Alma Mater Studiorum · Università di Bologna

**SCUOLA DI SCIENZE**
**Corso di Laurea in Matematica**

# LA METRICA DI WASSERSTEIN

**Relatore:**
**Chiar.mo Prof.**
**Stefano Pagliarani**

**Presentata da:**
**Emanuele Fioriti**

**I Sessione**
**Anno Accademico 2020/2021**

# Introduzione

La metrica di Wasserstein è una distanza nell'insieme delle misure di probabilità, sviluppata all'interno della teoria del trasporto ottimale. Questa teoria si fonda sul problema di trovare il modo ottimale di trasportare masse, sorto per esigenze economiche pochi anni prima della rivoluzione francese. Pertanto, con ottimale si intendeva un trasporto che minimizzasse le spese, ma con il passare degli anni lo stesso problema fu riscontrato in altri ambiti, ad esempio la meccanica dei fluidi, o più recentemente la meteorologia con Figalli. Pertanto il significato della parola ottimale dipende dal contesto in cui sorge il problema, corrispondendo comunque all'esigenza di trovare il punto di minimo del costo del trasporto.

Una prima formulazione adeguata del problema si deve al matematico russo Leonid Kantorovich, poi perfezionata in questo modello: preso $\mathcal{X}$, l'insieme dei luoghi di partenza della massa, preso $\mathcal{Y}$, l'insieme dei luoghi d'arrivo, si stima una funzione $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$. Qui $c(x, y)$ indica il costo di trasporto di un'unità di massa dal punto $x$ al punto $y$, perciò $c$ è chiamata funzione costo (l'unità è qualcosa fissata a priori, ad esempio il chilogrammo, o il metro cubo). Siccome in generale ci sono diversi luoghi di partenza e diversi luoghi di arrivo, si distribuisce la massa con due misure di probabilità, $\mu$ su $\mathcal{X}$ e $\nu$ su $\mathcal{Y}$. Infine si considerano le misure di probabilità su $\mathcal{X} \times \mathcal{Y}$ che hanno $\mu, \nu$ come marginali, chiamate piani di trasporto. Si cerca il piano di trasporto ottimale, punto di minimo della funzione

$$\pi \to \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y),$$

che è appunto il costo del trasporto per il piano $\pi$.

L'obiettivo di questa tesi è partire da questo modello nel caso in cui la massa sia trasportata all'interno di uno spazio polacco ($\mathcal{Y} = \mathcal{X}$ e $\mathcal{X}$ è polacco) prendendo come funzione costo la sua distanza elevata ad un $p \geq 1$, per definire la metrica di Wasserstein e studiarne alcune proprietà topologiche.

Nel capitolo 1 daremo la definizione della metrica di Wasserstein, la prova che effettivamente soddisfi gli assiomi di una distanza, e definiremo lo spazio di Wasserstein, un sottoinsieme di misure di probabilità sul quale la metrica è sicuramente finita. Infine, dalla interpretazione economica del problema, daremo un'idea intuitiva di come il problema di minimizzazione delle spese si possa trasformare nel problema duale di massimizzazione dei profitti, dal quale ricaveremo la formula duale di Kantorovich.

Nel capitolo 2 rafforzeremo il concetto di convergenza debole tra distribuzioni all'interno dello spazio di Wasserstein, richiedendo delle ulteriori proprietà che dimostreremo essere equivalenti. L'obiettivo principale è quello di provare che questo nuovo concetto è equivalente alla convergenza per la metrica di Wasserstein, e per farlo utilizzeremo la definizione di tightness di un insieme di misure: proveremo che una successione di Cauchy nello spazio di Wasserstein è tight.

Infine, nel capitolo 3 dimostreremo che lo spazio di Wasserstein eredita la struttura di spazio polacco, provando che ogni misura al suo interno può essere approssimata da una combinazione finita di delta di Dirac. Per questo fine, richiameremo il concetto di variazione assoluta di una misura con segno, e proveremo che esiste una relazione tra questa e la metrica di Wasserstein.

# Indice

# Capitolo 1

# La metrica di Wasserstein e il suo ambiente ideale

Prima di iniziare con la trattazione, fissiamo alcuni concetti. Useremo sempre uno spazio polacco $(\mathcal{X}, d)$, ossia uno spazio metrico completo e separabile. Lo spazio sarà sempre dotato della sigma-algebra di Borel, cioè quella generata dagli aperti metrici. Date due misure di probabilità $\mu$ e $\nu$ su $\mathcal{X}$, chiamiamo $\Pi(\mu, \nu)$ l'insieme delle misure di probabilità nel prodotto $\mathcal{X} \times \mathcal{X}$ , dotato della sigma-algebra prodotto, che hanno $\mu$ e $\nu$ come marginali. L'insieme di tutte le misure di probabilità su $\mathcal{X}$ verrà indicato con $P(\mathcal{X})$. Dato $Y = (Y_1, ..., Y_m)$, un vettore aleatorio a valori in $\mathcal{X}^m$, denoteremo con $\mu(Y) = \mu(Y_1, ..., Y_m)$ la legge di Y.

**Definizione 1.** *Dato $(\mathcal{X}, d)$ uno spazio polacco, dato $p \in [1, +\infty)$, e date $\mu, \nu \in P(\mathcal{X})$, si definisce la metrica di Wasserstein di ordine p come:*

$$W_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \Big( \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p d\pi \Big)^{1/p}. \tag{1.1}$$

Questo è equivalente a:

$$W_p(\mu, \nu) = \inf_{Y,Z} \Big\{ E[d(Y, Z)^p]^{1/p} : \, \mu(Y) = \mu \, , \, \mu(Z) = \nu \Big\}.$$

Un teorema che ci garantisce l'esistenza di un piano di trasporto ottimale è il seguente.

**Teorema 1.** *Siano* $(\mathcal{X}, \mu)$, $(\mathcal{Y}, \nu)$ *spazi polacchi di probabilità, e sia* $c = c(x, y)$ *non negativa e misurabile. Allora esiste sempre un piano di trasporto tra* $\mu$ *e* $\nu$ *che minimizza la funzione :*

$$\Pi(\mu, \nu) \ni \pi \to \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi.$$

Quindi l'estremo inferiore nella formula (1.1) in realtà è un minimo. Per la dimostrazione di questo teorema rimandiamo al capitolo 4 di [1].

**Osservazione 1.** $W_p$ *può assumere il valore* $+\infty$*, pertanto saranno selezionate le misure di probabilità su* $\mathcal{X}$ *in modo da garantire valori finiti (si veda la Definizione 2).*

Nell'ipotesi che $W_p$ assuma valori finiti, vediamo ora che soddisfa gli assiomi di una distanza. Prima di tutto si ha bisogno del seguente lemma, la cui dimostrazione è triviale:

**Lemma 1** (di incollamento)**.** *Siano* $(Y_1, Y_2)$ *e* $(Z_1, Z_2)$ *vettori aleatori a valori in* $\mathcal{X}\times\mathcal{X}$*, tali che* $\mu(Y_2) = \mu(Z_1)$*. Allora esiste una terna di variabili aleatorie* $(X_1, X_2, X_3)$ *t.c.* $\mu(X_1, X_2) = \mu(Y_1, Y_2)$ *e* $\mu(X_2, X_3) = \mu(Z_1, Z_2)$*.*

Abbiamo quindi:

- Banalmente $W_p \geq 0$ e simmetrica.
- Date $\mu_1, \mu_2, \mu_3$, siano $(Y_1, Y_2)$ ottimale con marginali $\mu_1, \mu_2$ e $(Z_1, Z_2)$ ottimale con marginali $\mu_2, \mu_3$. Allora esistono $X_1, X_2, X_3$ come nel lemma di incollamento, e si ha:

  $$W_p(\mu_1, \mu_3) \leq E[d(X_1, X_3)^p]^{1/p} \leq E[(d(X_1, X_2) + d(X_2, X_3))^p]^{1/p}$$

  $$\leq E[d(X_1, X_2)^p]^{1/p} + E[d(X_2, X_3)^p]^{1/p} = W_p(\mu_1, \mu_2) + W_p(\mu_2, \mu_3)$$

  per ottimalità e per la disuguaglianza Minkowski.
- Se $W_p(\mu, \nu) = 0$, allora il piano di trasporto ottimale $\pi$ è concentrato sulla diagonale. Pertanto, per ogni boreliano $H$:

  $$\mu(H) = \pi(H^2) = \nu(H) \text{ e dunque } \mu = \nu.$$

**Definizione 2.** *Dato* $(\mathcal{X}, d)$*, il rispettivo spazio di Wasserstein di ordine* $p$ *è, dato un* $x_0 \in \mathcal{X}$ *arbitrario, il seguente spazio:*

$$P_p(\mathcal{X}) := \Big\{\mu \in P(\mathcal{X}) : \int_{\mathcal{X}} d(x, x_0)^p d\mu < +\infty\Big\}.$$

**Osservazione 2.** *Si può dimostrare che lo spazio non dipende dalla scelta di* $x_0$*.*

Vediamo che su questo spazio $W_p$ è finita. Sia $\pi \in \Pi(\mu, \nu)$ con $\mu, \nu \in P_p(\mathcal{X})$. Siccome

$$d(x, y)^p \leq 2^{p-1}[d(x, x_0)^p + d(x_0, y)^p] \tag{1.2}$$

affinchè $d(x, y)^p$ sia $\pi-$integrabile è sufficiente che lo sia $[d(x, x_0)^p + d(x_0, y)^p]$. Per linearità del valore atteso, è sufficiente che $d(x, x_0)^p$ sia $\mu-$integrabile e $d(y, x_0)^p$ sia $\nu-$integrabile, e questo è garantito perché $\mu, \nu \in P_p(\mathcal{X})$.

Vediamo qualche proprietà basilare.

- Applicando la disuguaglianza di Hölder si ha che:

  $$1 \leq p \leq q \Rightarrow W_p \leq W_q.$$

  In particolare $W_1$ è la più piccola di tutte.

- $W_p(\delta_x, \delta_y) = d(x, y)$ : questo perchè $\Pi(\delta_x, \delta_y) = \Big\{\delta_{(x,y)}\Big\}$.

## 1.1 La formula duale di Kantorovich

Nel caso $p = 1$, $W_1$ è anche chiamata metrica di Kantorovich-Rubinstein, e ha un'interessante formula duale:

$$W_1(\mu, \nu) = \sup\Big\{\int_{\mathcal{X}} \varphi d\nu - \int_{\mathcal{X}} \varphi d\mu : ||\varphi||_{Lip} \leq 1\Big\}.$$

Per arrivare a questa formulazione di $W_1$ è necessario trasformare il problema di minimizzazione del costo in un problema di massimizzazione dei profitti.

Questo nuovo problema, equivalente al problema iniziale sotto opportune ipotesi, è chiamato problema duale di Kantorovich. Le incognite di questo problema sono delle "funzioni prezzo": date $\mu$ e $\nu$, non cerchiamo un piano di trasporto, ma una funzione che massimizzi, sotto alcuni vincoli, una quantità il cui calcolo non ha bisogno di considerare lo spazio prodotto.
Per definire il problema duale è utile dare un'idea intuitiva attraverso una situazione tipica in cui può essere usato. Potremmo pensare di essere un'azienda che si occupa di gestire trasporti di beni alimentari a Bologna. Siamo ingaggiati da fornai e bar per trasportare cornetti per la colazione. Indichiamo con $\mathcal{X}$ l'insieme dei fornai, con $\mathcal{Y}$ l'insieme dei bar, con la coppia ordinata $(x, y)$ il consorzio tra il fornaio $x$ e il bar $y$. Siccome i fornai hanno produzioni diverse, e i bar hanno richieste diverse, i nostri esperti modellano le produzioni con una misura di probabilità $\mu$ su $\mathcal{X}$, e le richieste con $\nu$ su $\mathcal{Y}$. Infine stimiamo una funzione costo $c$: $c(x, y)$ sarà il costo stimato per trasportare un'unità di cornetti per il consorzio $(x, y)$.
Dato questo modello, supponiamo di fare società con una compagnia esterna, che si offre di occuparsi solo del trasporto dei cornetti in quanto sono in grado di trasportare con un minor costo. La compagnia dovrà comprare i cornetti dai fornai per poi rivenderli ai bar, semplificando anche il metodo di pagamento di queste attività. Sia $\psi(x)$ il prezzo al quale compra un'unità di cornetti dal fornaio $x$, e sia $\phi(y)$ il prezzo al quale vende un'unità al bar $y$. Ciò che il consorzio $(x, y)$ pagherà ora per il trasporto di un'unità è $\phi(y) - \psi(x)$, e questo sarà anche parte del profitto della nostra società. Per essere più competitivi nel mercato, è necessario che ogni consorzio paghi meno di quanto facesse prima: la nostra società dovrà imporre i prezzi in modo tale che:

$$\phi(y) - \psi(x) \leq c(x, y), \ \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

A questo punto possiamo focalizzarci sul massimizzare i profitti: cerchiamo una coppia di prezzi competitiva soluzione del problema:

$$\sup \Big\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) : \ \ \phi(y) - \psi(x) \leq c(x, y) \Big\},$$

chiamato problema duale di Kantorovich, imponendo inoltre $\phi \in L^1(\mathcal{Y})$ e $\psi \in L^1(\mathcal{X})$. Il vincolo di competitività lega i due problemi dalla chiara disuguaglianza:

$$\sup_{\phi-\psi\leq c}\Big\{\int_{\mathcal{Y}}\phi(y)d\nu(y)-\int_{\mathcal{X}}\psi(x)d\mu(x)\Big\}\leq\inf_{\pi\in\Pi(\mu,\nu)}\Big\{\int_{\mathcal{X}\times\mathcal{Y}}c(x,y)d\pi(x,y)\Big\}.$$

Se riuscissimo a trovare una terna $\phi,\psi,\pi$ per cui ci sia l'uguaglianza, avremmo una coppia di prezzi ottimale ed un piano di trasporto ottimale. Tornando all'esempio, è nell'interesse della nostra società avere il maggior prezzo di vendita e il minor prezzo di acquisto possibili, senza perdere competitività. In tal senso, vorremmo che:

$$\phi(y)=\inf_{x\in\mathcal{X}}\Big(\psi(x)+c(x,y)\Big)\quad e\quad \psi(x)=\sup_{y\in\mathcal{Y}}\Big(\phi(y)-c(x,y)\Big). \tag{1.3}$$

In questo modo, è impossibile migliorare i prezzi senza perdere competitività, e si può dimostrare che una volta presi due prezzi competitivi, questi possono essere migliorati con dei prezzi come in (1.3).
Dunque possiamo costruire $\phi$ partendo da $\psi$, che rimane l'unica incognita del problema duale di Kantorovich. Ma non possiamo prendere una funzione qualunque. Infatti, avendo $\psi$, e calcolando $\phi$ come in (1.3), non è garantita la seconda uguaglianza in (1.3). Sembra logico allora dare le seguenti definizioni.

**Definizione 3.** *Siano $\mathcal{X}$ e $\mathcal{Y}$ due insiemi, e $c:\mathcal{X}\times\mathcal{Y}\to(-\infty,+\infty]$. Data una funzione $\psi:\mathcal{X}\to\mathbb{R}\cup\{+\infty\}$, la sua c - trasformata è la funzione definita come:*

$$\psi^c(y):=\inf_{x\in\mathcal{X}}\Big(\psi(x)+c(x,y)\Big),$$

*e il suo c - subdifferenziale è l'insieme dato da:*

$$\partial_c\psi:=\Big\{(x,y)\in\mathcal{X}\times\mathcal{Y}:\ \psi^c(y)-\psi(x)=c(x,y)\Big\}.$$

*Data una funzione $\phi:\mathcal{Y}\to\mathbb{R}\cup\{-\infty\}$, la sua c - trasformata è:*

$$\phi^c(x):=\sup_{y\in\mathcal{Y}}\Big(\phi(y)-c(x,y)\Big).$$

**Definizione 4.** *Siano* $\mathcal{X}$ *e* $\mathcal{Y}$ *due insiemi, e* $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$. *Una funzione* $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ *è detta* $c$ *- convessa se, non è identicamente* $+\infty$, *ed esiste* $f : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ *t.c.*

$$\psi(x) = \sup_{y \in \mathcal{Y}} \Big( f(y) - c(x,y) \Big) = f^c(x).$$

*D'altra parte, una* $\phi : \mathcal{Y} \to \mathbb{R} \cup \{-\infty\}$ *è detta* $c$ *- concava se, non è identicamente* $-\infty$, *ed esiste* $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ *per cui* $\phi = \psi^c$.

**Osservazione 3.** *Si può dimostrare che una funzione* $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ *è* $c$ *- convessa se e solo se* $(\psi^c)^c = \psi$.

Per una prova si veda la Proposizione 5.8 di [1]. Adesso abbiamo tutti gli strumenti per enunciare il seguente teorema, dal quale ricaveremo la formula duale di Kantorovich come caso particolare. Per la prova rimandiamo al Teorema 5.10 di [1].

**Teorema 2.** *Siano* $(\mathcal{X}, \mu)$ *e* $(\mathcal{Y}, \nu)$ *due spazi polacchi di probabilità e sia* $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ *una funzione costo inferiormente semicontinua t.c.:*

$$c(x,y) \geq a(x) + b(y),$$

*per qualche* $a \in L^1(\mu)$ *e* $b \in L^1(\nu)$ *a valori reali superiormente semicontinue ( è sufficiente* $c \geq 0$ *). Allora c'è dualità:*

$$\begin{aligned} &\min_{\pi \in \Pi(\mu,\nu)} \Big\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\pi(x,y) \Big\} \\ = &\sup_{\phi \in L^1(\nu), \psi \in L^1(\mu), \phi - \psi \leq c} \Big\{ \int_{\mathcal{Y}} \phi(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \Big\} \\ = &\sup_{\psi \in L^1(\mu), \psi\ c-convessa} \Big\{ \int_{\mathcal{Y}} \psi^c(y) d\nu(y) - \int_{\mathcal{X}} \psi(x) d\mu(x) \Big\}. \end{aligned} \tag{1.4}$$

*Se inoltre* $c$ *e il costo ottimale sono finiti, e vale la disuguaglianza:*

$$\int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\mu(x) d\nu(y) < +\infty \tag{1.5}$$

*allora, sia il problema iniziale sia quello duale hanno soluzione: nella formula (1.4) il sup in realtà è un max.*

Ricaviamo ora la formula duale di Kantorovich.

**Corollario 1.** *Sia $(\mathcal{X}, d)$ uno spazio polacco, siano $\mu, \nu \in P_1(\mathcal{X})$. Allora:*

$$W_1(\mu, \nu) = \max \Big\{ \int_{\mathcal{X}} \varphi d\nu - \int_{\mathcal{X}} \varphi d\mu : ||\varphi||_{Lip} \leq 1 \Big\}.$$

***DIM.*** Dato che $d$ è non negativa e continua, le ipotesi del teorema sono soddisfatte e c'è dunque dualità. Vediamo che una funzione $\varphi$ è $d$ - convessa se e solo se è Lipschitziana di costante minore o uguale di 1.
Ricordiamo prima due disuguaglianze utili.

- $\forall x_1, x_2, x_3,\ |d(x_1, x_2) - d(x_1, x_3)| \leq d(x_2, x_3)$, si ricava con la disuguaglianza triangolare.
- dal fatto che $\forall g, h,\ \sup(g + h) \leq \sup(g) + \sup(h)$ si può ricavare che

  $$\forall g, h, t,\ |\sup(g - t) - \sup(h - t)| \leq \sup |g - h| = ||g - h||_\infty,$$

  dalla quale, per t identicamente nulla, si ha:

  $$\forall g, h,\ |\sup(g) - \sup(h)| \leq ||g - h||_\infty.$$

Supponiamo $\varphi$ $d$ - convessa. Allora esiste $f : \mathcal{X} \to \mathbb{R}$ t.c.

$$\varphi(x) = \sup_{y \in \mathcal{X}} \Big( f(y) - d(x, y) \Big).$$

Siano $x_1, x_2 \in \mathcal{X}$,

$$|\varphi(x_1) - \varphi(x_2)| = |\sup_{y \in \mathcal{X}} \Big( f(y) - d(x_1, y) \Big) - \sup_{y \in \mathcal{X}} \Big( f(y) - d(x_2, y) \Big)|$$

$$\leq ||f(y) - d(x_1, y) - f(y) + d(x_2, y)||_\infty = \sup_{y \in \mathcal{X}} |d(x_2, y) - d(x_1, y)|$$

$$\leq \sup_{y \in \mathcal{X}} d(x_1, x_2) = d(x_1, x_2).$$

Sia ora $\varphi$ Lipschitziana, $||\varphi||_{Lip} \leq 1$. Allora $\varphi(x) \leq \varphi(y) + d(x,y), \forall x, y$. Dunque,

$$\varphi(x) = \inf_{y \in \mathcal{X}} \Big( \varphi(y) + d(x,y) \Big),$$

cioè $\varphi(x) = \varphi^d(x)$. Adesso, $\varphi = \varphi^d$, $\varphi^d = (\varphi^d)^d$, dunque $\varphi = (\varphi^d)^d$ e quindi $\varphi$ è $d$ - convessa (si veda l'Osservazione 3). Abbiamo anche dimostrato che, in tal caso, la funzione coincide con la sua trasformata. La dualità diventa:

$$W_1(\mu,\nu) = \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x,y) d\pi(x,y) = \sup \Big\{ \int_{\mathcal{X}} \varphi^d d\nu - \int_{\mathcal{X}} \varphi d\mu : \ \varphi \, \text{è} \, d - convessa \Big\}$$

$$= \sup \Big\{ \int_{\mathcal{X}} \varphi d\nu - \int_{\mathcal{X}} \varphi d\mu : \ ||\varphi||_{Lip} \leq 1 \Big\},$$

che è appunto la formula duale di Kantorovich per $W_1$.
Inoltre, avendo $\mu, \nu \in P_1(\mathcal{X})$, abbiamo dimostrato precedentemente che vale la disuguaglianza in (1.5), e quindi l'estremo superiore nella formula duale è un massimo. $\square$

# Capitolo 2

# La convergenza debole secondo Wasserstein

Richiamiamo innanzitutto che una successione $(\mu_k)_{k\in\mathbb{N}}$ in $P(\mathcal{X})$ converge debolmente ad una $\mu \in P(\mathcal{X})$ se, per ogni funzione $\varphi$ continua e limitata, si ha:

$$\int \varphi d\mu_k \to \int \varphi d\mu.$$

In tal caso scriveremo $\mu_k \xrightarrow{d} \mu$.

**Osservazione 4.** *Questo tipo di convergenza risulta più debole rispetto alla distanza $W_p$. Per essere più espliciti, $W_p$ è solo inferiormente semicontinua rispetto alla topologia indotta dalla convergenza debole. Cioè, se $\mu_k \xrightarrow{d} \mu$, e $\nu_k \xrightarrow{d} \nu$, allora:*

$$W_p(\mu, \nu) \leq \liminf_{k\to+\infty} W_p(\mu_k, \nu_k).$$

Per una prova si veda il capitolo 4 di [1]. Inoltre, come vedremo nel Teorema 3, la convergenza per $W_p$ implica quella debole. Pertanto, in $P_p(\mathcal{X})$, rafforziamo il concetto della convergenza debole con una condizione aggiuntiva, con l'obiettivo di rendere equivalenti le due nozioni.

**Definizione 5.** *Diremo che una successione $(\mu_k)_{k\in\mathbb{N}}$ in $P_p(\mathcal{X})$ converge ad una $\mu \in P_p(\mathcal{X})$ (e scriveremo $\mu_k \to \mu$) se è soddisfatta una delle seguenti condizioni equivalenti per un $x_0 \in \mathcal{X}$ (e poi per tutti):*

*1.* $\mu_k \xrightarrow{d} \mu$ *e*

$$\int d(x,x_0)^p d\mu_k \to \int d(x,x_0)^p d\mu.$$

*2.* $\mu_k \xrightarrow{d} \mu$ *e*

$$\limsup_{k\to+\infty} \int d(x,x_0)^p d\mu_k \leq \int d(x,x_0)^p d\mu.$$

*3.* $\mu_k \xrightarrow{d} \mu$ *e*

$$\lim_{R\to+\infty} \limsup_{k\to+\infty} \int_{d(x,x_0)\geq R} d(x,x_0)^p d\mu_k = 0.$$

*4.* $\forall \varphi \in C(\mathcal{X})$ *t.c.* $|\varphi(x)| \leq T(1+d(x,x_0)^p)$, *per una* $T \in \mathbb{R}$, *si ha:*

$$\int \varphi d\mu_k \to \int \varphi d\mu.$$

***Proviamo che queste richieste sono equivalenti.*** $1 \Rightarrow 2$, e $4 \Rightarrow 1$ sono banali.

$2 \Rightarrow 3$: per ogni $R$ esiste una funzione $f_R$ a supporto nel disco di centro $x_0$ e raggio $R$, continua, identicamente pari a 1 sul disco di raggio $\frac{R}{2}$, che poi tende a 0 in modo decrescente. Chiamiamo $D$ il disco di raggio $R$. Allora vale:

$$\int_{D^c} d(x,x_0)^p d\mu_k = \int_{\mathcal{X}} d(x,x_0)^p d\mu_k - \int_D d(x,x_0)^p d\mu_k \leq$$

$$\int_{\mathcal{X}} d(x,x_0)^p d\mu_k - \int_D f_R(x) d(x,x_0)^p d\mu_k = \int_{\mathcal{X}} d(x,x_0)^p d\mu_k - \int_{\mathcal{X}} f_R(x) d(x,x_0)^p d\mu_k.$$

Passando al $\limsup$ otteniamo:

$$\limsup_{k\to\infty} \int_{D^c} d(x,x_0)^p d\mu_k \leq \limsup_{k\to\infty} \int_{\mathcal{X}} d(x,x_0)^p d\mu_k - \limsup_{k\to\infty} \int_{\mathcal{X}} f_R(x) d(x,x_0)^p d\mu_k \leq$$

$$\int_{\mathcal{X}} d(x,x_0)^p d\mu - \int_{\mathcal{X}} f_R(x) d(x,x_0)^p d\mu.$$

Dato che $f_R$, per $R \to +\infty$, converge alla funzione costante pari a 1, per convergenza dominata abbiamo:

$$\lim_{R\to+\infty} \limsup_{k\to+\infty} \int_{D^c} d(x,x_0)^p d\mu_k \leq \int_{\mathcal{X}} d(x,x_0)^p d\mu - \int_{\mathcal{X}} d(x,x_0)^p d\mu = 0,$$

e dunque si ha la tesi.

$3 \Rightarrow 4$: notiamo innanzitutto che è sufficiente dimostrarlo per $\varphi \geq 0$. Sia come prima $D$ il disco di centro $x_0$ e raggio $R$. Allora, per ogni $R$,

$$\limsup_{k\to+\infty} |\int_{\mathcal{X}} \varphi d\mu_k - \int_{\mathcal{X}} \varphi d\mu| \leq$$

$$\limsup_{k\to+\infty} |\int_{D} \varphi d\mu_k - \int_{D} \varphi d\mu| + \limsup_{k\to+\infty} |\int_{D^c} \varphi d\mu_k - \int_{D^c} \varphi d\mu|.$$

Analizziamo il secondo addendo: questo è minore o uguale di:

$$\limsup_{k\to+\infty} \int_{D^c} \varphi d\mu_k + \int_{D^c} \varphi d\mu \leq$$

$$\limsup_{k\to+\infty} \int_{D^c} (T + Td(x,x_0)^p) d\mu_k + \int_{D^c} \varphi d\mu \leq$$

$$T \limsup_{k\to+\infty} \int_{D^c} d(x,x_0)^p d\mu_k + \int_{D^c} \varphi d\mu + T \limsup_{k\to+\infty} \mu_k(D^c),$$

che, per $R \to +\infty$, tende a zero (per $\mu_k(D^c)$, si noti che per $R > 1$ è maggiorabile con l'integrale su $D^c$ di $d(x,x_0)^p$). Dunque:

$$\limsup_{k\to+\infty} |\int_{\mathcal{X}} \varphi d\mu_k - \int_{\mathcal{X}} \varphi d\mu| \leq \lim_{R\to+\infty} \limsup_{k\to+\infty} |\int_{D} \varphi d\mu_k - \int_{D} \varphi d\mu|.$$

Fissato $R$, possiamo penalizzare $\varphi$ con una funzione $f$ che coincide con $\varphi$ su $D$. Pertanto,

$$\limsup_{k\to+\infty} |\int_{D} \varphi d\mu_k - \int_{D} \varphi d\mu| \leq \limsup_{k\to+\infty} |\int_{D} f d\mu_k - \int_{D} f d\mu| \leq$$

$$\limsup_{k\to+\infty} |\int_{\mathcal{X}} f d\mu_k - \int_{\mathcal{X}} f d\mu| + \limsup_{k\to+\infty} |\int_{D^c} f d\mu_k - \int_{D^c} f d\mu|.$$

Il primo addendo è uguale a zero per la convergenza debole. Dunque c'è da considerare come si comporta il secondo per $R$ che va all'infinito. Questo è minore o uguale a:

$$\limsup_{k\to+\infty} \int_{D^c} f d\mu_k + \int_{D^c} f d\mu \leq$$

$$T \limsup_{k\to+\infty} \int_{D^c} d(x,x_0)^p d\mu_k + \int_{D^c} f d\mu + T \limsup_{k\to+\infty} \mu_k(D^c),$$

che, per $R \to +\infty$, tende a 0. $\square$

Come speravamo, la tipologia di convergenza appena definita è equivalente alla convergenza per $W_p$.

**Teorema 3.** *Siano date una successione* $(\mu_k)_{k\in\mathbb{N}}$ *in* $P_p(\mathcal{X})$ *e* $\mu \in P_p(\mathcal{X})$. $\mu_k \to \mu$ *se e solo se* $W_p(\mu_k, \mu) \to 0$.

Prima di procedere con la dimostrazione, notiamo subito due importanti conseguenze e facciamo delle considerazioni.

- $W_p$ è continua su $P_p(\mathcal{X})$:
  se $\mu_k \to \mu$ e $\nu_k \to \nu$, allora $W_p(\mu_k, \nu_k) \to W_p(\mu, \nu)$.

- Se la distanza d è limitata, allora la convergenza secondo Wasserstein è equivalente all'usuale convergenza debole.

In più la convergenza debole può essere metrizzata in vari modi, ad esempio con la "**bounded Lipschitz distance**":

$$d_{bL}(\mu, \nu) := \sup\Big\{ \int \varphi d\mu - \int \varphi d\nu,\ \ ||\varphi||_\infty + ||\varphi||_{Lip} \leq 1 \Big\}.$$

E allora cosa rende conveniente l'utilizzo di $W_p$?

- Essendo un minimo è facilmente maggiorabile. Infatti, date due misure $\mu, \nu$, la costruzione di un qualsiasi piano di trasporto pone un limite dall'alto per $W_1(\mu, \nu)$.

- Per sua natura è molto utile nei problemi di trasporto ottimo.

- La formula duale di Kantorovich per $W_1$ è uno strumento molto utile: passare dalla formula originale a quella duale è spesso conveniente a livello tecnico.

- Incorpora molto della geometria dello spazio. Per esempio, come visto nel Capitolo 1, la mappa $x \mapsto \delta_x$ è un'immersione isometrica di $\mathcal{X}$ in $P_p(\mathcal{X})$, ma ci sono legami più profondi. Quest'esempio spiega a grandi linee il motivo per cui $(P_p(\mathcal{X}), W_p)$ si adatta bene a problemi dove geometria e convergenza debole sono coinvolti.

Per dimostrare il teorema abbiamo bisogno di alcuni strumenti.

**Definizione 6.** *Un insieme di misure di probabilità $M$ su $\mathcal{X}$ si dice tight se:*

$$\forall \varepsilon > 0 \, \exists K_\varepsilon \subset \mathcal{X} \, compatto : \, \forall \mu \in M si \, abbia \, che \, \mu(\mathcal{X} \backslash K_\varepsilon) < \varepsilon.$$

Se una successione $(\mu_k)_{k \in \mathbb{N}}$ è tight, per il teorema di Helly esistono $(\mu_{n_k})_{k \in \mathbb{N}}$ e $\mu \in P(\mathcal{X})$ t.c. $\mu_{n_k} \xrightarrow{d} \mu$. Per una dimostrazione si veda [2], dove è fatta per le distribuzioni, ma facilmente estendibile agli spazi polacchi.

**Lemma 2.** *Siano $\mathcal{X}$ e $\mathcal{Y}$ spazi polacchi, e siano $P$ e $Q$ due insiemi di misure di probabilità tight rispettivamente di $\mathcal{X}$ e $\mathcal{Y}$. Allora, l'insieme di tutti i piani di trasporto le cui marginali sono in $P$ e in $Q$ è tight nel prodotto.*

Per una prova si veda il lemma 4.4 di [1].

**Teorema 4.** *Siano $\mathcal{X}$ e $\mathcal{Y}$ spazi polacchi e sia $c = c(x, y)$ una funzione costo continua e limitata inferiormente. Date $\mu \in P(\mathcal{X})$ e $\nu \in P(\mathcal{Y})$, siano $(c_k)_{k \in \mathbb{N}}$ una successione di funzioni costo, $(\mu_k)_{k \in \mathbb{N}}, (\nu_k)_{k \in \mathbb{N}}$ successioni rispettivamente in $P(\mathcal{X})$ e $P(\mathcal{Y})$, e $(\pi_k)_{k \in \mathbb{N}}$ piani di trasporto ottimali rispetto a $c_k$ tra $\mu_k$ e $\nu_k$. Se:*

$$c_k \to c \; uniformemente,$$

$$\mu_k \xrightarrow{d} \mu \in P(\mathcal{X}),$$

$$\nu_k \xrightarrow{d} \nu \in P(\mathcal{Y}),$$

$$e \int c_k(x, y) d\pi_k(x, y) < +\infty, \;\; \forall k \in \mathbb{N},$$

*allora $\exists (\pi_{n_k})_{k \in \mathbb{N}}$ t.c. $\pi_{n_k} \xrightarrow{d} \pi \in \Pi(\mu, \nu)$. Quest'ultimo è ottimale rispetto a $c$ se*

$$\liminf_{k \to +\infty} \int c_k d\pi_k < +\infty.$$

Per una prova si veda il Teorema 5.20 di [1].

**Teorema 5** (di Prokhorov)**.** *Se $\mathcal{X}$ è uno spazio polacco, allora un insieme di misure di probabilità $M$ è tight se e solo se $M$ è precompatto nella topologia della convergenza debole. In particolare ogni insieme finito è tight.*

**Lemma 3.** *Sia $(a_n)_{n\in\mathbb{N}}$ una successione in uno spazio topologico $\mathcal{T}$. Se ogni sottosuccessione $(a_{n_k})_{k\in\mathbb{N}}$ ammette una sottosuccessione $(a_{n_{k_j}})_{j\in\mathbb{N}}$ che converge al medesimo $a \in \mathcal{T}$, allora anche $(a_n)_{n\in\mathbb{N}}$ converge ad $a$.*

***DIM.*** Per assurdo, esiste un intorno $U$ di $a$ t.c. $(a_n)_{n\in\mathbb{N}}$ non è definitivamente in $U$. Dunque si può creare una sottosuccessione $(a_{n_k})_{k\in\mathbb{N}}$ di elementi di $U^c$, e nessuna sottosuccessione $(a_{n_{k_j}})_{j\in\mathbb{N}}$ può convergere ad $a$, contraddicendo l'ipotesi. □

**Lemma 4.** *Se $(\mu_k)_{k\in\mathbb{N}}$ è una successione di Cauchy in $(P_p(\mathcal{X}), W_p)$ allora è tight.*

***DIM.*** Sia $(\mu_k)_{k\in\mathbb{N}}$ di Cauchy in $(P_p(\mathcal{X}), W_p)$. Siccome $W_p \geq W_1$, $(\mu_k)_{k\in\mathbb{N}}$ è di Cauchy anche per $W_1$. Sia $\varepsilon > 0$ e sia $N$ tale che per $k \geq N$ si abbia $W_1(\mu_N, \mu_k) < \varepsilon^2$. Allora per ogni $k$ esiste $j \in \{1, ..., N\}$ t.c. $W_1(\mu_j, \mu_k) < \varepsilon^2$ (per $k < N$ si prenda $j = k$). Siccome l'insieme $\{\mu_1, \cdots, \mu_N\}$ è tight, esiste un compatto $K$ per cui $\mu_j(\mathcal{X}\backslash K) < \varepsilon \; \forall j \in \{1, ..., N\}$. $K$ può essere coperto da un numero finito di palle $B(x_1, \varepsilon), ..., B(x_m, \varepsilon)$. Sia $U$ l'unione delle palle, e siano:

$$U_\varepsilon := \{x \in \mathcal{X} : d(x, U) < \varepsilon\} \subset B(x_1, 2\varepsilon) \cup B(x_m, 2\varepsilon);$$

$$\phi(x) := \Big(1 - \frac{d(x, U)}{\varepsilon}\Big)_+.$$

Notiamo che $\mathbf{1}_U \leq \phi \leq \mathbf{1}_{U_\varepsilon}$ e $\phi$ é Lipschitziana di costante $1/\varepsilon$. Usando anche la formula duale di Kantorovich vediamo che per ogni $k$, e per $j \leq N$ si ha:

$$\mu_k(U_\varepsilon) \geq \int \phi d\mu_k = \int \phi d\mu_j + \Big(\int \phi d\mu_k - \int \phi d\mu_j\Big) \geq \mu_j(U) - \frac{W_1(\mu_k, \mu_j)}{\varepsilon}.$$

Dato che $\mu_j(U) \geq \mu_j(K) \geq 1 - \varepsilon$ e esiste $j$ per cui $W_1(\mu_j, \mu_k) < \varepsilon^2$, si ha:

$$\mu_k(U_\varepsilon) \geq 1 - \varepsilon - \varepsilon = 1 - 2\varepsilon.$$

In sostanza abbiamo dimostrato che per ogni $\varepsilon > 0$ esistono finiti $x_i$, con $1 \leq i \leq m$ tale che ogni $\mu_k$ assegni a $Z = \bigcup B(x_i, 2\varepsilon)$ probabilità almeno

$1 - 2\varepsilon$. Il problema è che $Z$ potrebbe non essere compatto. Un rimedio potrebbe essere rimpiazzare $\varepsilon$ con $2^{-(\ell+1)}\varepsilon$ con $\ell \in \mathbb{N}$, quindi ci saranno $x_i$ , $1 \leq i \leq m(\ell)$ , t.c.

$$\mu_k(\mathcal{X} \backslash \bigcup B(x_i, 2^{-\ell}\varepsilon)) \leq 2^{-\ell}\varepsilon,$$

e definendo

$$S := \bigcap_{1 \leq p \leq +\infty} \bigcup_{1 \leq i \leq m(p)} D(x_i, \varepsilon 2^{-p})$$

si ha $\mu_k(\mathcal{X} \backslash S) < \varepsilon$. Inoltre S è compatto perchè chiuso e totalmente limitato per costruzione.

$\square$

***Dimostrazione del Teorema 3.*** Sia $(\mu_n)_{n \in \mathbb{N}}$ una successione convergente a $\mu$ per $W_p$. Allora è di Cauchy per $W_p$, e per il Lemma 4, la successione è tight. Dunque ammette una sottosuccessione $(\mu_{n_k})_{k \in \mathbb{N}}$ convergente debolmente ad una $\tilde{\mu} \in P(\mathcal{X})$. Dato che $W_p$ è inferiormente semicontinua rispetto alla convergenza debole (si veda l'Osservazione 4):

$$W_p(\tilde{\mu}, \mu) \leq \liminf_{k \to +\infty} W_p(\mu_{n_k}, \mu) = 0,$$

allora $\tilde{\mu} = \mu$, e per il Lemma 3 è necessario che $\mu_n \xrightarrow{d} \mu$. Vediamo se è soddisfatta una delle condizioni aggiuntive.
Prima di tutto ricordiamo una disuguaglianza sempre valida per una coppia $(a, b)$ di numeri reali positivi: per ogni $\varepsilon > 0$ esiste una costante $C_\varepsilon$ t.c.

$$(a + b)^p \leq (1 + \varepsilon)a^p + C_\varepsilon b^p.$$

Combinandola con la disuguaglianza triangolare, si ha che per ogni $x_0, x, y \in \mathcal{X}$ vale:

$$d(x_0, x)^p \leq (1 + \varepsilon)d(x_0, y)^p + C_\varepsilon d(x, y)^p.$$

Per ogni $n$, sia $\pi_n$ un piano di trasporto ottimale tra $\mu_n$ e $\mu$, e integriamo per $\pi_n$ la disuguaglianza ottenuta. Usando la proprietà delle marginali:

$$\int d(x_0, x)^p d\mu_n \leq (1 + \varepsilon) \int d(x_0, y)^p d\mu + C_\varepsilon \int d(x, y)^p d\pi_n.$$

Ma
$$\int d(x,y)^p d\pi_n = W_p(\mu_n,\mu)^p \to 0,$$
e dunque:
$$\limsup_{n\to+\infty} \int d(x_0,x)^p d\mu_n \leq (1+\varepsilon) \int d(x_0,y)^p d\mu.$$
E siccome vale per ogni $\varepsilon$ possiamo mandare $\varepsilon$ a zero ed ottenere:
$$\limsup_{n\to+\infty} \int d(x_0,x)^p d\mu_n \leq \int d(x,x_0)^p d\mu.$$
Che è una delle condizioni aggiuntive per la convergenza in $P_p(\mathcal{X})$.

Viceversa, sia $(\mu_n)_{n\in\mathbb{N}}$ convergente a $\mu$ in $P_p(\mathcal{X})$, e introduciamo $(\pi_n)_{n\in\mathbb{N}}$ piani di trasporto ottimali tra $\mu_n$ e $\mu$. Dal teorema di Prokhorov, $(\mu_n)_{n\in\mathbb{N}}$ è tight, anche $\{\mu\}$ è tight: per il Lemma 2 anche $(\pi_n)$ è tight, e a meno di prendere una sottosuccessione, si ha $\pi_n \xrightarrow{d} \pi$, dove $\pi \in \Pi(\mu,\mu)$. Il Teorema 4 ci assicura che $\pi$ è ottimale per $\mu$ e $\mu$ (si prendano $c_n = d^p$ per ogni $n$, il resto segue per ottimalità di $\pi_n$), e dunque è concentrato sulla diagonale e determinato solo da $\mu$. Pertanto, il Lemma 3 e il fatto che ogni sottosuccessione sia tight implicano che $\pi_n \xrightarrow{d} \pi$.
Siano ora $x_0 \in \mathcal{X}$ e $R > 0$. Notiamo che:
$$\mathbf{1}_{d(x,y)\geq R} \leq \mathbf{1}_{[d(x,x_0)\geq R/2 \, e \, d(x,x_0)\geq d(x,y)/2]} + \mathbf{1}_{[d(x_0,y)\geq R/2 \, e \, d(x_0,y)\geq d(x,y)/2]}.$$
Risulta:
$$[d(x,y)^p - R^p]_+ \leq d(x,y)^p \mathbf{1}_{[d(x,x_0)\geq R/2 \, e \, d(x,x_0)\geq d(x,y)/2]} +$$
$$d(x,y)^p \mathbf{1}_{[d(x_0,y)\geq R/2 \, e \, d(x_0,y)\geq d(x,y)/2]} \leq$$
$$2^p d(x,x_0)^p \mathbf{1}_{d(x,x_0)\geq R/2} + 2^p d(x_0,y)^p \mathbf{1}_{d(x_0,y)\geq R/2}.$$
Dunque:
$$W_p(\mu_n,\mu)^p = \int d(x,y)^p d\pi_n = \int [\min(d(x,y),R)]^p d\pi_n + \int [d(x,y)^p - R^p]_+ d\pi_n$$
$$\leq \int [\min(d(x,y),R)]^p d\pi_n + 2^p \int_{d(x,x_0)\geq R/2} d(x,x_0)^p d\pi_n + 2^p \int_{d(x_0,y)\geq R/2} d(x_0,y)^p d\pi_n$$

$$= \int [\min(d(x,y),R)]^p d\pi_n + 2^p \int_{d(x,x_0)\geq R/2} d(x,x_0)^p d\mu_n + 2^p \int_{d(x_0,y)\geq R/2} d(x_0,y)^p d\mu.$$

Siccome $\pi_n \xrightarrow{d} \pi$, il primo termine tende a zero, mentre:

$$\limsup_{n\to+\infty} W_p(\mu_n,\mu)^p \leq 2^p \Big( \lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{d(x,x_0)\geq R/2} d(x,x_0)^p d\mu_n$$

$$+ \lim_{R\to+\infty} \limsup_{n\to+\infty} \int_{d(x_0,y)\geq R/2} d(x_0,y)^p d\mu \Big) = 0.$$

È uguale a zero perchè $\mu_n \to \mu$ (la condizione numero 3). $\square$

# Capitolo 3

# Proprietà topologiche dello spazio di Wasserstein

Per dimostrare che lo spazio di Wasserstein è uno spazio polacco, è conveniente utilizzare il concetto di variazione assoluta di una misura con segno, che ha un ruolo molto importante nella teoria della misura. Qui porremo l'attenzione sul fatto che è un modo di confrontare le misure di probabilità, e troveremo una relazione con la metrica di Wasserstein che risulterà di grande aiuto nel provare la separabilità di $P_p(\mathcal{X})$.

Una misura con segno è una funzione d'insiemi con le stesse proprietà delle misure, ma cade la proprietà della non negatività. Per il Teorema di decomposizione di Hahn (per una prova si veda il Teorema 6.14 di [3]), per ogni misura con segno $\mu$ su uno spazio misurabile $\mathcal{X}$ esistono due insiemi misurabili $P$ ed $N$ tali che:

- $P \cup N = X$ e $P \cap N = \emptyset$.
- $\mu(E) > 0 \; \forall E \subset P$, $E$ misurabile.
- $\mu(E) \leq 0 \; \forall E \subset N$, $E$ misurabile.

$P$ verrà chiamato insieme positivo, mentre $N$ insieme negativo. Definendo $\mu_+(E) := \mu(P \cap E)$ e $\mu_-(E) := -\mu(N \cap E)$, con $E$ misurabile, queste risultano essere delle misure non negative, e fanno in modo che $\mu = \mu_+ - \mu_-$, mentre la misura $|\mu| := \mu_+ + \mu_-$ é chiamata variazione assoluta di $\mu$, e il suo valore massimo $||\mu|| := |\mu|(X)$ é detta variazione totale di $\mu$. Date invece due misure $\mu$ e $\nu$, queste si possono confrontare con la variazione totale: $\mu - \nu$ è una misura con segno, e dunque $||\mu - \nu||$ è ben definita, ed è anche una distanza nello spazio delle misure. Solo la disuguaglianza triangolare non è banale: se $a$ e $b$ sono misure, $P_a$, $N_a$ sono gli insiemi positivi e negativi per $a$, $P$ ed $N$ quelli per $a + b$, per ogni boreliano $H$:

$$|a + b|(H) = (a + b)(H \cap P) - (a + b)(H \cap N) =$$

$$(a(H \cap P) - a(H \cap N)) + (b(H \cap P) - b(H \cap N)),$$

ma influiscono le parti positive e negative di $a$. Su $H \cap P$, :

$$a(H \cap P) = a(H \cap P \cap P_a) + a(H \cap P \cap N_a) \leq a(H \cap P \cap P_a) \leq a(H \cap P_a),$$

mentre su $H \cap N$:

$$a(H \cap N) = a(H \cap N \cap P_a) + a(H \cap N \cap N_a) \geq a(H \cap N \cap N_a) \geq a(H \cap N_a).$$

Dunque $a(H \cap P) - a(H \cap N) \leq |a|(H)$, lo stesso vale per $b$. Allora,

$$|a + b|(H) \leq |a|(H) + |b|(H),$$

e questo implica la disuguaglianza triangolare per la variazione totale. Dunque, comunque si prenda una terna $\mu, \nu, \tau$ di misure:

$$|\mu - \nu| \leq |\mu - \tau| + |\tau - \nu|. \tag{3.1}$$

Quindi la metrica di Wasserstein e la variazione sono due modi diversi di confrontare le misure di probabilità, ed essendo $W_p$ definita come un minimo, ci si potrebbe aspettare un limite dall'alto. Infatti $W_p$ può essere controllata dalla variazione assoluta pesata, nel senso specificato dal seguente teorema.

**Teorema 6.** *Dato $(\mathcal{X}, d)$ uno spazio polacco, date $\mu, \nu \in P(\mathcal{X})$, dati un $x_0 \in \mathcal{X}$ e $p, q > 1$ t.c. $\frac{1}{p} + \frac{1}{q} = 1$, vale la disuguaglianza:*

$$W_p(\mu, \nu) \leq 2^{\frac{1}{q}} \Big( \int d(x, x_0)^p d|\mu - \nu| \Big)^{1/p},$$

*e la quantità a destra è anch'essa una distanza, in particolare la disuguaglianza triangolare segue da* (3.1)

**Osservazione 5.** *Nel caso in cui* $p = 1$*, la disuguaglianza diventa:*

$$W_1(\mu, \nu) \leq \int d(x, x_0) d|\mu - \nu|,$$

*e se il diametro di* $\mathcal{X}$ *è limitato da* $D \in \mathbb{R}$*, allora:*

$$W_1(\mu, \nu) \leq D||\mu - \nu||.$$

***DIM.*** Sia $a = (\mu - \nu)_+(\mathcal{X}) = (\mu - \nu)_-(\mathcal{X})$, e sia $\pi$ la misura su $\mathcal{X} \times \mathcal{X}$ definita come:

$$\pi := \gamma + \frac{1}{a}(\mu - \nu)_+ \otimes (\mu - \nu)_-,$$

dove $\gamma$ è la misura che ha come marginali $(\mu - (\mu - \nu)_+, \mu - (\mu - \nu)_+)$ ed è concentrata sulla diagonale. Notiamo che $\pi$ è un piano di trasporto tra $\mu$ e $\nu$:

$$\pi(\mathcal{X} \times \mathcal{X}) = \gamma(\mathcal{X} \times \mathcal{X}) + \frac{1}{a}a^2 = \mu(\mathcal{X}) - (\mu - \nu)_+(\mathcal{X}) + a = 1.$$

Sia $H \subset \mathcal{X}$ un boreliano:

$$\pi(H \times \mathcal{X}) = \mu(H) - (\mu - \nu)_+(H) + \frac{1}{a}(\mu - \nu)_+(H)a = \mu(H).$$

$$\pi(\mathcal{X} \times H) = \mu(H) - (\mu - \nu)_+(H) + (\mu - \nu)_-(H) = \mu(H) - (\mu - \nu)(H) = \nu(H),$$

quindi $\pi$ ha $\mu$ e $\nu$ come marginali. Ora, usando la definizione di $W_p$, quella di $a$, la disuguaglianza in (1.2), risulta:

$$W_p(\mu, \nu)^p \leq \int d(x, y)^p d\pi = \frac{1}{a} \int d(x, y)^p d(\mu - \nu)_+(x) d(\mu - \nu)_-(y)$$

$$\leq \frac{2^{p-1}}{a} \int [d(x, x_0)^p + d(x_0, y)^p] d(\mu - \nu)_+(x) d(\mu - \nu)_-(y)$$

$$\leq 2^{p-1} \int d(x, x_0)^p d[(\mu - \nu)_+ + (\mu - \nu)_-](x) = 2^{p-1} \int d(x, x_0)^p d|\mu - \nu|(x).$$

Dunque, siccome $\frac{1}{q} = \frac{p-1}{p}$, abbiamo:

$$W_p(\mu, \nu) \leq 2^{\frac{1}{q}} \Big( \int d(x, x_0)^p d|\mu - \nu|(x) \Big)^{1/p}.$$

$\square$

Adesso abbiamo tutti gli strumenti per dimostrare che lo spazio di Wasserstein eredita la struttura di spazio polacco. Il fatto che sia uno spazio metrico è già stato mostrato, quindi rimane da provare la completezza e la separabilità. In particolare vedremo che il suo sottoinsieme denso e numerabile è formato da combinazioni finite di delta di Dirac a coefficienti razionali.

**Teorema 7.** *Siano $(\mathcal{X}, d)$ uno spazio polacco e $p \in [1, +\infty)$, allora lo spazio $(P_p(\mathcal{X}), W_p)$ è anch'esso uno spazio polacco, ed ogni misura di probabilità può essere approssimata da una successione di misure di probabilità a supporto finito.*

***DIM.*** Proviamo prima la separabilità. Sia $D$ un sottoinsieme denso numerabile di $\mathcal{X}$ e sia $Q$ l'insieme delle misure di probabilità del tipo:

$$\sum a_j \delta_{x_j},$$

dove gli $a_j$ sono razionali e gli $x_j$ sono elementi di $D$ in numero finito. Mostriamo che $Q$ è denso in $P_p(\mathcal{X})$.
Siano $\varepsilon > 0$ e $x_0$ un elemento arbitrario di $D$. Se $\mu \in P_p(\mathcal{X})$, allora poniamo:

$$\ell := \int_{\mathcal{X}} d(x, x_0)^p d\mu < +\infty.$$

Allora, la funzione sui boreliani data da:

$$\gamma(H) := \frac{1}{\ell} \int_H d(x, x_0)^p d\mu$$

è una misura di probabilità, e per il Teorema di Prokhorov $\{\gamma\}$ è tight. Dunque esiste $K \subset \mathcal{X}$ compatto per cui:

$$\gamma(\mathcal{X} \backslash K) \leq \frac{\varepsilon^p}{\ell},$$

e quindi:

$$\int_{\mathcal{X} \backslash K} d(x, x_0)^p d\mu \leq \varepsilon^p.$$

Se $\ell = 0$ segue lo stesso risultato. Ora, ricopriamo $K$ con una famiglia finita di palle metriche: $B(x_k, \varepsilon/2)$, con $1 \leq k \leq N$, $x_k \in D$. Definiamo:

$$T_k := B(x_k, \varepsilon) \backslash \bigcup_{j<k} B(x_j, \varepsilon).$$

Questi sono insiemi disgiunti che ricoprono $K$. Definiamo $f$ su $\mathcal{X}$ t.c.

$$f(T_k \cap K) := \{x_k\} \quad f(\mathcal{X} \backslash K) := \{x_0\}.$$

Allora, per ogni $x \in K$, $d(x, f(x)) \leq \varepsilon$, e dunque:

$$\int d(x, f(x))^p d\mu \leq \varepsilon^p \int_K d\mu + \int_{\mathcal{X}\backslash K} d(x, x_0)^p d\mu \leq 2\varepsilon^p.$$

Se chiamiamo $\mu_f$ la legge di $f$, allora abbiamo dimostrato che $W_p(\mu, \mu_f) \leq 2\varepsilon^p$, e ovviamente $\mu_f$ può essere scritta come $\sum a_j \delta_{x_j}$, $0 \leq j \leq N$. Quindi $\mu$ può essere approssimata da una combinazione di delta di Dirac, e vediamo che i coefficienti possono essere sempre rimpiazzati da coefficienti razionali per un piccolo errore: dal Teorema 6 si ha

$$W_p\Big(\sum_{j \leq N} a_j \delta_{x_j}, \sum_{j \leq N} b_j \delta_{x_j}\Big) \leq 2^{1/q} [\max_{k,l} d(x_k, x_l)] \sum_{j \leq N} |a_j - b_j|^{1/p},$$

e questo errore può essere ridotto il più possibile con una buona scelta dei coefficienti razionali $b_j$.

Proviamo ora la completezza. Sia $(\mu_n)_{n \in \mathbb{N}}$ di Cauchy in $(P_p(\mathcal{X}), W_p)$. Per il Lemma 4, la successione è tight, e ammette dunque una sottosuccessione $(\mu_{n_k})_{k \in \mathbb{N}}$ che converge debolmente ad una $\mu \in P(\mathcal{X})$. Ma, siccome $W_p$ è inferiormente semicontinua rispetto alla convergenza debole:

$$W_p(\delta_{x_0}, \mu)^p = \int d(x, x_0)^p d\mu \leq \liminf_{k \to +\infty} W_p(\delta_{x_0}, \mu_{n_k})^p = \liminf_{k \to +\infty} \int d(x, x_0)^p d\mu_{n_k} < +\infty,$$

quindi $\mu \in P_p(\mathcal{X})$. Data $(\mu_j)_{j \in \mathbb{N}}$ un'altra sottosuccessione,

$$W_p(\mu, \mu_j) \leq \liminf_{k \to +\infty} W_p(\mu_{n_k}, \mu_j) \ \ \forall j \in \mathbb{N},$$

e in particolare:

$$\limsup_{j \to +\infty} W_p(\mu, \mu_j) \leq \limsup_{k,j \to +\infty} W_p(\mu_{n_k}, \mu_j) = 0.$$

Questo implica $W_p(\mu, \mu_j) \to 0$. Dato che $(\mu_n)_{n\in\mathbb{N}}$ è di Cauchy e ha una sottosuccessione convergente $(\mu_j)_{j\in\mathbb{N}}$, l'intera successione deve convergere a $\mu$. $\square$

# Bibliografia

[1] Cédric Villani. Optimal Transport, old and new, *Springer,* 2009

[2] Andrea Pascucci. Teoria della probabilità, *Springer,* 2020

[3] Walter Rudin. Real and complex analysis. *McGraw-Hill*, terza edizione, 1987