Distributionally Robust Optimization with Data Geometry

Jiashuo Liu, Jiayun Wu, Bo Li, Peng Cui

Department of Computer Science and Technology
Tsinghua University

Over-Pessimism Problem of DRO

- The objective function of DRO:

$$\min_\theta \sup_{Q\in\mathcal{P}(P_{tr})} \mathbb{E}_Q[\ell(f_\theta(X),Y)]$$

  - $\mathcal{P}(P_{tr})$ is the distribution set defined via some distance metric as:

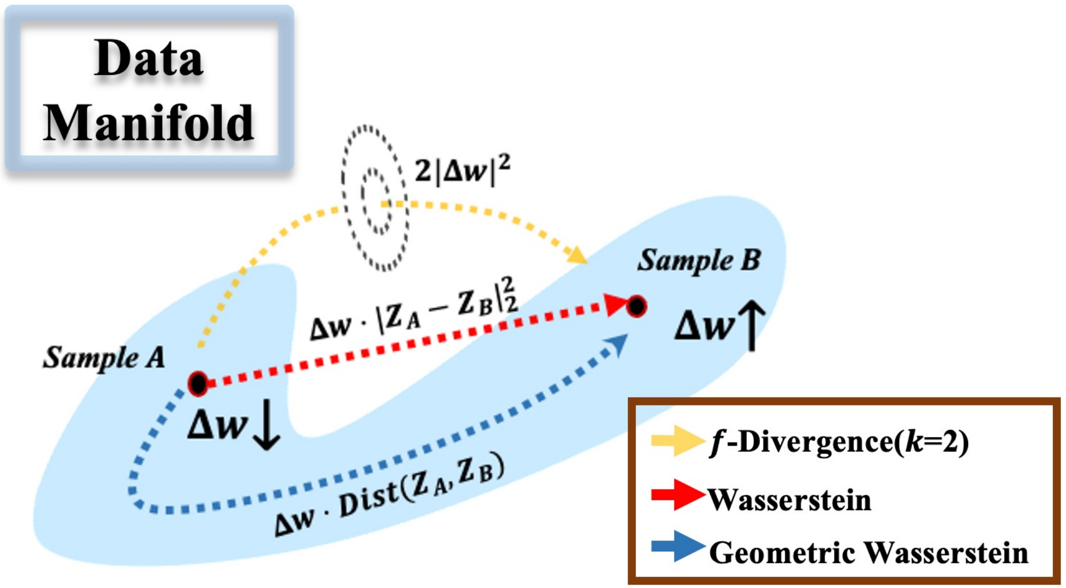
$$\mathcal{P}(P_{tr}) = \{Q: Dist(Q,P_{tr}) \le \rho\}$$

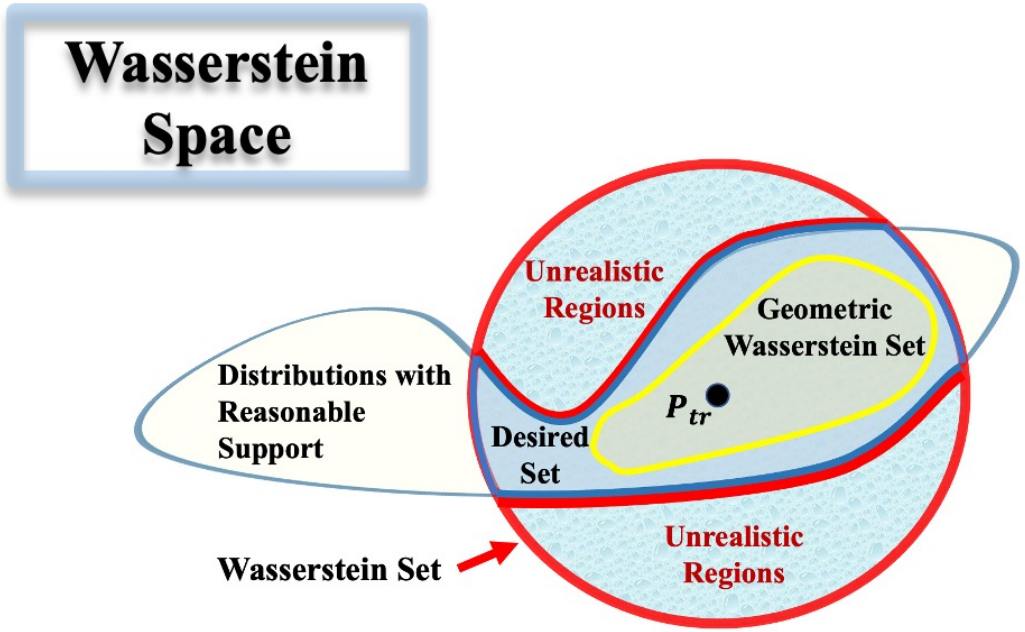
- When the testing distribution is included in $\mathcal{P}(P_{tr})$, the testing performance is guaranteed.
- When the distribution set $\mathcal{P}(P_{tr})$ is overwhelmingly large, the learned model will predict with ***low-confidence***.

→ Over-Pessimism

What Caused the Over-pessimism?

—— *from the distance metric perspective*





Leverage the data geometry to form a more reasonable distribution set.

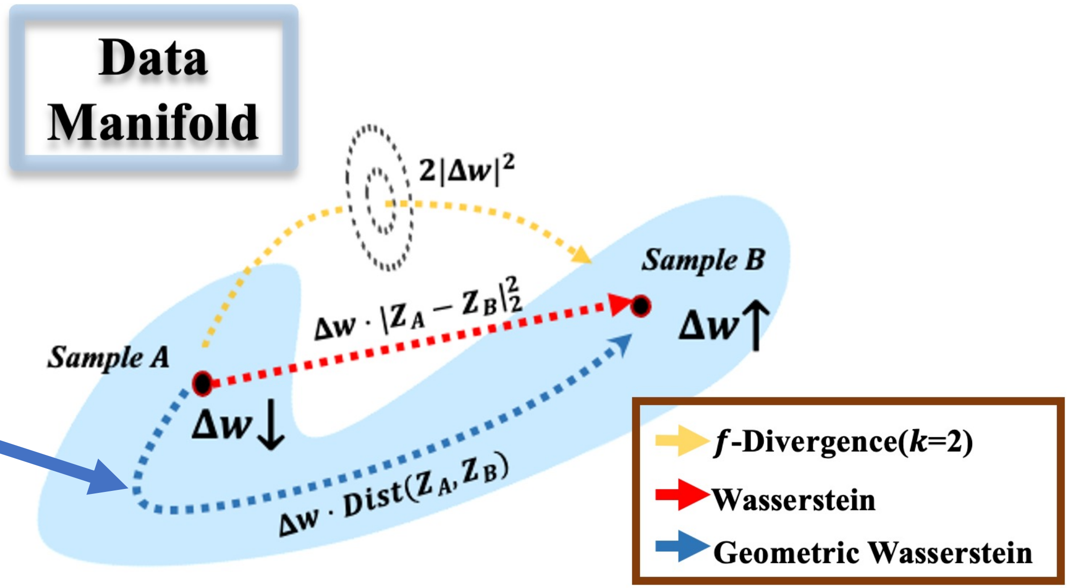
Geometric Wasserstein Distance

Definition 3.1 (Discrete Geometric Wasserstein Distance $\mathcal{GW}_{G_0}(\cdot,\cdot)$ [4]). *Given a finite graph G_0, for any pair of distributions $p^0, p^1 \in \mathscr{P}_o(G_0)$, define the Geometric Wasserstein Distance:*

$$\mathcal{GW}^2_{G_0}(p^0,p^1) := \inf_v \left\{ \int_0^1 \frac{1}{2} \sum_{(i,j)\in E} \kappa_{ij}(p) v_{ij}^2 \, dt \;\middle|\; \frac{dp}{dt} + div_{G_0}(pv) = 0, p(0) = p^0, p(1) = p^1 \right\}, \quad (2)$$

where $v \in \mathbb{R}^{n\times n}$ denotes the velocity field on G_0, p is a continuously differentiable curve $p(t)$: $[0,1] \to \mathscr{P}_o(G_0)$, and $\kappa_{ij}(p)$ is a pre-defined interpolation function between p_i and p_j.

The density transfers smoothly along the data manifold.



NEURAL INFORMATION PROCESSING SYSTEMS

Geometric Wasserstein DRO

- Objective function:

$$\theta^* = \arg\min_{\theta\in\Theta} \sup_{P:\mathcal{GW}^2_{G_0}(\hat{P}_{tr},P)\le\epsilon} \left\{ \mathcal{R}_n(\theta,p) = \sum_{i=1}^{n} p_i \ell(f_\theta(x_i),y_i) - \beta \sum_{i=1}^{n} p_i \log p_i \right\}.$$

- Sample weights updating:

$$\frac{dp_i}{dt} = \sum_{j:(i,j)\in E} w_{ij}\kappa_{ij}(\ell_i - \ell_j) + \beta \sum_{j:(i,j)\in E} w_{ij}\kappa_{ij}(\log p_j - \log p_i),$$

Algorithm 1 Geometric Wasserstein Distributionally Robust Optimization (GDRO)

Input: Training Dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^n$, learning rate α_θ, gradient flow iterations T, entropy term β, manifold representation G_0 (learned by kNN algorithm from D_{tr}).
Initialization: Sample weights initialized as $(1/n, \dots, 1/n)^T$. Predictor's parameters initialized as $\theta^{(0)}$.

for $i = 0$ **to** Epochs **do**
 1. Simulate gradient flow for T time steps according to Equation 5∼6 to learn an approximate worst-case probability weight p^T.
 2. $\theta^{(i+1)} \leftarrow \theta^{(i)} - \alpha_\theta \nabla_\theta (\sum_i p_i^T \ell_i(\theta))$
end for

Jiashuo Liu

- Page: ljsthu.github.io
- Email: liujiashuo77@gmail.com
- Twitter: @liujiashuo77
- WeChat: jiashuo200819