EDPs via Fluxo de Gradiente em Espaços de Wasserstein

Autor: Davi Sales Barreira

Sumário

Ideia Geral

O espaço de Wasserstein se trata de um espaço métrico de medidas de probabilidade embutido com a métrica de Wasserstein.

Um Fluxo de Gradiente é um sistema de equações onde a evolução do sistema se dá através da descida de gradiente.

A ideia geral dessa apresentação é mostrar como algumas EDPs podem ser reformuladas em termos de um Fluxo de Gradiente em um espaço de Wasserstein. Apresentaremos como reformular a equação de calor, porém, esse método é mais geral, sendo aplicável para muitas outras EDPs.

Por que interpretar EDPs como Fluxo de Gradiente em Wasserstein?

1. Estética. Veremos que é uma bela interpretação que permite entender as EDPs de outro ponto de vista;
2. Reformulação permite utilizar outros ferramentais para demonstrar, por exemplo, taxas de convergência, existência e unicidade;
3. Esquema de discretização de fluxos de gradiente como algoritmo para aproximar soluções fracas para as EDPs.

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

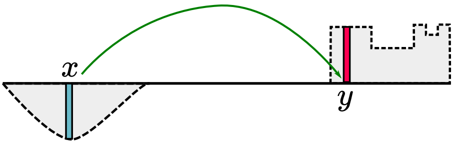

Figure 1: Massa não pode ser separada.

Kantorovich Problem - Relaxação do problema original de Monge.

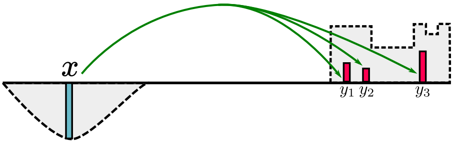

Figure 2: Massa pode ser separada.

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:

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

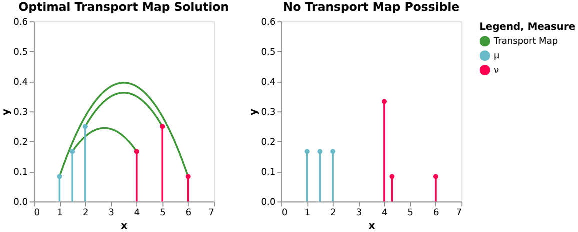

Figure 3: Exemplo de dois problemas de Transporte Ótimo.

Definition (Acoplamento)

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**:

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

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

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

Funções ϕ, ψ são chamdas de **Potenciais de Kantorovich**.

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.

É possível mostrar que a distância de Wasserstein W_p é de fato uma métrica no espaço de probabilidade $\mathcal{P}_p(\Omega)$.
Além disso, ela metriza a convergência fraca de medidas de probabilidade, i.e., sejam $\mu_n, \mu \in \mathcal{P}_p(\mathbb{R}^n)$, então

$$\mu_n \rightharpoonup \mu \iff W_p(\mu_n, \mu) \to 0. \tag{6}$$

Outro aspecto que temos que abordar são as condições de existência das soluções.

Theorem (Existência de Planos de Transporte)

Sejam X e Y espaços métricos poloneses (complete and separável). Dados $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ e $c : X \times Y \to [0, +\infty]$, se c for inferiormente semi-continua, então (KP) possui solução.

Theorem (Existência de Mapas de Transporte)

Seja $\Omega \subset \mathbb{R}^n$ compacto, e $c(x, y) = h(x - y)$ com h estritamente convexa. Dado $\mu \ll \lambda$, e $\partial\Omega$ negligenciável. Então, existe solução para o problema de transporte de Monge, e além disso

$$T(x) = x - (\nabla h)^{-1}(\nabla \phi(x)). \tag{7}$$

Onde $\phi(x)$ é o potencial de Kantorovich e T é o mapa de transporte ótimo.

O problema original de Transporte Ótimo é formulado de forma que o transporte ocorre de forma "instantânea", porém, é possível partir de premissas mais fundamentais, e encontrar uma formulação dinâmica para o problema, onde a solução não é mais um plano, mas sim uma curva $\gamma : [0, 1] \to \mathcal{P}(\Omega)$, com $\gamma(0) = \mu$ e $\gamma(1) = \nu$.

Baseado nessa formulação dinâmica, é possível provar o seguinte resultado:

Theorem

(Benamou-Brenier Formula) Sejam $\rho_0, \rho_1 \in \mathcal{P}_p(\Omega)$ para $p > 1$. Então, a seguinte caracterização é válida para espaços com a distância p-Wasserstein:

$$\frac{1}{p} W_p^p(\rho_0, \rho_1) = \inf_{(\rho, v)} \left\{ \int_0^1 \int_\Omega |v(t,x)|^p \, d\rho_t(x) dt : \textit{Condições B.B} \right\}.$$

Onde as condições de Benamou-Brenier são

$$(\rho_t)_{t\in[0,1]} \textit{ Absolutamente Contínua no espaço de Wasserstein,}$$
$$(\rho_t, v_t)_{t\in[0,1]} \textit{é solução fraca de } \partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0$$
$$\rho(0,\cdot) = \rho_0, \; \rho(1,\cdot) = \rho_1.$$

Seja uma função $F : \mathbb{R}^n \to \mathbb{R} \in C^1$, e $x_0 \in \mathbb{R}^n$, onde queremos descobrir $x(t)$ que resolve o seguinte sistema de equações:

$$\begin{cases} x'(t) = -\nabla F(x(t)), \ t > 0, \\ x(0) = x_0. \end{cases} \tag{8}$$

A solução $x(t)$ do sistema acima será uma curva iniciando em x_0 e se movendo na direção de menor gradiente, ou seja, a solução é dada pelo famoso algoritmo de descida de gradiente. Em outras palavras, a solução $x(t)$ caracteriza um fluxo de gradiente.

Esse problema é simples quando estamos em espaços de dimensão finita e com funções diferenciáveis, porém, torna-se mais interessante e complexo quando começamos a considerar espaços de dimensão infinita como $\mathcal{P}_2(\mathbb{R}^n)$. Neste cenário, temos que repensar, por exemplo, a ideia de gradiente, já que não está mais claro que seria o gradiente quando $x(t) = \rho_t \in \mathcal{P}_2(\mathbb{R}^n)$. Além disso, F não é mais uma função de $\mathbb{R}^n$ em $\mathbb{R}$, mas um funcional atuando em medidas de probabilidade.

Dadas condições sob F, é possível provar, por exemplo, que as soluções são únicas.

Theorem

Seja $F : \mathbb{R}^n \to \mathbb{R}$ **convexa**, $x_0 \in \mathbb{R}^n$, *e* x_1 *e* x_2 *duas soluções do fluxo de gradiente. Então,*

$$|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|, \ \forall t > 0. \tag{9}$$

Logo, a solução do sistema de equações é única.

É possível obter resultados para condições menos restritivas que convexidade em F, como, por exemplo, usando λ-convexidade. Além disso, também podemos trocar a condição de ∇F por ∂F (subdiferencial), tendo assim ainda mais generalidade.

Outra propriedade relevante dos fluxos de gradiente é que eles podem ser caracterizados por meio do chamado *Esquema de Minimização de Movimento*, sendo resolvidos por meio de uma discretização temporal.

O *Esquema de Minimização de Movimento* é definido pela seguinte iteração:

$$x_{k+1}^{\tau} \in \mathsf{argmin}_x F(x) + \frac{|x - x_k^{\tau}|^2}{2\tau}. \tag{10}$$

A primeira vista, o esquema de minimização acima pode parecer contra-intuitivo, entretanto, supondo que F é derivável, sabemos que a solução de (10) é obtida quando $\nabla(F(x) + \frac{|x - x_k^\tau|^2}{2\tau}) = 0$, assim,

$$-\nabla F(x_{k+1}^\tau) = \frac{x_{k+1}^\tau - x_k^\tau}{\tau}. \tag{11}$$

Ou seja, esse esquema de minimização é o famoso esquema implícito de Euler. Lembre-se da diferença entre o esquema implícito e o explícito de Euler:

(Euler Implícito)
$$x_{k+1}^\tau = x_k^\tau - \tau \nabla F(x_{k+1}^\tau) \tag{12}$$

(Euler Explicito)
$$x_{k+1}^\tau = x_k^\tau - \tau \nabla F(x_k^\tau) \tag{13}$$

É possível provar que para $\tau \to 0$, podemos interpolar os pontos (x_k^τ) para obter uma solução que converge para a solução $x(t)$ de (8).

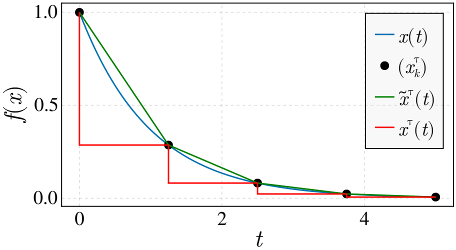

Figure 4: Exemplo de aproximação da solução $x(t)$.

Queremos agora estender essa ideia do fluxo de gradiente para espaços os espaços de 2-Wasserstein, e vamos assumir também que nossas medidas de probabilidade são absolutamente contínuas em relação a Lebesgue. Vamos assim alterar nosso Esquema de Minimização de Movimento para

$$\rho_{k+1}^{\tau} \in \mathrm{argmin}_{\rho} F(\rho) + \frac{W_2^2(\rho, \rho_k^{\tau})}{2\tau}, \tag{14}$$

Onde F agora não é mais uma função de $\mathbb{R}^n \to \mathbb{R}$, mas sim um funcional $F : P_2(\Omega) \to \mathbb{R}$, e.g. $F(\rho) = \int_{\Omega} f(\rho(x))dx$.

$$\rho_{k+1}^{\tau} \in \text{argmin}_{\rho} F(\rho) + \frac{W_2^2(\rho, \rho_k^{\tau})}{2\tau} \tag{15}$$

O problem de minimização acima é em um espaço de funções, onde buscamos a medida de probabilidade ρ que minimiza. Assim, utilizaremos a ideia de **primeira variação** proveniente do Cálculo de Variações. Seja $G : \mathcal{P}(\Omega) \to \mathbb{R}$ um funcional, chamaremos $\frac{\delta G}{\delta \rho}(\rho)$ a primeira variação de G, caso exista uma função única (a menos de uma constante), tal que

$$\frac{d}{d\varepsilon} G(\rho + \varepsilon\chi)|_{\varepsilon=0} = \int \frac{\delta G}{\delta \rho}(\rho) d\chi, \tag{16}$$

para toda perturbação χ. Note a perturbação deve satisfazer $\int d\chi = 0$ e além disso, deve existir pelo menos $\varepsilon \in [0, \varepsilon_0]$, tal que $\rho + \varepsilon\chi \in \mathcal{P}(\Omega)$.

Análogamente ao caso em $\mathbb{R}$, se $\frac{\delta G}{\delta}(\rho^*) = 0$(ou constante), temos assim uma possível solução para o problema de otimização. Se provamos, por exemplo, que nosso funcional é contínuo em algum sentido, como em convergência fraca de probabilidade, e que é convexo. Teremos então existência e unicidade para esse problema de otimização, com ρ^* sendo a função que minimiza.

Se nosso funcional é $F(\rho) = \int_\Omega f(\rho(x))dx$, onde $f : \mathbb{R} \to \mathbb{R}$ é convexa e superlinear, teremos que

$$\frac{\delta F}{\delta \rho}(\rho) = f'(\rho) \tag{17}$$

.

Além disso, podemos provar que para o funcional $\rho \mapsto W_2(\rho, \nu)$, temos que

$$\frac{\delta W_2(\cdot, \nu)}{\delta \rho}(\rho_0) = \phi. \tag{18}$$

Lembrando que queremos minimiza a equação abaixo

$$\rho_{k+1}^{\tau} \in \text{argmin}_{\rho} F(\rho) + \frac{W_2^2(\rho, \rho_k^{\tau})}{2\tau}. \quad (19)$$

Temos então que no ponto de mínimo

$$\frac{\delta F}{\delta \rho}(\rho_{k+1}^{\tau}) + \frac{\phi}{\tau} = \text{const.} \quad (20)$$

.

Para $\mu \ll \lambda$, com Ω compacto, temos que nosso espaço 2-Wasserstein tem sempre um mapa $T(x) = x - \nabla\phi(x)$ (apresentamos na sessão de existência). Logo

$$-\boldsymbol{v}(x) := \frac{T(x) - x}{\tau} = -\frac{\phi(x)}{\tau} = \nabla(\frac{\delta F}{\delta \rho}(\rho))(x). \quad (21)$$

Pela Equação de Benamou-Brenier, temos então que

$$\partial_t \rho - \nabla \cdot (\rho \nabla(\frac{\delta F}{\delta \rho}(\rho))(x) = 0. \quad (22)$$

Finalmente, chegamos onde queríamos desde o começo.
Se nosso funcional é $F(\rho) = \int_\Omega f(\rho(x))dx$, onde $f : \mathbb{R} \to \mathbb{R}$ é convexa e superlinear, teremos que

$$\frac{\delta F}{\delta \rho}(\rho) = f'(\rho) \tag{23}$$

.
Mais ainda, se $f(t) = t \log t$, temos que $f'(t) = 1 + \log t$ e que $\nabla(f'(\rho)) = \frac{\nabla \rho}{\rho}$.
E assim, chegamos na equação do calor:

$$\partial_t \rho - \nabla \cdot (\rho \nabla (\frac{\delta F}{\delta \rho}(\rho))(x) = \partial_t \rho - \Delta \rho - 0. \tag{24}$$

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

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

[5] 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.

[6] 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.

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

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

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

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

[11] 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.

[12] 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`.

[13] 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.

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

[15] 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.

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

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

[18] 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).

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

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