i a_i \right| \right) \right] \leq \\
\leq \mathbb{E}_\varepsilon \left[ \sup_{a \in \mathcal{F}(x_1^n)} \left( e^{\frac{\lambda}{n} \sum \varepsilon_i a_i} + e^{-\frac{\lambda}{n} \sum \varepsilon_i a_i} \right) \right] &\leq E_\varepsilon \left[ \sum_{a \in \mathcal{F}(x_1^n)} e^{\frac{\lambda}{n} \sum \varepsilon_i a_i} + e^{-\frac{\lambda}{n} \sum \varepsilon_i a_i} \right] = \\
= \sum_{a \in \mathcal{F}(x_1^n)} E_\varepsilon \left[ e^{\frac{\lambda}{n} \sum \varepsilon_i a_i} + e^{-\frac{\lambda}{n} \sum \varepsilon_i a_i} \right] =_1 2 \sum_{a \in \mathcal{F}(x_1^n)} E_\varepsilon \left[ e^{\frac{\lambda}{n} \sum \varepsilon_i a_i} \right] &=_2 \sum_{a \in \mathcal{F}(x_1^n)} \prod_{i=1}^n E_\varepsilon \left[ e^{\frac{\lambda}{n} \varepsilon_i a_i} \right] = \\
= 2 \sum_{a \in \mathcal{F}(x_1^n)} \prod_{i=1}^n \frac{e^{\frac{\lambda}{n} a_i} + e^{-\frac{\lambda}{n} a_i}}{2} \leq_1 2 \sum_{a \in \mathcal{F}(x_1^n)} \prod_{i=1}^n \exp\left( \frac{\lambda^2 a_i^2}{2n^2} \right) &= 2 \sum_{a \in \mathcal{F}(x_1^n)} \exp\left( \frac{\lambda^2}{2n^2} ||a||^2 \right) \leq_2 \\
\leq_2 2 \left| \mathcal{F}(x_1^n) \right| \exp\left( \frac{\lambda^2}{2n^2} r^2 \right) &\leq_3 2(n+1)^\mu \exp\left( \frac{\lambda^2}{2n^2} r^2 \right)
\end{aligned}$$

dove

- $=_1$ deriva dal fatto che $\varepsilon_i$ e $-\varepsilon_i$ hanno la stessa distribuzione
- $=_2$ deriva dal fatto che $\varepsilon_1, \dots, \varepsilon_n$ sono tra loro indipendenti
- $\leq_1$ si ottiene dal seguente sviluppo in serie di Taylor

$$\frac{e^x + e^{-x}}{2} = \sum_{k=0}^\infty \frac{x^{2k}}{(2k)!} \leq \sum_{k=0}^\infty x^2 k 2^k k! = e^{\frac{x^2}{2}}$$

- $\leq_2$ abbiamo posto

$$r = \max_{a \in \mathcal{F}(x_1^n)} ||a||$$

- $\leq_3$ $\mathcal{F}$ ha discriminante polinomiale di ordine $\nu$

Abbiamo dunque provato che

$$\exp\left(\lambda \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right]\right) \leq 2(n+1)^\mu \exp\left(\frac{\lambda^2}{2n^2} r^2\right) \tag{3}$$

Applicando il logaritmo e dividendo per $\lambda$ entrambi i termini di (3) si ha

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq \frac{1}{\lambda} \ln\left(2(n+1)^\nu\right) + \frac{r^2}{2n^2}\lambda = \frac{A}{\lambda} + B\lambda \tag{4}$$

Poichè $\lambda$ è un parametro libero lo scelgo affinchè il membro di sinistra sia minimo.
Una semplice verifica mostra che il minimo si ha per $\lambda = 2\sqrt{AB}$, riscrivendo (4) per tale valore otteniamo

$$\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq_1 D(x_1^n) \sqrt{\frac{2\ln\left(2(n+1)^\nu\right)}{n}} \leq_2 2\sqrt{\frac{\nu \ln(n+1)}{n}}$$

dove

- $\leq_1$ abbiamo usato che

$$r = \max_{a \in \mathcal{F}(x_1^n)} ||a|| = \sup_{f \in \mathcal{F}} \sqrt{\sum_{i=1}^n f(x_i)} \quad \Rightarrow \quad \frac{r}{\sqrt{n}} = D(x_1^n)$$

- $\leq_2$ poichè $\nu \geq 1$ otteniamo $\ln 2 \leq \ln\left((n+1)^\nu\right)$ e dunque

$$\ln\left(2(n+1)^\nu\right) = \ln 2 + ln\left((n+1)^\nu\right) \leq 2\ln\left((n+1)^\nu\right) = 2\nu \ln(n+1)$$

□

**Corollario 4.8.** *Sia $\mathcal{F}$ una classe di funzioni su $\chi$ uniformemente limitata da $b$ con discriminante polinomiale di ordine $\nu$. Allora*

$$\mathcal{R}_n(\mathcal{F}) \leq 4b\sqrt{\frac{\nu \ln(n+1)}{n}}$$

*Dimostrazione.* Sotto la condizione di uniforme limitatezza ho $D(x_1^n) \leq b$.
Rimpiazzando nel lemma $x_1^n$ con $(X_1, \ldots, X_n)$ e calcolando il valore atteso rispetto a $X$ ad entrambi i membri ottengo la tesi (il membro di destra non è aleatorio).

□

**Corollario 4.9.** *Se $\mathcal{F}$ è una classe di funzioni uniformemente limitate da $b$ con discriminante polinomiale di ordine $\nu$ allora $\forall \delta > 0$ si ha*

$$\mathbb{P}\left(||\mathbb{P}_n - P_X||_{\mathcal{F}} < \delta + 2b\sqrt{\frac{\nu \ln(n+1)}{n}}\right) \geq 1 - 2\exp\left(-\frac{n\delta^2}{2b^2}\right)$$

*ed inoltre $\mathcal{F}$ è di Glivenko-Cantelli in senso forte.*

*Dimostrazione.* La tesi segue dal Corollario 4.8 e dal Teorema 4.5.

“

# 5 Dimensione VC

**Definizione 5.1.** Sia $\chi$ un insieme

$$\mathcal{B}(\chi) = \{f \;\; \chi \to \{0,1\}\} = \{ \text{ funzioni booleane su } \chi\}$$

*Osservazione* 14. Esiste una bigezione tra $\mathcal{P}(\chi)$ e $\mathcal{B}(\chi)$ definita mandando il sottoinsieme $S$ nella sua funzione caratteristica.

**Definizione 5.2.** Sia $\mathcal{F} \subseteq \mathcal{B}(\chi)$. Diremo che $\Lambda \subseteq \chi$ è **shattered** (**frantumato**) da $\mathcal{F}$ se ogni funzione $g \in \mathcal{B}(\Lambda)$ è la restrizione a $\Lambda$ di una funzione $f \in \mathcal{F}$

*Osservazione* 15. $\Lambda$ è frantumato da $\mathcal{F}$ se

$$\mathcal{B}(\Lambda) = \left\{ f_{|\Lambda} \mid f \in \mathcal{F} \right\} =: \mathcal{F}_{|\Lambda}$$

*Osservazione* 16. Sia $\Lambda = \{x_1, \ldots, x_n\}$ finito.

$$\Lambda \text{ è frantumato da } \mathcal{F} \quad \Leftrightarrow \quad |\mathcal{F}(x_1^n)| = 2^n$$

**Definizione 5.3.** Dato $\mathcal{F} \subseteq \mathcal{B}(\chi)$ definiamola sua **dimensione VC** come la massima cardinalità di $\Lambda \subseteq \chi$ frantumato da $\mathcal{F}$. Se tale massimo non esiste diremo che la sua dimensione VC è $+\infty$

**Lemma 5.1.** *Se $d = VC(\mathcal{F})$ allora*

$$|\mathcal{F}| \geq 2^d$$

*Dimostrazione.* Sia $\Lambda \subseteq \chi$ l'insieme frantumato che realizza il massimo. Essendo $\Lambda$ frantumato la mappa

$$\psi : \; \mathcal{F} \to \mathcal{B}(\chi) \qquad \psi(f) = f_{|\Lambda}$$

è suriettiva e dunque $2^d = |\mathcal{B}(\Lambda)| \leq |\mathcal{F}|$ $\square$

**Lemma 5.2** (di Sauer–Shelah)**.** *Sia $\mathcal{F} \subseteq \mathcal{B}(\chi)$. Se $|\chi| = n$ e $d = VC(\mathcal{F})$ vale*

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \leq (n+1)^d$$

*Dimostrazione.* Proviamo solo la seconda disuguaglianza (non è originariamente parte del lemma).
Dalla formula del binomio di Newton:

$$(n+1)^d = \sum_{k=0}^{d} \binom{d}{k} n^k$$

Ma

$$\binom{d}{k} n^k = \frac{d(d-1)\cdots(d-k+1)}{k!} n^k \geq \frac{n^k}{k!} \geq \frac{n(n-1)\cdots(n-k+1)}{k!} = \binom{n}{k}$$

**Proposizione 5.3.** *Sia $\mathcal{F} \subseteq \mathcal{B}(\chi)$ con $d = VC(\mathcal{F}) < \infty$. Allora $\mathcal{F}$ ha discriminate polinomiale di ordine $d$*

*Dimostrazione.* Fissiamo una $n$-upla $x_1^n \in \chi^n$ e supponiamo che sia formata da elementi distinti. Siccome gli elementi sono distinti c'è una biezione naturale tra $\mathcal{F}(x_1^n)$ e $f_{|\Lambda}$ con $\Lambda = \{x_1, \dots, x_n\}$. Dunque

$$|\mathcal{F}(x_1^n)| = \left|\mathcal{F}_{|\Lambda}\right| \leq (n+1)^{VC(\mathcal{F}_{|\Lambda})}$$

dove l'ultima disuguaglianza è il lemma di Sauer–Shelah.
Per concludere proviamo che $\forall \Lambda \subseteq \chi$ si ha $VC\left(\mathcal{F}_{|\Lambda}\right) \leq VC(\mathcal{F})$.
Supponiamo per assurdo che $VC\left(\mathcal{F}_{|\Lambda}\right) > VC(\mathcal{F})$ allora deve esistere $A \subseteq \Lambda$ frantumato da $\mathcal{F}_{|\Lambda}$ con $|A| > VC(\mathcal{F})$ quindi

$$\mathcal{B}(A) = \left\{g_{|A} \mid g \in \mathcal{F}_{|\Lambda}\right\} \quad \Rightarrow \quad \mathcal{B}(A) = \left\{g_{|A} \mid g \in \mathcal{F}\right\}$$

infatti se $g \in \mathcal{F}_{|_A}$ allora $g = f_{|\Lambda}$ con $f \in \mathcal{F}$.

Abbiamo provato che $A$ è frantumato da $\mathcal{F}$ il che è assurdo.

Se $x_1, \dots, x_n$ non sono distinti, sia $y_1, \dots, y_k$ la sotto-upla estratta massimale che contiene elementi distinti. Allora esiste una biezione naturale tra $\mathcal{F}(x_1^n)$ e $\mathcal{F}(y_1^k)$ quindi

$$|\mathcal{F}(x_1^n)| = \left|\mathcal{F}(y_1^k)\right| \leq (k+1)^{VC(\mathcal{F})} \leq (n+1)^{VC(\mathcal{F})}$$

dove la prima disuguaglianza deriva dal caso precedente.

□

*Osservazione* 17. La proposizione non garantisce l'ottimalità dell'ordine.

Come conseguenza immediata della Proposizione 5.3 e del Corollario 4.8 otteniamo

**Proposizione 5.4.** *Sia $\mathcal{F} \subseteq \mathcal{B}(\chi)$ con $VC(\mathcal{F}) < \infty$. Allora*

$$R_n(\mathcal{F}) \leq 2\sqrt{\frac{VC(\mathcal{F})\ln(n+1)}{n}}$$

*Dimostrazione.* Infatti le funzioni in $f$ sono booleane e quindi limitate uniformemente da 1 □

In realtà vale una stima molto più forte (che non dimostremo)

**Teorema 5.5.**

$$R_n(\mathcal{F}) \leq 2C\sqrt{\frac{VC(\mathcal{F})}{n}}$$

*Osservazione* 18. $\mathcal{R}_n(\mathcal{F})$ è un oggetto probabilistico molto complesso: è un valore atteso della soluzione di un problema di ottimizzazione random. Il membro destro della disuguaglianza precedente non ha nulla di probabilistico e in molti casi si calcola molto facilmente.

# 6 Statistical Learning

L'obbiettivo della statistical learning è quello di stimare, guardando i dati, una funzione $T : \chi \to \mathbb{R}$ detta **funzione target**. Ovvero se $X_1, \dots, X_n$ è un campione casuale in $\chi$ con distribuzione $P_X$, voglio approssimare $T$ tramite l'osservazione del **training data**: $(X_1, T(X_1)), \dots, (X_n, T(X_n))$.

*Osservazione* 19. Spesso nelle applicazioni, $T \in \mathcal{B}(\chi)$

**Esempio 6.1.** *Supponiamo che siano sufficienti d misure per determinare se una persona ha o meno il diabete, ovvero, supporremo che esista una funzione*

$$T : \mathbb{R}^d \to \{0,1\} \qquad T(k_1, \dots, k_d) = 1 \quad \Leftrightarrow \quad \text{la persona con dati}(k_1, \dots, k_d) \text{ è diabetica}$$

*Nella pratica, non conoscendo $T$, vogliamo trovare una sua rappresentazione approssimata. Per farlo fisso una classe di funzioni $\mathcal{F}$ e cercherò in $\mathcal{F}$ la mia approssimazione. In tal caso $\mathcal{F}$ prende il nome di* ***spazio ipotesi****.*

**Definizione 6.1.** Il **rischio** associato a $f \in \mathcal{F}$ è definito come

$$R(f) = \mathbb{E}\left[(f(X) - T(X))^2\right]$$

e denotiamo come

$$f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$$

dove abbiamo assunto che la mappa che manda $f$ nel suo rischio abbia un punto di minimo.

**Definizione 6.2.** Il **rischio empirico** di $f \in \mathcal{F}$ è definito da

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - T(X_i))^2$$

e denotiamo con

$$f_n^\star = \arg\min_{f \in \mathcal{F}} R_n(f)$$

*Osservazione* 20. Notiamo che essendo la funzione $T$ sconosciuta non è possibile calcolare $f^\star$ e tantomeno $R(f^\star)$ mentre, avendo a disposizione il training set, posso calcolare esplicitamente la mappa che associa a $f$ il suo rischio empirico e minimizzandola (assumendo che abbia minimo) calcolare $f_n^\star$.
Da quanto appena detto, segue che la domanda opportuna da farsi non è stimare o calcoare l'errore di $f_n^\star$ ma piuttosto occorre capire quanto errore si aggiunge rispetto all'errore minimo $R(f^\star)$ scegliendo come "predictor" $f_n^\star$

**Definizione 6.3.** Il **rischio in eccesso** è dato da $R(f_n^\star) - R(f^\star)$

**Proposizione 6.2.** *Supponiamo che $\mathcal{F} \subseteq \mathcal{B}(\chi)$ e $T \in \mathcal{B}(\chi)$ allora vale*

$$0 \leq R(f_n^\star) - R(f^\star) \leq 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| = 2 \left|\left|\mathbb{P}_n - P_X\right|\right|_{\mathcal{L}}$$

*dove*

$$\mathcal{L} = \left\{(f-T)^2 \mid f \in \mathcal{F}\right\} = \{1_{f \neq T} \mid f \in \mathcal{F}\}$$

*Dimostrazione.* Chiamiamo

$$\varepsilon = \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$$

Dalla definizione di $\varepsilon$ e usando che $f_n^\star \in \mathcal{F}$ otteniamo

$$|R_n(f_n^\star) - R(f_n^\star)| \leq \varepsilon \quad \Rightarrow \quad R(f_n^\star) \leq R_n(f_n^\star) + \varepsilon$$

Dalla definizione di $\varepsilon$ e usando che $f^\star \in \mathcal{F}$ otteniamo

$$|R_n(f^\star) - R(f^\star)| \leq \varepsilon \quad \Rightarrow \quad R_n(f^\star) \leq R(f^\star) + \varepsilon$$

dunque combinando le due disuguaglianze e usando la definizione di $f_n^\star$ come punto di minimo della funzione $R_n(\cdot)$ si ha

$$R(f_n^\star) \leq R_n(f_n^\star) + \varepsilon \leq R_n(f^\star) + \varepsilon \leq R(f^\star) + 2\varepsilon$$

Notiamo inoltre che

$$R(f^\star) = \min_{f \in \mathcal{F}} R(f) \leq R(f_n^\star) \quad \Rightarrow \quad R(f_n^\star) - R(f^\star) \geq 0$$

Proviamo ora l'ultima uguaglianza della tesi

$$\varepsilon = \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_i) - T(x_i))^2 - \mathbb{E}\left[(f(X) - T(X))^2\right] \right| =$$

$$= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f-T)^2(X_i) - \mathbb{E}\left[(f-T)^2(X)\right] \right| = \sup_{h \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^n h(X_i) - \mathbb{E}[h(X)] \right|$$

Ma l'ultimo termine è proprio $||\mathbb{P}_n - P_X||_{\mathcal{L}}$

□

Per stimare il rischio in eccesso possiamo stimare la complessità della classe $\mathcal{L}$ con quella di $\mathcal{F}$ utilizzando la

**Proposizione 6.3.** *Sia $\mathcal{L}$ come nella Proposizione 6.2 allora*

$$\mathcal{R}_n(\mathcal{L}) \leq \mathcal{R}(\mathcal{F}) + \frac{1}{\sqrt{n}}$$

*Dimostrazione.* Data una funzione $g \in \mathcal{B}(\chi)$, definiamo

$$\tilde{g} : \chi \to \{-1, 1\} \qquad \tilde{g}(x) = \begin{cases} 1 & \text{se } g(x) = 1 \\ -1 & \text{se } g(x) = 0 \end{cases}$$

Allora

$$\mathcal{R}_n(\mathcal{L}) = E_{X,\varepsilon}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(f(X_i) - T(x_i)\right)^2 \right| \right]$$

Ora

$$\left(f(X_i) - T(x_i)\right)^2 = 1_{f(X_i) \neq T(X_i)} = 1_{\tilde{f}(X_i) \neq \tilde{T}(X_i)} = \frac{1 - \tilde{f}(X_i)\tilde{T}(X_i)}{2}$$

e dunque

$$\mathcal{R}_n(\mathcal{L}) = \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \left(1 - \tilde{f}(X_i)\tilde{T}(X_i)\right) \right| \right]$$

Ora

$$Z(X_1,\dots,X_n) = \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\left(1-\tilde{f}(X_i)\tilde{T}(X_i)\right)\right|\right] \leq$$
$$\leq \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] + \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\tilde{f}(X_i)\tilde{T}(X_i)\right|\right]$$

Ma (vedi Proposizione 4.3 (iii))

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \leq \frac{1}{\sqrt{n}}$$

inoltre, $\tilde{T}(X_1),\dots,\tilde{T}(X_n)$ sono numeri fissati in $\{-1,1\}$ e dunque i vettori $\left(\varepsilon_i\tilde{T}(X_i)\right)_{i=1}^n$ e $(\varepsilon_i)_{i=1}^n$ sono ugualmente distribuiti. Quindi

$$Z(X_1,\dots,X_n) \leq \frac{1}{\sqrt{n}} + \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\tilde{f}(X_i)\right|\right] =_1 \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(2f(X_i)-1)\right|\right] \leq$$
$$\leq 2\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] + \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\right|\right] \leq \frac{1}{\sqrt{n}} + 2\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right]$$

Prendendo il valore atteso rispetto a $(X_1,\dots,X_n)$ otteniamo la tesi.

□

**Corollario 6.4.** *Supponiamo che $\mathcal{F}\subseteq\mathcal{B}(\chi)$ abbia dimensione VC finita e $T\in\mathcal{B}(\chi)$ allora*

$$\mathbb{E}\left[R(f_n^\star)-R(f^\star)\right] \leq 8\sqrt{\frac{VC(\mathcal{F})\ln(n+1)}{n}} + \frac{2}{\sqrt{n}}$$

*Dimostrazione.* Usando la Proposizione 6.2 e il Teorema 4.3 otteniamo ha

$$\mathbb{E}\left[R(f_n^\star)-R(f^\star)\right] \leq \mathbb{E}\left[||\mathbb{P}_n - P_X||\right]_{\mathcal{L}} \leq 2\mathcal{R}(\mathcal{L})$$

Dalla Proposzione 6.3 si ha

$$\mathcal{R}_n(\mathcal{L}) \leq \mathcal{R}(\mathcal{F}) + \frac{1}{\sqrt{n}}$$

e stimando la complessità di Rademacher con la dimensione VC (Proposizione 5.4) otteniamo la tesi □

In realtà abbiamo un teorema più forte che dice

**Teorema 6.5** (Rischio in eccesso via dimensione VC)**.** *Supponiamo che la classe $\mathcal{F}$ abbia $1\leq VC(\mathcal{F})<\infty$. Allora*

$$\mathbb{E}\left[R(f_n^\star)-R(f^\star)\right] \leq C\sqrt{\frac{VC(\mathcal{F})}{n}}$$

*dove $C$ è una costante assoluta calcolabile*

# 7 Variabili aleatorie gaussiane e subgaussianne

## 7.1 Variabili aleatorie normali

Ricordiamo alcune propietà sulle variabili aleatorie normali senza fornire dimostrazioni (se non alcuni cenni)

**Proposizione 7.1.** *Se $X \sim \mathcal{N}(0,1)$ allora per ogni $t \geq 0$ vale*

$$\mathbb{P}(X \geq t) \leq \frac{1}{t\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

**Definizione 7.1.** La funzione $\Gamma$ è definita come

$$\Gamma : (0,+\infty) \to (0,+\infty) \qquad \Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} \, \mathrm{d}x$$

**Proposizione 7.2.** *Sia $X \sim \mathcal{N}(0,1)$ allora valgono i seguenti fatti*

*(i) Per ogni $t \geq 0$ vale*

$$\mathbb{P}(|X| > t)) \leq 2e^{-\frac{t^2}{2}}$$

*(ii) Per ogni $p \in \mathbb{N}$, $X$ ha momento di ordine $p$ finito e vale*

$$||X||_{L^p} = \sqrt{2} \frac{\Gamma\left(\dfrac{1+p}{2}\right)}{\Gamma\left(\dfrac{1}{2}\right)}$$

*(iii) $||X||_{L^p} = O(\sqrt{p})$ per $p \to +\infty$*

*(iv) Per ogni $\lambda \in \mathbb{R}$ vale*

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

*Dimostrazione.* (cenni)

(i) Se $t \geq 1$ usando la Proposizione 7.1 e notando che $\sqrt{\frac{2}{\pi}} \leq 1$ si ottiene

$$P(X \geq t) \leq e^{-\frac{t^2}{2}}$$

si conclude usando la simmetria della distribuzione normale standard.
Se $0 \leq t < 1$ allora notiamo che il termine di destra è maggiore di 1 e dunque la disuguaglianza è banalmente vera (il membro di sinistra è una probabilità)

(ii) Segue da semplici cambi di variabili negli integrali.

(iii) Si usa la formula generalizzata di Stirling per $\Gamma(z)$ per $z >> 1$

## 7.2 Variabili aleatorie subgaussiane

**Proposizione 7.3.** *Sia $X$ v.a. reale. Allora i seguenti fatti sonno equivalenti*

*(i)* $\exists k_1 > 0$ *tale che*

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{k_1^2}\right)$$

*(ii)* $\exists k_2 > 0$ *tale che*

$$||X||_{L^p} \leq k_2 \sqrt{p}$$

*(iii)* $\exists k_3 > 0$ *tale che*

$$\mathbb{E}\left[\exp\left(\lambda^2 X^2\right)\right] \leq \exp\left(k_3^2 \lambda^2\right) \quad \forall |\lambda| \leq \frac{1}{k_3}$$

*(iv)* $\exists k_4 > 0$ *tale che*

$$\mathbb{E}\left[\exp\left(\frac{X^2}{k_4}\right)\right] \leq 2$$

*Inoltre se $X$ è centrata, i fatti precedenti sono equivalenti a*

*(v)* $\exists k_5 > 0$ *tale che*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(k_5^2 \lambda^2\right) \quad \forall \lambda \in \mathbb{R}$$

*In aggiunta, esistono delle costanti assolute calcolabili $C_{ij}$ tali che*

$$K_i \leq C_{ij} K_j \quad \forall i, j = 1, \ldots, 4, (5)$$

*Osservazione* 21. Dalla Proposizione 7.2 sappiamo che se $X \sim \mathcal{N}(0,1)$ allora valgono le condizioni (i),(ii) e (v) (e dunque anche le altre)

*Osservazione* 22. Se la v.a. non è centrata la propietà (v) non può valere.
Usando la disuaglianza di Jensen (Teorema 1.6) con $\varphi(t) = e^{\lambda t}$ otteniamo

$$\mathbb{E}[\exp(\lambda X)] \geq e^{\lambda \mathbb{E}[X]} \quad \forall \lambda \in \mathbb{R}$$

Ora sviluppando in un intorno di 0 (e dunque per $\lambda$ piccolo) si ha

$$e^{\lambda \mathbb{E}[X]} = 1 + \lambda \mathbb{E}[X] + o(\lambda)$$

$$e^{k_5^2 \lambda^2} = 1 + o(\lambda)$$

Dunque se valesse (v) si avrebbe

$$\cancel{1} + \lambda \mathbb{E}[X] + o(\lambda) \leq \cancel{1} + o(\lambda)$$

il che è assurdo $\square$

**Definizione 7.2.** Data $X$ v.a. reale, definiamo la sua semi**norma subgaussiana** come

$$||X||_{\psi_2} = \inf\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}$$

con la convenzione $\inf \emptyset = +\infty$

**Definizione 7.3.** Una v.a. reale $X$ è detta subgaussiana se $||X||_{\psi_2} < +\infty$

*Osservazione* 23. Dall'Osservazione 21 otteniamo che $X \sim \mathcal{N}(0,1)$ è subgaussiana.

**Lemma 7.4.** *Sia $X$ v.a. reale.*

$$||X||_{\psi_2} = 0 \quad \Leftrightarrow \quad X = 0 \ q.c.$$

*Dimostrazione.* $\Leftarrow$ Se $X = 0$ q.c. allora

$$\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\} = \left\{t > 0 \mid \mathbb{E}\left[e^0\right] \leq 2\right\} = (0, +\infty)$$

e dunque $||X||_{\psi_2} = \inf\limits_{t \in (0,+\infty)} t = 0$.

$\Rightarrow$ Dimostriamo la tesi in maniera contronominale.
Supponiamo che $\mathbb{P}(X = 0) \neq 1$, dunque $\exists a > 0$ con $\mathbb{P}(|X| > a) > 0$. Fissato $t > 0$ allora

$$\mathbb{E}\left[\exp\frac{X^2}{t^2}\right] \geq \mathbb{P}(|X| > a)\exp\left(\frac{a^2}{t^2}\right) \to +\infty \quad \text{per } t \downarrow 0$$

Dunque $\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}$ non può contenere $(0, \varepsilon)$ per un certo $\varepsilon > 0$ da cui $||X||_{\psi_2} > 0$.

□

**Lemma 7.5.** *Se $||X||_{\psi_2} \in (0, +\infty)$ allora*

$$||X||_{\psi_2} = \min\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}$$

*Dimostrazione.* Per definizione di norma subgaussiana

$$||X||_{\psi_2} = \inf\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\} = t_\star$$

dunque esiste una successione $t_n \downarrow t_\star$ tale che

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t_n^2}\right)\right] \leq 2$$

Poichè $t_n$ è decrescente si ha che

$$\exp\left(\frac{X^2}{t_n^2}\right) \leq \exp\left(\frac{X^2}{t_{n+1}^2}\right)$$

e dunque per il Teorema di Beppo-Levi si ha

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t_\star^2}\right)\right] = \lim_{n \to +\infty} \mathbb{E}\left[\exp\left(\frac{X^2}{t_n^2}\right)\right] \leq 2$$

ovvero $t_\star \in \left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}$ ed è dunque un minimo.

□

**Teorema 7.6.** *L'insieme delle v.a. subgaussiane definite su un fissato spazio di probabilità (con l'identificazione $X \sim Y$ se $X = Y$ q.o) sono uno spazio vettoriale.*
*La norma subgaussiana rende tale spazio uno spazio di Banach.*

*Dimostrazione.* Denotiamo con $\mathcal{S}$ l'insieme delle v.a. subgaussiane quozientato per la relazione di essere uguali quasi certamente. Proveremo che $\mathcal{S}$ è normato dalla norma subgaussiana (che dunque è una norma), non proveremo che è un Banach.

- Dal Lemma 7.4 sappiamo che $||X||_{\psi_2} = 0 \quad \Leftrightarrow \quad X \sim 0$
- Sia $a \in \mathbb{R}$ e $X \in \mathcal{S}$

  $$||aX||_{\psi_2} = \inf\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{a^2X^2}{t^2}\right)\right] \leq 2\right\} = |a| \inf\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{s^2}\right)\right] \leq 2\right\} = |a| \, ||X||_{\psi 2}$$

  dove la seconda uguaglianza deriva dal cambio di variabili $t = |a| \, s$
- Siano $X, Y \in \mathcal{S}$, proviamo che vale la disuguaglianza triangolare.
  Possiamo supporre $X, Y \not\sim 0$ infatti se $X \sim 0$ allora otteniamo $X + Y \sim Y$ e la tesi segue banalmente.
  Supponiamo dunque che $||X||_{\psi_2}, ||Y||_{\psi_2} \in (0, +\infty)$ e quindi per la Proposizione 7.5 si ha

  $$a = ||X||_{\psi_2} = \min\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq 2\right\}$$

  $$b = ||Y||_{\psi_2} = \min\left\{t > 0 \mid \mathbb{E}\left[\exp\left(\frac{Y^2}{t^2}\right)\right] \leq 2\right\}$$

  Sia $\varphi(t) = e^{t^2}$ allora poichè tale funzione è convessa si ha

  $$\varphi\left(\frac{X+Y}{a+b}\right) \leq \frac{a}{a+b}\varphi\left(\frac{X}{a}\right) + \frac{b}{a+b}\varphi\left(\frac{Y}{b}\right)$$

  e dunque

  $$\mathbb{E}\left[\exp\left(\frac{X+Y}{a+b}\right)^2\right] = \mathbb{E}\left[\varphi\left(\frac{X+Y}{a+b}\right)\right] \leq \frac{a}{a+b}\mathbb{E}\left[\varphi\left(\frac{X}{a}\right)\right] + \frac{b}{a+b}\mathbb{E}\left[\varphi\left(\frac{Y}{b}\right)\right] \leq$$
  $$\leq \frac{a}{a+b}2 + \frac{b}{a+b}2 \leq 2$$

  e dunque per definizione di norma subgaussiana si ha

  $$||X+Y||_{\psi_2} \leq a + b = ||X||_{\psi_2} + ||Y||_{\psi_2}$$

$\square$

*Osservazione* 24. Sia $X \not\sim 0$ subgaussiana allora $||X||_{\psi_2}$ è la migliore costante $k_4$ nella Proposizione 7.3 ed inoltre per la seconda parte della stessa dimostrazione, sappiamo che esiste una costante assoluta tale che

$$P(|X| > t) \leq 2\exp\left(-C\frac{t^2}{||X||_{\psi_2}}\right)$$
$$||X||_{L^p} \leq C \, ||X||_{\psi_2} \sqrt{p} \text{ con } p > 1$$

e aggiungendo $X$ centrata ottengo

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(C \, ||X||^2_{\psi_2} \lambda^2\right) \quad \forall \lambda \in \mathbb{R}$$

**Proposizione 7.7.** *Se $X$ è una v.a. subgaussiana allora*

$$||X - \mathbb{E}[X]||_{\psi_2} \leq C\,||X||_{\psi_2}$$

*dove $C$ è una costante assoluta.*

*Dimostrazione.* Essendo la norma subgaussiana una norma si ha

$$||X - \mathbb{E}[X]||_{\psi_2} \leq ||X||_{\psi_2} + ||\mathbb{E}[X]||_{\psi_2} = ||X||_{\psi_2} + |\mathbb{E}[X]|\,||1||_{\psi_2} = ||X||_{\psi_2} + \tilde{C}\,|\mathbb{E}[X]|$$

Ora

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|] = ||X||_{L^1} \leq \hat{C}\,||X||_{\psi_2}$$

dove per l'ultima disuguaglianza abbiamo usato che per $X$ subgaussiana vale

$$||X||_{L^p} \leq \hat{C}\,||X||_{\psi_2}\sqrt{p}$$

con $p = 1$.

□

**Proposizione 7.8.** *Siano $X_1, \ldots, X_n$ v.a. subgaussiane indipendenti e centrate. Allora*

$$\left|\left|\sum_{i=1}^n X_i\right|\right|_{\psi_2}^2 \leq C\sum_{i=1}^n ||X_i||_{\psi_2}^2$$

*dove $C$ è una costante assoluta.*

*Dimostrazione.*

$$\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \leq \prod_{i=1}^n \exp\left(\tilde{c}\lambda^2\,||X_i||_{\psi_2}^2\right) =$$
$$= \exp\left(\tilde{C}\lambda^2\sum_{i=1}^n ||X_i||_{\psi_2}^2\right)$$

dove l'ultima disuguaglianza deriva dall'Osservazione 24.
Abbiamo provato che $\sum X_i$ (che è centrata) soddisfa la condizione (v) della Proposizione 7.3 con

$$k_5 = \tilde{C}\sum_{i=1}^n ||X_i||_{\psi_2}^2$$

Ora dalla seconda parte della Proposizione 7.3 sappiamo che

$$k_4^2 \leq \hat{C}k_5^2 \leq \tilde{C}\hat{C}\sum_{i=1}^n ||X_i||_{\psi_2}^2$$

si conclude ricordando che la norma subgaussiana è il migliore $k_4$

□

## 7.3 Disuguaglianza di Hoeffding

**Teorema 7.9** (Disuguaglianza di Hoeffding per v.a. subgaussiane)**.** *Siano* $X_1, \dots, X_n$ *v.a. subgaussian indipendenti e centrate. Allora* $\forall t \geq 0$ *vale*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq 2\exp\left(-c\frac{t^2}{\sum_{i=1}^{n} ||X_i||_{\psi_2}^2}\right)$$

*Dimostrazione.* Essendo l'insieme delle v.a. subgaussiane (con l'usuale identificazione q.c.) uno spazio vettoriale normato, $\sum X_i$ è subgaussiana e dunque per la Proposizione 7.3 si ha

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| > t\right) \leq 2\exp\left(-c\frac{t^2}{||\sum_{i=1}^{n} X_i||_{\psi_2}^2}\right)$$

la tesi segue applicando la disuguaglianza della proposizione precedente.

□

**Corollario 7.10.** *Siano* $X_1, \dots, X_n$ *v.a. subgaussiane indipendenti e centrate e* $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ *(non tutti nulli). Allora* $\forall t > 0$ *vale*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \alpha_i X_i\right| > t\right) \leq 2\exp\left(-C\frac{t^2}{K^2\,||\alpha||_2^2}\right)$$

*dove*

- *$C$ è una costante assoluta*
- $K = \max ||X_i||_{\psi_2}$

*Dimostrazione.* Applicando la disuguaglianza di Hoeffding alle variabili $\alpha_1 X_1, \dots, \alpha_n X_n$ (anch'esse centrate, subgaussiane e indipendenti) otteniamo

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \alpha_i X_i\right| > t\right) \leq 2\exp\left(-C\frac{t^2}{\sum_{i=1}^{n} ||\alpha_i X_i||_{\psi_2}^2}\right)$$

e poichè $||\cdot||_{\psi_2}$ è una norma si ha

$$\sum_{i=1}^{n} ||\alpha_i X_i||_{\psi_2}^2 = \sum_{i=1}^{n} \alpha_i^2\,||X_i||_{\psi_2}^2 \leq K^2\,||\alpha||_2^2$$

da cui la tesi.

□

*Osservazione* 25. Se $X_1, \dots, X_n$ sono subgaussiane indipendenti posso applicare la disuguaglianza di Hoeffding alle v.a . centrate $X_1 - \mathbb{E}[X_1], \dots, X_n - \mathbb{E}[X_n]$ e usare la seguente proposizione per stimare la norma subgaussiana della v.a. centrata

**Teorema 7.11** (Disuguaglianza di Hoeffding per v.a. di Rademacher)**.** *Siano* $X_1, \dots, X_n$ *delle v.a. di Rademacher indipendenti e* $\alpha \in \mathbb{R}^n$*. Allora* $\forall t \geq 0$ *vale*

$$\mathbb{P}\left(\sum_{i=1}^{n} \alpha_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\,||a||_2^2}\right)$$

*Dimostrazione.* Fissato $\lambda > 0$ si ha

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n \alpha_i X_i \geq t\right) &= \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^n \alpha_i X_i\right) \geq e^{\lambda t}\right) \leq_1 \\
&\leq_1 e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n \alpha_i X_i\right)\right] = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\left[\exp\left(\lambda \alpha_i X_i\right)\right] = \\
&= e^{-\lambda t}\prod_{i=1}^n \frac{1}{2}\left(e^{\lambda\alpha_i} + e^{-\lambda\alpha_i}\right) = e^{-\lambda t}\prod_{i=1}^n \cosh(\lambda\alpha_i) \leq_2 \\
&\leq_2 e^{-\lambda t}\prod_{i=1}^n \exp\left(\frac{\lambda^2\alpha_i^2}{2}\right) = \exp\left(-\lambda t + \frac{\lambda}{2}\left|\left|\alpha\right|\right|_2^2\right) = \exp(f(\lambda))
\end{aligned}$$

dove

- $\leq_1$ è la disuguaglianza di Markov.
- $\leq_2$ si ottiene usando $\cosh(x) \leq e^{\frac{x^2}{2}}$

Ora

$$\arg\min_{\lambda} f(\lambda) := \lambda_{opt} = \frac{t}{\left|\left|a\right|\right|_2^2}$$

e dunque si ha

$$\mathbb{P}\left(\sum_{i=1}^n \alpha_i X_i \geq t\right) \leq e^{f(\lambda_{opt})} = \exp\left(-\frac{t^2}{2\left|\left|a\right|\right|_2^2}\right)$$

□

**Corollario 7.12** (Disuguaglianza di Hoeffding bilatera per v.a. di Rademacher)**.** *Siano $X_1, \dots, X_n$ delle v.a. di Rademacher indipendenti e $\alpha \in \mathbb{R}^n$. Allora $\forall t \geq 0$ vale*

$$\mathbb{P}\left(\left|\sum_{i=1}^n \alpha_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\left|\left|a\right|\right|_2^2}\right)$$

*Dimostrazione.* Siccome $X_i \sim -X_i$ si ha

$$\mathbb{P}\left(\sum_{i=1}^n \alpha_i X_i \leq t\right) = \mathbb{P}\left(\sum_{i=1}^n \alpha_i (-X_i) \geq t\right) = \mathbb{P}\left(\sum_{i=1}^n \alpha_i X_i \geq t\right)$$

Si conclude notando che

$$\left\{\left|\sum_{i=1}^n \alpha_i X_i\right| \geq t\right\} \subseteq \left\{\sum_{i=1}^n \alpha_i X_i \geq t\right\} \cup \left\{\sum_{i=1}^n \alpha_i X_i \leq t\right\}$$

□

Notiamo che se $Y \sim Bernoulli\left(\frac{1}{2}\right)$ allora $2Y - 1$ sono variabili di Rademacher dunque possiamo riformulare i risultati precedenti come

**Corollario 7.13.** *Sia $Y_1, \dots, Y_n$ v.a. di Bernoulli $\left(\frac{1}{2}\right)$ indipendenti. Sia $a \in \mathbb{R}^n$ allora vale*

$$\mathbb{P}\left(\sum_{i=1}^n a_i\left(Y_i - \frac{1}{2}\right) > t\right) \leq \exp\left(-2\frac{t^2}{\left|\left|a\right|\right|_2^2}\right)$$
$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i\left(Y_i - \frac{1}{2}\right)\right| > t\right) \leq 2\exp\left(-2\frac{t^2}{\left|\left|a\right|\right|_2^2}\right)$$

*Dimostrazione.* Se poniamo $X_i = 2Y_i - 1$ allora $X_1, \ldots, X_n$ sono variabili di Rademacher indipendenti. Notiamo che

$$\sum_{i=1}^{n} a_i \left(Y_i - \frac{1}{2}\right) = \frac{1}{2} \sum_{i=1}^{n} a_i X_i$$

e dunque applicando la disuaglianza di Hoffding per le v.a. di Rademacher si ottiene

$$\mathbb{P}\left(\sum_{i=1}^{n} a_i \left(Y_i - \frac{1}{2}\right) > t\right) = \mathbb{P}\left(\sum_{i=1}^{n} a_i X_i > 2t\right) \geq \exp\left(-\frac{2t^2}{||a||_2^2}\right)$$

Per la disuguaglianza bilatera basta osservare chre $Y_i - \frac{1}{2}$ sono v.a. simmetriche.

□

**Teorema 7.14** (Disuguaglianza di Hoeffding per v.a. limitate)**.** *Siano $X_1, \ldots, X_n$ v.a. limitate e indipendenti con $X_i \sim [m_i, M_i]$. Allora $\forall t \geq 0$ vale*

$$\mathbb{P}\left(\sum_{i=1}^{N} (X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (M_i - m_i)^2}\right)$$

*Osservazione* 26 (Potenziamento di algoritmi randomizzati)*.* Supponiamo di avere un algoritmo randimizzato che da risposte corrette ad un problema di decisione con probabilità di $\frac{1}{2} + \delta$. Se facciamo girare l'algoritmo un numero $N$ dispari di volte e scegliamo come decisione finale quella più frequente otteniamo che $\forall \varepsilon \in (0, 1)$ si ha

$$N > \frac{1}{2\delta^2} \ln\left(\frac{1}{\varepsilon}\right) \quad \Rightarrow \quad \mathbb{P}(\text{ la decisione finale è corretta}) \geq 1 - \varepsilon$$

*Dimostrazione.* Segue dalla disuaglianza di Hoeffding applicata a v.a. limitate.
Siano $X_1, \ldots, X_N$ v.a. di Bernoulli di parametro $\frac{1}{2} - \delta$ allora

$$\mathbb{P}(\text{ la decisione finale è sbagliata}) = \mathbb{P}\left(\sum_{i=1}^{N} X_i \geq \frac{N+1}{2}\right)$$

Ora

$$\left\{\sum_{i=1}^{N} X_i \geq \frac{N+1}{2}\right\} = \left\{\sum_{i=1}^{N} X_i - \left(\frac{1}{2} - \delta\right) \geq \frac{N+1}{2} - N\left(\frac{1}{2} - \delta\right)\right\}$$

e dunque appplicando la disuaglianza di Hoeffding con $t = \frac{1}{2} + \delta N$ si ha

$$\mathbb{P}\left(\sum X_i \geq \frac{M+1}{2}\right) \leq \exp\left(-2\frac{\left(\frac{1}{2} + \delta N\right)^2}{N}\right) \leq \exp\left(-2\frac{\delta^2 N^2}{N}\right) = \exp\left(-2\delta^2 N\right)$$

Ora una semplice verifica, prova che se vale la relazione tra $N$ e $\varepsilon$ si ha la tesi.

□

# 8 Norme di matrici

## 8.1 Norma operatoriale per matrici

**Definizione 8.1.** Sia $A \in \mathbb{R}^{m\times n}$ allora definiamo la norma operatoriale di $A$ come

$$||A|| = \sup_{x\in\mathbb{R}^n} \frac{||Ax||_2}{||x||_2} = \max_{\substack{x\in\mathbb{R}^n\\ ||x||=1}} ||Ax||_2$$

**Lemma 8.1.** *Sia $A \in \mathbb{R}^{m\times n}$ allora vale*

$$||A|| = \max_{\substack{x\in S^{n-1}\\ y\in S^{m-1}}} \langle Ax,\ y\rangle$$

**Lemma 8.2.** *Sia $A \in \mathbb{R}^{m\times n}$ e $\varepsilon \in (0,1)$ allora se $N$ è un $\varepsilon-$ net di $S^{n-1}$ si ha*

$$\sup_{x\in N} ||Ax||_2 \leq ||A|| \leq \frac{1}{1-\varepsilon} ||Ax||_2$$

*Dimostrazione.* La prima disuguaglianza segue banalmente dalla definizione infatti $N \subseteq S^{n-1}$. Proviamo la seconda disuaglianza.
Per definizione di norma operatoriale

$$\exists x \in S^{n-1} \qquad ||Ax||_2 = ||A||$$

Per definizione di $\epsilon-$net, esiste $x_0 \in N$ tale che $||x - x_0||_2 \leq \varepsilon$ dunque otteniamo

$$||Ax - Ax_0||_2 =\leq ||A||\,\varepsilon$$

e utilizzando la disuguaglianza triangolare si ha

$$||Ax_0||_2 \geq ||Ax||_2 - ||Ax - Ax_0||_2 \geq ||A|| - \varepsilon\,||A|| = (1-\varepsilon)\,||A||$$

e dunque otteniamo

$$||A|| \leq \frac{1}{1-\varepsilon} ||Ax_0||_2 \leq \frac{1}{1-\varepsilon} \sup_{x'\in N} ||Ax'||_2$$

□

**Lemma 8.3.** *Sia $A \in \mathbb{R}^{m\times n}$. Dato $\varepsilon \in \left[0, \frac{1}{2}\right)$, siano $N$ e $M$ $\varepsilon-$net rispettivamente di $S^{n-1}$ e $S^{m-1}$ vale*

$$\sup_{\substack{x\in N\\ y\in M}} \langle Ax,\ y\rangle \leq ||A|| \leq \frac{1}{1-2\varepsilon} \sup_{\substack{x\in N\\ y\in M}} \langle Ax,\ y\rangle$$

*Dimostrazione.* Per il Lemma 8.1 si ha

$$||A|| = \sup_{\substack{x\in S^{n-1}\\ y\in S^{m-1}}} \langle Ax,\ y\rangle$$

e per il Teorema di Weistrass tale sup è un massimo, dunque

$$\exists x \in S^{n-1} \quad \exists y \in S^{m-1} \qquad ||A|| = \langle Ax,\ y\rangle$$

Ora per definizione di $\varepsilon-$net otteniamo che

$$\exists x_0 \in N \qquad ||x - x_0||_2 \leq \varepsilon$$
$$\exists y_0 \in M \qquad ||y - y_0||_2 \leq \varepsilon$$

Dunque si ha

$$\begin{aligned}|\langle Ax,\ y\rangle - \langle Ax_0,\ y_0\rangle| &= |\langle A(x-x_0),\ y\rangle - \langle Ax_0,\ y-y_0\rangle| \leq \\ &\leq ||A||\,||x-x_0||_2\,||y||_2 + ||A||\,||x_0||_2\,||y-y_0||_2\end{aligned}$$

Ricordando che $||y||_2 = ||x_0||_2 = 1$ otteniamo la tesi.

□

## 8.2 Norme di matrici subgaussiane

**Teorema 8.4.** *Sia $A \in \mathbb{R}^{m\times n}$ con entrate indipendenti, subgaussiane e centrate. Allora $\forall t > 0$ vale*

$$\mathbb{P}\left(||A|| \leq CK\left(\sqrt{m}+\sqrt{n}+t\right)\right) \leq 1-2e^{-t^2}$$

*dove $C$ è una costante assoluta e $K = \max ||A_{ij}||_{\psi_2}$*

*Dimostrazione.* La dimostrazione procede per passi.

- Fissato $\varepsilon = \frac{1}{4}$, siano $N$ e $M$ $\varepsilon-$net ottimali rispettivamente di $S^{n-1}$ e $S^{m-1}$. Per il Lemma precedente, vale

$$||A|| \leq 2\max_{\substack{x\in N\\ y\in M}}\langle Ax,\ y\rangle \tag{5}$$

- Fissati $x, y$ si ha

$$\langle Ax,\ y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_iy_j$$

  e quindi per la Proposizione 7.8 otteniamo

$$\begin{aligned}||\langle Ax,\ y\rangle||_{\psi_2}^2 \leq C\sum_{i=1}^{n}\sum_{i=1}^{n}||A_{ij}x_iy_j||_{\psi_2}^2 = C\sum_{i=1}^{n}\sum_{j=1}^{n}x_i^2y_j^2\,||A_{ij}||_{\psi_2}^2 \leq\\ \leq CK^2\sum_{i=1}^{n}x_i^2\sum_{j=1}^{n}y_j^2 = CK^2\end{aligned}$$

  dove abbiamo utilizato che $||x||_2 = ||y||_2 = 1$.
  Per l'Osservazione 24 si ha

$$\mathbb{P}\left(\langle Ax,\ y\rangle \geq u\right) \leq 2\exp\left(-\frac{cu^2}{K^2}\right)$$

- $$\begin{aligned}\mathbb{P}\left(\max_{\substack{x\in N\\ y\in M}}\langle Ax,\ y\rangle \geq u\right) = \mathbb{P}\left(\bigcup_{\substack{x\in N\\ y\in M}}\{\langle Ax,\ y\rangle \geq u\}\right) \leq\\ \leq \sum_{\substack{x\in N\\ y\in M}}\mathbb{P}\left(\langle Ax,\ y\rangle \geq u\right) \leq 9^{n+m}2\exp\left(-\frac{cu^2}{K^2}\right)\end{aligned} \tag{6}$$

  dove per l'ultima disuguaglianza abbiamo usato che il numero d'impacchettamento di $S^{k-1}$ è stimato dall'alto da $\left(1+\frac{2}{\varepsilon}\right)^k$

Dunque usando (5) si ha

$$\mathbb{P}\left(||A|| \leq CK\left(\sqrt{m}+\sqrt{n}+t\right)\right) \leq_1 \mathbb{P}\left(2\max_{\substack{x\in N\\ y\in M}}\langle Ax,\ y\rangle \leq CK\left(\sqrt{m}+\sqrt{n}+t\right)\right)$$

Ora usando (6) otteniamo

$$\begin{aligned}\mathbb{P}\left(2\max_{\substack{x\in N\\ y\in M}}\langle Ax,\ y\rangle \geq CK\left(\sqrt{m}+\sqrt{n}+t\right)\right) \leq\\ \leq 9^{n+m}2\exp\left(-c\frac{C^2}{4}\left(\sqrt{n}+\sqrt{m}+t\right)^2\right) \leq 9^{n+m}2\exp\left(-c\frac{C^2}{4}\left(n+m+t^2\right)\right)\end{aligned}$$

e per $C$ abbastanza grande possiamo supporre

$$9^{n+m} \exp\left(-c\frac{C^2}{4}(n+m)\right) \leq 1$$

$$\exp\left(-c\frac{C^2}{4}t^2\right) \leq e^{-t^2}$$

da cui la tesi.

□

**Corollario 8.5.** *Sia $A \in \mathbb{R}^{n\times n}$ una matrice simmetrica con entrate subgaussiane e centrate. Inoltre se le entrate della parte triangolare superiore sono indipendenti vale*

$$\mathbb{P}\left(||A|| \leq CK\left(\sqrt{n}+t\right)\right) \geq 1 - 4e^{-t^2}$$

*dove $C$ è una costante assoluta e $K = \max ||A_{ij}||_{\psi_2}$*

*Dimostrazione.* Chiamata $A^+$ la parte triangolare superiore e $A^-$ la parte strettamente triangolare inferiore si ha che $A = A^+ + A^-$ ed inoltre entrambe le matrici soddisfano le ipotesi del Teorema precedente da cui

$$\mathbb{P}\left(\left|\left|A^+\right|\right| \geq \frac{C}{2}\left(\sqrt{n}+t\right)\right) \leq 2e^{-t^2}$$

$$\mathbb{P}\left(\left|\left|A^-\right|\right| \geq \frac{C}{2}\left(\sqrt{n}+t\right)\right) \leq 2e^{-t^2}$$

Usando la disuguaglianza triangolare si ha

$$||A|| \geq D \quad \Rightarrow \quad \left(\left|\left|A^+\right|\right| \geq \frac{D}{2} \text{ o } \left|\left|A^+\right|\right| \geq \frac{D}{2}\right)$$

e dunque

$$\mathbb{P}\left(||A|| \geq CK\left(\sqrt{n}+t\right)\right) \leq \mathbb{P}\left(\left|\left|A^+\right|\right| \geq \frac{C}{2}\left(\sqrt{n}+t\right)\right) + \mathbb{P}\left(\left|\left|A^-\right|\right| \geq \frac{C}{2}\left(\sqrt{n}+t\right)\right) \leq$$
$$\leq 4e^{-t^2}$$

□

# 9 Teoria perturbativa di matrici simmetriche

Data una matrice $S \in \mathbb{R}^{n\times n}$ simmetrica, posso orfinare i suoi $n$ autovalori reali come

$$\lambda_1(S) \geq \cdots \lambda_n(S)$$

**Teorema 9.1** (Teorema min-max di Courant-Fisher)**.** *Sia $S \in \mathbb{R}^{n\times n}$ una matrice simmetrica allora*

$$\lambda_i(S) = \max_{\substack{E\subseteq\mathbb{R}^n \\ \dim E = i}} \min_{\substack{x\in E \\ ||x||_2=1}} \langle Sx,\ x\rangle$$

**Corollario 9.2** (Disuguaglianza di Weyl)**.** *Siano $S, T \in \mathbb{R}^{n\times n}$ due matrici simmetriche. Allora vale*

$$|\lambda_i(S) - \lambda_i(T)| \leq ||S - T|| \quad \forall i = 1, \ldots, n$$

**Definizione 9.1.** Siano $x, y \in \mathbb{R}^n$. Definiamo l'**angolo** tra i due vettori come

$$\angle(x, y) = \arccos\left(\frac{\langle x,\ y\rangle}{||x||_2\,||y||_2}\right)$$

**Teorema 9.3** (di Davis-Kohan)**.** *Siano $S, T \in \mathbb{R}^{n\times n}$ due matrici simmetriche. Supponiamo che $\lambda_i(S)$ sia semplice e definiamo*

$$\delta = \min_{j\neq i} |\lambda_j(S) - \lambda_i(S)|$$

*Allora fissati due autovettori unitari $v_i(S)$ e $v_i(T)$ relativi agli autovalori $\lambda_i(S)$ e $\lambda_i(T)$ vale*

$$\sin\left(\angle(v_i(T), v_i(S))\right) \leq 2\frac{||T - S||}{\delta}$$

**Proposizione 9.4.** *La disuaglianza DK implica che $\exists \theta \in \{-1, 1\}$ tale che*

$$||v_i(S) - \theta v_i(T)||_2 \leq \sqrt{8}\frac{||T - S||}{\delta}$$

# 10 Cluster Analysis

## 10.1 Two blocks models

**Definizione 10.1.** Sia $V = C_1 \cup C_2$ con $|C_1| = |C_2| = \frac{n}{2}$ definiamo con $G(n,p,q) = (V,E)$ il grafo random tale che

$$\forall \{x,y\} \subseteq V \qquad \mathbb{P}(\{x,y\} \in E) = \begin{cases} p & \text{se } x,y \in C_1 \text{ o } x,y \in C_2 \\ q & \text{altrimenti} \end{cases}.$$

Cerchiamo un algoritmo che data una realizzazione di $G(n,p,q)$ riesca ad individuare i due sottoinsiemi.

**Lemma 10.1.** *Detta $A$ la matrice di adiacenza della realizzazione di un grafo $G(n,p,q)$ allora posta*

$$A = D + R \text{ dove } D = \mathbb{E}[A]$$

*otteniamo che $rk(D)$ con autovalori*

$$\lambda_1(D) = \frac{p+q}{2}n \qquad \lambda_2(D) = \frac{p-q}{2}n$$

*Inoltre i vettori $v_1(D) = \begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}^T$ e $v_2(D) = (v_2(D)_i)$ con $v_2(D)_i = \begin{cases} 1 & \text{se } i \in C_1 \\ -1 & \text{se } i \in C_2 \end{cases}$ sono rispettivamente autovettori.*

*Dimostrazione.* Possiamo supporre che $V = \{1, \ldots, n\}$ e a meno di cambiare l'ordine dei vertici (e dunque eseguire un cambio di base), $C_1 = \left\{1, \ldots, \frac{n}{2}\right\}$. Dunque esiste una matrice invertibile $U$ tale che

$$UDU^{-1} = \tilde{D} = \left( \begin{array}{c|c} P & Q \\ \hline Q & P \end{array} \right) \text{ dove } P = p \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \; Q = q \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix} \in \mathbb{R}^{\frac{n}{2} \times \frac{n}{2}}$$

Ora $\tilde{D}$ è formata da copie di due colonne linearmente indipendenti e ha dunque rango 2 ed inoltre si mostra facilmente che

$$\tilde{D} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{p+q}{2} n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$\tilde{D} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} = \frac{p-q}{2} n \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

□

Spinti da questo Lemma, costruiamo l'algoritmo:

**Input:** $A$ matrice di adiacenza
**Output:** partizione dei vertici del grafo in $\mathcal{C}$ e $\mathcal{C}^\star$
$v_2(A) \leftarrow$ autovettore (di norma 1) relativo al secondo autovalore di $A$ ;
$\mathcal{C} \leftarrow \{i \in V \mid v_2(A)_i > 0\}$;
$\mathcal{C}^\star \leftarrow \{i \in V \mid v_2(A)_i \leq 0\}$;

**Algorithm 1:** Spectral clustering algorithm

**Lemma 10.2.** *Esiste una costante assoluta $C_0$ tale che $\forall n \geq 1$ si ha*

$$\mathbb{P}\left(||R|| \leq C_0\sqrt{n}\right) \geq 1 - 4e^{-n}$$

*Dimostrazione.* Utilizzando il Corollario 8.5 con $t = \sqrt{n}$ otteniamo che

$$\mathbb{P}\left(||R|| \leq 2\tilde{C}K\sqrt{n}\right) \geq 1 - 4e^{-n} \tag{7}$$

dove $\tilde{C}$ è una costante assoluta, mentre $K = \max ||R_{ij}||_{\psi_2}$. Notiamo che $||R_{ij}||_\infty \leq 1$ essendo le entrate v.a. di Bernoulli centrate e quindi (per una stima vista METTI RIF) si ha

$$||R_{ij}||_{\psi_2} \leq \frac{||R_{ij}||_\infty}{\ln 2} \leq 1$$

dunque la tesi segue da (7) ponendo $C = 2\tilde{C}$.

□

**Definizione 10.2.** Data una partizione dei vertici $\mathcal{C}, \mathcal{C}^\star$ dei vertici definiamo

$$\Delta = \min\{|C_1 \triangle \mathcal{C}|, |C_1 \triangle \mathcal{C}^\star|\}$$

*Osservazione* 27. Poichè

$$C_1 \triangle \mathcal{C} = (C_1 \cap \mathcal{C}^\star) \cup (C_2 \cap \mathcal{C})$$

$|C_1 \triangle \mathcal{C}|$ conta il numero di vertici classificati male se $\mathcal{C}$ approssima $C_1$ e $\mathcal{C}^\star$ approssima $C_2$. La $\Delta$ definita sopra conta, dunque, la bontà dell'approssimazione.

**Teorema 10.3.** *Siano $p, q \in (0, 1)$ con $p > q$ e definiamo*

$$\mu = \min\left\{q, \frac{p-q}{2}\right\}$$

*Allora con probabilità $\geq 1 - 4e^{-n}$ lo spectral clustering algorithm fornisce come output una partizione $\mathcal{C}, \mathcal{C}^\star$ dei vertici tale che*

$$\Delta \leq \frac{C}{\mu^2}$$

*dove $C$ è una costante assoluta.*

*Dimostrazione.* Applicando la Proposizione 9.4 alle matrici $S = D$ e $T = A$ con $i = 2$, otteniamo che $\exists \theta \in \{-1, 1\}$ tale che

$$||v_2(D) - \theta v_2(A)||_2 \leq \sqrt{8}\frac{||D - A||}{\mu n} = \frac{\sqrt{8}}{\mu n}||R|| \tag{8}$$

infatti

$$\delta = \min_{j \neq 2}\{|\lambda_j(D) - \lambda_2(D)|\} = \min\{|\lambda_2(D)|, |\lambda_1(D) - \lambda_2(D)|\} = \min\left\{q, \frac{p-q}{2}\right\}n = \mu n$$

Moltiplicando (8) per $\sqrt{n}$ otteniamo

$$||u_2(D) - \theta u_2(A)||_2 \leq \frac{\sqrt{8}}{\mu n} \cdot \frac{||R||}{\sqrt{n}} \tag{9}$$

dove $u_2(D)$ è l'autovettore del Lemma 10.1 e $u_2(A) = \sqrt{n} v_2(A)$.
Se ci mettiamo nell'evento $\{||R|| \leq C_0\sqrt{n}\}$ che per il Lemma 10.2 ha probabilità $\geq 1 - 4e^{-n}$, l'equazione (9) diventa

$$||u_2(D) - \theta u_2(A)||_2 \leq \frac{\sqrt{8}C_0}{\mu} = \frac{\sqrt{C}}{\mu}$$

Proviamo adesso che

$$\Delta \leq ||u_2(D) - \theta u_2(A)||_2^2$$

il che conclude la dimostrazione.

- $\theta = 1$.

$$j \in C_1 \cap \mathcal{C}^\star \quad \Rightarrow \quad (u_2(D)_i = +1 \text{ e } u_2(A)_i \leq 0) \quad \Rightarrow \quad (u_2(D)_i - u_2(A)_i)^2 \geq 1$$

$$j \in C_2 \cap \mathcal{C} \quad \Rightarrow \quad (u_2(D)_i = -1 \text{ e } u_2(A)_i \geq 0) \quad \Rightarrow \quad (u_2(D)_i - u_2(A)_i)^2 \geq 1$$

  e dunque

$$\begin{aligned}\Delta \leq |C_1 \triangle \mathcal{C}| = |C_1 \cap \mathcal{C}^\star| + |C_2 \cap \mathcal{C}| = \sum_{j \in (C_1 \cap \mathcal{C}^\star) \cup (C_2 \cap \mathcal{C})} 1 \leq \\ \leq \sum_{j=1}^{n} (u_2(D)_i - u_2(A)_i)^2 = ||u_2(D) - \theta u_2(A)||_2^2\end{aligned}$$

- $\theta = -1$.
  Come sopra ma consideriamo $C_1 \triangle \mathcal{C}^\star$

□

## 10.2 Cluster Analysys

**Definizione 10.3.** Dato un insieme $\mathcal{X} = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^n$, una **funzione di dissimilarità** è una qualsiasi funzione $d$ tale che

- $d(x, y) \geq 0$
- $d(x, y) = d(y, x)$

Nel seguito, per alleggerire la notazione, scriveremo

$$d(x_i, x_j) =: d_{ij}$$

**Definizione 10.4.** Fissato un naturale $K$, una suddivisione in $K$ **cluster** è una qualsiasi mappa surgettiva

$$C: \{1, \dots, N\} \to \{1, \dots, K\}$$

sotto l'identificazione data dalla relazione di equivalenza

$$C \sim C' \quad \Leftrightarrow \quad \exists \sigma \in S_K \;\; C = \sigma C'$$

**Definizione 10.5.** Dato una suddivisione $C$ e una funzione di dissimilarità, definiamo la **loss function** come

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{i=1 \\ C(i)=k}}^{n} \sum_{\substack{j=1 \\ C(j)=k}}^{N} d_{ij}$$

Definiamo, inoltre, la quantià

$$B(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{i=1 \\ C(i)=k}}^{n} \sum_{\substack{j=1 \\ C(j)\neq k}}^{n} d_{ij}$$

*Osservazione* 28. La funziondi di loss $W$ misura la dispersione all'interno del cluster (*within-clusters point scatter*), mentre la funzione $B$ misura la dispersione tra i punti di cluster differenti (*between-clusters point scatter*).

Il nostro obbiettivo è trovare il clustering che minimizza la funzione di loss oppure quello che massimizza la funzione $B$. Il seguente lemma prova che è equivalente richiedere una tra le due condizioni

**Lemma 10.4.** *Per ogni cluster $C$ vale*

$$B(C) + W(C) = \frac{1}{3} \sum_{i=1}^{n} \sum_{j=1}^{N} d_{ij}$$

*Dimostrazione.*

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} = \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{i=1 \\ C(i)=k}}^{n} \left( \sum_{\substack{j=1 \\ C(j)=k}}^{n} d_{ij} + \sum_{\substack{j=1 \\ C(j)\neq k}}^{n} d_{ij} \right) = W(C) + B(C)$$

□

## 10.3 K-means

Consideriamo un caso particolare in cui $d(x,y) = ||x-y||_2^2$ e d'ora in avanti, per alleggerire la notazione toglieremo il pedice alla norma.

**Lemma 10.5.** *Siano $X, Y$ v.a. reali i.i.d.. Allora*

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

*Dimostrazione.*

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

□

**Corollario 10.6.** *Nel caso in cui $d(x,y) = ||x-y||^2$ si ha*

$$W(C) = \sum_{i=1}^{K} N_k \sum_{\substack{i=1 \\ C(i)=k}}^{N} ||x_i - \overline{x}_k||^2$$

*dove*

$$N_k = \left|C^{-1}(\{k\})\right|$$

$$\overline{x}_k = \frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} x_i$$

*Dimostrazione.* Denotiamo $x_i = (x_{i1}, \dots, x_{iN})$. Fissiamo $k, \alpha$.
Consideriamo $X$ a valori $\{x_{i\alpha} \mid C(i) = k\}$ che assume i valori con equal probabilità (assumiamo i valori $x_{i\alpha}$ tra loro distinti, senno la probabilità proporzionale al numero delle occorrenze) allora se $Y \sim X$ si ha

$$Var(X) = \frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} (x_{i\alpha} - \overline{x}_{k\alpha})^2$$

$$\frac{1}{2}\mathbb{E}[(X-Y)^2] = \frac{1}{2N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} \sum_{\substack{j=1 \\ C(j)=k}}^{N} (x_{i\alpha} - x_{j\alpha})^2$$

e dunque per il Lemma precedente

$$\frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} (x_{i\alpha} - \overline{x}_{k\alpha})^2 = \frac{1}{2N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} \sum_{\substack{j=1 \\ C(j)=k}}^{N} (x_{i\alpha} - x_{j\alpha})^2$$

Moltiplicando per $N_k^2$ l'equazione precedente, sommando per $\alpha = 1, \dots, N$ e $k = 1, \dots, K$ otteniamo la tesi.

□

Ricordando che

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \min_{a \in \mathbb{R}} \mathbb{E}[(X-a)^2]$$

il Corollario precedente ci fornisce la seguente caratterizzazione variazionale: $\forall k, \alpha$

$$\frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} (x_{i\alpha} - \overline{x}_{k\alpha})^2 = \min_{a \in \mathbb{R}} \left\{ \frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}}^{N} (x_{i\alpha} - a)^2 \right\} \tag{10}$$

e dunque

$$\frac{1}{N_k} \sum_{\substack{i=1 \\ C(i)=k}} ||x_i - \overline{x}_k||^2 = \min_{a \in \mathbb{R}^n} \left\{ \sum_{\substack{i=1 \\ C(i)=k}}^{N} ||x_i - a||^2 \right\}$$

Introduciamo una funzione loss estesa (ha più entrate)

$$W_e(C, m_1, \dots, m_K) = \sum_{k=1}^{K} N_k \sum_{\substack{i=1 \\ C(i)=k}}^{N} ||x_i - m_k||^2$$

Per (10), fissata $C$

$$\min_{m_1, \dots, m_k \in \mathbb{R}^n} W_e(C, m_1, \dots, m_K) = W_e(C, \overline{x}_1, \dots, \overline{x}_K) = W(C)$$

dunque se $C_\star = \arg\min W(C)$ allora $C_\star$ è la prima entrata del minimizzante di $W_e$. Per trovare il minimizzante di $W$, trovo quello di $W_e$.

L'algoritmo di $K$-means è un algoritmo discendente e si struttura come segue

**Input:** $\mathcal{X} = \{x_1, \dots, x_N\} \subseteq \mathbb{R}^n$ e $d$ funzione di dissimilarità
**Output:** Suddivisione $C$
**for** $i = 1, \dots, K$ **do**
  $j \leftarrow$ `random`$(1, \dots, N)$;
  $m_i \leftarrow x_j$
**end**
**while** $C_{new} \neq C$ **do**
  $C \leftarrow G_{new}$;
  **for** $i$=1,... $N$ **do**
    $j \leftarrow \text{argmin}_{k=1,\dots,K} ||x_i - m_k||$;
    $C_{new}(i) \leftarrow j$
  **end**
  **for** $k$= 1, ..., $K$ **do**
    $m_k \leftarrow$ `baricentro di` $C_{new}^{-1}(k)$
  **end**
**end**

**Algorithm 2:** $k$-means algorithm

## 10.4 Unormalize Laplacian spectral cluster

**Definizione 10.6.** Dato un grafo pesato $G = (V, E, W)$ definiamo **laplaciano** di $G$

$$L = D - W \text{ dove } D = diag(d_1, \ldots, d_N) \;\; d_i = \sum_{j=1}^{N} w_{ij}$$

Il nome del Laplaciano segue dalla seguente

*Osservazione* 29. Consideriamo una griglia toroidale in $\mathbb{R}^n$ e

$$w_{ij} = \begin{cases} 1 & \text{se } i, j \text{ lato della griglia} \\ 0 & \text{altrimenti} \end{cases}$$

Se identifichiamo la funzione $f : V \to \mathbb{R}$ con il vettore di $R^N$

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}$$

allora

$$(Lf)_i = (Df)_i - (Wf)_i = d_{ii} f_i - \sum_{j=1}^{N} w_{ij} f_j = 2n f_i - \sum_{\substack{j \\ \text{adiacente ad } i}} f_j j = \sum_{\substack{j \\ \text{adiacennte ad} i}} (f_j - f_i)$$

e dunque

$$(Lf)(x_i) = \sum_j f(x_j) - f(x_i) = \sum_{\substack{j \\ ||x_i - x_j|| = 1}} f(x_i) - f(x_j) =$$

$$\sum_{\alpha=1}^{n} f(x_i + e_\alpha) + f(x_i - e_\alpha) - 2f(x_i)$$

ora l'ultima espressione è la discretizzazione del laplaciano.

**Proposizione 10.7.** *Il laplaciano $L$ gode delle seguenti proprietà*

*(i)* $\forall f \in \mathbb{R}^N$

$$f \cdot Lf = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left(f_i - f_j\right)^2$$

*(ii) $L$ è una matrice simmetrica semidefinita positiva*

*(iii) Siano $A_1, \ldots, A_M$ le componenti connesse di $G$. Allora $0$ è un autovalore di $L$ con molteplicità $M$. Inoltre, il nucleo di $L$ è generato dai vettori $1_{A_1}, \ldots, 1_{A_M}$ dove*

$$\left(1_{A_k}\right)_i = \begin{cases} 1 & \textit{se } i \in A_k \\ 0 & \textit{altrimenti} \end{cases}$$

*Dimostrazione.*

(i)
$$\begin{aligned} f\cdot Lf &= \sum_i f_i(Lf)_i = \sum_{i,j} f_i L_{ij} f_j = \sum_i d_{ii} f_i^2 - \sum_{i,j} w_{ij} f_i f_j = \\ &= \sum_i f_i^2 \left(\sum_j w_{ij}\right) - \sum_{i,j} w_{ij} f_i f_j = \\ &= \frac{1}{2}\left(\sum_{i,j} f_i^2 w_{ij} + \sum_{i,j} f_j^2 w_{ji}\right) - \frac{1}{2}\left(2\sum_{i,j} w_{ij} f_i f_j\right) = \frac{1}{2}\sum_{ij} w_{ij}\left(f_i - f_j\right)^2 \end{aligned}$$

dove abbiamo usato la simmetria di $W$

(ii) Poichè $D, W$ sonno simmetriche anche $L = D - W$ lo è. Ricordando che $w_{ij} \geq 0$ si ottiene che $L$ è semidefinita positiva.

(iii) Se $f \in Ker(f)$ allora $f \cdot Lf = 0$. Ora

$$f \cdot Lf = 0 \quad \Leftrightarrow \quad \sum_{ij} w_{ij}\left(f_i - f_j\right)^2 = 0 \quad \Leftrightarrow \quad \forall i,j \; w_{ij} > 0 \quad f_i = f_j$$

Ovvero se $\{i, j\}$ è un lato di $G$, $f_i = f_j$ e dunque $f$ deve essere costante sulle componenti connesse. Abbiamo provato che

$$f \in KerL \quad \Rightarrow \quad f \text{ costante su } A_k \; k = 1, \ldots, M$$

Viceversa, se $f$ è costante sulle componenti connesse si ha $Lf = 0$ essendo

$$(Lf)_i = \sum_j w_{ij}(f_i - f_j)$$

*Osservazione* 30. Siano $v_1, \ldots, v_n$ una base di $KerL$ e $V = \begin{pmatrix} v_1 & \ldots & v_n \end{pmatrix}$. Diremo che $x \equiv y$ se la riga $x$-esima di $V$ è uguale alla riga $y$-esima di $V$. Allora

$$x \equiv y \quad \Leftrightarrow \quad x, y \text{ stanno nella stessa componente connessa}$$

*Dimostrazione.* Sia $h_i = \frac{1_{A_i}}{|A_i|}$.
$\Leftarrow$ Se $x, y$ stanno nella stessa componente connessa allora $(h_i)_x = (h_i)_y$ per ogni $i = 1, \ldots, M$. Ma $\{h_1, \ldots, h_M\}$ sono una base del nucleo e dunque $\exists b_{ij}$ con

$$v_i = \sum_j b_{ij} h_j$$

da cui la tesi.
$\Rightarrow$ $\forall k = 1, \ldots, M$ fisso $x_k \in A_k$. Dato $w \in \mathbb{R}^N$ definisco $\tilde{w} \in \mathbb{R}^M$ dove

$$\tilde{w}_k = w_{x_k}$$

Notiamo che i vettori $\tilde{v}_1, \ldots, \tilde{v}_M$ sono linearmente indipendenti (se non lo fossero, non lo sarebbero nemmeno i $v_i$) e dunque la matrice

$$V = \begin{pmatrix} \tilde{v}_1 & \ldots & \tilde{v}_M \end{pmatrix} \in R^{M \times M}$$

non ha righe uguali. Abbiamo dunque provato che se $V$ ha due righe uguali allora le righe corrispondono ad indici nella stessa componente connessa.

**Input:** $G$ grafo con $M$ componenti connesse.
**Output:** Componenti connesse $C_1, \ldots, C_M$
$L \leftarrow$ laplaciano di $G$;
**for** $i = 1, \ldots, M$ **do**
| $u_i \leftarrow i-$esimo autovettore normalizzato ortogonalae a $u_j$ con $j < i$
**end**
$U \leftarrow (u_1, \ldots, u_M)$;
$(\tilde{C}_1, \ldots, \tilde{C}_m) \leftarrow$ M-means(U);
**for** $i = 1, \ldots M$ **do**
| $C_i \leftarrow \left\{ j \in [N] \mid u_j \in \tilde{C}_i \right\}$
**end**

**Algorithm 3:** Unnormalize Laplacian spectral cluster algorithm

Dalla proposizione e dall'osservazione segue l'Algoritmo 3

**Definizione 10.7.** Dato un insieme $\mathcal{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^n$, una **funzione di similarità** è una qualsiasi funzione $s$ tale che

- $s(x, y) \in \mathbb{R}$
- $s(x, y) = s(y, x)$

Nel seguito, per alleggerire la notazione, scriveremo

$$s_{ij} := s(x_i, x_j)$$

**Esempio 10.8.** *Alcune possibili scelte per tale funzione sono*

1. *Fissato* $\varepsilon > 0$,
$$s_{ij} = 1_{||x_i - x_j|| \leq \varepsilon}$$
2. *Fissato* $K$
$$s_{ij} = 1_{x_i \text{ è tra i primi k-primi vicini (in senso euclideo) a } x_j \text{ e viceversa}}$$
3. *Fissato* $K$
$$s_{ij} = 1_{x_i \text{ è tra i primi k-primi vicini (in senso euclideo) a } x_j \text{ o viceversa}}$$
4. *Fissato* $\sigma > 0$
$$s_{ij} = \exp\left( -\frac{||x_i - x_j||^2}{2\sigma^2} \right)$$

Dato un insieme di dati e una funzione di simmilarità, possiamo definire un grafo pesato $G = (V, E, P)$ dove

$$V = \{1, \ldots, N\}$$
$$\{i, j\} \in E \quad \Leftrightarrow \quad w_{ij} = f(s_{ij}) > 0$$

con $f : \mathbb{R} \to [0, +\infty)$.

*Osservazione* 31. Tipicamente quando si usano le funzioni di similarità di tipo 1. 2. o 3. la funzione $f$ è l'identità. Se consideriamo una funzione tipo 4. e $f$ l'identità otterrei un grafo completo. In questo caso si fissa $\delta > 0$ e $f(z) = 1_{z > \delta}$

# 11 Vettori aleatori

## 11.1 Vettori isotropi

**Definizione 11.1.** Dato un vettore aleatorio $X$ in $\mathbb{R}^n$ definiamo la sua **matrice di covarianza** $\Sigma(X)$ in $R^{n\times n}$ come

$$\Sigma(X)_{ij} = \mathbb{E}[X_i X_j]$$

Diremo che $X$ è **isotropo** se $\Sigma(X) = I_n$

**Lemma 11.1.** *Sia $X$ vettore aleatorio isotropo allora*

*(i)*

$$\forall x \in \mathbb{R}^n \textit{ deterministico} \quad \mathbb{E}[\langle X,\, x\rangle^2] = ||x||^2$$

*(ii)*

$$\mathbb{E}[||X||^2] = n$$

*(iii) Se $Y$ è vettore aleatorio isotropo e indipendente da $X$ vale*

$$\mathbb{E}[\langle X,\, Y\rangle^2] = n$$

*Dimostrazione.*

(i)

$$\mathbb{E}[\langle X,\, x\rangle^2] = \mathbb{E}\left[\left(\sum_{i=1}^n X_i x_i\right)^2\right] = \sum_{i=1}^n \sum_{j=1}^2 \mathbb{E}[X_i x_i X_j x_j] =$$
$$= \sum_{i=1}^n \sum_{j=1}^n x_i x_j \mathbb{E}[X_i X_j] = \sum_{i=1}^n x_i^2 = ||x||^2$$

infatti essendo $X$ isotropo $\mathbb{E}[X_i X_j] = \delta_{ij}$

(ii)

$$\mathbb{E}[||X||^2] = \sum_{i=1}^n \mathbb{E}[X_i^2] = n$$

(iii) Usando la proprietà di speranza condizionata

$$\mathbb{E}[\langle X,\, Y\rangle^2] = \mathbb{E}\left[\mathbb{E}[\langle X,\, Y\rangle^2 \;|Y]\right] =_1 \mathbb{E}[||Y||^2] = n$$

dove $=_1$ deriva dal fatto che $X$ è isotropo per $\mathbb{P}(\cdot|\, Y)$ infatti essendo $X \amalg Y$ otteniamo

$$\mathbb{E}[X_i X_j \;|Y] = \mathbb{E}[X_i X_j] = \delta_{ij}$$

**Definizione 11.2.** Diremo che $X \sim Unif(\sqrt{n}S^{n-1})$ se $X$ è un vettore aleatorio con legge proporzionalle alla misura di Hausdorff $(n-1)$-dimensionale su $S^{n-1}$.

**Lemma 11.2.** *Valgono i seguenti risultati*

- $X \sim Unif(\sqrt{n}S^{n-1})$ *è isotropo.*
- *Se $X_1, \ldots, X_n$ sono v.a. di Rademacher indipendenti allora il vettore $X = (X_1, \ldots, X_n)$ è isotropo.*

*Dimostrazione.*

- Poichè la distribuzione è uniforme

  $$\mathbb{E}[X_1^2] = \dots \mathbb{E}[X_n^2]$$

  e dunque

  $$\mathbb{E}[X_i^2] = \frac{1}{n}\mathbb{E}\left[\sum_{j=1}^{n} X_j^2\right] = \frac{1}{n}\mathbb{E}[||X||^2] = 1$$

  infatti $X \in \sqrt{n}S^{n-1}$

  $$\mathbb{E}[X_iX_j] = \mathbb{E}[X_i(-X_j)] = -\mathbb{E}[X_iX_j] \quad \Rightarrow \quad \mathbb{E}[X_iX_j] = 0$$

  dove abbiamo usato che essendo la distribuzione uniforme, $(X_i, X_j) \sim (X_i, -X_j)$

- $\mathbb{E}[X_i^2] = \mathbb{E}[1] = 1$.
  Siano $i \neq j$ allora

  $$\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\mathbb{E}[X_j] = 0$$

  essendo le v.a. di Rademacher centrate e indipendenti.

*Osservazione* 32. Se $O \in O(n)$ e $X \sim Unif(\sqrt{S}^{n-1})$ allora $OX \sim Unif(\sqrt{n}S^{n-1})$

## 11.2 Vettori e matrici gaussiane

**Definizione 11.3.** Un vettore aleatorio $X$ è detto **gaussiano** standard se ha entrate i.i.d normali standard.
Un vettore aleatorio $Y$ è detto gaussiano se $\exists A \in R^{n\times d}$,$b \in \mathbb{R}^n$ deterministico e $X$ gaussiano standard d-dimensionale tale che $Y = AX + b$

**Proposizione 11.3.** *Se $Y = AX + b$ è gaussiano allora*

$$\mathbb{E}[Y] = b$$

$$Cov(Y) = AA^T$$

*In particolare, un vettore gaussiana standard è isoptropo.*

**Proposizione 11.4.** *Sia $X$ un vettore gaussiano in $\mathbb{R}^n$ allora la sua funzione generatrice*

$$M_X(t) = \exp\left(\langle t,\, b\rangle + \frac{1}{2}\langle t,\, \Gamma t\rangle\right)$$

*dove*

$$b = \mathbb{E}[X] \qquad \Gamma = Cov(X)$$

*Osservazione* 33. Dal Teorema 1.3, se $X$ e $Y$ sono vettori gaussiani con stessa media e stessa matrice di covarianza, allora $X \sim Y$. Dunque denotiamo la legge di un vettore di media $b$ e covarianza $\Gamma$ con $\mathcal{N}(b, \Gamma)$.

**Proposizione 11.5.** *Sia $\Gamma$ invertibile e $X \sim \mathcal{N}(b, \Gamma)$ allora*

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^n\sqrt{\det\Gamma}}\exp\left(-\frac{1}{2}\langle (x-b),\, \Gamma^{-1}(x-b)\rangle\right)$$

**Definizione 11.4.** Una matrice aleatoria $G$ è detta gaussiana se ha entrate i.i.d. normali standard.

**Proposizione 11.6.** *Sia $G$ una matrice $n \times n$ gaussiana standard. Dette $G_1, \ldots, G_n$ le sue colonne si ha*

$$\mathbb{P}\left(G_1, \ldots, G_n \text{ base di } \mathbb{R}^n\right) = 1$$

*Dimostrazione.* Basta provare che $\mathbb{P}(\det G = 0) = 0$. Ricordiamo che

$$\det G = \sum_{\sigma\in S_n}(-1)^{P(\sigma)}G_{1,\sigma(1)}\cdots G_{n,\sigma(n)} := F\left((G_{ij})_{i,j=1}^n\right)$$

Ora poichè

$$\Gamma = \{x = (x_{i,j})_{i=1}^n \mid F(x) = 0\}$$

ha misura di Lebesgue nulla otteniamo la tesi.

∎

# 12 Ampiezza sferica e gaussiana

**Definizione 12.1.** Dato $T \subseteq \mathbb{R}^n$ la sua **ampiezza sferica** è data da

$$w_S(T) = \mathbb{E}\left[\sup_{t \in T} \langle t,\, \theta \rangle\right]$$

dove $\theta \sim Unif(S^{n-1})$

**Definizione 12.2.** Dato $T \subseteq \mathbb{R}^n$ la sua **ampiezza gaussiana** è data da

$$w(T) = \mathbb{E}\left[\sup_{t \in T} \langle t,\, g \rangle\right]$$

dove $g \sim \mathcal{N}(0, I_n)$

**Lemma 12.1.** *Sia $g \sim \mathcal{N}(0, I_n)$ allora $\frac{g}{||g||}$ e $||g||$ sono v.a. indipenennti.*
*Inoltre $\frac{g}{||g||} \sim Unif(S^{n-1})$ mentre $||g||$ è una v.a. continua con densità della forma $ce^{-\frac{r^2}{2}}r^{n-1}$*

*Dimostrazione.* Poichè $g \sim \mathcal{N}(0, I_n)$

$$\mathbb{E}[h(G)] = \frac{1}{(\sqrt{2\pi})^n} \int_{\mathbb{R}^n} e^{-\frac{||x||^2}{2}} h(x)\, \mathrm{d}x = \frac{1}{(\sqrt{2\pi})^n} \int_{S^{n-1}} \mathrm{d}s(\theta) \int_0^{+\infty} \mathrm{d}r r^{n-1} e^{-\frac{r^2}{2}} h(r\theta)$$

Calcolando tale valore atteso per $h(g) = h_1\left(\frac{g}{||g||}\right) h_2(||g||)$ otteniamo la tesi.

$\square$

**Proposizione 12.2.** *Sia $T \subseteq \mathbb{R}^n$ allora*

$$w(T) = \mathbb{E}[||g||_2] w_S(T)$$

*dove $g \sim \mathcal{N}(0, I_n)$*

*Dimostrazione.*

$$\begin{aligned} w(T) =& \mathbb{E}\left[\sup_{t \in T} \langle g,\, t \rangle\right] = E\left[\sup_{t \in T} \langle ||g|| \frac{g}{||g||},\, t \rangle\right] = \mathbb{E}\left[||g|| \sup_{t \in T} \langle \frac{g}{||g||},\, t \rangle\right] =_1 \\ &=_1 \mathbb{E}[||g||] \mathbb{E}\left[\sup_{t \in T} \langle \frac{g}{||g||},\, t \rangle\right] =_2 w_S(T) \end{aligned}$$

dove $=_1$ deriva dall'indipendenza delle due v.a. mentre $=_2$ dal fatto che $\frac{g}{||g||}$ è uniforme sulla sfera.

$\square$

**Lemma 12.3.** *Esiste una costante assoluta $C$ tale che*

$$\sqrt{n} - C \leq \mathbb{E}[||g||] \leq \sqrt{n}$$

*Dimostrazione.* Proviamo solo l'upper bound.

$$\mathbb{E}[||g||] \leq_{C-S} \mathbb{E}\left[||g||^2\right]^{\frac{1}{2}} = \mathbb{E}\left[\sum_{i=1}^n g_1^2\right]^{\frac{1}{2}} = \left(\sum_{i=1}^n \mathbb{E}\left[g_1^2\right]\right)^{\frac{1}{2}} \leq \sqrt{n}$$

**Proposizione 12.4.** *Siano $T, S \subseteq \mathbb{R}^n$ allora valgono le seguenti affermazioni*

*(i)* $w(T) \in [0, +\infty]$

*(ii)* $w(T) \leq \infty \quad \Leftrightarrow \quad T$ *limitato*

*(iii)* $\omega(OT + y) = w(T)$ *per ogni* $O \in O(n)$ *e* $y \in \mathbb{R}^n$

*(iv)* $w(T) = w(conv(T))$

*(v)* $w(T + S) = w(T) + w(S)$

*(vi)* $w(aT) + |a|\, w(T)$ *per ogni* $a \in \mathbb{R}$

*(vii)* $\omega(T) = \frac{1}{2} w(T - T)$

*(viii)* $\frac{1}{\sqrt{2\pi}} diam(T) \leq w(T) \leq \frac{\sqrt{n}}{2} diam(T)$

*Dimostrazione.* Nel seguito della dimostrazione $g \sim \mathcal{N}(0, I_n)$

(i) Sia $t_0 \in T$ allora

$$w(T) = \mathbb{E}\left[\sup_{t \in T} \langle g,\, t\rangle\right] \geq \mathbb{E}[\langle g,\, t_0\rangle]$$

Ma poichè $g \sim -g$ otteniamo

$$\mathbb{E}[\langle g,\, t_0\rangle] = \mathbb{E}[\langle -g,\, t_0\rangle] = -\mathbb{E}[\langle g,\, t_0\rangle] = 0$$

(ii) Proviamo solo $\Leftarrow$. Per C.S. si ha $\langle g,\, t\rangle \leq ||g||\, ||t||$ e dunque posto $C(T) = \sup ||t||$ si ha

$$w(T) = \mathbb{E}\left[\sup_{t \in T} \langle g,\, t\rangle\right] \leq C(T)\mathbb{E}[||g||] < \infty$$

(iii) Proviamo l'invarianza per l'azione di $O(n)$. Sia $O \in O(n)$ allora

$$w(OT) = \mathbb{E}\left[\sup_{t \in T} \langle g,\, Ot\rangle\right] = \mathbb{E}\left[\sup_{t \in T} \langle O^T g,\, t\rangle\right]$$

ora $O^T g \sim g$ e dunque la tesi.
Proviamo l'invarianza per traslazioni

$$w(T + y) = \mathbb{E}\left[\sup_{t \in T} \langle g,\, t + y\rangle\right] = \mathbb{E}\left[\sup_{t \in T} \langle g,\, t\rangle\right] + \mathbb{E}[\langle g,\, y\rangle]$$

Ora ripercorrendo quanto fatto nel punto (i) si prova che il secondo addendo è nullo

(iv) Poichè $T \subseteq conv(T)$ usando la monotonia dell'estremo superiore ho

$$w(T) \leq w(conv(T))$$

Per l'altra disuguaglianza notiamo che se $\sum \lambda_i = 1$ con $\lambda_i \geq 0$ e $x_1, \ldots, x_n \in T$ allora

$$\langle g,\, \sum_{i=1}^m \lambda_i x_i\rangle = \sum_{i=1}^m \lambda_i \langle g,\, x_i\rangle \leq \max_i \langle g,\, x_i\rangle \sum_{i=1}^m \lambda_i = \max_i \langle g,\, x_i\rangle \leq \sup_{t \in T} \langle g,\, t\rangle$$

Dunque

$$w(conv(T)) = \mathbb{E}\left[\sup_{m \geq 1} \sup_{\substack{\lambda_1, \ldots, \lambda_m \geq 0 \\ \sum \lambda_i = 1}} \sup_{x_1, \ldots, x_m \in T} \langle g,\, \sum \lambda_i x_i\rangle\right] \leq \mathbb{E}\left[\sup_{t \in T} \langle g,\, t\rangle\right] = w(T)$$

(v)
$$w(T+S) = \mathbb{E}\left[\sup_{\substack{t\in T\\ s\in S}}\langle g,\, t+s\rangle\right] = \mathbb{E}\left[\sup_{t\in T}\langle g,\, t\rangle + \sup_{s\in S}\langle g,\, s\rangle\right] = w(T)+w(S)$$

(vi) Sia $a>0$ allora
$$w(aT) = \mathbb{E}\left[\sup_{t\in T}\langle g,\, at\rangle\right] = a\mathbb{E}\left[\sup_{t\in T}\langle g,\, t\rangle\right] = aw(T)$$
Il caso $a<0$ si prova osservando che $g\sim -g$ e dunque $w(T)=w(-T)$

(vii) Segue applicando il punto precedente a $T=T$ e $S=-T$

(viii) Per il punto precedente
$$w(T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g,\, x-y\rangle\right]$$
Ora se $a,b\in T$ si ha
$$w(T) \geq \frac{1}{2}\mathbb{E}\left[\max\langle g,\, a-b\rangle, \langle g,\, b-a\rangle\right] = \frac{1}{2}\mathbb{E}\left[|\langle g,\, b-a\rangle|\right]$$
Ma ora
$$\langle g,\, a-b\rangle = \sum_{i=1}^{n} g_i(a_i-b_i) \sim \mathcal{N}\left(0, \sum(a_i-b_i)^2\right) = \mathcal{N}\left(0, ||a-b||^2\right)$$
Abbiamo dunque
$$w(T) \geq \frac{1}{2}\,||a-b||\,\mathbb{E}[|Z|] \text{ con } Z\sim\mathcal{N}(0,1)$$
passando al sup in $a,b$ otteniamo l'upper bound.
Per il lower bound
$$w(T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g,\, x-y\rangle\right] \leq \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}||g||\,||x-y||\right] \leq \frac{1}{2}\mathbb{E}[||g||\,diam(T)]$$
Ricordando che $\mathbb{E}[||g||]\leq\sqrt{n}$ la tesi

□

**Esempio 12.5.** *Posta*
$$B_p^n = \left\{x\in\mathbb{R}^n \mid ||x||_p = 1\right\}$$
*si ha*
$$w\left(S^{n-1}\right) = \mathbb{E}\left[\sup_{t\in S^{n-1}}\langle g,\, t\rangle\right] = \mathbb{E}[||g||] \in [\sqrt{n}-C, \sqrt{n}]$$
$$w\left(B_2^n\right) = \mathbb{E}\left[\sup_{t\in B_2^n}\langle g,\, t\rangle\right] = \mathbb{E}[||g||_2]$$
$$w\left(B_\infty^n\right) = \mathbb{E}\left[||g||_1\right] = \sqrt{\frac{2}{\pi}}n$$
$$w\left(B_1^n\right) = \mathbb{E}\left[\sup_{t\in B_2^n}\langle g,\, t\rangle\right] = \mathbb{E}[||g||_\infty]$$

*Osservazione* 34. Vale il seguente risultato
$$\exists C_1, C_2 \quad c_1\sqrt{\log n} \leq \mathbb{E}[||g||_\infty] \leq C_2\log n$$

# 13 Recovery problem

Andiamo a presentare un problema che va sotto il nome di **recovery**. Supponiamo di disporre di $K \subseteq \mathbb{R}^n$ un insieme limitato e siamo interessati a trovare un certo vettore $x \in K$. Vogliamo trovare $x$ avendo a disposizione solo $m$ osservazioni gaussiane di tipo lineare su $x$ cioè conosco

$$y_1 = \langle a^{(1)},\, x\rangle, \dots, y_m = \langle a^{(m)},\, x\rangle$$

dove $a^{(1)}, \dots, a^{(m)}$ sono vettori gaussiani standard indipendenti (in senso probabilistico).
Posto $A = \begin{pmatrix} a^{(1)} \\ \vdots \\ a^{(m)} \end{pmatrix}$ io conosco $y = Ax$.

La filosofia è cercare di trovare il minimo valore di $m$ per cui se stimo $x$ con una generica soluzione $\hat{x}$ del sistema

$$\begin{cases} \hat{x} \in K \\ y = A\hat{x} \end{cases} \tag{11}$$

allora $x$ e $\hat{x}$ sono “vicini".

Per formalizzare tale problema occorre introdurre la Grassmanniana

**Definizione 13.1.** Siano $1 \leq m \leq n$ allora la **varietà grasmaniana** $G_{n,m}$ è la famiglia dei sottospazi $m$-dimensionale di $\mathbb{R}^n$. Definiamo su $G_{n,m}$ la metrica

$$d(E, F) = ||P_E - P_F||$$

dove $||\cdot||$ è la norma operatoriale per operatori lineari e $P_E$ è la proiezione ortogonale su $E$.

**Definizione 13.2.** Una distribuzione è detta uniforma su $G_{n,m}$ se la sua legge è invariante sotto l'azione del gruppo ortogonale

**Lemma 13.1.** *Sia $n \geq m$ e $A \in \mathbb{R}^{n\times m}$ una matrice gaussiana standard. Allora*

$$KerA \sim\ Unif(G_{n,n-m})$$

*Dimostrazione.* Poniamo $A' = \begin{pmatrix} A^1 & \dots & A^m \end{pmatrix}$ dove $A^i$ è la $i$-esima colonna di $A$, allora $A'$ è una matrice gaussiana standard $m \times m$ e quindi per la Proposizione 11.6 è quasi certamente invertibile. $rank(A) = m$ e dunque $\dim KerA = n - m$. Abbiamo dunque provato $Ker(A) \in G_{n,n-m}$. Per concludere basta provare che $OKer(A) \sim Ker(A)$ per ogni $O \in O(n)$.

$$y \in KerA \quad\Rightarrow\quad Ay = 0 \quad\Rightarrow\quad AO^TOy = 0 \quad\Rightarrow\quad Oy \in Ker(AO^T)$$
$$y \in Ker(AO^T) \quad\Rightarrow\quad AO^Ty = 0 \quad\Rightarrow\quad O^Ty \in KerA \quad\Rightarrow\quad y = O(O^Ty) \in OKer(A)$$

Le implicazione precedenti provano che $OKer(A) = Ker(AO^T) \sim Ker(A)$ infatti $A \sim AO^T$. $\square$

**Teorema 13.2** ($M^\star$-bound)**.** *Sia $K \subseteq \mathbb{R}^n$ limitato e sia $1 \leq m \leq n$. Se $E \sim Unif(G_{n,n-m})$ allora*

$$\mathbb{E}[diam(K \cap E)] \leq C_0 \frac{w(K)}{\sqrt{n}}$$

*dove $C_0 = 2\sqrt{2\pi}$*

**Teorema 13.3** (Estimation from gaussian linear observation-feasability problem). *Sia $1 \leq m \leq n$ e $K \subseteq \mathbb{R}^m$ limitato. Sia $A \in \mathbb{R}^{m\times n}$ una matrice gaussiana standard. Se $\hat{x}, x \in K$ tali che $A\hat{x} = Ax$ allora*

$$\mathbb{E}\left[||x-\hat{x}||_2\right] \leq 2C_0\frac{w(K)}{\sqrt{n}}$$

*dove $C_0 = 2\sqrt{2\pi}$*

*Dimostrazione.* Applicando $M^\star$-bound rimpiazzando $K$ con $K-K$ ottengo

$$\mathbb{E}\left[diam((K-K)\cap E)\right] \leq C_0\frac{w(K-K)}{\sqrt{n}} = 2C_0\frac{w(K)}{\sqrt{n}} \tag{12}$$

dove $E = KerA \sim Unif(G_{n,n-m})$ Ora poichè $x, \hat{x} \in K$ si ha $x - \hat{x} \in K-K$ e poichè $Ax = A\hat{x}$ si ha $x-\hat{x} \in KerA = E$. Notando che anche $0 \in (K-K)\cap E$ otteniamo

$$||x-\hat{x}||_2 = ||(x-\hat{x})-0||_2 \leq diam((K-K)\cap E)$$

Prendendo il valore atteso in entrambi i membri ed utilizzando (12) la tesi.

□

*Osservazione* 35. Se vogllio che l'errore in media sia più piccolo di $\delta$, basta prendere

$$m \geq \frac{C_0^2}{4\delta^2}w(K)^2$$

ovvero l'ordine dell'$m$ ottimale è lo stesso ordine di $w(K)^2$

*Osservazione* 36. Se $K$ è convesso, noti $(K, A, y)$ la soluzione del sistema (11) è numericamente migliore rispetto al caso non convesso.
Poichè sappiamo che $w(K) = w(conv(K))$ e $x \in K \subseteq conv(K)$ posso applicare la teoria sostituendo a $K$ il suo inviluppo convesso senza peggiorare le stime

## 13.1 Recovery problem con rumore

Considereremo un problema simile al caso precedente ma supporremmo che le $m$ misurazioni gaussiane siano affette da rumore.

**Teorema 13.4** ($M^\star$-bound generalizzato). *Sia $T \subseteq \mathbb{R}^n$ limitato e $A \in \mathbb{R}^{m\times n}$ matrice gaussiana standard. Fissato $\varepsilon > 0$ poniamo*

$$T_\varepsilon = \left\{t \in T \mid \frac{1}{m}||At||_{L^1} \leq \varepsilon\right\}$$

*allora vale*

$$\mathbb{E}\left[\sup_{t\in T_\varepsilon}||t||_2\right] \leq \sqrt{\frac{8\pi}{m}}\mathbb{E}\left[\sup_{t\in T}|\langle g,\, t\rangle|\right] + \varepsilon\sqrt{\frac{\pi}{2}}$$

**Proposizione 13.5.** *Il Teorema precedente generalizza l'$M^\star$-bound*

*Dimostrazione.* Si applichi $M^\star$-bound generalizzato sostituendo $T$ con $T-T$ e con $\varepsilon = 0$. Allora

$$T_0 = \{t \in T-T \ \ ||At||_{L^1} = 0\} = (T-T)\cap KerA \supseteq (T\cap KerA) - (T\cap KerA)$$

dove per l'ultima inclusione abbiamo usato che $KerA$ è chiuso per differenze.
Allora

$$\mathbb{E}\left[\sup_{t\in T_0}||t||\right] \geq \mathbb{E}\left[\sup_{t,t'\in T\cap KerA}||t-t'||\right] = \mathbb{E}\left[diam(T\cap KerA)\right]$$

Notando che

$$\mathbb{E}\left[\sup_{t\in T-T}|\langle g,\, t\rangle|\right] = \mathbb{E}\left[\sup_{t\in T-T}\langle g,\, t\rangle\right] = w(T-T) = 2w(T)$$

si ha la tesi.

□

**Teorema 13.6** (Feasability problem con rumore)**.** *Sia $K \subseteq \mathbb{R}^n$ limitato, $A \in \mathbb{R}^{m\times n}$ gaussiana standard, $\varepsilon > 0$ fissato e $\mu \in \mathbb{R}^m$ tale che $\frac{1}{m}||\mu||_{L^1} \leq \varepsilon$. Sia $x \in K$ e $y = Ax + \mu$ allora preso $\hat{x} \in K$ con $\frac{1}{m}||A\hat{x} - y||_{L^1} \leq \varepsilon$ vale*

$$\mathbb{E}\left[||x-\hat{x}||\right] \leq 4\sqrt{2\pi}\frac{w(K)}{\sqrt{m}} + \varepsilon\sqrt{2\pi}$$

*Dimostrazione.* Applicando $M^\star$-bound generalizzato con $T = K - K$ e sostituendo $\varepsilon$ con $2\varepsilon$ si che

$$\mathbb{E}\left[\sup_{t\in T_{2\varepsilon}}||t||\right] \leq \sqrt{\frac{8\pi}{m}}\mathbb{E}\left[\sup_{t\in K-K}|\langle g,\, t\rangle|\right] + \varepsilon\sqrt{2\pi}$$

Ora con considerazioni simili a quelle usate nella dimostrazione della Proposizione precedente si ha $\mathbb{E}\left[\sup_{t\in K-K}|\langle g,\, t\rangle|\right] = 2w(T)$ e la tesi segue provando che $x - \hat{x} \in T_{2\varepsilon}$ infatti

$$\frac{1}{m}||Ax - A\hat{x}||_{L^1} = \frac{1}{m}||y - \mu - A\hat{x}||_{L^1} \leq \frac{1}{m}||y - A\hat{x}||_{L^1} + \frac{1}{m}||\mu||_{L^1} \leq 2\varepsilon$$

□

# 14 Compressione di dati

Prima di enunciare il Lemma di Johnson-Linderstrauss, alla base dell'algoritmo di compressione illustrato in questa sezione, occorre presentare alcuni risultati preliminari

**Teorema 14.1** (di concentrazione per funzioni Lipschitziane sulla sfera). *Sia* $X \sim Unif(\sqrt{n}S^{n-1})$ *e* $f : \sqrt{n}S^{n-1} \to \mathbb{R}$ *una funzione lipschitziana.*

*(i)*
$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq C\,||f||_{Lip}$$

*(ii) Se, inoltre,* $p > 0$ *e* $f \geq 0$ *allora*
$$||f(X) - ||f(X)||_{L^p}||_{\psi_2} \leq C(p)\,||f||_{Lip}$$
*con* $C(p)$ *indipendente da* $f$*.*

*dove* $||f||_{Lip}$ *è la migliore costante* $c = c(f)$ *tale che* $|f(x) - f(y)| \leq c\,||x - y||$

*Dimostrazione.* Non dimostriamo il primo punto, osserviamo solamente che essendo la funzione $f$ lipschitzianna e la sfera compatti, la funzione $f$ è limitata e quindi integrabile.

Proviamo (ii). Dalla disuguaglianza triangolare e dal punto (i) si ha

$$\begin{aligned} ||f(X) - ||f(X)||_{L^p}||_{\psi_2} &\leq ||f(X) - \mathbb{E}[f(X)]||_{\psi_2} + ||\mathbb{E}[f(X)] - ||f(X)||_{p}\,||_{\psi_2} \leq \\ &\leq C\,||f||_{Lip} + |\mathbb{E}[f(X)] - ||f(X)||_{L^p}|\,||1||_{\psi_2} \end{aligned}$$

Ora poichè $f \geq 0$ allora $||\mathbb{E}[f(X)]||_{L^p} = |\mathbb{E}[f(X)]| = \mathbb{E}[f(X)]$ e quindi

$$|\mathbb{E}[f(X)] - ||f(X)||_{L^p}| = |\,||\mathbb{E}[f(X)]||_{L^p} - ||f(X)||_{L^p}\,| \leq ||\mathbb{E}[f(X)] - f(X)||_{L^p}$$

Ora usando l'Osservazione 24 si ha

$$||f(X) - \mathbb{E}[f(X)]||_{L^p} \leq C\sqrt{p}\,||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq C\sqrt{p}\,||f||_{Lip}$$

da cui la teis

Possiamo riformulare il Teorema precedente per $X \sim Unif(S^{n-1})$ e otteniamo la

**Proposizione 14.2.** *Sia* $X \sim Unif(S^{n-1})$ *e* $f : S^{n-1} \to \mathbb{R}$ *una funzione lipschitziana. Allora*

$$||f(X) - \mathbb{E}[f(X)]||_{\psi_2} \leq \frac{C}{\sqrt{n}}\,||f||_{Lip} \tag{13}$$

*Inoltre se* $f \geq 0$ *vale*

$$||f(X) - ||f(X)||_{L^p}||_{\psi_2} \leq \frac{C(p)}{\sqrt{n}}\,||f||_{Lip} \tag{14}$$

*Dimostrazione.* Sia
$$g : \sqrt{n}S^{n-1} \to \mathbb{R} \quad g(x) = f\left(\frac{x}{\sqrt{n}}\right)$$

allora $g$ è lipschitziana e vale $||g||_{Lip} = \frac{||f||_{Lip}}{\sqrt{n}}$. Poniamo $Y = \sqrt{n}X$ si ha $Y \sim Unif(\sqrt{n}S^{n-1})$ e $f(X) = g(Y)$. Applicando i due risultati a $g$ la tesi segue.

□

*Osservazione* 37 (Chiave per il Lemma successivo). Sia $z \in S^{n-1}$ e sia $\tilde{P} : \mathbb{R}^n \to \mathbb{R}^m$ la proiezione sulle prime $m$ componenti. Se $Z \sim Unif(S^{n-1})$ allora

$$||Pz|| \sim ||\tilde{P}Z||$$

**Lemma 14.3.** *Sia $\varepsilon$ e $P = P_E$ con $E \sim Unif(G_{n,m})$. Fissato $z \in \mathbb{R}^n$ vale*

*(i)*

$$\mathbb{E}\left[||Pz||^2\right]^{\frac{1}{2}} = \sqrt{\frac{m}{n}}\,||z||$$

*(ii) con probabilità $\geq 1 - 2\exp(-c\varepsilon^2 m)$ si ha*

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\,||z|| \leq ||Pz|| \leq (1+\varepsilon)\sqrt{\frac{m}{n}}\,||z|| \tag{15}$$

*Dimostrazione.* Per omogeneità basta provare la tesi per $z \in S^{n-1}$.

(i) Usando l'Osservazione precedente,

$$\mathbb{E}\left[||\tilde{P}Z||^2\right] = \mathbb{E}\left[Z_1^2 + \cdots + Z_m^2\right] = m\mathbb{E}[Z_1^2]$$

ma

$$1 = \mathbb{E}\left[||Z||^2\right] = \mathbb{E}\left[Z_1^2 + \cdots + Z_n^2\right] = n\mathbb{E}\left[Z_1^2\right]$$

da cui la tesi

(ii) La funzione

$$f : S^{n-1} \to \mathbb{R} \qquad f(x) = ||(x_1, \ldots, x_m)||$$

è lipschitziana con $||f||_{Lip} = 1$, usando (14) e l'Osservazione 24 otteniamo che esiste una costante universale $C_1$ con

$$\mathbb{P}\left(|f(Z) - ||f(Z)||_{L^2}| \geq t\right) \leq 2\exp\left(-C_1 n t^2\right)$$

ma $f(Z) = \left|\left|\tilde{P}Z\right|\right|$ e $||f(Z)||_{L^2} = \sqrt{\frac{m}{n}}$.

Prendendo $t = \varepsilon\sqrt{\frac{m}{n}}$ si ha la tesi

□

**Teorema 14.4** (Lemma di Johnson-Linderstrauss)**.** *Sia $\chi \subseteq \mathbb{R}^n$ di cardinalità $N$ e $\varepsilon > 0$. Sia inoltre $m \in \mathbb{N}$ tale che*

$$n \geq m \geq \frac{C}{\varepsilon^2}\log N$$

*Sia $E \sim Unif(G_{n,m})$ e posta $Q = \sqrt{\frac{n}{m}}P_E$ con $P_E$ proiezione ortogonale su $E$, con probabilità di almeno $1 - 2\exp(-c\varepsilon^2 m)$ vale*

$$(1-\varepsilon)\,||x-y|| \leq ||Qx - Qy|| \leq (1+\varepsilon)\,||x-y||$$

*dove $C, c$ sono costanti assolute positive.*

*Dimostrazione.* Dalla definizione di $Q$, posso riscrivere (15) come

$$(1-\varepsilon)\,||z|| \leq ||Qz|| \leq (1+\varepsilon)\,||z|| \tag{16}$$

Ora

$$\begin{aligned}\mathbb{P}\left((16) \text{ non valga } \forall z = x-y \text{ con } x,y \in \chi\right) &\leq \sum_{x,y\in\chi} \mathbb{P}\left((16) \text{ non valga per } z = x-y\right) = \\ &= N^2 2\exp(-c\varepsilon^2 m) = 2\exp\left(2\log N - c\varepsilon^2 m\right) \leq 2\exp\left(2m\frac{\varepsilon^2}{C} - c\varepsilon^2 m\right) = \\ &= 2\exp\left(-\varepsilon^2 m\left(c - \frac{2}{C}\right)\right)\end{aligned}$$

prendendo $\frac{2}{C} = \frac{c}{2}$ si ha

$$\mathbb{P}\left((16) \text{ non valga } \forall z = x-y\ x,y\in\chi\right) \leq 2\exp\left(-\varepsilon^2 m\frac{c}{2}\right)$$

Ora rinominando $\frac{c}{2}$ con $c$ e passando al complementare otteniamo la tesi.

$\square$

*Osservazione* 38. Per generare $E \sim Unif(G_{n,m})$ basta prendere $E = span(v_1,\dots,v_m)$ dove $v_1,\dots,v_m$ sono i.i.d. gaussiani standard

*Osservazione* 39. $Q(\chi)$ è uno spazio $m$-dimensionale: invece di memorizzare $\chi$ (quindi $nN$ entrate), memorizzo $Q(\chi)$ (quindi $mN$)

# 15 Esercizi

**Esercizio 15.1.** *Sia $K = [0, a]$ con $a > 0$ dotato della distanza euclidea. Calcolare il suo numero di ricoprimento e impacchettamento.*

*Dimostrazione.* L'intuizione geometrica ci suggerisce che

$$N(k, d, \varepsilon) = \lceil \frac{a}{2\varepsilon} \rceil$$

Un $\varepsilon$-net con tale cardinalità esiste (si prendono i punti $\varepsilon, 3\varepsilon, ...$ e se $a$ non è un multiplo di $2\varepsilon$ prendiamo ancha $a$), proviamo che non ne possono esistere di cardinalità minore.
Sia $A$ un $\varepsilon$-net di $K$ allora

$$K \subseteq \bigcup_{x\in A} [x - \varepsilon, x + \varepsilon] \quad \Rightarrow \quad l(K) \leq l\left(\bigcup_{x\in A} [x - \varepsilon, x + \varepsilon]\right) \leq \sum_{x\in A} l([x - \varepsilon, x + \varepsilon]) \leq 2\varepsilon\, |A|$$

Da cui $|A| \geq \dfrac{a}{2\varepsilon}$ e poichè la cardinalità è un numero naturale, possiamo prendere la parte intera superiore.

Sia $A \subseteq [0, a]$ un insieme $\varepsilon$-separato di cardinalità $n$. Allora poichè ogni punto deve essere a distanza $\varepsilon$ si ha

$$(n - 1)\varepsilon < a \quad \Rightarrow \quad n < \frac{a}{\varepsilon} + 1 \tag{17}$$

Andiamo a distinguere due differenti casi

- Se $a$ è un multiplo di $\varepsilon$ allora la disuguaglianza (17) diventa $n \leq \frac{a}{\varepsilon}$.
  Siano
  $$x_k = k(\varepsilon + \delta) \quad \text{con } k = 0, 1, \ldots \frac{a}{\varepsilon} - 1 \text{ e } \delta > 0 \text{piccolo}$$
  Allora $\{x_k\}$ è un $\varepsilon$-separato di cardinalità massima. Il numero d'impacchettamento vale $\frac{a}{\varepsilon}$
- Supponiamo che $a$ non sia un multiplo di $\varepsilon$. Essendo la cardinalità un numero naturale, la disuguaglianza (17) diventa $n \leq \left\lfloor \frac{a}{\varepsilon} \right\rfloor + 1$. Se prendiamo
  $$x_k = k(\varepsilon + \delta) \quad \text{con } k = 0, 1, \ldots \frac{a}{\varepsilon} \text{ e } \delta > 0 \text{piccolo}$$
  otteniamo che $\{x_k\}$ è un insieme $\varepsilon$-separato massimale e dunque il numero d'impacchettamento vale $\left\lfloor \frac{a}{\varepsilon} \right\rfloor + 1$

**Esercizio 15.2.**

- *Supponiamo che la distanza su $T$ sia indotta da una norma allora*
  $$d(x, y) \leq \varepsilon \quad \Leftrightarrow \quad B\left(x, \frac{\varepsilon}{2}\right) \cap B\left(y, \frac{\varepsilon}{2}\right) \neq \emptyset$$
- *Se la distanza non è indotta da una norma vale solamente una delle due implicazioni*

*Dimostrazione.*

- $\Leftarrow$ Sia $z \in B\left(x, \frac{\varepsilon}{2}\right) \cap B\left(y, \frac{\varepsilon}{2}\right)$ allora
  $$d(x, y) \leq d(x, z) + d(z, y) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$
  Si noti che non abbiamo usato nessuna propietà dedotta dalla norma

- $\Rightarrow$ Poniamo $z = \dfrac{x+y}{2}$ allora

$$d(x,z) = ||x-z|| = \left|\left|x - \frac{x+y}{2}\right|\right| = \frac{||x||-y}{2} \leq \varepsilon 2$$

  e similmente con $y$. Dunque $z$ appartiene all'intersezione delle due palle.

Come controesempio possiamo prendere $\varepsilon = 1$ con $K$ qualsiasi dotato della metrica discreta:

$$d(x,y) = \begin{cases} 1 & \text{se } x \neq y \\ 0 & \text{se } x = y \end{cases}$$

**Esercizio 15.3.**

- *Mostrare che non vale l'implicazione*

$$K \subseteq L \quad \Rightarrow \quad N(K,d,\varepsilon) \leq N(L,d,\varepsilon)$$

- *Mostrare che vale l'implicazione*

$$K \subseteq L \quad \Rightarrow \quad P(K,d,\varepsilon) \leq P(L,d,\varepsilon)$$

*Dimostrazione.*

- Consideriamo i sottoinsiemi di $\mathbb{R}$: $K_1 = \{-1,1\}$ e $K_2 = \{-1,0,1\}$.
  Chiaramente $N(K_1,1) = 2$ infatti un 1-net deve essere formato da punti dello spazio. Ora una palla centrata in un punto di raggio 1 non contiene l'altro punto (sono a distanza 2). L'insieme $\{0\}$ è un 1-net per $K_2$ dunque $N(K_2,1) \leq 1$ e dunque $N(K_2,1) = 1$ essendo $N(K,1) \geq 1$ per qualsiasi $K$.
- L'implicazione segue ricordando la definizione di $P$ e notando che ogni $\epsilon$-separato per $K$ è un $\epsilon$-separato per $L$.

□

**Esercizio 15.4.** *Verificare che nel Teorema di Glivenko-Cantelli valga*

$$\mathbb{F}_n(t) - F(t) \geq \mathbb{F}_n(t_{i-1}) - F(t_{i-1}) - \varepsilon \text{ se } t_{i-1} \leq t < t_i$$

*Dimostrazione.* Si usa che $\mathbb{F}_n$ e $F$ sono funzioni crescenti con limite a sinistra e dunque

$$\mathbb{F}_n(t) - F(t) \geq \mathbb{F}_n(t_{i-1}) - F(t_i^-)$$

si conclude ricordando le proprietà della partizione $(t_i)$.

□

**Esercizio 15.5.** *Supponendo che osservando il campione aleatorio $X_1, \ldots, X_{100}$ rileviamo i seguenti valori $-2, 5, 8, 3$ con frequenza $20, 25, 40, 15$. Determinare la misura empirica e la funzione di ripartizione empirica associate a tale realizzazione.*

*Dimostrazione.* Dalla definizione di misura empirica si ha

$$\mathbb{P}_n = \frac{1}{100}\left(20\delta_{-2} + 15\delta_3 + 25\delta_5 + 40\delta_8\right)$$

e poichè $\mathbb{F}_n(t) = \mathbb{P}_n((-\infty, t])$ si ha

$$\mathbb{F}_n(t) = \begin{cases} 0 & \text{se } t < -2 \\ \frac{20}{100} & \text{se } -2 \leq t < 3 \\ \frac{35}{100} & \text{se } 3 \leq t < 5 \\ \frac{60}{100} & \text{se } 5 \leq t < 8 \\ 1 & \text{se } t \geq 8 \end{cases}$$

**Esercizio 15.6.** *Consideriamo un campione aleatorio* $(X_i)_{i=1}^n$ *di v.a. reali (non conosciamo la distribuzione della popolazione). Usando la disuguaglianza DKW determinare di quale ampiezza* $n$ *dobbiamo prendere il campione affinchè, con probabilità almeno* 90%*, la funzione di ripartizione empirica differisca in norma uniforme al più* 0.05 *dalla funzione di ripartizione della distribuzione della popolazione.*

*Dimostrazione.* L'esercizio ci chiede di trovare $n$ minimale affinchè valga

$$\mathbb{P}\left( \, ||\mathbb{F}_n - F||_\infty \leq 0.05\right) \geq 0.9$$

*Dimostrazione.* La tesi equivale a provare che

$$\mathbb{P}\left( \, ||\mathbb{F}_n - F||_\infty > 0.05\right) \geq 0.1$$

Se troviamo $n$ e $x$ tali che

$$2e^{-2x^2} = 0.10 \qquad \frac{x}{\sqrt{n}} = \frac{1}{20}$$

otteniamo la tesi.
Svolgendo i conti otteniamo che $n$ deve essere almeno 600

**Esercizio 15.7.** *Sia* $\mathcal{A}_1 = \left\{G \in \mathcal{A} \mid \int |t| \, \mathrm{d}G(t) < \infty \right\}$ *allora il funzionale*

$$\gamma: \, \mathcal{A}_1 \to \mathbb{R} \quad \gamma(G) = \int t \, \mathrm{d}G(t)$$

*non è continua.*
*La funzione diventa continua se restringiamo il dominio a* $\mathcal{A}_0 = \{G \; [a,b] \to [0,1]\} \cap \mathcal{A}$

*Dimostrazione.* Sia

$$X_n = \begin{cases} 0 & \text{con probabilità} 1 - \frac{1}{n} \\ n^2 & \text{con probabilità} \frac{1}{n} \end{cases}$$

e $G_n$ la sua funzione di ripartizione. Allora

$$\gamma(G_n) = \mathbb{E}[G_n] = n^2 \frac{1}{n} = n \to +\infty$$

Notiamo perè che posta

$$G(t) = \begin{cases} 0 & \text{se } t < 0 \\ 1 & \text{se } t \geq 0 \end{cases}$$

si ha $G \in \mathcal{A}$ ( $\mathrm{d}G = \delta_0$) e $||G - G_n||_\infty \to 0$.

Il secondo punto si fa utilizzando la definizione di integrale di *Stinges(??)*

$$\int_a^b t \, \mathrm{d}G(t) = \lim_{n \to +\infty} \sum_{k=0}^{n} \left(a + \frac{k}{n}\right) \left( G\left(a + \frac{k}{n}\right) - G\left(a + \frac{k-1}{n}\right)\right) = \cdots = bG(b) - \int_a^b G(t) \, \mathrm{d}t$$

**Esercizio 15.8.** *Tale esercizio consiste nel dimostrare il seguente*

**Lemma 15.1.** *Sia $\mathcal{G} = \{g : \chi \to \mathbb{R}\}$ e $X$ una v.a. a valori in $\chi$*

$$\sup_{g \in \mathcal{G}} \mathbb{E}[g(X)] \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} |g(X)|\right]$$

*Dimostrazione.* Sia $g \in \mathcal{G}$ allora vale

$$g \leq \sup_{f \in \mathcal{G}} |g|$$

infatti $g \leq |g|$ e $g \in \mathcal{G}$. Applicando il valore atteso otteniamo

$$\mathbb{E}[g(X)] \leq \mathbb{E}\left[\sup_{f \in \mathcal{G}} |f|(X)\right]$$

Applicando l'estremo superiore per $g \in \mathcal{G}$ a sinistra otteniamo la tesi □

**Esercizio 15.9.** *Supponiamo che $P_X$ sia senza atomi. Dimostrare che la complessità di Rademacher della classe $\mathcal{S} = \{1_S \mid S \text{ finito}\}$ è maggiore di $\frac{1}{2}$.*

*Dimostrazione.* Poichè $P_X$ non ha atomi allora con probabilità 1, un campione $X_1, \dots, X_n$ ha elementi distinti. Fissata un'estrazione e dati $\varepsilon_1, \dots, \varepsilon_n$ allora

- Se nelle v.a. di Rademacher $+1$ ha maggior frequenza, l'insieme che massimizza è

  $$S = \{X_i \,|\varepsilon_i = 1\}$$

  allora

  $$\frac{1}{n}\sum \varepsilon_i 1_S(X_i) = \frac{|\{i \mid \varepsilon_i = 1\}|}{n} \geq \frac{1}{2} \cdot \frac{n}{2} = \frac{1}{2}$$

- Se nelle v.a. di Rademacher $-1$ ha maggior frequenza, l'insieme che massimizza è

  $$S = \{X_i \,|\varepsilon_i = -1\}$$

  allora

  $$\left|\frac{1}{n}\sum \varepsilon_i 1_S(X_i)\right| = \frac{|\{i \mid \varepsilon_i = -1\}|}{n} \geq \frac{1}{2} \cdot \frac{n}{2} = \frac{1}{2}$$

□

**Esercizio 15.10.** *Calcolare la complessità di Rademacher dell'insime*

$$A = \{(2, 3, -1),\ (1, 1, 0),\ (3, -2, 1)\}$$

*Dimostrazione.* Ricordiamo che $R_3(A) = \mathbb{E}[Z_A]$ con $Z_A = \sup_{a \in A} |[| \langle a,\, \varepsilon\rangle]$.
Notiamo che

| $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ | $\lvert\langle a_1,\, \varepsilon\rangle\rvert$ | $\lvert\langle a_2,\, \varepsilon\rangle\rvert$ | $\lvert\langle a_3,\, \varepsilon\rangle\rvert$ | $Z_A$ |
|---|---|---|---|---|
| $(1, 1, 1)$ | 4 | 4 | 2 | 4 |
| $(1, 1, -1)$ | 6 | 2 | 0 | 6 |
| $(1, -1, 1)$ | 2 | 0 | 6 | 6 |
| $(1, -1, -1)$ | 1 | 0 | 4 | 4 |

e sfruttando la simmetria otteniamo che

$$\mathbb{E}[Z_A] = \frac{1}{8} \sum_{\varepsilon_1 = \pm 1} \sum_{\varepsilon_2 = \pm 1} \sum_{\varepsilon_3 = \pm 1} Z_A(\varepsilon_1, \varepsilon_2, \varepsilon_3) = \frac{1}{8} \cdot 2\,(4 + 6 + 6 + 4) = 5$$

□

**Esercizio 15.11.** *Calcolare la dimensione VC delle seguenti classi*

- *Coppie di intervalli in* $\mathbb{R}$
- *Cerchi in* $\mathbb{R}^2$
- *Rettangoli in* $\mathbb{R}^2$

*Dimostrazione.*

- Si prova con una semplice verifica che 4 punti sono sempre frantumati. Mentre se un insieme contiene 5 punti non può essere frantumato. Sia $\Lambda \supseteq \{x_1 < \cdots < x_5\}$ allora la funzione
$$f(x_1) = f(x_3) = f(x_5) = 1 \text{ e } f(x_2) = f(x_4) = 0$$
non si può realizzare come restrizione di una funzione caratteristica di due intervalli.

- Si prova con una semplice verifica che 3 punti disposti ai vertici di un triangolo equilatero sono frantumati. Proviamo che un insieme di 4 punti non è mai frantumato.
Siano $\Lambda = \{a, b, c, d\} \subseteq \mathbb{R}^2$ un insieme di cardinalità 4 e poniamo
$$m = \min \{|A| \mid A \subseteq \Lambda \;\; \Lambda \subseteq conv(A)\}$$
dove $conv(A)$ è il più piccolo convesso che contiene $A$

  - $m \neq 1$. Infatti se, per assurdo, $m = 1$ si avrebbe, ad esempio, $\Lambda \subseteq conv(\{a\}) = \{a\}$ e dunque $\Lambda$ consterebbe di un solo punto.
  - $m = 2$. In questo caso i punti sono allineati e possiamo identificare $\Lambda$ come un sottoinsieme di $\mathbb{R}$ e assumendo $a < b < c < d$ la funzione
  $$f(a) = f(d) = 1 \qquad f(b) = f(c) = 0$$
  non si realizza come restrizione di indicatrici di cerchi (se esistesse si avrebbe $a, d \in C$ cerchio e dunque per convessità anche $b \in C$)
  - $m = 3$. Possiamo assumere, senza perdità di generalità che $a, b, c$ siano tre vertici di un triangolo (non degenere) e $d \in conv(\{a, b, c\})$. Dalla convessità dei cerchi la funzione
  $$f(a) = f(b) = f(c) = 1 \qquad f(d) = 0$$
  non si può realizzare come restrizione di un'indicatrice di un cerchio.
  - $m = 4$. I quattro punti sono i vertici di un quadrilatero (non degenere). Se chiamo $\alpha, \beta.\gamma, \delta$ gli angoli rispettivamente in $a, b, c, d$ si ottiene
  $$\alpha + \beta + \gamma + \delta = 360°$$
  dunque posso assumere, a meno di permutare gli indici, che $\alpha + \gamma \leq 180°$. Proviamo che la funzionne
  $$f(a) = f(c) = 1 \qquad f(b) = f(d)$$
  non può essere indotta dall'indicatrice di un cerchio.
  Supponiamo per assurdo che esista un cerchio $C$ tale che $a, c \in C$ e $b, d \notin C$. Consideriamo il quadrilatero $a', b, c', d'$ il quadrilatero che si ottiene spostando i vertici $a, c$ lungo $ac$ in modo che $a', c' \in \partial C$.
  Poichè $a, c \in C$, se chiamiamo $\alpha', \beta', \delta', \gamma'$ gli angoli del nuovo quadrilatero si ottiene
  $$\alpha' \leq \alpha \quad \gamma' \leq \gamma$$

Siano $b', d' \in \partial C$ i punti che si ottengono spostando $b, d$ lungo $bd$ e chiamiamo $\alpha'', \beta'', \gamma'', \delta''$ gli angoli di $a', b', c', d'$ allora poichè $b, d \notin C$ si ottiene

$$\alpha'' < \alpha' \quad \gamma'' < \gamma'$$

Ora essendo $a', b', c', d'$ iscritto in una circonferenza deve accadere che

$$\alpha'' + \gamma'' = 180°$$

Il che è assurdo poichè

$$\alpha'' + \gamma'' < \alpha + \gamma \leq 180°$$

- La dimensione $VC$ è 4 (banale verifica)

**Esercizio 15.12.** *Sia $\mathcal{F} = \{1_S \mid S \subseteq \mathbb{R}^2$ cerchio$\}$.*
*Trovare $n \in \mathbb{N}$ tale che*

$$\mathbb{P}\left(||\mathbb{P}_n - P_X||_{\mathcal{F}} < 0.5\right) \geq 0.90$$

*Come cambia tale $n$ se si vuole uno scarto minore di 0.2.*
*Si assuma che le costanti assolute valgano tutte $e^2$.*

*Dimostrazione.* Poichè $VC(\mathcal{F}) = 3$ dalla Proposizione 5.4

$$\mathbb{P}\left(||\mathbb{P}_n - P_X||_{\mathcal{F}} < 2c\sqrt{\frac{VC(\mathcal{F})}{n}} + \delta\right) \geq 1 - 2\exp\left(-\frac{\delta^2 n}{2}\right)$$

e dunque specializzando al nostro caso si ha

$$\mathbb{P}\left(||\mathbb{P}_n - P_X||_{\mathcal{F}} < 2e^2\sqrt{\frac{3}{n}} + \delta\right) \geq 1 - 2\exp\left(-\frac{\delta^2 n}{2}\right)$$

Ora imponendo che

$$1 - 2\exp\left(-\frac{\delta^2 n}{2}\right) > 0.9$$

$$2e^2\sqrt{\frac{3}{n}} + \delta \leq 0.5$$

si ottiene che il miglior $n \in \mathbb{N}$ è 3146.

**Esercizio 15.13.** *Siamo interessati a risolvere tramite apprendimento statistico un problema di classificazione. Disponiamo di un campione casuale composto da variabili aleatorie a valori nell'intervallo $[0, 1]$. Supponiamo di usare come spazio ipotesi $F$ le funzioni caratteristiche di intervalli $[a, b]$. Descrivere il protocollo per approssimare con qualche funzione in $F$ la funzione target determinando quanto grande deve essere il campione per avere*

- *Un rischio eccessivo non superiore a 0.3 con probabilit'a almeno del 70%*
- *Un valore atteso di rischio eccessivo non superiore a 0.3*

*Eventuali costanti assolute devono essere poste pari a 2).*

*Dimostrazione.* Dalla Proposizione 6.2 si ha

$$R(f_n^\star) - R(f^\star) \leq 2 \, ||\mathbb{P}_n - P||_{\mathcal{L}}$$

ed inoltre utlizzando il Teorema 5.5 e la Proposizione 6.3 otteniamo

$$R_n(\mathcal{L}) \leq R_n(\mathcal{F}) + \frac{1}{\sqrt{n}} \leq 2C\sqrt{\frac{VC(\mathcal{F})}{n}} + \frac{1}{\sqrt{n}}$$

Ora la dimensione VC della classe degli intervalli vale 1, usando il Teorema 4.5 otteniamo

$$\mathbb{P}\left( ||\mathbb{P}_n - P||_{\mathcal{L}} < \delta + \frac{2C+1}{\sqrt{n}} \right) \geq 1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$$

Dunque

$$\mathbb{P}\left( R(f_n^\star) - R(f^\star) < 2\delta + 2\frac{2C+1}{\sqrt{n}} \right) \geq P\left( ||\mathbb{P}_n - P||_{\mathcal{L}} < \delta + \frac{2C+1}{\sqrt{n}} \right) \geq 1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$$

Dunque per risolvere l'esercizio basta

$$\begin{cases} 1 - 2\exp\left(-\frac{n\delta^2}{2}\right) \geq 0.7 \\ 2\delta + 2\frac{2C+1}{\sqrt{n}} \leq 0.3 \end{cases} \quad \Rightarrow \quad \begin{cases} \delta \geq \sqrt{-\frac{2\ln(0.15)}{n}} \\ 2\sqrt{\frac{-2\ln(0.15)}{n}} + 2\frac{2C+1}{\sqrt{n}} \leq 0.3 \end{cases} \quad \Rightarrow \quad n \geq \frac{\left(\sqrt{-2\ln(0.15)} + 5\right)^2}{18}$$

**Esercizio 15.14.** *Sia $X$ v.a. gaussiana standard. Calcolare $E\left[e^{cX^2}\right]$ con $c \in \mathbb{R}$. Usare il risultato ottenuto per verificare direttamente le proprietà (iii) e (iv) della Proposizione 7.3 che caratterizza le variabili aleatorie subgaussiane. In particolare, calcolare esattamente $||X||_{\psi_2}$*

*Dimostrazione.*

$$\mathbb{E}\left[cX^2\right] = \frac{1}{\sqrt{2\pi}} \int \exp\left(cx^2 - \frac{1}{2}x^2\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int \exp\left(-\frac{1}{2}(1-2c)\,x^2\right) = \begin{cases} +\infty & \text{se } 1 - 2c \geq 0 \\ \frac{1}{\sqrt{1-2c}} & \end{cases}$$

Per la proprietà (iii) usiamo il calcolo di sopra per $c = \lambda^2$ ottenendo che per $c$ piccoli vale

$$\mathbb{E}\left[\exp(\lambda^2 X^2)\right] = \frac{1}{\sqrt{1-2c}} \leq 1 + 2c \leq e^{3c}$$

dunque la proprietà.
Per (iv) notiamo che per $t \geq \sqrt{2}$ vale

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] = \frac{1}{\sqrt{1 - \frac{2}{t^2}}} \leq 2 \quad \Leftrightarrow \quad t \geq \sqrt{\frac{8}{3}} \geq \sqrt{2}$$

Abbiamo dunque provato che $||X||_{\psi_2} = \sqrt{\frac{8}{3}}$

**Esercizio 15.15.** *Calcolare la norma subgaussiana di $X \sim$ Bernoulli$(p)$*

*Dimostrazione.* Sia $p \neq 0$ allora vale

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] = pe^{\frac{1}{t^2}} + 1 - p \leq 2 \quad \Leftrightarrow \quad t^2 \geq \left(\ln\left(\frac{1+p}{p}\right)\right)^{-1}$$

da cui

$$||X||_{\psi_2} = \ln\left(\frac{1+p}{p}\right)^{-\frac{1}{2}}$$

□

**Esercizio 15.16.** *Calcolare la norma subgaussiana di $X$ v.a. di Rademacher.*

*Dimostrazione.* Notando che $X^2 = 1$ quasi certamente otteniamo che

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] = \exp\left(\frac{1}{t^2}\right) \leq 2 \quad \Leftrightarrow \quad t \geq \frac{1}{\sqrt{\ln 2}}$$

□

**Esercizio 15.17.** *Sia $X$ v.a. limitata allora è subgaussiana*

*Dimostrazione.*

$$\mathbb{E}\left[\exp\left(\frac{X^2}{t^2}\right)\right] \leq \exp\left(\frac{||X||_\infty^2}{t^2}\right) \leq 2 \quad \Leftrightarrow \quad t^2 \geq \frac{||X||_\infty^2}{\ln 2}$$

dove $||\cdot||_\infty$ è la norma del sup-essenziale.
Abbiamo dunque che

$$||X||_{\psi_2} \leq \frac{||X||_\infty}{\sqrt{\ln 2}} < \infty$$

**Esercizio 15.18.** *Provare che le seguenti variabili aleatorie non sono subgaussiane:*

1. *Poisson*
2. *Esponenziale*
3. *Pareto*
4. *Cauchy*

*Dimostrazione.*

1. Sia $X \in \mathbb{N}$ allora

$$\begin{aligned}\mathbb{P}(|X| \geq n) = \mathbb{P}(X \geq n) > \mathbb{P}(X = n) = e^{-\gamma}\frac{\gamma^n}{n!} \geq e^{-\gamma}\frac{\gamma^n}{n^n} = \\ = \exp\left(-\gamma + n\ln\gamma - n\ln n\right)\end{aligned} \tag{18}$$

Se, per assurdo, $X$ fosse subgaussiana, dalla Proposizione 7.3 si avrebbe che esiste $C > 0$ tale che

$$\mathbb{P}(|X| \geq n) \leq \exp\left(\ln 2 - cn^2\right) \tag{19}$$

Mettendo insieme le disequazioni (18) e (19) si ottiene

$$\exp\left(-\gamma + n\ln\gamma - n\ln n\right) \leq \exp\left(\ln 2 - cn^2\right)$$

che è assurdo per $n$ grande.

2. Se $t \geq 0$ allora

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(X \geq t) = e^{-\gamma t} \tag{20}$$

Se, per assurdo, $X$ fosse subgaussiana, dalla Proposizione 7.3 si avrebbe che esiste $C > 0$ tale che

$$\mathbb{P}(|X| \geq t) \leq 2e^{-ct^2} \tag{21}$$

Mettendo insieme le disequazioni (20) e (21) si ottiene

$$e^{-\gamma t} \leq 2e^{-ct^2}$$

che è assurdo per $t$ grande

3. Ricordiamo che la v.a. di Pareto di parametri $x_m$ e $\alpha$ ha valori in $[x_m, \alpha)$ con densità $f(x) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$ dunque se $t \geq x_m$ allora

$$P(|X| \geq T) = P(X > t) = \int_t^{+\infty} f(x)\,\mathrm{d}x = x_m^\alpha t^{-\alpha}$$

e dunque non esiste un $K$ tale che $x_m^\alpha t^{-\alpha} \leq 2e^{-\frac{t^2}{k_1^2}}$

4. Ricordiamo che la v.a. di Cauchy ha valori in $\mathbb{R}$ con densità $f(x) = \frac{1}{\pi(x^2+1)}$ dunque

$$P(X \geq t) = \int_t^{+\infty} \frac{1}{\pi(x^2+1)} = \frac{1}{2} - \frac{\arctan(t)}{\pi}$$

ora non esistono costanti tali per cui $\frac{1}{2} - \frac{\arctan(t)}{\pi} \leq 2e^{-\frac{t^2}{k_1^2}}$

**Esercizio 15.19.** *Provare che $||X||_{\psi_2} \leq C(\sigma + |\mu|)$ se $X \sim \mathcal{N}(\mu, \sigma^2)$ dove $C$ è una costante assoluta.*

*Dimostrazione.* Notiamo che se $X \sim \mathcal{N}(\mu, \sigma^2)$ allora $X \sim \sigma Z + \mu$ la tesi segue usando che $||\cdot||_{\psi_2}$ è una norma.

**Esercizio 15.20.** *Riformulare il Teorema di Hoeffding per v.a. di Rademacher indipendenti considerando ora una famiglia di v.a. di Bernoulli di parametro $\frac{1}{2}$ indipendenti.*

*Dimostrazione.* Questo risultato è stato dimostrato nel Corollario 7.13.

**Esercizio 15.21.** *Lancio una moneta onesta N volte. Provare che la probabilità di avere almeno $\frac{3}{4}N$ teste è limitata dall'alto da $e^{-\frac{N}{8}}$.*

*Dimostrazione.* Consideriamo le v.a.

$$X_i = \begin{cases} 1 & \text{se all'}i\text{-esimo lancio esce testa} \\ 0 & \text{altrimenti} \end{cases}$$

Allora $X_1, \dots, X_N$ sono Bernoulli di parametro $\frac{1}{2}$ e applicando la disuguaglianza di Hoeffding otteniamo

$$\mathbb{P}\left(\sum_{i=1}^N X_i \geq \frac{3}{4}N\right) = \mathbb{P}\left(\sum_{i=1}^N \left(X_i - \frac{1}{2}\right) \geq \frac{3}{4}N - \frac{N}{2}\right) \leq e^{-\frac{N}{8}}$$

□

**Esercizio 15.22.** *Usando il risultato ottenuto nell'Osservazione 26, se l'algoritmo dà risposta corretta con probabilità del 70% determinare il minimo numero di iterazioni N affichè con probabilità almeno 90% la decisione presa (con majority vote) sia corretta. Consideriamo lo stesso problema se prendiamo 95% invece di 90%*

**Esercizio 15.23.** *Tale Esercizio è stato dimostrato come Lemma 8.3.*

**Esercizio 15.24.** *Sia $X$ vettore gaussiano standard e $O \in O(n)$. Provare che $OX$ è gaussiano standard.*

*Dimostrazione.* $OX$ è gaussiano essendo immagine lineare di un vettore gaussiano. Per la Proposizione 11.3, inoltre, $OX \sim \mathcal{N}(0, \Gamma)$ con $\Gamma = OO^T = I_n$

**Esercizio 15.25.** *Questo esercizio è stato dimostrato come Proposizione 11.6*

**Esercizio 15.26.** *Sia $G$ matrice gaussiana standard $n \times n$. Allora, fissato $O \in O(n)$, le matrici $OG$ e $GO$ sono entrambi matrici gaussiane standard.*

*Dimostrazione.* Proviamo che le entrate di $OG$ sono v.a. i.i.d. gaussiane standard.

Se $G = \begin{pmatrix} G_1 & \dots & G_n \end{pmatrix}$ consideriamo il vettore $X = \begin{pmatrix} G_1 \\ \vdots \\ G_n \end{pmatrix}$.

Per l'esercizio precedente il vettore

$$\tilde{O}X = \begin{pmatrix} O & & \\ & \ddots & \\ & & O \end{pmatrix} X$$

è gaussiano standard e dunque ha entrate i.i.d. gaussiane standard.
La tesi segue notando che $OG$ e $\tilde{O}X$ hanno le stesse entrate.

□

**Esercizio 15.27.** *Consideriamo $\chi = [0,1]$ con $P_X \sim \mathcal{U}([0,1])$. Proviamo che la classe di funzioni*

$$\mathcal{F} = \{1_S \mid S \subseteq [0,1] \text{ finito } \}$$

*non è di Glivenko-Cantelli.*
*Data $f \in \mathcal{F}$ si ha*

$$P_X(f) = \int_0^1 f(x) P_X(\,\mathrm{d}x) = 0$$

*infatti $P_X$ da massa nulla ai singoli punti e $f$ non è nulla solo su singoli punti.*
*Notiamo che se $f = 1_{\{X_1,\dots,X_n\}}$ allora $\mathbb{P}_n(f) = 1$ e quindi*

$$||\mathbb{P}_n - P_X||_{\mathcal{F}} = \sup_{g \in \mathcal{F}} |\mathbb{P}_n(g) - P_X(g)| \geq 1 - 0 = 1$$

*e dunque non può essere di Glivenko-Cantelli*

**Esercizio 15.28.** *Sia $X \sim \mathcal{N}(0,1)$. Calcolare esplicitamente $\mathbb{E}\left[\exp(cX^2)\right]$ e verificare direttamente le propietà (iii) e (iv).*

*Dimostrazione.*

$$\mathbb{E}[\exp(cX^2)] = \frac{1}{\sqrt{2\pi}} \int e^{cx^2} e^{-\frac{1}{2}x^2} \,\mathrm{d}x = \begin{cases} \frac{1}{\sqrt{1-2c}} & \text{se } 1 - 2c > 0 \\ +\infty & \text{altrimenti} \end{cases}$$

Per la verifica della proprietà (iii) si usa la formula di prima con $c = \lambda^2$ ottenendo

$$\mathbb{E}[\exp(\lambda^2 X^2)] = \frac{1}{\sqrt{1-2c}} \leq 1 + 2c \leq e^{3c} \text{ per } c \text{ opportunamente piccolo}$$

Un semplice calcolo prova che $||X||_{\psi_2} = \sqrt{2}$

**Esercizio 15.29.** *Sia $T \subseteq \mathbb{R}^n$ allora*

$$w(T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y \in T} |\langle g,\, x - y\rangle|\right]$$

*Dimostrazione.*

$$\begin{aligned}W(T) =&\frac{1}{2}w(T-T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\langle g,\, x-y\rangle\right] = \\ =&\frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}\max\left\{\langle g,\, x-y\rangle, \langle g,\, y-x\rangle\right\}\right] = \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T}|\langle g,\, x-y\rangle|\right]\end{aligned}$$

**Esercizio 15.30.** *Sia $T \subseteq \mathbb{R}^n$ con $T = -T$ allora*

$$w(T) = \mathbb{E}\left[\sup_{t\in T}|\langle g,\, t\rangle|\right]$$

*Dimostrazione.*

$$w(T) = \mathbb{E}\left[\sup_{t\in T}\langle g,\, t\rangle\right] = \mathbb{E}\left[\sup_{t\in T}\max\left\{\langle g,\, t\rangle, \langle g,\, -t\rangle\right\}\right] = \mathbb{E}\left[\sup_{t\in T}|\langle g,\, t\rangle|\right]$$

□

**Esercizio 15.31.** *La classe di funzioni $\mathcal{F} = \left\{1_{[-\infty,t]} \mid t \in \mathbb{R}\right\}$ ha discriminante polinomiale di ordine* 1

*Dimostrazione.* Fissata una successione $x_1^n$ e data $f = 1_{(-\infty,t]}$ a meno di permutare gli indici in modo che la successione sia ordinata in maniera crescente si ha

$$(f(x_1), \ldots, f(x_n)) = (1, \ldots, 1, 0, \ldots, 0)$$

e dunque se $\psi$ è la permutazione richiesta otteniamo

$$\psi\left(\mathcal{F}(x_1^n)\right) = \{(0, \ldots, 0), (1, 0, \ldots, 0), \ldots, (1, \ldots, 1)\}$$

e dunque $|\mathcal{F}(x_1^n)| = n+1$ □