Capítulo 4

Distâncias de Wasserstein

4.1 Definição e propriedades iniciais

A distâncias de Wasserstein-p é definida no conjunto das medidas de Radon com p-momento finito, *i.e.* dado um ponto qualquer $x_0 \in \mathcal{X}$ do espaço Polonês ambiente, definimos

$$\mathscr{P}_p(\mathcal{X}) \stackrel{\text{def.}}{=} \left\{ \mu \in \mathscr{P}(\mathcal{X}) : M_p(\mu) \stackrel{\text{def.}}{=} \int_{\mathcal{X}} d_{\mathcal{X}}(x, x_0) \mathrm{d}\mu(x) < +\infty \right\}. \tag{4.1}$$

Para todo $p \geq 1$, condição que $\mu, \nu \in \mathscr{P}_p(\mathcal{X})$ garante que o problema de Kantorovitch com custo dado por $c(x, y) = d^p_{\mathcal{X}}(x, y)$ tem valor finito e por isso podemos definir a seguinte quantidade

$$W_p(\mu, \nu) \stackrel{\text{def.}}{=} \min_{\gamma \in \Pi(\mu,\nu)} \left(\int_{\mathcal{X}\times\mathcal{X}} d^p_{\mathcal{X}}(x, y) \mathrm{d}\gamma(x, y) \right)^{1/p} = \| d_{\mathcal{X}}(\cdot, \cdot) \|_{L^p(\gamma)}. \tag{4.2}$$

Para provar que essa quantidade define uma distância, precisamos do *teorema de disintegração*, que nada mais é que a existência de densidades de probabilidade condicional. Essa é uma questão não trivial em teoria de probabilidade, mas para medidas de probabilidade borelianas num espaço polonês, a esperança condicional sempre existe.[1]

Teorema 4.1.1 (Disintegração)**.** *Seja $\gamma \in \mathscr{P}(\mathcal{X} \times \mathcal{Y})$ e seja $\nu \stackrel{\text{def.}}{=} (\pi_{\mathcal{Y}})_\sharp \pi$ a marginal em $\mathcal{Y}$. Então existe uma família de probabilidades*

$$\{\gamma_y\}_{y\in\mathcal{Y}} \subset \mathscr{P}(\mathcal{X})$$

tal que:

1. *para toda função boreliana limitada $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,*

$$\int_{\mathcal{X}\times\mathcal{Y}} \varphi(x, y) \mathrm{d}\gamma(x, y) = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} \varphi(x, y) \mathrm{d}\gamma_y(x) \right) \mathrm{d}\nu(y). \tag{4.3}$$

Além disso, nós escrevemos $\gamma = \gamma_y \otimes \nu(\mathrm{d}y)$.

[1] Ver "Multidimensional Diffusion Processes"- Daniel W.Stroock, S. R. Srinivasa Varadhan, 1997, Teorema 1.1.6

2. *para ν-q.t.p. y, γ_y é uma medida de probabilidade;*

3. *a família $y \mapsto \gamma_y$ é mensurável no sentido fraco.*

Demonstração. Como $\mathcal{X}$ e $\mathcal{Y}$ são poloneses, o espaço produto $\mathcal{X} \times \mathcal{Y}$ também é polonês. Logo existe uma probabilidade condicional regular de γ em relação à aplicação

$$\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}.$$

Isto significa que existe uma família $\{\gamma_y\}_{y \in \mathcal{Y}}$ de medidas de probabilidade em $\mathcal{X}$ tal que, para toda função boreliana limitada φ,

$$\mathbb{E}_\gamma[\varphi(X, Y) \mid Y = y] = \int_{\mathcal{X}} \varphi(x, y) \mathrm{d}\gamma_y(x) \quad \text{para } \nu\text{-q.t.p. } y.$$

Integrando essa identidade em relação a ν, obtemos exatamente (4.3). A mensurabilidade fraca segue da construção padrão das probabilidades condicionais em espaços métricos. A unicidade vale pelo teorema de unicidade da esperança condicional. □

A disintegração fornece uma interpretação canônica de planos de transporte como famílias de medidas condicionais. Ela permite provar o seguinte resultado estrutural.

Lema 4.1.1 (Lema de colagem)**.** *Sejam $\mu, \nu, \lambda \in \mathscr{P}(\mathcal{X})$ e*

$$\gamma_{12} \in \Pi(\mu, \nu), \qquad \gamma_{23} \in \Pi(\nu, \lambda).$$

Então existe $\gamma \in \mathscr{P}(\mathcal{X}^3)$ tal que

$$(\pi_1, \pi_2)_\sharp \gamma = \gamma_{12}, \qquad (\pi_2, \pi_3)_\sharp \gamma = \gamma_{23}.$$

No enunciado anterior, $\pi_i(x_1, x_2, x_3) = x_i$, para $i \in \{1, 2, 3\}$.

Demonstração. Aplicamos o Teorema 4.1.1 aos dois planos.

Primeiro, disintegramos γ_{12} em relação à segunda variável:

$$\gamma_{12}(\mathrm{d}x, \mathrm{d}y) = \gamma_{12,y}(\mathrm{d}x) \otimes \nu(\mathrm{d}y),$$

onde $\gamma_{12,y} \in \mathscr{P}(\mathcal{X})$ para ν-q.t.p. y.

De modo análogo, disintegramos γ_{23} em relação à primeira variável:

$$\gamma_{23}(\mathrm{d}y, \mathrm{d}z) = \gamma_{23,y}(\mathrm{d}z) \otimes \nu(\mathrm{d}y),$$

com $\gamma_{23,y} \in \mathscr{P}(\mathcal{X})$.

Definimos então uma medida $\pi \in \mathscr{P}(\mathcal{X}^3)$ por

$$\gamma(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z) \stackrel{\text{def.}}{=} \gamma_{12,y}(\mathrm{d}x) \otimes \gamma_{23,y}(\mathrm{d}z) \otimes \nu(\mathrm{d}y).$$

Pela fórmula de disintegração, para toda função teste boreliana limitada φ,

$$\int \varphi(x, y) \mathrm{d}(\pi_{\mathcal{X}}, \pi_{\mathcal{Y}})_\sharp \gamma = \int_{\mathcal{Y}} \left(\int_{\mathcal{X}} \varphi(x, y) \mathrm{d}\gamma_{12,y}(x) \right) \mathrm{d}\nu(y),$$

o que mostra que $(\pi_{\mathcal{X}}, \pi_{\mathcal{Y}})_\sharp \gamma = \gamma_{12}$. O mesmo argumento vale para $(\pi_{\mathcal{Y}}, \pi_3)_\sharp \gamma = \gamma_{23}$. □

Proposição 4.1.1. *A quantidade W_p é uma distância no espaço $\mathscr{P}_p(\mathcal{X})$.*

Demonstração. Para provar que W_p precisamos demonstrar os pontos seguintes:

1. $0 \leq W_p(\mu, \nu) = W_p(\nu, \mu)$, para todo par $\mu, \nu \in \mathscr{P}_p(\mathcal{X})$;
2. $W_p(\cdot, \cdot)$ satisfaz a desigualdade triangular;
3. $W_p(\mu, \nu) = 0$ se, e somente se, $\mu = \nu$.

O ponto 1. segue diretamente da definição. Para provar o item 2., tomemos três medidas $\mu, \nu, \lambda \in \mathscr{P}_p(\mathcal{X})$. Seja $\gamma_{\mu,\lambda} \in \Pi(\mu, \lambda)$ e $\gamma_{\lambda,\nu} \in \Pi(\lambda, \nu)$ planos de transporte ótimo para $W_p(\mu, \lambda)$ e $W_p(\lambda, \nu)$. Usando o lema de colagem, tome um plano de transporte $\gamma \in \mathscr{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{X})$ tais que $(\pi_{1,2})_\sharp = \gamma_{\mu,\lambda}$ e $(\pi_{2,3})_\sharp = \gamma_{\lambda,\nu}$.

Podemos usar a desigualdade triangular em $L^p(\gamma)$, de modo que

$$W_p(\mu, \nu) \leq \|d_{\mathcal{X}}(x_1, x_3)\|_{L^p(\gamma)} \leq \|d_{\mathcal{X}}(x_1, x_2)\|_{L^p(\gamma)} + \|d_{\mathcal{X}}(x_2, x_3)\|_{L^p(\gamma)}. \tag{4.4}$$

Pela otimalidade de $\gamma_{\mu,\lambda}$, pela condição de marginais de γ, temos que $\|d_{\mathcal{X}}(x_1, x_2)\|_{L^p(\gamma)} = W_p(\mu, \lambda)$. Similarmente, temos $\|d_{\mathcal{X}}(x_2, x_3)\|_{L^p(\gamma)} = W_p(\lambda, \nu)$. A desigualdade triangular segue.

Para provar o ponto 3, note que se $\mu = \nu$, claramente o mapa de transporte $T = \mathrm{id}$ atinge a cota inferior $0 \leq \|d_{\mathcal{X}}(X, X)\|_{L^p(\Omega,\mathbb{P})}$ e portanto é ótimo. Para verificar a afirmação conversa, note que se

$$0 = W_p^p(\mu, \nu) = \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^p(x, y) \mathrm{d}\gamma,$$

então γ é concentrada no gráfico da aplicação identidade. Logo, pela Proposição 3.3.1, segue que $\gamma = (\mathrm{id}, \mathrm{id})_\sharp \mu$, e portanto $\nu = \mu$. □

Proposição 4.1.2. *Para todos $1 \leq q \leq p$, temos que*

$$W_1(\mu, \nu) \leq W_q(\mu, \nu) \leq W_p(\mu, \nu),$$

e se $\mathcal{X}$ é limitado, temos a desigualdade reversa

$$W_p(\mu, \nu) \leq (\operatorname{diam} \mathcal{X})^{\frac{p-1}{p}} W_1(\mu, \nu)^{1/p}.$$

Demonstração. A prova segue da desigualdade de Jensen. Dados $q < p$, tome γ um plano de transporte ótimo para $W_p^p(\mu, \nu)$. Como $q < p$, a função $t \mapsto t^{p/q}$ é convexa e portanto

$$W_q(\mu, \nu)^{p/q} \leq \left(\int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^q(x, y) \mathrm{d}\gamma \right)^{p/q} \leq \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^p(x, y) \mathrm{d}\gamma = W_p^p(\mu, \nu).$$

A primeira estimativa segue tomando a potência $1/p$ em ambos os lados.

Para a segunda, tome agora γ um plano de transporte ótimo para $W_1(\mu, \nu)$. Então, como $d_{\mathcal{X}}(x, y) \leq \operatorname{diam} \mathcal{X} < +\infty$ para todo par de pontos (x, y), temos que

$$W_p^p(\mu, \nu) \leq \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}^p(x, y) \mathrm{d}\gamma \leq (\operatorname{diam} \mathcal{X})^{p-1} \int_{\mathcal{X} \times \mathcal{X}} d_{\mathcal{X}}(x, y) \mathrm{d}\gamma = (\operatorname{diam} \mathcal{X})^{p-1} W_1(\mu, \nu).$$

Novamente, a estimativa segue tomando a potência $1/p$ em ambos os lados. □

A Proposição 4.1.2 motiva um melhor entendimento da distância de Wasserstein-1. De fato, o caso particular $p = 1$ admite uma caracterização mais fina da formulação dual

$$W_p^p(\mu, \nu) = \sup_{\varphi} \int_{\mathcal{X}} \varphi \mathrm{d}\mu + \int_{\mathcal{X}} \varphi^c \mathrm{d}\nu,$$

onde relembramos que φ^c é transformada c de φ e o supremo pode ser tomado entre funções φ que já sejam c-concavas, isto é, já são a transformada c de uma outra função. No caso $c = d_{\mathcal{X}}$, funções c-concavas são 1-Lipschitz. Isso nos dá a seguinte simplicação da fórmula de dualidade, que é chamada muitas vezes de dualidade de *Kantorovitch-Rubistein*

Proposição 4.1.3. *Seja $\varphi \in \mathscr{C}^0(\mathcal{X})$, e considere o custo $c(x, y) = d_{\mathcal{X}}(x, y)$, então $\varphi^c(x) \stackrel{\text{def.}}{=} \inf_{y \in \mathcal{X}} d_{\mathcal{X}}(x, y) - \varphi(y)$ é 1-Lipschitz. Por outro lado, se φ é 1-Lipschitz, então $\varphi^c = -\varphi$.*

Consequentemente, temos a seguinte fórmula de dualidade para a distância de Wasserstein-1: para todo par $\mu, \nu \in \mathscr{P}_1(\mathcal{X})$ temos que

$$W_1(\mu, \nu) = \sup_{f \text{ 1-Lip}} \int_{\mathcal{X}} f \mathrm{d}(\mu - \nu).$$

Demonstração. Para todo $x \in \mathcal{X}$, e todo $\varepsilon > 0$, existe pela definição de ínfimo, um y_x tal que

$$d_{\mathcal{X}}(x, y_x) - \varphi(y_x) < \varphi^c(x) + \varepsilon.$$

Desse modo, dados $x, z \in \mathcal{X}$, tome y_x como à cima e note que

$$\begin{aligned} \varphi^c(z) - \varphi^c(x) &\leq d_{\mathcal{X}}(z, y_x) - \varphi(y_x) - (d_{\mathcal{X}}(x, y_x) - \varphi(y_x)) - \varepsilon = d_{\mathcal{X}}(z, y_x) - d_{\mathcal{X}}(x, y_x) - \varepsilon \\ &\leq d_{\mathcal{X}}(x, z) - \varepsilon. \end{aligned}$$

Trocando os papeis de x e z obtemos que

$$|\varphi^c(z) - \varphi^c(x)| \leq d_{\mathcal{X}}(x, z) - \varepsilon,$$

e fazendo $\varepsilon \to 0$, obtemos que φ^c é 1-Lipschitz.

Por outro lado, tomando $x = y$ no ínfimo definindo $\varphi^c(y)$, temos que $\varphi^c(y) \leq -\varphi(y)$. Dado $\varepsilon > 0$, tome x tal que $d_{\mathcal{X}}(x, y) - \varphi(x) \leq \varphi^c(y) + \varepsilon$. Como φ é 1-Lipschitz segue que

$$-\varphi(y) \leq d_X(x, y) - \varphi(x) \leq \varphi^c(y) + \varepsilon.$$

Como $\varepsilon > 0$ é arbitrário, o resultado segue.

A segunda afirmação é uma consequência direta das condições de otimalidade obtidas com a transformada c e o resultado anterior. $\square$

A fórmula de dualidade para $W_1(\mu, \nu)$

$$W_1(\mu, \nu) = \sup_{f \text{ 1-Lipschitz}} \int_{\mathcal{X}} f \mathrm{d}(\mu - \nu) \tag{4.5}$$

exemplifica a propriedade fundamental das distâncias de Wasserstein, de ser equivalente à convergência estreita de medidas de probabilidade. O leitor atento perceberá que o conjunto onde o supremo é tomado, das funções 1-Lipschitz, é estritamente menor que das funções $\mathscr{C}_b(\mathcal{X})$ e portanto o supremo à cima ir para 0 não implica diretamente que

$$\int_{\mathcal{X}} f \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f \mathrm{d}\mu$$

para toda função $f \in \mathscr{C}_b(\mathcal{X})$.

Pelo Lema de aproximação de funções s.c.i., podemos usar funções Lipschitz e limitadas, junto de argumentos de convergência monótona para provar o resultado seguinte.

Lema 4.1.2. *Seja $\mathcal{X}$ um espaço polonês e $(\mu_n)_{n\in\mathbb{N}} \subset \mathscr{P}(\mathcal{X})$. Então $\mu_n \xrightarrow[n\to\infty]{} \mu$ se, e somente se,*

$$\int_{\mathcal{X}} f \mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f \mathrm{d}\mu, \quad \textit{para todo função } f \textit{ 1-Lipschitz.}$$

4.2 Propriedades topológicas de $(\mathscr{P}_p, W_p)$

Agora queremos entender como se comporta o espaço $\mathscr{P}_p(\mathcal{X})$ quando munido da distância W_p. Boa parte das propriedades topológicas de $(\mathscr{P}_p(\mathcal{X}), W_p)$ são herdadas do espaço ambiente $(\mathcal{X}, d_{\mathcal{X}})$.

A completude do espaço $(\mathscr{P}_p(\mathcal{X}), W_p)$ pode ser obtida a partir da dualidade de Kantorovitch-Rubistein em W_1.

Teorema 4.2.1. *Se o espaço $(\mathcal{X}, d_{\mathcal{X}})$ é completo, então o espaço $(\mathscr{P}_p(\mathcal{X}), W_p)$ também é completo para todo $p \geq 1$.*

Demonstração. Dado $p \geq 1$, seja $(\mu_n)_{n\in\mathbb{N}} \subset \mathscr{P}_p(\mathcal{X})$ uma sequência de Cauchy em $(\mathscr{P}_p(\mathcal{X}), W_p)$. Vamos provar que toda sequência de Cauchy em $(\mathscr{P}_p(\mathcal{X}), W_p)$ é tight, e portanto pré-compacta na topologia estreita pelo Teorema de Prokohorov.

Pela Proposição 4.1.2, a sequência $(\mu_n)_{n\in\mathbb{N}}$ também é de Cauchy em $(\mathscr{P}_1(\mathcal{X}), W_1)$. Logo, para todo $\varepsilon > 0$, existe $N \in \mathbb{N}$ tal que para todo $n \geq N$, temos que

$$W_1(\mu_n, \mu_N) < \varepsilon^2.$$

O conjunto $(\mu_i)_{i=1}^N$ sendo finito, ele é compacto, e portanto existe um conjunto compacto $K \subset \mathcal{X}$ tal que

$$\mu_i(\mathcal{X} \setminus K) < \varepsilon \text{ para todo } i = 1, \ldots, N.$$

Pela definição de compacidade, existe uma quantidade finita de pontos $(x_j)_{j=1}^m$ tais que

$$K \subset U \overset{\text{def.}}{=} \bigcup_{j=1}^m B(x_j, \varepsilon).$$

Seja $U_\varepsilon \overset{\text{def.}}{=} \{x \in \mathcal{X} : d_{\mathcal{X}}(x, U) < \varepsilon\} \subset \bigcup_{j=1}^m B(x_j, 2\varepsilon)$, a ε-vizinhança de U. Logo existe uma função Lipschitz $\varphi : \mathcal{X} \to [0,1]$ tal que $1_U \leq \varphi \leq 1_{U_\varepsilon}$. De fato, basta tomar

$$\varphi(\cdot) \overset{\text{def.}}{=} \left(1 - \frac{\operatorname{dist}(\cdot, U)}{\varepsilon}\right)_+.$$

Note que como a função distância é 1-Lipschitz, a função φ é $\frac{1}{\varepsilon}$-Lipschitz. Logo, essa função φ pode ser usada na fórmula de dualidade de Kantorovitch-Rubistein para obter, para todo $n \geq N$, que

$$\begin{aligned}\mu_n(U_\varepsilon) &\geq \int_{\mathcal{X}} \varphi \mathrm{d}\mu_n = \int_{\mathcal{X}} \varphi \mathrm{d}\mu_N + \int_{\mathcal{X}} \varphi \mathrm{d}(\mu_n - \mu_N) \\ &\geq \mu_N(U) - \frac{1}{\varepsilon} W_1(\mu_n, \mu_N) \geq \mu_N(K) - \varepsilon > 1 - 2\varepsilon.\end{aligned}$$

Isso não implica a compacidade da sequência $(\mu_n)_{n\in\mathbb{N}}$ pois o conjunto U_ε não é compacto. No entanto, dado $\varepsilon > 0$ podemos repetir esse mesmo argumento e obter uma sequência de pontos $(x_i)_{i\in\mathbb{N}}$ tal que para todo $k \in \mathbb{N}$ e todo $n \geq N$ tenhamos que

$$\mu_n \left(\mathcal{X} \setminus \bigcup_{i=1}^{N(k)} B(x_i, 2^{-k}\varepsilon) \right) < 2^{-k}\varepsilon.$$

Podemos então definir o conjunto compacto $K \stackrel{\text{def.}}{=} \bigcap_{k=1}^{\infty} \bigcup_{i=1}^{N(k)} \overline{B(x_i, 2^{-k}\varepsilon)}$. Logo, para todo $n \geq N$

$$\mu_n(\mathcal{X} \setminus K) \leq \sum_{k\in\mathbb{N}} \mu_n \left(\mathcal{X} \setminus \bigcup_{i=1}^{N(k)} B(x_i, 2^{-k}\varepsilon) \right) < \sum_{k\in\mathbb{N}} 2^{-k}\varepsilon = \varepsilon.$$

Além disso, o conjunto K é fechado, como interseção numerável de conjuntos fechados, e totalmente limitado. Logo, como em espaços métricos completos, qualquer conjunto é compacto se e somente se é completo e totalmente limitados, segue que K é compacto.

Isso implica que a sequência $(\mu_n)_{n\in\mathbb{N}}$ é tight e portanto pré-compacta na topologia estreita. Seja $\mu \in \mathscr{P}(\mathcal{X})$ um ponto de acumulação da sequência $(\mu_n)_{n\in\mathbb{N}}$ na topologia estreita. Como as distâncias de Wasserstein são semi-contínuas inferiormente nessa topologia e $(\mu_n)_{n\in\mathbb{N}}$ é de Cauchy em W_p, segue que

$$\limsup_{n\to\infty} W_p(\mu_n, \mu) \leq \limsup_{n\to\infty} \liminf_{m\to\infty} W_p(\mu_n, \mu_m) = \lim_{n,m\to\infty} W_p(\mu_n, \mu_m) = 0.$$

□

Exercício 4.1. A prova do Teorema 4.2.2 é muito mais simples quando o espaço ambiente $\mathcal{X}$ é $\mathbb{R}^d$. Por quê? Refaça essa prova nesse caso.

Para provar que $(\mathscr{P}_p(\mathcal{X}), W_p)$ é um espaço polonês, falta provar que é separável. Para isso, vamos definir o seguinte subconjunto de $\mathscr{P}(\mathcal{X})$:

$$\mathscr{D} \stackrel{\text{def.}}{=} \left\{ \sum_{i=1}^{N} a_i \delta_{x_i} : \begin{array}{c} a_i \in [0,1] \cap \mathbb{Q} \text{ para todo } i = 1, \ldots, N, \\ \sum_{i=1}^{N} a_i = 1, \quad N \in \mathbb{N} \end{array} \right\}, \tag{4.6}$$

onde $(x_i)_{i\in\mathbb{N}}$ é um subconjunto denso de $\mathcal{X}$, que existirá sempre que o espaço ambiente $\mathcal{X}$ for ele mesmo separável. Como a união enumerável de conjuntos enumeráveis é enumerável, $\mathscr{D}$ é um subconjunto enumerável de $\mathscr{P}(\mathcal{X})$.

Teorema 4.2.2. *Se o espaço $(\mathcal{X}, d_{\mathcal{X}})$ é separável, então o espaço $(\mathscr{P}_p(\mathcal{X}), W_p)$ também é separável para todo $p \geq 1$.*

Demonstração. Seja $\mu \in \mathscr{P}_p(\mathcal{X})$, pelo teorema de Ulam para todo $n \in \mathbb{N}$ existe um compacto K_n tal que

$$\int_{\mathcal{X}\setminus K_n} (1 + d_{\mathcal{X}}^p(x, x_0))\mathrm{d}\mu(x) < \frac{1}{n}.$$

Dado um conjunto $(x_i)_{i\in\mathbb{N}}$ enumerável e denso $\mathcal{X}$, pela definição de compacidade, para todo $r > 0$, existe um número finito de pontos $(x_i)_{i=1}^{N_n}$ tal que

$$K_n \subset \bigcup_{i=1}^{N_n} B_r(x_i).$$

Com um argumento clássico, podemos construir uma família finita de conjuntos disjuntos cobrindo K_n, basta tomar

$$U_1 \overset{\text{def.}}{=} B_r(x_1), \quad U_{i+1} \overset{\text{def.}}{=} U_i \setminus B_r(x_{i+1}).$$

Defina agora

$$\bar{\mu}_n \overset{\text{def.}}{=} \bar{a}_0\delta_{x_0} + \sum_{i=1}^{N_n} \bar{a}_i\delta_{x_i}, \quad \bar{a}_i \overset{\text{def.}}{=} \mu(U_i) \text{ for } 1 \leq i \leq N_n,$$

e $\bar{a}_0 = 1 - \sum_{i=1}^{N_n} \bar{a}_i$.

Desse modo, temos que

$$W_p^p(\mu, \bar{\mu}_n) \leq \int_{\mathcal{X}\setminus K_n} d_{\mathcal{X}}^p(x, x_0)\mathrm{d}\mu(x) + \sum_{i=1}^{N_n} \int_{U_i} d_{\mathcal{X}}^p(x, x_i)\mathrm{d}\mu(x) \leq \frac{1}{n} + r^p \sum_{i=1}^{N_n} \mu(U_i) \leq \frac{1}{n} + r^p.$$

Tomando $r = 1/n$, podemos então escolher $\mu_n \in \mathscr{D}$ arbitrariamente próximo de $\bar{\mu}_n$, por exemplo $W_p(\mu_n, \bar{\mu}_n) < 1/n$, com uma escolha apropriada de pesos a_i suficientemente próximos de $\bar{a}_i$. Desse modo, usando a desigualdade triangular temos que

$$W_p(\mu, \mu_n) \leq W_p(\mu, \bar{\mu}_n) + W_p(\mu_n, \bar{\mu}_n) \xrightarrow[n\to\infty]{} 0.$$

Disso temos que $\mathscr{P}_p(\mathcal{X})$ é separável na topologia induzida por W_p. □

4.3 A topologia induzida por W_p

Agora que definimos as distâncias de Wasserstein-p, queremos entender qual é a topologia que elas induzem no espaço $\mathscr{P}_p(\mathcal{X})$. Vamos provar que uma sequência $(\mu_n)_{n\in\mathbb{N}} \subset \mathscr{P}_p(\mathcal{X})$ converge para $\mu \in \mathscr{P}_p(\mathcal{X})$ na métrica W_p se, e somente se, $\mu_n \xrightharpoonup[n\to\infty]{} \mu$ e os momentos de ordem p da sequência convergem para o momento de ordem p de μ

$$M_p(\mu_n) = \int_{\mathcal{X}} d_{\mathcal{X}}^p(x, x_0)\mathrm{d}\mu_n(x) \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} d_{\mathcal{X}}^p(x, x_0)\mathrm{d}\mu(x) = M_p(\mu).$$

Teorema 4.3.1. *Seja $(\mu_n)_{n\in\mathbb{N}} \subset \mathscr{P}_p(\mathcal{X})$ e $\mu \in \mathscr{P}_p(\mathcal{X})$. Então, $W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$ se, e somente se, $\mu_n \xrightharpoonup[n\to\infty]{} \mu$ e $M_p(\mu_n) \xrightarrow[n\to\infty]{} M_p(\mu)$.*

Demonstração. Vamos primeiro provar que a convergência em W_p implica a convergência fraca e a convergência dos momentos. Fixado um ponto $x_0 \in \mathcal{X}$, note que

$$M_p(\mu_n) = \int_{\mathcal{X}} d_{\mathcal{X}}^p(x, x_0)\mathrm{d}\mu_n(x) = W_p^p(\mu_n, \delta_{x_0}).$$

Logo, a desigualdade triangular implica que

$$\left| M_p(\mu_n)^{1/p} - M_p(\mu)^{1/p} \right| = |W_p(\mu_n, \delta_{x_0}) - W_p(\delta_{x_0}, \mu)| \leq W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0,$$

o que implica a convergência dos momentos.

Para provar a convergência fraca, recorde para que uma sequência $(\mu_n)_{n\in\mathbb{N}}$ convirja fracamente para μ, é suficiente provar que $\int_{\mathcal{X}} f\mathrm{d}\mu_n \xrightarrow[n\to\infty]{} \int_{\mathcal{X}} f\mathrm{d}\mu$ para toda função Lipschitz f. Para uma sequência $(\mu_n)_{n\in\mathbb{N}}$ convergindo para μ em W_p, usando a fórmula de dualidade de Kantorovitch-Rubistein e o fato de que $W_1 \leq W_p$, temos que

$$\left| \int_{\mathcal{X}} f\mathrm{d}\mu_n - \int_{\mathcal{X}} f\mathrm{d}\mu \right| = \mathrm{Lip}(f) W_1(\mu_n, \mu) \leq \mathrm{Lip}(f) W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0.$$

A convergência fraca segue.

Para provar a afirmação conversa, trataremos primeiro o caso onde as medidas μ_n e μ são concentradas em um conjunto limitado K. Nesse caso, pela Proposição 4.1.2, as distâncias W_p e W_1 são equivalentes em $\mathscr{P}(K)$. Logo basta provar que $W_1(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$.

Seja uma sequência $(f_n)_{n\in\mathbb{N}}$ de funções 1-Lipschitz ótimas para a formulação dual de $W_1(\mu_n, mu)$. Podemos assumir sem perda de generalidade que $f_n(x_0) = 0$ para um ponto fixo $x_0 \in K$. Dessa forma, as funções f_n são 1-Lipschitz e uniformemente limitadas em K, logo pelo Teorema de Ascoli-Arzelà, existe uma subsequência $(f_{n_k})_{k\in\mathbb{N}}$ convergindo uniformemente para uma função f 1-Lipschitz.

Disso, segue que

$$\begin{aligned} W_1(\mu_{n_k}, \mu) &= \int_{\mathcal{X}} f_{n_k}\mathrm{d}\mu_{n_k} - \int_{\mathcal{X}} f_{n_k}\mathrm{d}\mu \\ &= \underbrace{\int_{\mathcal{X}} f\mathrm{d}\mu_{n_k} - \int_{\mathcal{X}} f\mathrm{d}\mu}_{\xrightarrow[k\to\infty]{} 0} + \underbrace{\int_{\mathcal{X}} (f_{n_k} - f)\mathrm{d}\mu_{n_k} - \int_{\mathcal{X}} (f_{n_k} - f)\mathrm{d}\mu}_{\leq 2\|f_{n_k} - f\|_\infty \xrightarrow[k\to\infty]{} 0}. \end{aligned}$$

O primeiro termo converge para 0 pela convergência fraca, enquanto que o segundo converge para 0 pela convergência uniforme de f_{n_k} para f. Logo, a subsequência $W_1(\mu_{n_k}, \mu)$ converge para 0. Repetindo esse argumento para qualquer subsequência de μ_n, temos que toda subsequência admite uma nova subsequência que converge para μ na distância W_1. Segue da propriedade de Urysohn que toda a sequência $W_1(\mu_n, \mu)$ converge para 0.

Usando a desigualdade $W_p(\mu_n, \mu) \leq C W_1(\mu_n, \mu)^{1/p}$, o mesmo vale para $W_p(\mu_n, \mu)$.

No caso geral, considere um ponto $x_0 \in \mathcal{X}$ e defina a sequência de medidas não negativas dada por

$$\sigma_n \overset{\text{def.}}{=} (1 + d_{\mathcal{X}}^p(\cdot, x_0))\,\mu_n, \quad \sigma \overset{\text{def.}}{=} (1 + d_{\mathcal{X}}^p(\cdot, x_0))\,\mu.$$

Pelo teorema de Portmanteau, segue que $\sigma_n \xrightarrow[n\to\infty]{} \sigma$ pois para todo conjunto aberto A, é fácil de verificar que $\sigma(A) \le \liminf_{n\to\infty} \sigma_n(A)$. Dado um $\varepsilon > 0$, tome um compacto $K \subset \mathcal{X}$ tal que $\sigma_n(\mathcal{X} \setminus K), \sigma(\mathcal{X} \setminus K) < \varepsilon$. Consideremos agora

$$\mu_{K,n} \overset{\text{def.}}{=} \mu_n \llcorner K + (1 - \mu_n(K))\delta_{x_0}, \quad \mu_K \overset{\text{def.}}{=} \mu \llcorner K + (1 - \mu(K))\delta_{x_0}.$$

Assumindo sem perdas de generalidade que $x_0 \in K$, temos que as medidas $\mu_{K,n}$ e μ_K são concentradas em K.

Também pelo teorema de Portmanteau, podemos provar que $\mu_{K,n} \xrightarrow[n\to\infty]{} \mu_K$. De fato, para todo conjunto aberto A, se $x_0 \in A$, temos que

$$\begin{aligned}\mu_K(A) &= \mu(A \cap K) + (1 - \mu(K)) = \mu(A \cup (\mathcal{X} \setminus K))\\ &\le \liminf_{n\to\infty} \mu_n(A \cup (\mathcal{X} \setminus K))\\ &= \liminf_{n\to\infty} \mu_{K,n}(A).\end{aligned}$$

Se $x_0 \notin A$, temos uma estimação análoga.

Dessa forma, aplicando a desigualdade triangular, temos que

$$W_p(\mu_n, \mu) \le W_p(\mu_n, \mu_{K,n}) + W_p(\mu_{K,n}, \mu_K) + W_p(\mu_K, \mu).$$

O termo do meio converge para 0 pela primeira parte da prova, já que as medidas são concentradas em K e $\mu_{K,n} \xrightarrow[n\to\infty]{} \mu_K$. Para estimar os outros termos, consideremos apenas $W_p(\mu, \mu_K)$, que pode ser estimado com o plano de transporte simples dado por

$$\gamma_K \overset{\text{def.}}{=} (\mathrm{id}, \mathrm{id})_\sharp(\mu \llcorner K) + (\mathrm{id}, x_0)_\sharp(\mu \llcorner \mathcal{X} \setminus K).$$

Disso, segue que

$$\begin{aligned}W_p^p(\mu, \mu_K) &\le \int_{\mathcal{X}\times\mathcal{X}} d_{\mathcal{X}}^p(x,y)\mathrm{d}\gamma_K(x,y) = \int_{\mathcal{X}\setminus K} d_{\mathcal{X}}^p(x, x_0)\mathrm{d}\mu(x)\\ &\le \sigma(\mathcal{X} \setminus K) < \varepsilon.\end{aligned}$$

Concluímos então que

$$\limsup_{n\to\infty} W_p(\mu_n, \mu) \le 2\varepsilon + \limsup_{n\to\infty} W_p(\mu_{K,n}, \mu_K) = 2\varepsilon.$$

Como $\varepsilon > 0$ é arbitrário, a prova está concluída. □

Observação 4.3.1. Note na prova anterior que quando as medidas são concentradas em um conjunto limitado, a convergência fraca já implica a convergência em W_p. Isso se explica pelo fato de que $d_{\mathcal{X}}^p(\cdot, x_0) \in \mathscr{C}_b(K)$.

Exercício 4.2. Assim como no Exercício 4.1, a prova do Teorema 4.3.1 é muito mais simples quando o espaço ambiente $\mathcal{X}$ é $\mathbb{R}^d$. Por quê? Refaça essa prova nesse caso.

4.4 A lei dos grandes números de Glivenko-Cantelli

Com os resultados que já conhecíamos sobre a convergência estreita de medidas de probabilidade, podemos demonstrar a lei dos Grandes Números de Glivenko-Cantelli para medidas empíricas, também conhecido como o Teorema fundamental da estatística, que incrementa a lei forte dos grandes números para variáveis aleatórias reais. Por outro lado, usando a caracterização de convergência na distância W_p, também podemos provar convergência em W_p com probabilidade 1.

Primeiramente, estabelecemos o seguinte Lema, que fornece um conjunto enumerável de funções teste, que implica a convergência estreita. Do mesmo jeito que diminuímos o conjunto $\mathscr{C}_b(\mathcal{X})$ para o conjunto $\mathrm{Lip}_b(\mathcal{X})$ para verificar a convergência estreita de uma sequência de medidas de probabilidade, podemos diminuir esse conjunto de funções teste ainda mais para um conjunto enumerável.

Lema 4.4.1 (Separador enumerável para a topologia estreita)**.** *Seja $\mathcal{X}$ um espaço polonês. Então existe uma família enumerável de funções Lipschitz $\mathcal{F} \subset \mathrm{Lip}(\mathcal{X})$ tal que uma sequência $(\mu_n)_{n\in\mathbb{N}} \xrightarrow[n\to\infty]{} \mu$ se, e somente se*

$$\int_{\mathcal{X}} f \mathrm{d}\mu_n \to \int_{\mathcal{X}} f \mathrm{d}\mu \text{ para toda } f \in \mathcal{F}.$$

Demonstração. Podemos construir o conjunto $\mathcal{F}$ com funções construídas como: sejam $\mathscr{D}$ um subconjunto denso e enumerável de $\mathcal{X}$ e f da forma

$$f(x) = \inf\left\{q_i + p_i d_{\mathcal{X}}(x,y) \wedge 1 : \begin{matrix} q_i, p_i \in \mathbb{Q} \cap [-1,1] \text{ for } i = 1,\dots,N \\ y \in \mathscr{D},\ N \in \mathbb{N} \end{matrix}\right\}.$$

□

Proposição 4.4.1 (Lei dos Grandes Números de Glivenko-Cantelli para medidas empíricas)**.** *Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade, e seja $(X_i)_{i\in\mathbb{N}}$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas, com valores em $\mathcal{X}$ e lei $\mu \in \mathscr{P}(\mathcal{X})$. Defina a medida empírica*

$$\mu_N \stackrel{\text{def.}}{=} \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}.$$

Então, $W_p(\mu_n, \mu) \xrightarrow[n\to\infty]{} 0$ $\mathbb{P}$-quase certamente.

Demonstração. Pelo resultado anterior, existe um conjunto enumerável de funções $\mathcal{F}$ para o qual verificar a convergência das integrais já garante a convergência estreita das medidas. Assim, para cada $f \in \mathcal{F}$, temos que $(f(X_i))_{i\in\mathbb{N}}$ é uma sequência i.i.d. de variáveis aleatórias. Por outro lado

$$\int_{\mathcal{X}} f \mathrm{d}\mu_N = \frac{1}{N}\sum_{i=1}^{N} f(X_i)$$

é a média de variáveis independentes e identicamente distribuídas com esperança $\mathbb{E}[f(X_1)] = \int_{\mathcal{X}} f \mathrm{d}\mu$. Pela Lei Forte dos Grandes Números, para cada $f \in \mathcal{F}$, existe um conjunto Ω_f de $\mathbb{P}$-probabilidade 1 onde

$$\frac{1}{N}\sum_{i=1}^{N} f(X_i(\omega)) \xrightarrow[N\to\infty]{} \int_{\mathcal{X}} f \mathrm{d}\mu \text{ para todo } \omega \in \Omega_f.$$

Como $\mathcal{F}$ é enumerável, o conjunto $\Omega \stackrel{\text{def.}}{=} \bigcap_{f \in \mathcal{F}} \Omega_f$ tem probabilidade 1, portanto a convergência quase certamente vale simultaneamente para todo $f \in \mathcal{F}$. Pelo resultado anterior, isso implica $\mu_N \xrightarrow[N \to \infty]{} \mu$ quase certamente.

Similarmente, a função $x \mapsto d^p_{\mathcal{X}}(x, x_0) \in L^1(\mu)$, já que $\mu \in \mathscr{P}_p(\mathcal{X})$, logo também segue da lei dos grandes números que

$$M_p(\mu_N) = \frac{1}{N} \sum_{i=1}^{N} d^p_{\mathcal{X}}(X_i, x_0) \xrightarrow[N \to \infty]{} \mathbb{E}\left[d^p_{\mathcal{X}}(X_1, x_0)\right] = M_p(\mu)$$

em um conjunto de $\mathbb{P}$-probabilidade 1.

Tomando a interseção dos conjuntos de probabilidade 1 referentes à convergência dos momentos e associados à cada $f \in \mathcal{F}$, ainda temos um conjunto de probabilidade 1, e nesse conjunto, pela caracterização de convergência na distância de Wasserstein, temos que $W_p(\mu_N, \mu) \xrightarrow[N \to +\infty]{} 0$. □