The Monge Gap: A Regularizer to Learn All Transport Maps

Theo Uscidda Marco Cuturi
presenter: Shen Yuan

中国人民大學 RENMIN UNIVERSITY OF CHINA 高瓴人工智能学院 Gaoling School of Artificial Intelligence

Outline

Outline

Monge formulation

Given a compact subset $\Omega \subset \mathbb{R}^d$, a continuous cost function $c : \Omega \times \Omega \to \mathbb{R}$ and two probability distributions $\mu, \nu \in \mathcal{P}(\Omega)$, the Monge problem is to find $T : \Omega \to \Omega$ that push-forward μ onto ν, which minimizes the averaged displacement cost:

$$W_c(\mu, \nu) := \inf_{T\sharp\mu=\nu} \int_\Omega c(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\mu(\boldsymbol{x}) \quad (1)$$

c-OT means any solution to 1 between μ and ν.

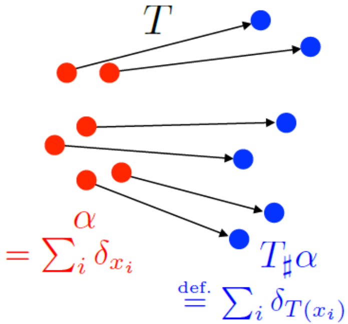

Figure 1: Push-forward of measures.

Kantorovich formulation

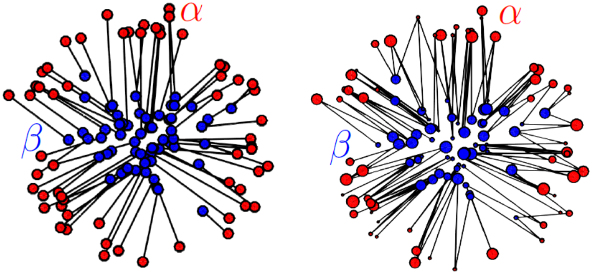

Figure 2: Comparison of transport maps and generic couplings.

Kantorovich formulation

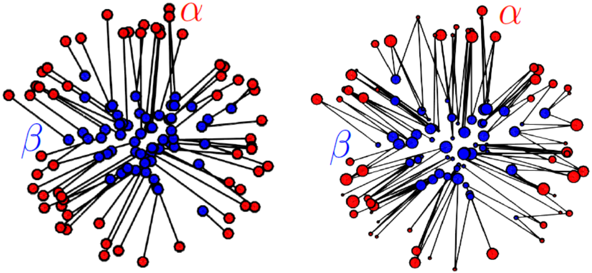

Figure 2: Comparison of transport maps and generic couplings.

$$W_c(\mu, \nu) := \min_{\pi \in \prod(\mu,\nu)} \int \int_{\Omega \times \Omega} c(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\pi(\boldsymbol{x}, \boldsymbol{y}) \tag{2}$$

An optimal coupling $\pi^\star = (\mathrm{Id}, T^\star)\sharp\mu$ always exists.

Entropic regularization

For empirical measures $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}, \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^n \delta_{\boldsymbol{y}_j}$ and $\varepsilon > 0$, $\boldsymbol{C} = [c(\boldsymbol{x}_i, \boldsymbol{y}_j)]_{ij}$,

$$W_{c,\varepsilon}(\hat{\mu}_n, \hat{\nu}_n) := \min_{\boldsymbol{P} \in U_n} \langle \boldsymbol{P}, \boldsymbol{C} \rangle - \varepsilon H(\boldsymbol{P}) \tag{3}$$

where $U_n = \{\boldsymbol{P} \in \mathbb{R}_+^{n \times n}, \boldsymbol{P}\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n, \boldsymbol{P}^T\mathbf{1}_n = \frac{1}{n}\mathbf{1}_n\}$ and $H(\boldsymbol{P}) = -\sum_{i,j=1}^n \boldsymbol{P}_{ij}\log(\boldsymbol{P}_{ij})$.

Outline

The Monge Gap

Given a cost c and a reference measure $\rho \in \mathcal{P}$, the Monge gap of a vector field $T : \Omega \to \Omega$ is defined as:

$$\mathcal{M}_\rho^c(T) := \int_\Omega \mathrm{c}(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\rho(\boldsymbol{x}) - W_\mathrm{c}(\rho, T\sharp\rho) \tag{4}$$

The Monge Gap

Given a cost c and a reference measure $\rho \in \mathcal{P}$, the Monge gap of a vector field $T : \Omega \to \Omega$ is defined as:

$$\mathcal{M}_\rho^c(T) := \int_\Omega \mathrm{c}(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\rho(\boldsymbol{x}) - W_\mathrm{c}(\rho, T\sharp\rho) \tag{4}$$

- For any vector field T, $\mathcal{M}_\rho^c(T) \geq 0$.
- T is a c-OT map between ρ and $T\sharp\rho \Leftrightarrow \mathcal{M}_\rho^c(T) = 0$

The Monge Gap

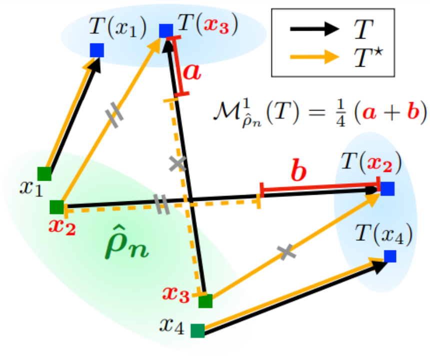

Figure 1. Sketch of the Monge Gap $\mathcal{M}^1_{\hat{\rho}_n}(T)$ instantiated with the euclidean cost $c(\cdot,\cdot) = \|\cdot - \cdot\|_2$, where $\hat{\rho}_n$ is a discrete measure supported on four points. Because the OT map $T^\star$ between $\hat{\rho}_n$ and $T\sharp\hat{\rho}_n$ does not coincide with T (notably on on points x_2, x_3), the Monge gap here is positive, and equal to differences in lengths that amount to $(a + b)/4$ in the plot.

Consistency of the Monge Gap

Lemma 3.2(Consistency). Given empirical measures $\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}$, provided that T is continuous, it almost surely holds

$$\lim_{n \to + \inf} \mathcal{M}^c_{\hat{\rho}_n}(T) = \mathcal{M}^c_{\rho}(T) \tag{5}$$

Consistency of the Monge Gap

Lemma 3.2(Consistency). Given empirical measures $\hat{\rho}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{\boldsymbol{x}_i}$, provided that T is continuous, it almost surely holds

$$\lim_{n \to +\inf} \mathcal{M}^c_{\hat{\rho}_n}(T) = \mathcal{M}^c_{\rho}(T) \tag{5}$$

$$\begin{aligned} \mathcal{M}^c_{\rho}(T) &:= \int_{\Omega} c(\boldsymbol{x}, T(\boldsymbol{x}))\mathrm{d}\rho(\boldsymbol{x}) - W_c(\rho, T\sharp\rho) \\ \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T) &:= \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{x}_i, T(\boldsymbol{x}_i)) - W_{c,\varepsilon}(\hat{\rho}_n, T\sharp\hat{\rho}_n) \end{aligned} \tag{6}$$

Relation to Cyclical Monotonicity

For any $n \in \mathbb{N}$, any set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \times \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} \subset \Gamma$ and permutation $\sigma \in \mathcal{S}_n$, a set $\Gamma \subset \Omega \times \Omega$ is c-CM(Cyclical Monotonicity) if

$$\sum_{i=1}^{n} c(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq \sum_{i=1}^{n} c(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) \tag{7}$$

Relation to Cyclical Monotonicity

For any $n \in \mathbb{N}$, any set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \times \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\} \subset \Gamma$ and permutation $\sigma \in \mathcal{S}_n$, a set $\Gamma \subset \Omega \times \Omega$ is c-CM(Cyclical Monotonicity) if

$$\sum_{i=1}^{n} c(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq \sum_{i=1}^{n} c(\boldsymbol{x}_i, \boldsymbol{y}_{\sigma(i)}) \tag{7}$$

With $\boldsymbol{y}_i := T(\boldsymbol{x}_i)$, the Monge gap estimator using permutations is:

$$\mathcal{M}^c_{\hat{\rho}_n}(T) = \frac{1}{n} \sum_{i=1}^{n} c(\boldsymbol{x}_i, T(\boldsymbol{x}_i)) - \min_{\sigma \in \mathcal{S}_n} \frac{1}{n} \sum_{i=1}^{n} c(\boldsymbol{x}_i, T(\boldsymbol{x}_{\sigma(i)})) \tag{8}$$

The cyclical monotonicity of that set is equivalent to the optimality of T.

Properties of the Monge Gap

Proposition 3.3. Let $\mu, \nu \in \mathcal{P}(\Omega)$ such that $\mathrm{Spt}(\mu) \subset \mathrm{Spt}(\rho)$, and a map T s.t. $T\sharp\mu = \nu$. Then $\mathcal{M}_\rho^c(T) = 0$ implies that T is a c-OT map between μ and ν.

Outline

Using directly the Monge gap as a regularizer

Given a loss function defined through a divergence Δ, a regularization weight $\lambda_{\mathrm{MG}} > 0$ is introduced:

$$\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta) := \Delta(T_\theta \sharp \mu, \nu) + \lambda_{\mathrm{MG}} \mathcal{M}_\rho^c(T_\theta) \tag{9}$$

Gradient of Monge Gap

According to the Danskin (1967) Theorems, $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}$ is differentiable and its gradient reads:

$$\nabla_\theta \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta) = \sum_{i,j=1}^{n} \left(\frac{1}{n}\delta_{ij} - \boldsymbol{P}^\varepsilon_{ij}\right) \nabla_\theta c(\boldsymbol{x}_i, T_\theta(\boldsymbol{x}_j)) \tag{10}$$

Gradient of Monge Gap

According to the Danskin (1967) Theorems, $\mathcal{M}^c_{\hat{\rho}_n,\varepsilon}$ is differentiable and its gradient reads:

$$\nabla_\theta \mathcal{M}^c_{\hat{\rho}_n,\varepsilon}(T_\theta) = \sum_{i,j=1}^{n} (\frac{1}{n}\delta_{ij} - \boldsymbol{P}^\varepsilon_{ij}) \nabla_\theta c(\boldsymbol{x}_i, T_\theta(\boldsymbol{x}_j)) \tag{10}$$

Since $\boldsymbol{P}^\varepsilon \in U_n, \forall i, j, 0 \leq \boldsymbol{P}^\varepsilon_{ij} \leq 1/n$, so:

$$\begin{cases} (1/n)\delta_{ij} - \boldsymbol{P}^\varepsilon_{ij} \geq 0 & \text{if} \quad i = j \\ (1/n)\delta_{ij} - \boldsymbol{P}^\varepsilon_{ij} \leq 0 & \text{if} \quad i \neq j \end{cases} \tag{11}$$

Handling Costs with Structure

For cost $c(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$ with h striclty convex, the map has structure, as a known functional depending on h^* applied to the gradient a dual potential. A parametrized vector field F_θ is introduced to model directly the dual potential gradient $\nabla\psi^\star$:

$$T_\theta : \boldsymbol{x} \mapsto \boldsymbol{x} - \nabla h^* \circ F_\theta(\boldsymbol{x}) \tag{12}$$

The final loss function

The regularizer penalizes the asymmetry of $\mathrm{Jax}_{\boldsymbol{x}}F$ for $\boldsymbol{x} \sim \rho$:

$$\mathcal{C}_\rho(F) = \mathbb{E}_{X\sim\rho}\left[\|\mathrm{Jac}_X F - \mathrm{Jac}_X^T F\|_2^2\right] \tag{13}$$

$$\min_{\theta\in\mathbb{R}^p} \mathcal{L}(\theta) := \Delta((I_d - \nabla h^* \circ F_\theta)\sharp\mu, \nu) + \lambda_{\mathrm{MG}}\mathcal{M}_\rho^c(I_d - \nabla h^* \circ F_\theta) + \lambda_{\mathrm{cons}}\mathcal{C}_\rho(F_\theta) \tag{14}$$

The selection of the regularization weights $(\lambda_{\mathrm{MG}}, \lambda_{\mathrm{cons}})$

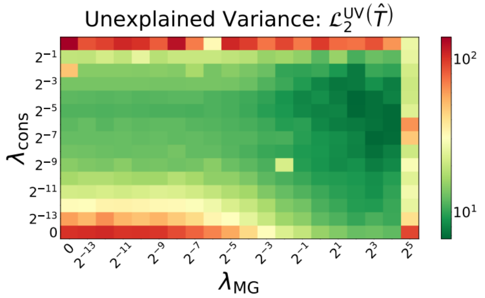

Figure 5. Heatmap showing the influence of the Monge gap $\mathcal{M}_\mu^2$ and the conservative regularizer $\mathcal{C}_\mu^2$, when learning the Monge map for the ℓ_2^2 cost between Korotin et al. (2021) benchmark pair of dimension $d = 32$. For each pair of regularization weights $(\lambda_{\mathrm{MG}}, \lambda_{\mathrm{cons}})$ on a regular grid. we report the unexplained variance $\mathcal{L}_2^{\mathrm{UV}}(\hat{T})$ provided by the the estimated map $\hat{T}$.

$$\mathcal{L}_2^{\mathrm{UV}}(\hat{T}) := 100 \frac{\mathbb{E}_\mu[\|\hat{T}(X) - T^*(X)\|^2]}{\mathrm{Var}_\nu(X)} \quad (15)$$

Outline

Experiments

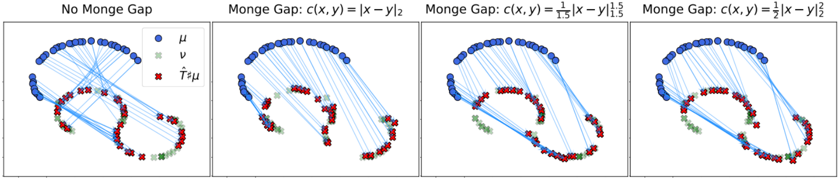

Figure 2. Fitting of transport maps between synthetic measures μ, ν in dimension $d = 2$, with the same fitting loss $\Delta = W_{2,\varepsilon}$ but Monge gap $\mathcal{M}_\mu^c$ instantiated with various costs c. We also fit an MLP without Monge gap, minimizing only the fitting loss. For $c(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2$, we use the method for generic costs §4.1, directly parameterizing T_θ as an MLP and using $\lambda_{\mathrm{MG}} = 5$. For $c(\mathbf{x}, \mathbf{y}) = \frac{1}{1.5}\|\mathbf{x} - \mathbf{y}\|_{1.5}^{1.5}$ and $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, since they have the form $c(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$ with h strictly convex and kwown Legendre transform h^*, we use the method for costs with structure §4.2. Accordingly, we parameterize $T_\theta = \mathrm{I}_d - \nabla h^* \circ F_\theta$ with an MLP F_θ and penalize lack of conservativity with $\mathcal{C}_\mu$. Moreover, we use $\lambda_{\mathrm{MG}} = 1$ and $\lambda_{\mathrm{cons}} = 0.01$.

Experiments

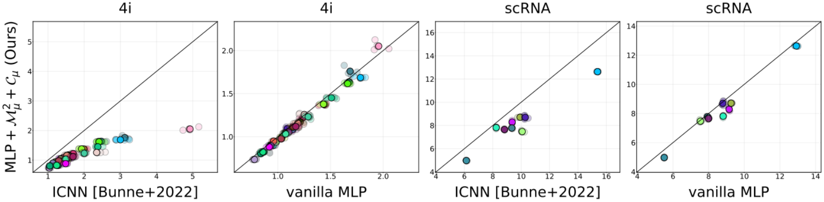

Figure 3. Fitting of a transport map $\hat{T}$ to predict the responses of cells populations to cancer treatments, on 4i and scRNA datasets, providing respectively 34 and 9 treatment responses. For each profiling technology and each treatment, we compare the predictions of a MLP trained with Monge gap $\mathcal{M}_\mu^2(F)$ + conservative regularizer $\mathcal{C}_\mu$ to those provided by a vanilla MLP (trained without regularization), and a gradient-ICNN learned via the neural dual formulation (Makkuva et al., 2020). We measure predictive performance using the Sinkhorn divergence between a batch of unseen (test) treated cells and a batch of unseen control cells mapped with $\hat{T}$, see§ 6.4 and Appendix B.5 for details. Each scatter plot displays points $z_i = (x_i, y_i)$ where y_i is the divergence obtained by our method and x_i that of the other baseline, on all treatments. A point below the diagonal $y = x$ refers to an experiment in which our methods outperforms the baseline. To each treatment, we assign a color and plot 5 runs, along with their mean (the brighter point).

Experiments

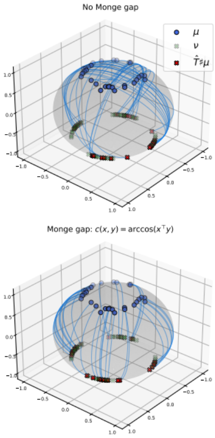

Figure 4. Fitting of transport maps betwen synthetic measures on the 2-sphere. In both cases, we parameterize the map as $T_\theta = F_\theta/\|F_\theta\|_2$ where F_θ is an MLP, and we use $\Delta = W_{\ell_2^2,\varepsilon}$ as fitting loss. On the upper plot, we do not use any regularizer while on the lower plot we regularize with the Monge gap instantiated for the geodesic cost $c(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x}^\top \mathbf{y})$ and use $\lambda_{\text{MG}} = 1$.

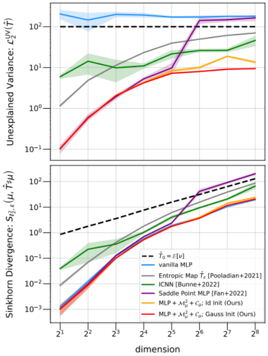

Figure 6. Performances of Monge gap-based learning and baselines on estimating the ground-truth maps between each pair of Gaussian mixtures μ, ν in dimension $d \in \{2, 4, 8, ..., 256\}$ of the Korotin et al. (2021) benchmark. We report both Sinkhorn divergence $S_{\ell_2^2,\varepsilon}(\hat{T}\sharp\mu, \nu)$ and the unexplained variance $\mathcal{L}_2^{\text{UV}}(\hat{T})$ averaged over 5 fittings.

Outline

Conclusion

- This paper provides a new strategy to train optimal transport maps.
- The regularizer adapts to any cost c, but requires defining a reference measure ρ.