FGV EMAp

Transporte Ótimo para Redes Neurais

Autor: Davi Sales Barreira

Sumário

Ideia Geral

Transporte Ótimo (OT) é uma área da matemática que estuda o problema de transportar "massa" em uma configuração para outra enquanto se miniza o custo de transporte.

Apesar de parecer um problema bastante específico, a ideia de se transportar objetos de maneira ótima é bastante ubíqua e possui diversas utilidades.

OT tem aparecido em diversas aplicações recentes de Machine Learning, como: **transfer learning, clustering, redução de dimensionalidade, modelos generativos**, entre outros.

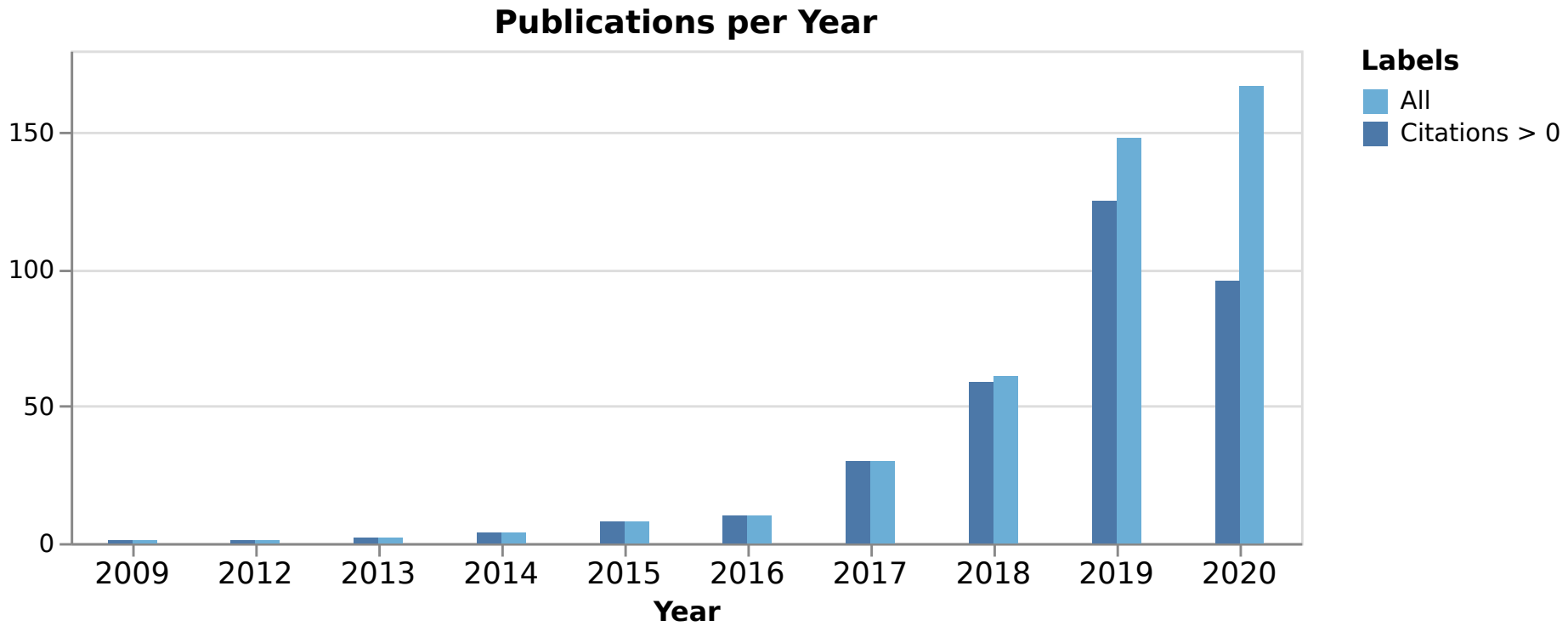

Figure 1: Gráficos com a evolução do número de publiações relacionadas a Transporte Ótimo com Machine Learning [21].

Ideia Geral

A solução de um problema de Transporte Ótimo sempre resulta em dois subprodutos, o **plano (mapa)** ótimo de transporte e o **custo mínimo** para realizar o transporte.

A maioria das aplicações em ML utiliza o custo mínimo para definir uma métrica de distância (e.g. Wasserstein). Porém, existem aplicações como Transfer Learning que utlilizam os mapas ótimos [21].

Nesta apresentação vamos focar na aplicação mais celebrada de OT em redes neurais, as chamadas **Wasserstein Generative Neural Networks** [3].

Problema de Monge - Qual a maneira ótima de transporta massa de uma configuração para outra?

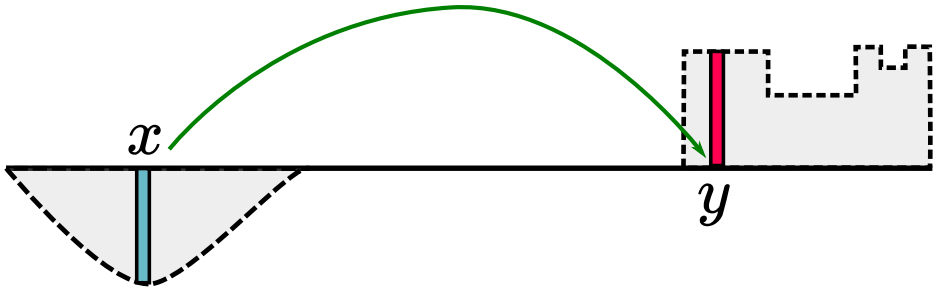

Figure 2: Massa não pode ser separada.

Kantorovich Problem - Relaxação do problema original de Monge.

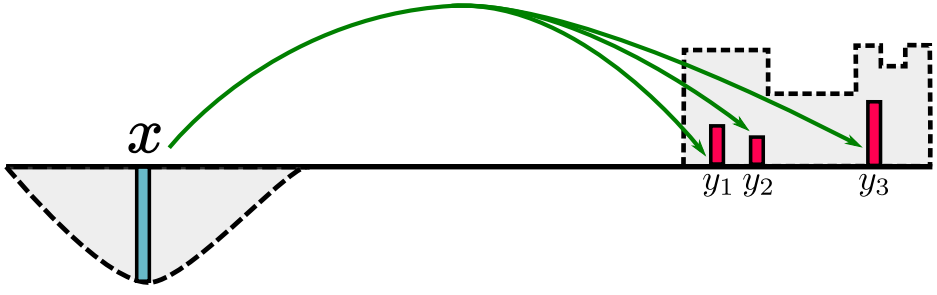

Figure 3: Massa pode ser separada.

Definition (Problema de Monge)

Dadas duas medidas de probabilidade $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ e uma função de custo $c : X \times Y \to [0, +\infty]$, resolva:

$$(MP) \qquad \inf \left\{ \int_X c(x, T(x)) d\mu \quad : \quad T_{\#}\mu = \nu \right\} \tag{1}$$

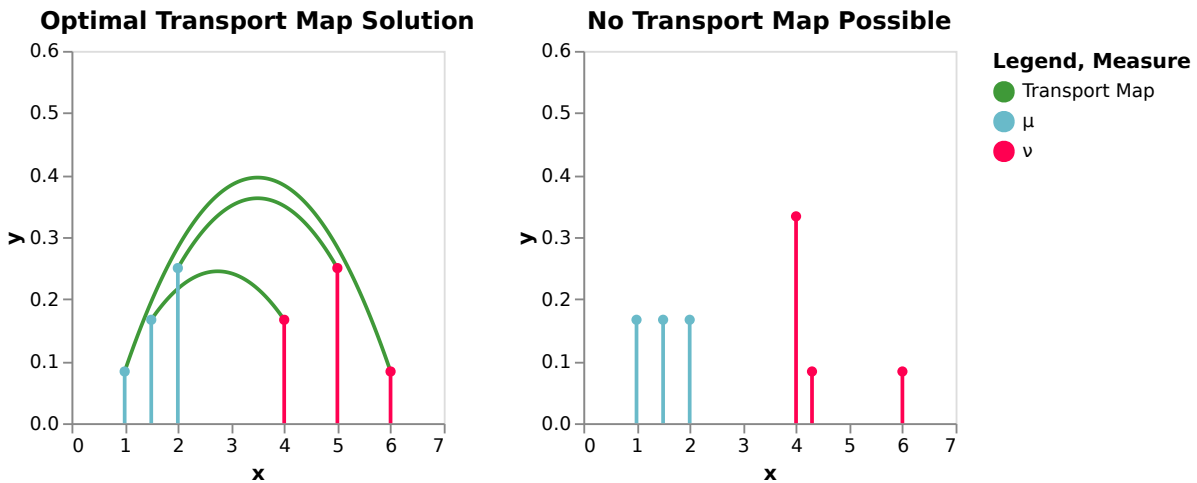

Figure 4: Exemplo de dois problemas de Transporte Ótimo.

Definition (Acoplamento (*Coupling*))

Sejam (X, μ) e (Y, ν) espaços de probabilidade. Para $\gamma \in \mathcal{P}(X \times Y)$, dizemos que γ é um acoplamento de (μ, ν) se $(\pi_X)_{\#}\gamma = \mu$ e $(\pi_Y)_{\#}\gamma = \nu$. Chamamos $\Pi(\mu, \nu)$ do conjunto de **Planos de Transporte**:

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) \ : \ (\pi_X)_{\#}\gamma = \mu \quad \text{and} \quad (\pi_Y)_{\#}\gamma = \nu\} \quad (2)$$

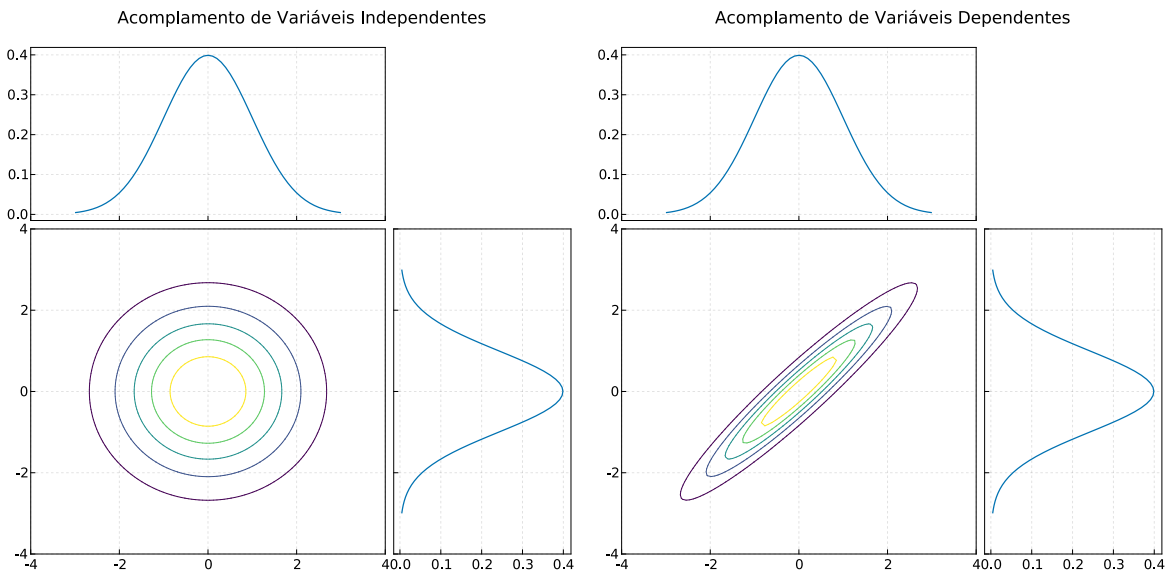

Figure 5: Exemplos de acomplamento.

Definition (Problema de Kantorovich)

Dadas duas medidas de probabilidade $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ e a função de custo $c : X \times Y \to [0, +\infty]$, resolva:

$$(KP) \qquad \inf \left\{ \int_{X \times Y} c(x, y) d\gamma \ : \ \gamma \in \Pi(\mu, \nu) \right\} \tag{3}$$

O Problema de Kantorovich tem uma formulação dual, que para certas condições de regularidade possui a mesma solução ótima que o problema primal (dualidade forte).

Definition (Problema Dual)

Dadas $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ e custo $c : X \times Y \to \mathbb{R}_+$. O Problema Dual é

$$(\mathrm{DP}) \quad \sup \left\{ \int_X \phi \, d\mu + \int_Y \psi \, d\nu \ : \phi \in C_b(X) \, , \psi \in C_b(Y) \, , \ \phi \oplus \psi \leq c \right\} \tag{4}$$

Funções ϕ, ψ são chamdas de **Potenciais de Kantorovich**. Essa formulação é utilizada nas Wasserstein GANs.

Definition (Distância de Wasserstein)

Seja (X, d) um espaço métrico polonês, com $c : X \times X \to \mathbb{R}$ tal que $c(x, y) = d(x, y)^p$, e $p \in [1, +\infty)$. Para $\mu, \nu \in \mathcal{P}_p(X)$, a distância de Wasserstein é dada por:

$$W_p(\mu, \nu) := \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times X} d(x, y)^p \, d\gamma \right)^{1/p} \tag{5}$$

$\mathcal{P}_p(X)$ é o espaço de medidas de probabilidade com p-ésimo momento.

A distância de Wasserstein preserva a geometria do espaço no qual está definida a medida.

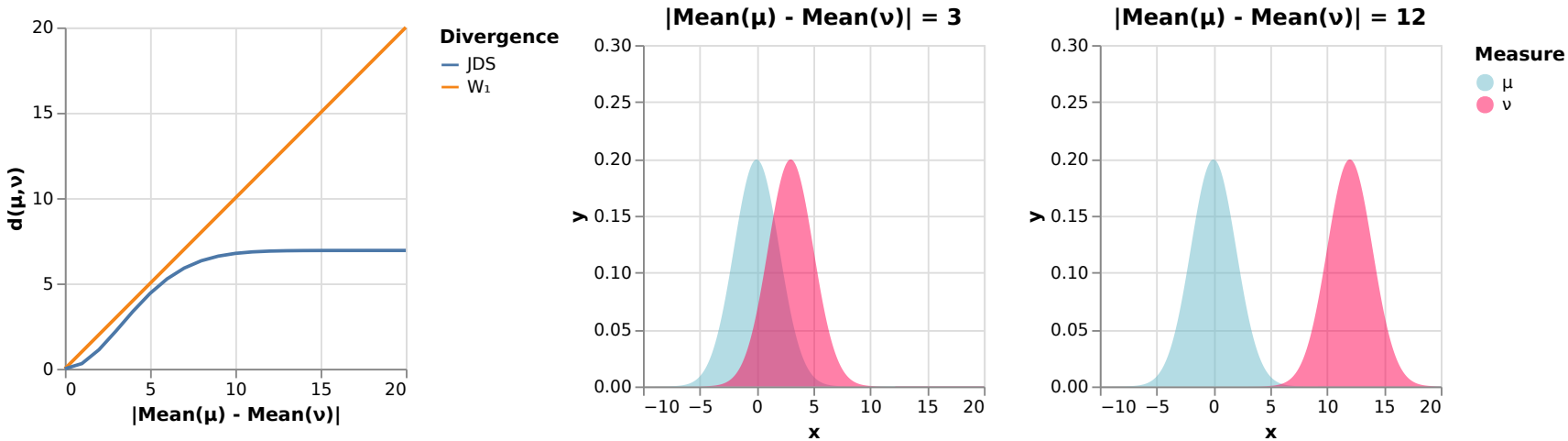

Figure 6: Comparação entre a distância de Wasserstein e Jensen-Shannon.

Dadas duas medidas de probabilidade μ e ν, como então computamos a distância de Wasserstein entre elas?

Calcular a distância de Wasserstein em espaços de alta dimensão (e.g. imagens) é bastante custoso.

Assim, variações da distância de Wasserstein foram desenvolvidas, como por exemplo:

1. Entropic Wasserstein;
2. Sliced Wasserstein.

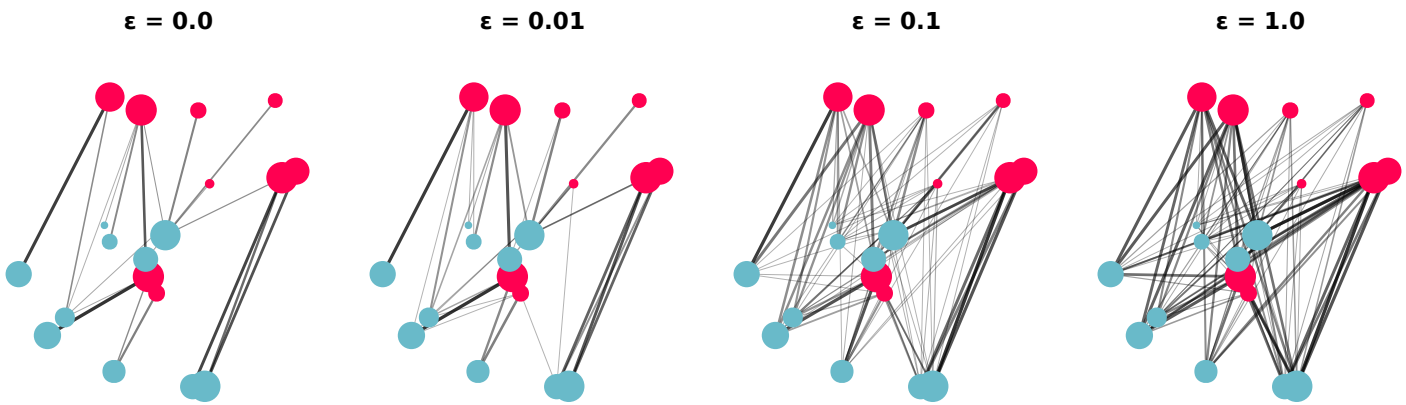

Figure 7: Exemplo de solução de OT com regularização entrópica [21].

Generative Adversarial Networks (GAN) foram originalmente introduzidas por Goodfellow et al. [13]. Essas redes são utilizadas com o objetivo de gerar dados sintéticos realísticos a partir de dados reais.

Figure 8: Faces geradas por GANs [1].

A ideia geral por trás das GANs é utilizar duas redes neurais competindo uma com a outra, sendo uma rede responsável por gerar amostras parecidas com os dados reais (*gerador*) , enquanto a outra busca identificar quando o dado é real ou sintético (*descriminador*).

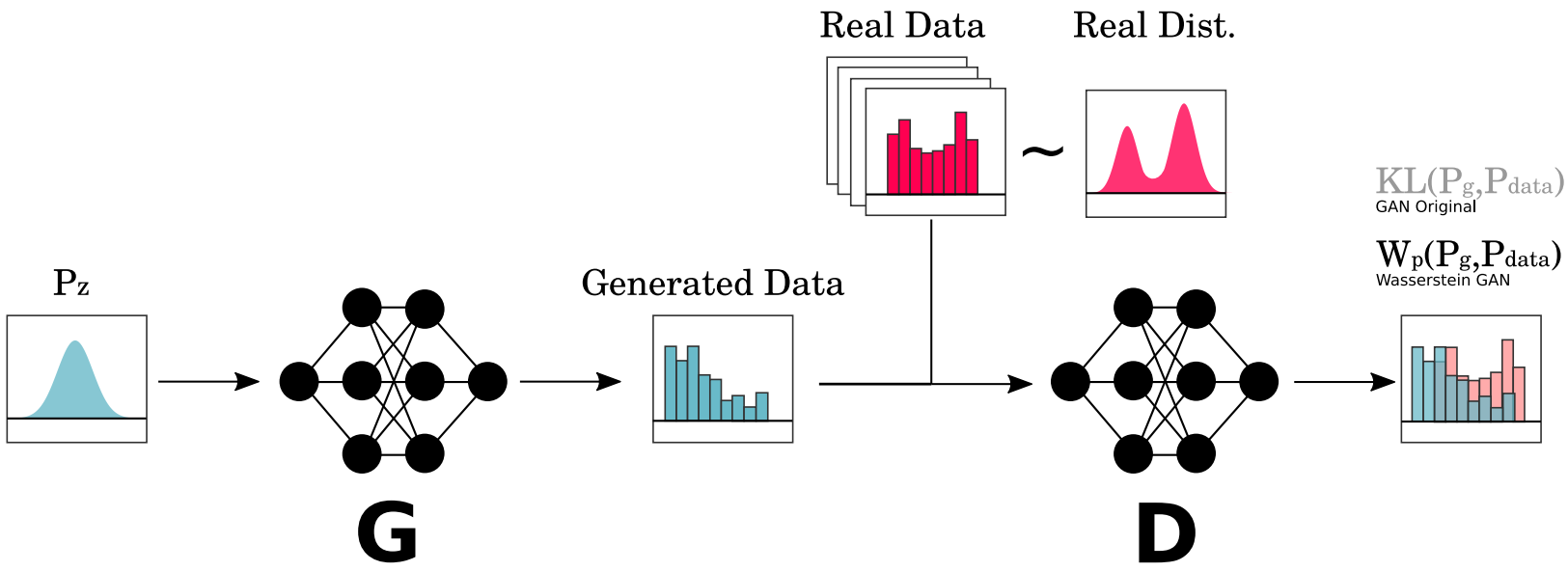

Figure 9: Generative Adversarial Network [21].

References I

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[2] Luigi Ambrosio, Elia Brué, and Daniele Semola. Lectures on optimal transport, 2021.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[4] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[5] R.I. Bot. *Conjugate duality in convex optimization*, volume 637. Springer Science & Business Media, 2009.

[6] G. Carlier and C. Poon. On the total variation wasserstein gradient flow and the tv-jko scheme. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:42, 2019.

[7] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[8] Bharath Bhushan Damodaran, Benjamin Kellenberger, R émi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[9] L.C. Evans and R.F. Gariepy. *Measure theory and fine properties of functions*. CRC press, 2015.

[10] Rémi FLAMARY. *Transport optimal pour l'apprentissage statistique*. PhD thesis, Télécom Paristech, 2019.

[11] Rémi Flamary. Optimal transport for machine learning. page 97, November 2019.

[12] David JH Garling. *Analysis on Polish spaces and an introduction to optimal transportation*, volume 89. Cambridge University Press, 2018.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[15] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

References II

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018.

[17] Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi. Wasserstein training of restricted boltzmann machines. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3718–3726. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/728f206c2a01bf572b5940d7d9a8fa4c-Paper.pdf`.

[18] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[19] R.T. Rockafellar. *Conjugate duality and optimization*. SIAM, 1974.

[20] R. Rossi and G. Savaré. Tightness, integral equicontinuity and compactness for evolution problems in banach spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 2(2):395–431, 2003.

[21] Davi Sales Barreira. *Optimal Transport for Machine Learning: Theory and Applications*. PhD thesis, 2021.

[22] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[23] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1): 87–154, 2017.

[24] user125646 (https://math.stackexchange.com/users/125646/user125646). How to show that the set of all lipschitz functions on a compact set x is dense in c(x)? Mathematics Stack Exchange. URL `https://math.stackexchange.com/q/665686`. URL:https://math.stackexchange.com/q/665686 (version: 2014-02-07).

[25] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[26] Larry Wasserman. Statistical methods for machine learning - lecture notes, 2018.