4 Una metrica per le misure di probabilità

L'equazione di Vlasov è di fatto un'equazione di Liouville non lineare, e per fare la teoria di esistenza e unicità servono delle metriche che tengono in conto del fatto che l'equazione di Liouville è l'equazione del trasporto delle misure.
Le usuali norme di funzioni non sono molto adatte; consideriamo per esempio A e B domini disgiunti di $\mathbb{R}^d$, di misura 1, e consideriamo la distanza in L_1 tra $\mathcal{X}_A$ e $\mathcal{X}_B$. Questa distanza è esattamente 2, indipendentemente da quanto siano lontani in $\mathbb{R}^d$ i due insiemi. Più in generale, $\|f-g\|_1$ misura la differenza dei valori tra f e g in tutti i possibili punti di $\mathbb{R}^d$, e poi li somma. Le distanze L_p operano nello stesso modo.
Abbiamo bisogno di una nozione di distanza che tenga conto anche della distanza "geometrica": definirò la distanza $1-$lipschitziana tra misure di probabilità, accennando anche al fatto che coincide con la distanza $1-$Wassertein, anche detta distanza di Kantorovich-Rubistein.

4.1 La metrica $1-$lipschitziana

Comincio con l'introdurre l'analogo della norma L_1 per le misure finite con segno[1]. Data una misura con segno μ, essa si decompone in $\mu = \mu^+ - \mu^-$ dove $\mu^\pm$ sono due misure positive a supporto disgiunto (per il Teorema di decomposizione di Hahn-Jordan). La norma di μ in variazione totale è

$$\|\mu\|_{VT} = \int \mu^+ + \int \mu^-.$$

Si può provare che questa è esattamente la norma nello spazio delle misure con segno considerato come duale delle funzioni continue e limitate:

$$\|\mu\|_{VT} = \sup_{\phi\in C_b,\, \|\phi\|_\infty\le 1} \left| \int \phi \,\mathrm{d}\mu \right|.$$

È facile osservare che si può rimuovere il modulo da questa definzione, infatti se $\psi = \phi\,\mathrm{sgn}(\int \phi\,\mathrm{d}\mu)$, si ha $\int \psi\,\mathrm{d}\mu = \left|\int \phi\,\mathrm{d}\mu\right|$. Ne segue che la **distanza in variazione totale** tra due misure di probabilità μ e ν si può ottenere come

$$\|\mu-\nu\|_{VT} = \sup_{\phi\in C_b,\, \|\phi\|_\infty\le 1} \int \phi\,\mathrm{d}\mu - \int \phi\,\mathrm{d}\nu.$$

Cambiando l'insieme delle funzioni ammissibili ϕ, si ottengono altre distanze, in particolare quella che stiamo cercando.
D'ora un poi mi limito a considerare misure di probabilità definite su un compatto K di $\mathbb{R}^d$, e indicherò l'insieme di queste misure con $\mathcal{P}(K)$.

Definizione di distanza $1-$lipschitziana
Indico con $\mathrm{Lip}_1 = \mathrm{Lip}_1(K)$ l'insieme delle funzioni lipschitziane da K in $\mathbb{R}$ con costante di Lipschitz pari a 1:

$$\mathrm{Lip}_1 = \{\phi \in C(K)| : \forall \mathbf{x},\mathbf{y}, |\phi(\mathbf{x})-\phi(\mathbf{y})| \le |\mathbf{x}-\mathbf{y}|\}$$

[1]Per essere rigorosi, una misura di uno spazio di Hausdorff è una misura di Borel se è una misura $\sigma-$additiva definita sulla $\sigma-$algebra dei boreliani (la più piccola $\sigma-$algebra che contiene gli aperti). Lo spazio delle misure di Borel finite e con segno su $\mathbb{R}^d$ coincide con il duale delle funzioni continue e limitate. Lo spazio delle misure boreliane con segno è sufficiente per fare probabilità in $\mathbb{R}^d$.

Date due misure μ, ν in $\mathcal{P}(K)$, la loro distanza $1-$lipschitziana è

$$D_1(\mu, \nu) = \sup_{\phi \in \mathrm{Lip}_1} \int \phi \, \mathrm{d}\mu - \int \phi \, \mathrm{d}\nu.$$

Per capire come funziona, consideramo il caso semplice $\mu = \delta_{\mathbf{x}_1}$, $\mu = \delta_{\mathbf{x}_2}$. Per ogni $\phi \in \mathrm{Lip}_1$,

$$\int \phi \, \mathrm{d}\mu - \int \phi \, \mathrm{d}\nu = \phi(\mathbf{x}_1) - \phi(\mathbf{x}_2) \leq |\mathbf{x}_1 - \mathbf{x}_2|$$

e si raggiunge l'uguaglianza segliendo come ϕ la funzione $\phi(\mathbf{x}) = |\mathbf{x} - \mathbf{x}_2|$. Dunque $D_1(\mu, \nu) = |\mathbf{x}_1 - \mathbf{x}_2|$. Abbiamo dunque definito una metrica che tiene in conto la distanza spaziale tra le due misure.
Come secondo esempio, consideriamo due mappe Φ^1 e Φ^2, una misura μ e le misure $\nu_i = \Phi^i_{\#}\mu$, $i = 1, 2$. Mostriamo che la distanza tra queste due misure è stimata dalla differenza tra le due mappe. Per ogni $\phi \in \mathrm{Lip}_1$,

$$\int \phi \, \mathrm{d}\nu_1 - \int \phi \, \mathrm{d}\nu_2 = \int (\phi(\Phi^1(\mathbf{x})) - \phi(\Phi^2(\mathbf{x}))\mu(\mathrm{d}\mathbf{x}) \leq \sup_{\mathbf{x}} |\Phi^1(\mathbf{x}) - \Phi^2(\mathbf{x})|$$

quindi $D_1(\Phi^1_{\#}\mu, \Phi^2_{\#}\mu)$ è controllato da $\|\Phi^1 - \Phi^2\|_\infty$.

Proposizione 4.1. *D_1 è una metrica.*

Dimostrazione.
• D_1 è ben definita.
Fissato $\mathbf{x}_0 \in K$, ogni funzione $1-$lipschitziana ϕ si può scrivere come $\phi(\mathbf{x}) = \phi(\mathbf{x}_0) + \psi(\mathbf{x})$, con $\psi(\mathbf{x}) = \phi(\mathbf{x}) - \phi(\mathbf{x}_0)$. La funzione ψ è ancora $1-$lipschitziana, inoltre, poiché K è compatto, $|\psi| \leq a$, dove a è il diametro di K. Ne segue che

$$\int \phi \, \mathrm{d}\mu - \int \phi \, \mathrm{d}\nu = \int \psi \, \mathrm{d}\mu - \int \psi \, \mathrm{d}\nu \leq 2a,$$

dunque D_1 esiste perche l'estremo superiore è limitato.

• D_1 è simmetrica.
È sufficiente scegliere $\phi = -\psi$, e passare all'estremo superiore su ψ.

• D_1 è semidefinita positiva.
Scegliendo $\phi \equiv 0$, si ottiene che $D_1 \geq 0$. Inoltre $D_1(\mu, \mu) = 0$.

• D_1 verifica la disuguaglianza triangolare.
Date $\mu, \nu, \rho \in \mathcal{P}(K)$ si ha

$$\int \phi(\mathrm{d}\mu - \mathrm{d}\nu) = \int \phi(\mathrm{d}\mu - \mathrm{d}\rho) + \int \phi(\mathrm{d}\rho - \mathrm{d}\nu) \leq D_1(\mu, \rho) + D_1(\rho, \nu).$$

Passando all'estremo superiore in ϕ si ottiene la tesi.

• D_1 è definita positiva.
Supponiamo che $D_1(\mu, \nu) = 0$. Allora per ogni $\phi \in \mathrm{Lip}_1$ deve valere $\int \phi \, \mathrm{d}\mu = \int \phi \, \mathrm{d}\nu$. Se così non fosse, si avrebbe $D_1(\mu, \nu) > 0$. Inoltre questa uguaglianza si estende a tutte le funzioni lipschitziane (basta dividere per la costante di Lipschitz). Sia ora α una funzione

continua. Provo che $\int \alpha(\mathrm{d}\mu - \mathrm{d}\nu) = 0$. Per ogni $\varepsilon > 0$ esiste una ψ_ε lipschitziana tale che $\|\alpha - \psi_\varepsilon\|_\infty < \varepsilon$ (per esempio si può invocare il teorema di Waierstrass). Dunque

$$\left|\int \alpha(\mathrm{d}\mu - \mathrm{d}\nu)\right| = \int |\alpha - \psi_\varepsilon|(\mathrm{d}\mu + \mathrm{d}\nu) \leq \varepsilon.$$

Poiché le misure sono identificate da come agiscono sulle funzioni continue, ne segue che μ e ν sono uguali.

$\square$

Consideriamo ora l'insieme delle misure di probabilità $\mathcal{P}(K)$ dotato della topologia della *convergenza debole nel senso delle misure*, cioè $\mu_n \rightharpoonup \mu$ se e solo se $\forall\, \phi \in C(K)$ si ha

$$\int \phi\, \mathrm{d}\mu_n \to \int \phi\, \mathrm{d}\mu.$$

Teorema 4.2. *D_1 è una metrica per la convegenza debole, inoltre $\mathcal{P}(K)$ è completo e compatto.*

Dimostrazione. Se $D_1(\mu_n, \mu)$, per ogni funzione test Lipschiziana ϕ, si ha

$$\left|\int \phi(\mathrm{d}\mu_n - \mathrm{d}\mu)\right| \leq L D_1(\mu_n, \mu) \to 0.$$

Sia α continua. Per ogni $\varepsilon > 0$ esiste ϕ_ε lipschitiziana tale che $\|\alpha - \phi_\varepsilon\|_\infty < \varepsilon$. Ne segue

$$\left|\int \alpha(\mathrm{d}\mu_n - \mathrm{d}\mu)\right| \leq \int |\alpha - \phi_\varepsilon|(\mathrm{d}\mu_n + \mathrm{d}\mu) + \left|\int \phi_\varepsilon(\mathrm{d}\mu_n - \mathrm{d}\mu)\right|.$$

Il primo termine è minore di 2ε per ogni n; a ε fissato il secondo tende a 0 in n, dunque è minore di ε per n sufficientemente grande. Se ne conclude che $\int \alpha(\mathrm{d}\mu_n - \mathrm{d}\mu) \to 0$, cioè $\mu_n \rightharpoonup \mu$.
Viceversa, suppongo che $\mu_n \rightharpoonup \mu$. In particolare, per ogni $\phi \in \mathrm{Lip}_1$,

$$\int \phi(\mathrm{d}\mu_n - \mathrm{d}\mu) \to 0.$$

Dalla definizione di D_1, che è un estremo superiore, si ha che deve esistere $\phi_n \in \mathrm{Lip}_1$ tale che

$$D_1(\mu_n, \mu) \leq \int \phi_n(\mathrm{d}\mu_n - \mathrm{d}\mu) + \frac{1}{n}.$$

Come abbiamo già fatto, possiamo sostiture a ϕ_n la funzione $\psi_n(\mathbf{x}) = \phi_n(\mathbf{x}) - \phi_n(\mathbf{x}_0)$, dove $\mathbf{x}_0$ è un qualunque fissato punto di K. La funzioni ψ_n sono $1-$lipschitziane e sono limitate dal diametro di K, dunque sono equicontinue ed equilimitate. Il teorema di Ascoli-Arzela permette di ottenere che a meno di sottosequenze ψ_n converge uniformemente a $\psi \in \mathrm{Lip}_1$. Indico ancora con n la sottosequenza. Si ha

$$D_1(\mu_n, \mu) \leq \int \psi_n(\mathrm{d}\mu_n - \mathrm{d}\mu) + \frac{1}{n} = \int \psi(\mathrm{d}\mu_n - \mathrm{d}\mu) + \int (\psi_n - \psi)(\mathrm{d}\mu_n - \mathrm{d}\mu) + \frac{1}{n}$$

Il primo termine va a 0 perché $\mu_n \rightharpoonup \mu$, il secondo perché è stimato da $2\|\psi_n - \psi\|_\infty$. Si ottiene che per la sottosuccessione estratta $D_1(\mu_n, \mu) \to 0$. Per concludere la prova bisogna dimostrare che tutta la successione tende a 0. L'argomento è il seguente. Per assurdo, se $D_1(\mu_n, \mu)$ non tendesse a 0, esisterebbe $\bar{\varepsilon} > 0$ e una sottosuccessione n_k tale che $D_1(\mu_{n_k}, \mu) \geq \bar{\varepsilon}$. Ripetendo l'argomento precedente, si otterrebbe che per una sottosuccessione di n_k il limite sarebbe 0. Questo è assurdo, dunque la successione tende a 0.

La prova della completezza è semplice: se μ_n è di Cauchy, allora per ogni $\phi \in \mathrm{Lip}_1$, $\int \phi \mu_n$ è di Cauchy. Passando al limite si ottiene un funzionale lineare sulle funzioni lipschitziane limitato nella norma dell'estremo superiore. Per densità si estende alle funzioni continue e dunque si identifica il limite con un funzionale lineare continuo, cioè una misura boreliana con segno μ. Per ogni n si ha che per ogni osservabile ϕ positivo $\int \phi \, \mathrm{d}\mu_n \geq 0$ e $\int \mathrm{d}\mu_n = 1$. Passando al limite in queste due realazioni si ottiene che μ è una misura di probabilità.

La compattezza (sequenziale) dello spazio delle misure per la convergenza debole si può ottenere da risultati di analisi funzionale, qui mi limito a dare un'indicazione del perché è vero con un esempio: sia μ_n una misura di probabilità sull'intervallo $[0, 1]$. La probabilità di $[0, 1/2]$ e quella di $(1/2, 1]$ sono due successioni positive limitate, dunque per sottosequenze convergono e la somma delle probabilità rimane 1 (non può disperdersi massa perché siamo su un compatto). Si può iterare questa procedura dividendo in quattro sottointervalli e ottenere una sottosequenza della sottosequenza per cui convergono le probabilità di $[0, 1/4]$, $(1/4, 1/2]$, $(1/2, 3/4]$, $(3/4, 1]$. Sempre passando a sottosequenze si possono anche anche le convergenze sulla misura degli estremi di questi intevalli. Iterando e utilizzando la sottosequenza diagonale, si trova una unica sottosequenza per cui esistono i limiti di $\mu_n([j/2^h, (j+1)/2^h])$, con $j = 0, \ldots 2^h - 1$, e degli analoghi intervalli chiusi o semichiusi. È facile convincersi che questi limiti definiscono una misura di probabilità.

$\square$

Termino questa sezione con una ossevazione: nel caso di misure definite su illimitati, la distanza $D_1(\mu, \nu)$ è infinita se $\int |\mathbf{x}| \, \mathrm{d}\mu(\mathrm{d}\mathbf{x})$ è finito e $\int |\mathbf{x}| \, \mathrm{d}\nu(\mathrm{d}\mathbf{x}) = +\infty$ (lo si mostri per esercizio). La convergenza debole è ancora metrizzabile, ma non con D_1.

Sunto

- La distanza D_1 è una metrica per $\mathcal{P}(K)$, con K compatto, che riesce a misurare gli effetti di un trasporto di massa.
- D_1 induce la topologia della convergenza debole nel senso delle misure su $\mathcal{P}(K)$, rispetto alla quale $\mathcal{P}(K)$ è sequenzialmente compatto.

4.2 Il problema di Monge-Kantorovich

Il problema di definire metriche che si comportanto bene rispetto al trasporto di misura è strettamento legato a questioni di ottimizzazione nella teoria analitica dei problemi di trasporto di massa. L'idea è che la distanza tra due misure di probabilità sia data da quanto costa trasportare l'una dell'altra. Questo problema si chiama problema di Monge, perché Monge se ne occupò cercando il modo più economico per trasportare palle di cannone da diversi depositi in diverse zone di impiego.

Definizione - il problema di Monge

Siano date due misure di probabilità μ e ν. Il problema di Monge consiste nel trovare la mappa Φ, se c'è, che realizza

$$\inf_{\Phi:\,\nu=\Phi_{\#}\mu} \int |\mathbf{x}-\Phi(\mathbf{x})|\mu(\mathrm{d}\mathbf{x})$$

Per esempio considero due misure singolari $\mu(\mathrm{d}\mathbf{x}) = \delta_{\mathbf{x}_1}(\mathrm{d}\mathbf{x})$, $\mu(\mathrm{d}\mathbf{x}) = \delta_{\mathbf{y}_1}(\mathrm{d}\mathbf{x})$, con $\mathbf{x}_1 \neq \mathbf{y}_1$. La distanza in variazione totale tra di esse è 2. Non è difficile comprendere che la migliore mappa di trasporto è la traslazione $\Phi_{\mathbf{h}}(\mathbf{x}) = \mathbf{x}+\mathbf{h}$, con $\mathbf{h} = \mathbf{y}_1 - \mathbf{x}_1$. È immediato verificare che $\Phi_{\mathbf{h}}\#\mu = \nu$, e che il costo è $|\mathbf{h}|$.
Poiché la misura di partenza è singolare, non è necessario specificare tutta Φ, ma solo trasportare la massa che è in $\mathbf{x}_1$ in $\mathbf{y}_1$. Generalizzo.

Definizione - matching bipartito euclideo
Siano date due $N-$uple di punti di $\mathbb{R}^d$, $\mathbf{x}_i$, $\mathbf{y}_i$, $i = 1,\dots N$. Sia $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i}(\mathrm{d}\mathbf{x})$ e sia $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{y}_i}(\mathrm{d}\mathbf{x})$. In problema di Monge consiste nel trovare il miglior matching bipartito, cioè la migliore permutazione σ degli indici che realizzi

$$\frac{1}{N}\inf_{\sigma} |\mathbf{x}_i - \mathbf{y}_{\sigma_i}|$$

Infatti le permutazioni di indici identificano le mappe che assegnano a ognuno dei punti $\mathbf{x}_i$ uno dei punti $\mathbf{y}_i$, in modo biiettivo, cioè realizzano il trasporto di μ in ν.

Si può però immaginare un altro modo di procedere, rompendo il vincolo di trasportare tutta la massa $1/N$ in $\mathbf{x}_i$ nel punto $\mathbf{y}_j$, ma ipotizzando di spostare parte della massa in vari punti di arrivo.

Definizione - il problema di Kantorovich discreto
Siano date due $N-$uple di punti di $\mathbb{R}^d$, $\mathbf{x}_i$, $\mathbf{y}_i$, $i = 1,\dots N$. Sia $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i}(\mathrm{d}\mathbf{x})$ e sia $\nu = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{y}_i}(\mathrm{d}\mathbf{x})$. Con il nome di **piano di trasporto** si intende una matrice $\{a_{ij}\}_{i,j=1}^N$ di elementi positivi tali che la somma su ogni riga e su ogni colonna sia $1/N$. Il problema di Kantorovich consiste nel trovare il piano di trasporto che realizza

$$\inf_{a} \sum_{i,j} a_{ij}|\mathbf{x}_i - \mathbf{y}_j|.$$

Il termine $a_{ij}|\mathbf{x}_i - \mathbf{y}_j|$ è il costo per spostare la parte di massa a_{ij} da $\mathbf{x}_i$ in $\mathbf{y}_j$.
Si può riformulare questo problema osservando che

$$\gamma(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) = \sum_{ij} a_{ij}\delta_{\mathbf{x}_i}(\mathrm{d}\mathbf{x})\delta_{\mathbf{y}_j}(\mathrm{d}\mathbf{y})$$

è una distribuzione congiunta di μ e ν,e tutte le distribuzioni congiunte di μ e ν sono di questo tipo. Un piano di trasporto è dunque una distribuzione congiunta, e il problema di Kantorovich consiste nel trovare

$$\inf_{\gamma} \int |\mathbf{x}-\mathbf{y}|\gamma(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y})$$

È un risultato classico dell'ottimizzazione convessa il fatto che il problema di Kantorovich è risolto da

$$\gamma(\mathrm{d}\mathbf{x},\mathrm{d}\mathbf{y}) = \sum_{i} \frac{1}{N}\delta_{\mathbf{x}_i}(\mathrm{d}\mathbf{x})\delta_{\mathbf{y}_{\sigma_i}}(\mathrm{d}\mathbf{y})$$

dove σ è la permutazione che risolve il problema di Monge.

Definizione - il problema di Kantorovich in generale
Siano date due misure μ e ν. Il problema di Kantorovich consiste nel trovare il piano di trasporto che realizza

$$\inf_{\gamma} \int |\mathbf{x} - \mathbf{y}| \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})$$

dove γ varia nell'insieme $\Gamma(\mu, \nu)$ delle misure congiunte di μ e ν.

In generale, il problema di Monge potrebbe non avere soluzione, per esempio se μ è una misura discreta su N punti e ν su $M > N$ punti, non esiste nessuna mappa Φ tale che $\nu = \Phi\#\mu$, mentre l'insieme dei piani di trasporto è sempre non vuoto, perché la distribuzione prodotto di μ e ν è una distribuzione congiunta. D'altra parte, se Φ trasporta μ in ν, la misura

$$\gamma = \mu \otimes (\Phi_{\#}\mu)$$

è un piano di trasporto.

Sunto

- Il problema di Kantorovich si può sempre formulare.
- Il costo secondo Kantorovich è minore o uguale al costo secondo Monge.
- Si può provare che se il problema di Monge ha soluzione, allora i due costi coincidono. In questo senso, il problema di Kantorovich è un **rilassamento** del problema di Monge

4.3 La distanza di Wassertein

Dato $p \in [1, +\infty)$, su $\mathcal{P}(K)$, si definisce le distanza p-Wasserstein come

$$W_p(\mu, \nu) = \inf_{\gamma} \left(\int |\mathbf{x} - \mathbf{y}|^p \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \right)^{1/p}.$$

La distanza $1-$Wasserstein prende anche il nome di distanza di Kantorovich–Rubinstein.

Teorema 4.3 (Dualità di Kantorovich)**.**

$$W_1(\mu, \nu) = \sup_{\varphi, \psi \in C(K): \varphi(x) + \psi(y) \leq |\mathbf{x} - \mathbf{y}|} \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu. \tag{15}$$

Questo teorema ha una interessante interpretazione "economica". La distanza W_1 rappresenta il più piccolo costo che per trasportare μ in ν. Data una γ, $\int |\mathbf{x} - \mathbf{y}|^p \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y})$ è il costo che si paga per trasportare μ in ν secondo il piano di trasporto γ.
Un altro modo di guardare a questo costo è misurare il costo per "raccogliere" $\mu(\mathrm{d}\mathbf{x})$ in $\mathbf{x}$, quantificato in $\phi(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$, e il costo di distribuire quanto raccolto secondo ν in $\mathbf{y}$ quantificato in $\psi(\mathbf{y})\mu(\mathrm{d}\mathbf{y})$. Affinché questo modo sia "conventiente", deve valere il vincolo $\phi(\mathbf{x}) + \psi(\mathbf{y}) \leq |\mathbf{x} - \mathbf{y}|$. In tal caso infatti

$$\int \phi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu = \int (\phi(\mathbf{x}) + \psi(\mathbf{y})\gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) \leq |\mathbf{x} - \mathbf{y}| \gamma(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}).$$

La dualità di Kantorovich asserisce che il massimo prezzo che si può chiedere per ritirare μ e ridistribuire in ν è proprio il costo minimo di trasporto.
Dalla disuguaglianza precedente si ottiene facilmente che il membro di destra della (15) è minore o uguale di W_1. Il viceversa non è semplice, una sua possibile prova richiedere la nozione di trasformata di Legendre su spazi di Banach, è disponibile sulle dispense del 2024/2025. Qui mi limito a dare un'idea del perché questo risultato è vero.
Userò le seguenti notazioni "compatte": date due misure di probabilità μ e ν, e date due fuzioni continue φ e ψ:

$$\mathrm{d}(\mu \otimes \nu) = \mu \otimes \nu(\mathrm{d}\mathbf{x}, \mathrm{d}\mathbf{y}) = \mu(\mathrm{d}\mathbf{x})\nu(\mathrm{d}\mathbf{y})$$
$$\varphi \oplus \psi = (\varphi \oplus \psi)(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) + \psi(\mathbf{y})$$

In particolare

$$\int \varphi \oplus \psi \ \mathrm{d}(\mu \otimes \nu) = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu.$$

Si noti che $\mathrm{d}(\mu \otimes \nu) \in \Gamma(\mu, \nu)$, infatti è la misura di probabilità congiunta che si ottiene pensando di estrarre $\mathbf{x}$ con μ, e $\mathbf{y}$ con ν, indipendenti. Inoltre indicherò con $c = c(\mathbf{x}, \mathbf{y})$ il costo $|\mathbf{x} - \mathbf{y}|$.
L'idea per provare la dualità è quella di usare due funzioni φ e ψ per esprimere il vincolo sulle distribuzioni congiunte. Sia $\mathcal{M}(K \times K)$ l'insieme delle misure con segno, cioè il duale di $C(K \times K)$, sia $\mathcal{M}^+$ il sottoinsieme delle misure positive. Evidentemente

$$\Gamma(\mu, \nu) \subset \mathcal{M}^+ \subset \mathcal{M}.$$

Consideriamo la differenza

$$\int \varphi \oplus \psi \ \mathrm{d}(\mu \otimes \nu) - \int \varphi \oplus \psi \ \mathrm{d}\gamma$$

per una generica $\gamma \in \mathcal{M}$. Si noti che questa differenza vale

$$\int \varphi(x)\mu(\mathrm{d}\mathbf{x}) + \int \psi(\mathbf{y})\nu(\mathrm{d}y) - \int (\varphi(\mathbf{x}) + \psi(\mathbf{y}))\gamma(\mathrm{d}x, \mathrm{d}y) =$$
$$\int \varphi(\mathbf{x})(\mu(\mathrm{d}\mathbf{x}) - \gamma^1(\mathrm{d}\mathbf{x})) + \int \psi(\mathbf{y})(\nu(\mathrm{d}\mathbf{y}) - \gamma^2(\mathrm{d}\mathbf{y}))$$

dove γ^1 e γ^2 sono le due marginali di γ. Se $\gamma \in \Gamma(\mu, \nu)$, al differenza è 0 qualunque siano φ, ψ. Al contrario, se $\gamma \notin \Gamma(\mu, \nu)$ $\mu \neq \gamma^1$ o $\nu \neq \gamma^2$ (o entrambi), e dunque esistono $\bar{\varphi}$, $\bar{\psi}$ su cui la differenza da un valore non nullo. Pertanto

$$\sup_{\varphi, \psi \in C(K)} \left(\int \varphi \oplus \psi \ \mathrm{d}(\mu \otimes \nu) - \int \varphi \oplus \psi \ \mathrm{d}\gamma \right) = \begin{cases} 0 & \text{se } \gamma \in \Gamma(\mu, \nu) \\ +\infty & \text{altrimenti} \end{cases}$$

infatti nel primo caso la differenza è nulla per ogni coppia di funzioni, nel secondo caso moltiplicando la coppia $(\bar{\varphi}, \bar{\psi})$ per una costante arbitrarie, opportuno, si ottiene un qualunque valore reale, pertanto il sup è infinito.
Nella definizione di W_1 posso considerare invece dell'estremo inferiore su $\gamma \in \Gamma$, all'estremo inferiore su $\gamma \in \mathcal{M}^+$, pagando però $+\infty$ se $\gamma \notin \Gamma$:

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \int c \, \mathrm{d}\gamma = \inf_{\gamma \in \mathcal{M}^+} \sup_{\varphi, \psi} \int c \, \mathrm{d}\gamma + \int \varphi \oplus \psi \ \mathrm{d}(\mu \otimes \nu) - \int \varphi \oplus \psi \ \mathrm{d}\gamma. \quad (16)$$

(ricordo che il costo c è $c = |\mathbf{x} - \mathbf{y}|$). Per una qualunque funzione F di due variabili si prova facilmente che in generale

$$\inf_{a \in A} \sup_{b \in B} F(a, b) \geq \sup_{b \in B} \inf_{a \in A} F(a, b).$$

In qualche caso vale l'uguaglianza, in particolare si dimostrsa che se F è definita su un compatto convesso, e se inoltre $F(a, b)$ è concava in a per ogni b fissato, ed e convessa in b per ogni a fissato, allora la coppia (a, b) che realizza gli estremi è la stessa indipendentemente dall'ordine, e individua un punto di sella di F. Nel nostro caso, convessità e concavità sono banali, perché la funzione in (16) è lineare nella coppia (φ, ψ) e dunque concava, e lineare in γ, ed è effettivamente definita su un convesso. Non è però soddisfatta la condizione di compattezza. Esistono estensioni di questo risultato che permettono di provare la tesi della scambiabilità degli estremi, che assumerò dunque vera senza altri dettagli.
Dunque

$$\inf_{\gamma \in \Gamma(\mu, \nu)} \int c \, \mathrm{d}\gamma = \sup_{\varphi, \psi} \inf_{\gamma \in \mathcal{M}^+} \int c \, \mathrm{d}\gamma + \int \varphi \oplus \psi \; \mathrm{d}(\mu \otimes \nu) - \int \varphi \oplus \psi \; \mathrm{d}\gamma$$

Noto che il calcolo dell'estremo inferiore riguarda solo i due termini in γ:

$$\inf_{\gamma \in \mathcal{M}^+} \int (c - \varphi \oplus \psi) \, \mathrm{d}\gamma$$

È evidente che se $\varphi \oplus \psi \leq c$ l'estremo inferiore è raggiunto su $\gamma = 0$, e vale 0. Al contrario, se esistono valori $(\mathbf{x}_0, \mathbf{y}_0)$ tali che $\varphi(\mathbf{x}_0) + \psi(\mathbf{y}_0) > |\mathbf{x}_0 - \mathbf{y}_0|$ basta concentrare γ su questi valori, con valori divergenti di massa per ottenere che l'inf è $-\infty$. Dunque l'estremo inferiore su $\mathcal{M}^+$ si ottiene dove la coppia di funzioni φ, ψ soddisfa il vincolo $\varphi \oplus \psi \leq c$. Abbiamo dunque provato che

$$W_1(\mu, \nu) = \sup_{\varphi, \psi \in C(K), \, \varphi \oplus \psi \leq c} \int \varphi \oplus \psi \; \mathrm{d}(\mu \otimes \nu).$$

Teorema 4.4 ($W_1 = D_1$)**.**
La distanza di Wasserstein W_1 è uguale alla distanza $1-$lipschitzianza D_1.

Dimostrazione. La prima osservazione è che se $\varphi \in \mathrm{Lip}_1$ la coppia di funzioni $\varphi(\mathbf{x})$, $\psi(\mathbf{y}) = -\varphi(\mathbf{y})$ verifica il vincolo richiesto, dunque

$$\sup_{\varphi, \psi : \varphi(\mathbf{x}) + \psi(\mathbf{y}) \leq |\mathbf{x} - \mathbf{y}|} \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu \geq \sup_{\varphi \in \mathrm{Lip}_1} \int \varphi \, (\mathrm{d}\mu - \mathrm{d}\nu),$$

e dunque abbiamo facilmente che $W_1 \geq D_1$. Dimostro la disuguaglianza opposta. Date φ e ψ con $\varphi(\mathbf{x}) + \psi(\mathbf{y}) \leq |\mathbf{x} - \mathbf{y}|$, osservo che

$$\varphi(\mathbf{x}) \leq \psi^*(\mathbf{x}) \coloneqq \inf_{\mathbf{y}} |\mathbf{x} - \mathbf{y}| - \psi(\mathbf{y}),$$

infatti per ogni $\mathbf{y}$, $\varphi(\mathbf{x}) \leq |\mathbf{x} - \mathbf{y}| - \psi(\mathbf{y})$, e dunque $\varphi(\mathbf{x})$ è minore dell'estremo inferiore in $\mathbf{y}$. Si ha dunque

$$\int \varphi \, \mathrm{d}\mu \leq \int \psi^* \, \mathrm{d}\mu$$

D'altra parte, per la definzione di Ψ^*, si ha

$$\psi^*(\mathbf{x}) + \psi(\mathbf{y}) \le |\mathbf{x} - \mathbf{y}|$$

da cui segue

$$\psi(\mathbf{y}) \le \psi^{**}(\mathbf{y}) \coloneqq \inf_{\mathbf{x}} |\mathbf{x} - \mathbf{y}| - \psi^*(\mathbf{x}).$$

Pertanto

$$\sup_{\varphi,\psi:\varphi(\mathbf{x})+\psi(\mathbf{y})\le|\mathbf{x}-\mathbf{y}|} \int \varphi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu = \sup_{\psi} \left(\int \psi^* \,\mathrm{d}\mu + \psi^{**} \,\mathrm{d}\nu \right)$$

Mostro che per costruzione ψ^* è una funzione $1-$lipschitziana, infatti per ogni $\mathbf{x}, \mathbf{x}_1$,

$$\begin{aligned} \psi^*(\mathbf{x}) &= \inf_{\mathbf{y}}(|\mathbf{x} - \mathbf{y}| - \psi(\mathbf{y})) \le \inf_{\mathbf{y}}(|\mathbf{x}_1 - \mathbf{y}| + |\mathbf{x} - \mathbf{x}_1| - \psi(\mathbf{y})) \\ &= \inf_{\mathbf{y}}(|\mathbf{x}_1 - \mathbf{y}| - \psi(\mathbf{y})) + |\mathbf{x} - \mathbf{x}_1| = \psi^*(\mathbf{x}_1) + |\mathbf{x} - \mathbf{x}_1|. \end{aligned}$$

Scambiando $\mathbf{x}$ con $\mathbf{x}_1$ e usando le due disuguaglianze, si ottine che $\psi^* \in \mathrm{Lip}_1$. Dunque

$$\sup_{\varphi,\psi:\varphi(\mathbf{x})+\psi(\mathbf{y})\le|\mathbf{x}-\mathbf{y}|} \int \varphi \,\mathrm{d}\mu + \int \psi \,\mathrm{d}\nu \le \sup_{\varphi\in\mathrm{Lip}_1} \int \varphi \,\mathrm{d}\mu + \varphi^* \,\mathrm{d}\nu.$$

La prova si conclude facilmente mostrando che se φ è una funzione $1-$lipschitziana, allora

$$\varphi^*(\mathbf{y}) = -\varphi(\mathbf{y}).$$

Dimostro quest'ultimo fatto.

$$\varphi^*(\mathbf{y}) = \inf_{\mathbf{x}}(|\mathbf{x} - \mathbf{y}| - \varphi(\mathbf{x})) = \inf_{\mathbf{x}}(|\mathbf{x} - \mathbf{y}| - (\varphi(\mathbf{x}) - \varphi(\mathbf{y})) - \varphi(\mathbf{y})$$

La funzione di cui si cerca l'estremo inferiore è non negativa, perché $|\mathbf{x}- \ vey| > \varphi(\mathbf{y}))-\varphi(\mathbf{y})$, d'altra parte è nulla in $\mathbf{x} = \mathbf{y}$, dunque l'estremo cercato è 0, e da questo fatto segue la tesi. □

Sunto

- La dualità di Kantorovich trasforma il problema del calcolo del minimo costo per spostare μ in ν, in un problema di massimo con vincoli.
- Nel caso di costo dato dalla distanza euclidea, la formulazione duale è equivalente al calcolo della distanza $1-$lipschitziana tra μ e ν.