Distributionally Robust Optimization with Data Geometry

Anonymous Author(s)
Affiliation
Address
email

Abstract

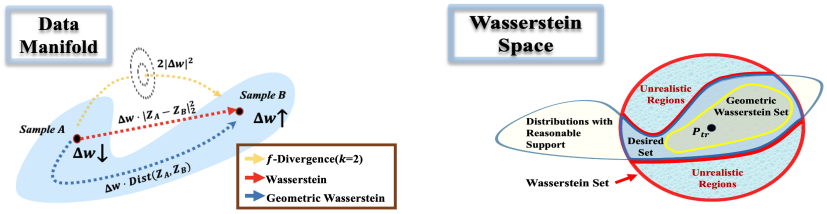
Distributionally Robust Optimization (DRO) serves as a robust alternative to empirical risk minimization (ERM), which optimizes for the worst-case distribution in an uncertainty set, typically specified by distance metrics including f-divergence and the Wasserstein distance. The metrics defined in the ostensible high dimensional space lead to exceedingly large uncertainty sets, resulting in the underperformance of most existing DRO methods. It has been well documented that high dimensional data approximately reside on low dimensional manifolds. In this work, to further constrain the uncertainty set, we incorporate data geometric properties into the design of distance metrics, resulting in our novel Geometric Wasserstein DRO (GDRO). Empowered by Gradient Flow, we derive a generically applicable approximate algorithm for the optimization of GDRO, and prove the bounded error rate of the approximation as well as the convergence of our algorithm. We also theoretically characterize the edge cases when certain existing DRO methods are degeneracy of GDRO. Extensive experiments justify the superiority of our GDRO to existing DRO methods in multiple settings with strong distributional shifts, and confirm that the uncertainty set of GDRO adapts to data geometry.

1 Introduction

Machine learning algorithms with empirical risk minimization often suffer from poor generalization performance under distributional shifts in real applications due to the widespread latent heterogeneity, domain shifts, and data selection bias, *etc*. It is demanded for machine learning algorithms to achieve uniformly good performances against potential distributional shifts, especially in high-stake applications. Towards this goal, distributionally robust optimization (DRO) [17, 14, 19, 3, 11, 9], stemming from the literature of robust learning, has been proposed and developed in recent years. It optimizes for the worst-case distribution within an uncertainty set $\mathcal{P}(P_{tr})$ lying around the training distribution P_{tr}. When the testing distribution P_{te} is contained in $\mathcal{P}(P_{tr})$, DRO could guarantee the generalization performance on P_{te}.

In principle, the effectiveness of DRO heavily depends on the rationality of its uncertainty set $\mathcal{P}(P_{tr})$ which is commonly formulated as a ball surrounding the training distribution endowed with a certain distance metric. An ideal uncertainty set should be constituted by all realistic distributions that may be encountered in test environments. However, the existing DRO methods adopting Wasserstein distance (i.e. WDRO methods [17, 19, 3, 11]) or f-divergence distance (i.e. f-DRO methods[14, 9]) tend to generate **over-flexible** uncertainty sets that incorporate unrealistic distributions far beyond the ideal uncertainty set [13, 11]. As such unrealistic distributions must violate the underlying predicting mechanism, they are prone to be the worst-case and attract much optimization energy in the DRO framework, making the learned model deviate from the true predicting mechanism.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Comparison of metrics on data manifold.

(b) Comparison of uncertainty sets.

Figure 1: Toy illustrations. Figure (a) illustrates the transportation paths and corresponding costs of the distance metrics. (b) depicts the uncertainty set of WDRO and GDRO compared with an ideal uncertainty set in Wasserstein space, where each point denotes a distribution.

Here we argue that the unrealistic distributions mentioned above originate from the distance metrics' inherent ignorance of data geometry, as illustrated in Figure 1. The Euclidean-norm transportation cost measured by the L_2-Wasserstein metric (Equation 2) leads to a straight-line transportation path as shown in Figure 1(a) (red dotted line) which deviates from the data manifold (blue region). Therefore, WDRO methods tend to create unrealistic samples beyond the underlying data manifold, resulting in unrealistic distributions. f-divergence can also be interpreted as a data geometry-independent measure of the transportation cost confined in the support of P_{tr}. Taking $\mathcal{X}_2$-divergence for example, the cost is a constant to transfer per unit of probability weights between samples, like a virtual tunnel (yellow dotted line in Figure 1(a)). In such a case, the noisy samples (e.g. outliers or samples with label noises) are more prone to be the worst-case and thus gather much larger weights than normal samples. The distribution with such a support is obviously unrealistic.

To mitigate the problem, it is demanded to introduce a new distance metric incorporating data geometry to further constrain the uncertainty set and avoid the undesired cases. As illustrated in Figure 1(a), considering the common assumption that data lie on a low-dimensional manifold[16, 20, 2], we expect the probability density transportation path (the blue dotted line) is **restricted within the data manifold** (the blue region). In this way, the resulted uncertainty set (i.e. the Geometric Wasserstein Set as shown in Figure 1(b)) could inherently exclude the distributions beyond the data manifold. Furthermore, it is harder to gather probability weights on isolated noisy samples, which also mitigate the undesired cases in f-DRO.

In this work, we propose a novel Geometric Wasserstein DRO (GDRO) method by exploiting the discrete Geometric Wasserstein distance [4] which measures the transportation cost of probability density along the geodesic in a metric space. As the Geometric Wasserstein distance does not enjoy an analytical expression, we derive an approximate algorithm utilizing Gradient Flow in the Finsler manifold endowed with Geometric Wasserstein Distance (in section 3.2). We further theoretically specify an exponentially vanishing error rate of our approximation as well as a $1/\sqrt{T}$ convergence rate of our algorithm, and characterize the edge cases when GDRO will degenerate to f-DRO or Wasserstein DRO (in section 3.3 and 3.4). Comprehensive experiments encompassing various distributional shifts, including sub-population shifts, domain shifts, and class difficulty shifts, validate the effectiveness of our proposed GDRO (in section 4). We also observe a lower Dirichlet Energy (i.e. higher smoothness) of GDRO's estimated worst-case distribution w.r.t the data manifold compared with existing DRO methods, justifying its adaptability to data geometry.

2 Preliminaries on Distributionally Robust Optimization

Notations. $X \in \mathcal{X}$ denotes the covariates, $Y \in \mathcal{Y}$ denotes the target, $f_\theta(\cdot) : \mathcal{X} \to \mathcal{Y}$ is the predictor parameterized by $\theta \in \Theta$. $P_{tr}(X, Y)$ and $P_{te}(X, Y)$ abbreviated with P_{tr} and P_{te} respectively represent the joint training distribution and test distribution. The random variable of data points is denoted by $Z = (X, Y) \in \mathcal{Z}$.

Distributionally Robust Optimization (DRO) is formulated as:

$$\theta^* = \arg\min_{\theta \in \Theta} \sup_{P \in \mathcal{P}(P_{tr})} \mathbb{E}_P[\ell(f_\theta(X), Y)] \tag{1}$$

where ℓ is a loss function and $\mathcal{P}(P_{tr}) = \{P : \mathit{Dist}(P, P_{tr}) \leq \epsilon\}$ characterizes the uncertainty set surrounding the training distribution restricted by a radius ϵ in terms of *Dist*, a distance metric

between probability distributions. The constraints for the uncertainty set distinguish various DRO methods. Most literature specifies *Dist* as the Wasserstein distance [17, 19, 3] or f-divergence [14, 9].

f-divergence DRO (*abbr.* f-DRO) f-divergence is defined as $D_f(P\|Q) = \int f(dP/dQ)dQ$, where $f(\cdot)$ is a convex function and $f(1) = 0$. Two typical instances of f-divergences are KL-divergence ($f(t) = t\log t$) and $\mathcal{X}_2$-divergence ($f(t) = (t-1)^2$). As shown in Figure 1(a), f-divergence ignores the data geometry and only cares about the probability of each sample (only dP, dQ occur). Since it is well-accepted that data lie on a low-dimensional manifold and adjacent data points have similar degrees of importance, data geometry information is crucial for a reasonable uncertainty set. Such ignorance induces many undesired cases in the f-divergence ball. For example, for heterogeneous data, one hopes to focus on some sub-population (e.g., put more weights on a group of data), but without data geometric information, the distribution in the f-divergence ball is prone to only focus on some isolated samples with higher noises (as shown in Figure 2(a)). Corresponding with our analysis, in Figure 2(b), we find the worst-case distribution found by f-DRO (with KL-divergence) has lower smoothness (larger Dirichlet Energy) w.r.t. data manifold.

Wasserstein DRO (*abbr.* WDRO) Despite the common definition of Wasserstein distance by Kantorovich, we shall introduce the Benamou-Brenier reformulation of L_2-Wasserstein metric which could be readily generalized to Geometric Wasserstein Distance[21]. For probability measures P and Q supported on $\mathcal{X}$, consider a transport plan $p(t,x) : ([0,1], \mathbb{R}^n) \mapsto \mathbb{R}$ continuous on $[0,1]$ which transforms the measure P into Q over the time interval $[0,1]$. Here $p(t,x)$ abbreviated with $p^t(x)$ denotes the probability density of x at time t and is an absolutely continuous probability measure. Then the L_2-Wasserstein metric could be defined as:

$$\mathcal{W}_2^2(P,Q) = \inf_{(p,v)\in\mathcal{V}} \left\{ \int_0^1 \left(\int_{\mathbb{R}^n} p^t(x)\|v^t(x)\|^2 dx \right) dt : \frac{\partial p^t}{\partial t} + \nabla\cdot(p^t v^t) = 0, p^0(\cdot) = P, p^1(\cdot) = Q \right\} \tag{2}$$

where $v : ([0,1], \mathbb{R}^n) \mapsto \mathbb{R}^n$ [1] is a time-dependent vector field satisfying the continuity function in Equation 2. Intuitively v is a velocity field continuously transporting masses to convert the density distribution from P to Q along a curve in the Wasserstein space. And the Wasserstein Distance measures the geodesic length which is shortest among all possible plans. The length of a curve in Wasserstein space is calculated in Equation 2 by integrating a total "kinetic energy" of the velocity field over the transportation process. WDRO is empowered with the flexibility to create samples beyond the training distribution, thus inspiring applications for adversarial robustness against restricted attacks[19]. But it causes an over-pessimistic estimation under strong distributional shifts[11] where mislabeling prevails for perturbed samples as demonstrated by our experiments(Figure 2(a)).

3 Proposed Method

We take as input a training dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^n$ and the corresponding manifold structure typically approximated by a weighted finite graph $G_0 = (V, E, w)$, where $V = \{1, \ldots, n\}$ is the vertex set, E is the edge set and $w = (w_{ij})_{i,j\in V}$ are the edge weights. In case the graph G_0 is not predefined, the k-nearest-neighbor (kNN) algorithm suffices for a fine approximation. For large-scaled datasets, we use the *NN-Descent* method [8] for approximate kNN graph construction, which has a low empirical complexity of $O(n^{1.14})$ and can be easily parallelized. In the following of this section, we first introduce the Geometric Wasserstein Distance and propose the overall objective of GDRO; then we derive an approximate algorithm for optimization utilizing Gradient Flow; finally, some theoretical properties are proved and connections with existing DRO methods are demonstrated.

3.1 Discrete Geometric Wasserstein Distance $\mathcal{GW}_{G_0}$ and GDRO

We firstly introduce the Discrete Wasserstein distance, which extends the Benamou-Brenier formulation of optimal transport problem in Equation 2 to a metric space. Since the manifold structure is merely accessible by a discrete graphical representation, one has to further discretize the Benamou-Brenier formula to describe the optimal transport on the graph. The first step is to define a discrete velocity field and its discrete divergence, which we mainly follow the construction by Chow *et al.*[4].

For a given weighted finite graph $G_0 = (V, E, w)$ with n nodes, a velocity field $v = (v_{ij})_{i,j\in V} \in \mathbb{R}^{n\times n}$ on G_0 is defined to be a skew-symmetric matrix on the edge set E such that $v_{ij} = -v_{ji}$ if $(i,j) \in E$. Given a probability set (simplex) supported on V:$\mathscr{P}(G_0) = \{(p_i)_{i=1}^n \in \mathbb{R}^n | \sum_{i=1}^n p_i =$

[1] v lies in the Lebesgue space of order 2 for the reference measure t and p_t.

$1, p_i \geq 0$, for any $i \in V\}$ and its interior denoted by $\mathscr{P}_o(G_0)$, one could thereafter define the divergence of pv as $\text{div}_{G_0}(pv) := -(\sum_{j \in V:(i,j)\in E} \sqrt{w_{ij}} v_{ij} \theta_{ij}(p))_{i=1}^n$ which is a vector in $\mathbb{R}^n$. The divergence vector is supposed to lie in the tangent space of $\mathscr{P}_o(G_0)$, summing over all the in-fluxes and out-fluxes along edges of a certain node, with each edge transporting a probability density $\sqrt{w_{ij}} v_{ij} \theta_{ij}(p)$. θ_{ij} is a predefined "cross section area" typically interpolated with the associated nodes' densities p_i, p_j. The direct approach is to take the algorithmic average such that $\theta_{ij}(p) = (p_i + p_j)/2$. However, to ensure the positiveness of p during optimization, we adopt the upwind interpolation:$\theta_{ij}(p) = \mathbb{I}(v_{ij} > 0)p_j + \mathbb{I}(v_{ij} \leq 0)p_i$. Now we are ready to define Geometric Wasserstein Distance as a generalization from Equation 2.

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:*

$$\mathcal{GW}^2_{G_0}(p^0, p^1) := \inf_v \left\{ \int_0^1 \frac{1}{2} \sum_{(i,j)\in E} \theta_{ij}(p) v_{ij}^2 dt : \frac{dp}{dt} + div_{G_0}(pv) = 0, p(0) = p^0, p(1) = p^1 \right\} \quad (3)$$

where the infimum is taken over all velocity fields on G_0, p is a continuously differentiable curve $p(t) : [0,1] \to \mathscr{P}_o(G_0)$, and $\theta_{ij}(p)$ is a pre-defined interpolation function between p_i and p_j.

Equation 3 resembles the Benamou-Brenie formulation of continuous L_2-Wasserstein distance in the sense that it integrates the transportation cost over an optimal transport process characterized by the distribution curve $p(t)$, except that the probability density could **only be transferred through the edges of a graph**, thus ensuring the transportation path to stay within the manifold (as the blue dotted line shown in Figure 1(a)) and inducing a smoother estimate of worst-case probability distribution w.r.t the data structure since weights are exchanged just between neighbors. Note that the **metric space** endowed with Geometric Wasserstein Distance $(\mathscr{P}_o(G_0), \mathcal{GW}_{G_0})$ is a Finsler structure[10].

Then we present the **overall objective** of Distributionally Robust Optimization with Geometric Wasserstein Distance (**GDRO**). Given the training dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^n$ and its empirical marginal distribution $\hat{P}_{tr} = \frac{1}{n}\sum_i \delta(x_i)$, along with a manifold structure represented by graph G_0, we intend to obtain a distributionally robust predictor parameterized by θ^* such that for certain $\epsilon > 0$:

$$\theta^* = \arg\min_{\theta\in\Theta} \sup_{P:\mathcal{GW}^2_{G_0}(\hat{P}_{tr},P)\leq\epsilon} \left\{ \mathcal{R}_n(\theta, p) = \sum_{i=1}^n p_i \ell(f_\theta(x_i), y_i) - \beta \sum_{i=1}^n p_i \log p_i \right\} \quad (4)$$

We add a minor entropy-regularization with a small β as proposed in the entropy-balancing literature[12] to avoid singular cases and ensure the convergence of our optimization in section 3.2. Since Equation 4 degenerates to a f-DRO as $\beta \to \infty$, the entropy regularized objective is an interpolation between GDRO and the optimization-friendly f-DRO. Owing to the Geometric Wasserstein Distance, the uncertainty set of GDRO excludes those distributions supported on points beyond the data manifold and the Geometric Wassserstein Ball is directional in Wasserstein space as it stretches along the data structure, as depicted in Figure 1(b).

3.2 Optimization

In this subsection, we derive the optimization algorithm for GDRO. For lack of an analytical expression of Geometric Wasserstein Distance, we give up providing an exact solution to Equation 4. Instead, we propose to solve the minimax objective with alternate optimization. For fixed probability weights p, the parameter θ could be optimized via gradient descents for $R_n(\theta, p)$ w.r.t. θ in parameter

Algorithm 1 Geometric Wasserstein Distributionally Robust Optimization (GDRO)

Input: Training Dataset $D = \{(x_i, y_i)\}_{i=1}^n$, learning rate α_θ, gradient flow iterations T, entropy term β, manifold representation G_0.

Initialization: Sample weights initialized as $(1/n, \ldots, 1/n)^T$. Predictor's parameters initialized as $\theta^{(0)}$.

for $i = 0$ **to** Epochs **do**

1. Simulate gradient flow for T time steps according to Equation 6∼7 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

space Θ. The inner supremum problem can be approximately solved via gradient ascents for $R_n(\theta, p)$ w.r.t. p in the Geometric Wasserstein space. And the cost measured by $\mathcal{GW}^2_{G_0}(\hat{P}_{tr}, \cdot)$ could be approximated with the length of the gradient flow, which is a curve in $(\mathscr{P}_o(G_0), \mathcal{GW}_{G_0})$.

Here we clarify some notations. $p : [0, T] \mapsto \mathscr{P}_o(G_0)$ denotes the continuous gradient flow, and the probability weight of the i-th sample at time t is abbreviated as $p_i(t)$. The time-discretized gradient flow corresponding with the time step τ is denoted as $\hat{p}_\tau : [0, T] \mapsto \mathscr{P}_o(G_0)$, and $\hat{p}_\tau(t)$ is abbreviated as $\hat{p}^t_\tau$. For the optimization, we adopt the time-discretized definition of Gradient Flow[21] for $-R_n(\theta, p)$ in the Geometric Wasserstein space as: (with the time step τ)

$$\hat{p}_\tau(t+\tau) = \arg \max_{p \in \mathscr{P}_o(G_0)} R_n(\theta, p) - \frac{1}{2\tau} \mathcal{GW}^2_{G_0}(\hat{p}_\tau(t), p) \tag{5}$$

When $\tau \to 0$, the time-discretized gradient flow $\hat{p}_\tau$ becomes the continuous one p. Note that Equation 5 describes the Gradient Flow as a steepest ascent curve locally optimizing for a maximal objective within an infinitesimal Geometric Wasserstein ball, and it coincides with the Lagrangian penalty problem of Equation 4. In theorem 3.1 we would prove that Equation 5 finds the exact solution to a **local** GDRO at each time step.

Following Chow *et al.*[4], the analytical solution to Equation 5 as $\tau \to 0$ could be derived as:

$$\frac{dp_i}{dt} = \sum_{j:(i,j)\in E} w_{ij}\theta_{ij}(\ell_i - \ell_j) + \beta \sum_{j:(i,j)\in E} w_{ij}\theta_{ij}(\log p_j - \log p_i) \tag{6}$$

where p_i denotes the time-dependent probability function of the i-th sample, and we take an upwind interpolation of θ: $\theta_{ij}(p) = \mathbb{I}(v_{ij} > 0)p_j + \mathbb{I}(v_{ij} \leq 0)p_i$, so that the probability density transferred on an edge equals the density from the origin node associated with the velocity field. The upwind interpolation guarantees that the probability weight p stays positive along the Gradient Flow in Equation 6. Then we discretize equation 6 with Forward Euler Method:

$$p_i(t+\alpha) = p_i(t) + \alpha dp_i(t)/dt \tag{7}$$

where α is a learning rate. For our algorithm, we control the maximum time step as $t \leq T$ in Equation 7 to approximately restrict the radius of the Geometric Wasserstein ball. We prove in theorem 3.2 that for the final time step $t = T$, the probability weights $p(T)$ learned by Equation 7 guarantees a **global** error rate e^{-CT} from the worst-case risk $R_n(\theta, p^*)$ constrained in an $\epsilon(\theta)$-radius ball where $\epsilon(\theta) = \mathcal{GW}_{G_0}(\hat{P}_{tr}, p(T))$ and $p^* = \arg\sup_p\{R_n(\theta, p) : \mathcal{GW}_{G_0}(\hat{P}_{tr}, p) \leq \epsilon(\theta)\}$. The result is similar to conventions in WDRO[19], which gives up providing a prescribed radius of its uncertainty set but turns to an approximation with a intermediate hyperparameter. Pseudo-code of the whole algorithm is shown in Algorithm 1. The whole derivations are in Appendix.

3.3 Theoretical Properties

We prove the equivalence of our Gradient-Flow based algorithm to a local GDRO problem and the bound of its global error rate. We first provide the robustness guarantee for the Lagrangian penalty problem in Equation 5.

Theorem 3.1 (Local Robustness Guarantees of Lagrangian Penalty Problem)**.** *For any $\tau > 0, t > 0$ and given θ, denote the solution of Equation 5 as $p^*(\theta) = \arg\sup_{p\in\mathscr{P}(G_0)} \mathcal{R}_n(\theta, p) - \frac{1}{2\tau}\mathcal{GW}^2_{G_0}(\hat{p}^t_\tau(\theta), p)$. Let $\epsilon_\tau(\theta) = \mathcal{GW}^2_{G_0}(\hat{p}^t_\tau(\theta), p^*(\theta))$, assume $\ell(\theta)$ is convex on θ, we have*

$$\min_{\theta\in\Theta} \sup_{p\in\mathscr{P}_o(G_0)} \mathcal{R}_n(\theta, p) - \frac{1}{2\tau}\mathcal{GW}^2_{G_0}(\hat{p}^t_\tau(\theta), p) = \min_{\theta\in\Theta} \sup_{p:\mathcal{GW}_{G_0}(\hat{p}^t_\tau(\theta), p)\leq\epsilon_\tau(\theta)} \mathcal{R}_n(\theta, p) \tag{8}$$

Theorem 3.1 proves that at each time step our Lagrangian penalty problem is an exact solution to a local GDRO within the $\epsilon_\tau(\theta)$-radius Geometric Wasserstein ball. It further shows that with $\tau \to 0$ in Equation 5, our gradient flow constantly finds the steepest descent direction. Then we theoretically analyze the global error rate brought by our approximate algorithm.

Theorem 3.2 (Global Error Rate Bound)**.** *Given the model parameter θ, denote the approximate worst-case by gradient descent in Equation 7 after time t as $p^t(\theta)$, and $\epsilon(\theta) = GW^2_{G_0}(\hat{P}_{tr}, p^t(\theta))$ denotes the distance between our approximation p^t and the training distribution $\hat{P}_{tr}$. Then denote the real worst-case distribution within the $\epsilon(\theta)$-radius discrete Geometric Wasserstein-ball as $p^*(\theta)$, that is,*

$$p^*(\theta) = \arg \sup_{p:\mathcal{GW}_{G_0}(\hat{P}_{tr}, p)\leq\epsilon(\theta)} \sum_{i=1}^n p_i\ell_i - \beta\sum_{i=1}^n p_i \log p_i \tag{9}$$

Here we derive the bound w.r.t. the error ratio of objective function $\mathcal{R}(p)$. For $\theta \in \Theta$, there exists $C > 0$ such that

$$\textit{Error Rate} = \left(\mathcal{R}(p^*) - \mathcal{R}(p^t)\right) / \left(\mathcal{R}(p^*) - \mathcal{R}(\hat{P}_{tr})\right) < e^{-Ct} \tag{10}$$

and when $t \to \infty$, Error Rate $\to 0$. The value of C depends on ℓ, β, n.

Theorem 3.2 theoretically characterizes 'how far' our approximation p^t is from the real worst-case p^* in terms of the drop ratio of the objective function $\mathcal{R}(p)$. At last we derive the convergence rate of our Algorithm 1.

Theorem 3.3 (Convergence of Algorithm 1)**.** *Denote the objective function for the predictor as:*

$$F(\theta) = \sup_{\mathcal{GW}_{G_0}(\hat{P}_{tr}, p) \leq \epsilon(\theta)} \mathcal{R}(\theta, p) \tag{11}$$

which is assumed as L-smooth and $\mathcal{R}(\theta, p)$ satisfies L_p-smoothness such that $\|\nabla_p R(\theta, p) - \nabla_p R(\theta, p')\|_2 \leq L_p \|p - p'\|_2$. $\epsilon(\theta)$ follows the definition in Theorem 3.2. Take a constant $\Delta_F \geq F(\theta^{(0)}) - \inf_\theta F(\theta)$ and set step size as $\alpha = \sqrt{\Delta_F/(LK)}$. For $t \geq T_0$ where T_0 is a constant, denote the upper bound of $\|p^t - p^*\|_2^2$ as γ and train the model for K steps, we have:

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K} \|\nabla_\theta F(\theta^{(k)})\|_2^2\right] - \frac{(1 + 2\sqrt{L\Delta_F/K})}{1 - 2\sqrt{L\Delta_F/K}} L_p^2 \gamma \leq \frac{2\Delta_F}{\sqrt{\Delta_F K} - 2L\Delta_F} \tag{12}$$

We assume a common smoothness of the objective function as in [19]. As $K \to \infty$, $\nabla_\theta F(\theta^{(k)})$ will achieve a square-root convergence only if γ is controlled by the exponentially vanishing error rate in Theorem 3.2. And the accuracy parameter γ remains a fixed effect on optimization accuracy

3.4 Connections with Conventional DRO Methods

In Theorem 3.4, we illustrate the connections of our GDRO with f-DRO.

Theorem 3.4 (Connection with f-DRO with KL-diveregence (KL-DRO).)**.** *Relax the discrete Geometric Wasserstein-ball regularization (set $\epsilon \to \infty$) and set the graph G_0 to a fully-connected graph, and then the solution of GDRO is equivalent to the following form of **KL-DRO**:*

$$\min_{\theta \in \Theta} \sup_{p: D_{KL}(p \| \hat{P}_{tr}) \leq \hat{\epsilon}(\theta)} \sum_{i=1}^{n} p_i \ell(f_\theta(x_i), y_i), \quad \textit{with } \hat{\epsilon}(\theta) = D_{KL}(p^*(\theta) \| \hat{P}_{tr}) \tag{13}$$

where $p^*(\theta) = \arg\max_p \sum_{i=1}^n p_i \ell(f_\theta(x_i), y_i) - \beta \sum_{i=1}^n p_i \log p_i$.

Remark (Connections with WDRO)**.** *Since conventional WDRO allows distributions to extend training support, our proposed GDRO is intrinsically different from WDRO. Intuitively, for infinite samples, if the graph G_0 is set to a fully-connected graph with edge weights $w_{ij} = \|z_i - z_j\|^2$ and β is set to 0, our GDRO resembles support-restricted version of WDRO.*

4 Experiments

In this section, we investigate the empirical performance of our proposed GDRO on different simulation and real-world datasets under varying kinds of distributional shifts , including subpopulation shifts, domain shifts, and class difficulty shifts. As for baselines, we compare with empirical risk minimization (ERM), WDRO and two typical f-DRO methods, including KL-DRO ($f(t) = t \ln t$) and $\mathcal{X}_2$-DRO ($f(t) = (t-1)^2$).

4.1 Simulation Data

As for the simulation data, we simulate domain shifts between training data and testing data via anti-causal effect and selection bias mechanism respectively. And we also investigate the influence brought by label noises to demonstrate that our GDRO could to some extent resist label noise.

1. Toy Example: Domain Shift via Anti-Causal Effect

Firstly, inspired by [1], we induce the domain shifts via the anti-causal effect as follows:

$$S \sim \mathcal{N}(0,1), \quad Y = \alpha_S S + S^2 + \mathcal{N}(0, 0.1), \quad V = \begin{cases} Y + \mathcal{N}(0,1) & \textit{with probability } 1-r \\ -0.1Y + \mathcal{N}(0,1) & \textit{with probability } r \end{cases} \tag{14}$$

Table 1: Results of the toy example with varying minor probability r. Each result is averaged over 10 runs, and the standard deviation is omitted since it is small for all methods.

Simulation 1: Toy Example (Domain Shift via Anti-Causal Effect)									
Minor Probability	$r = 0.01$			$r = 0.05$			$r = 0.1$		
	Mean_Error	Std_Error	Est_Error	Mean_Error	Std_Error	Est_Error	Mean_Error	Std_Error	Est_Error
ERM	7.144	4.260	3.999	5.594	3.210	2.970	4.521	2.444	2.230
WDRO	5.396	3.480	3.024	4.303	2.656	2.297	3.451	1.974	1.700
KL-DRO	2.672	1.145	1.134	2.678	1.148	1.194	2.531	1.086	1.157
$\mathcal{X}_2$-DRO	3.027	1.323	1.185	2.954	1.262	1.118	2.738	1.079	0.953
GDRO	**1.759**	**0.047**	**0.287**	**1.718**	**0.066**	**0.043**	**1.769**	**0.153**	**0.072**

where S serves as a stable feature with an unchanged relationship with Y (thus it should be used for prediction), but V is a spurious feature with changeable relationships with Y (thus one should avoid using it), and α_S is set to 5.0 for all data. For training data, we sample 10000 points, and design different settings with varying minor group ratios r. For testing data, we simulate 6 domains with strong shifts by varying $\alpha_V \in \{-3, -2, -1, 1, 2, 3\}$ and calculate the Mean_Error, Std_Error and parameter estimation error Est_Error for each method. From the results in Table 1, WDRO performs similarly to ERM, and two f-DRO methods (KL-DRO and $\mathcal{X}_2$-DRO) outperform ERM and WDRO. Such results verify our analysis that the Wasserstein uncertainty set cannot be large due to its over-flexibility, which greatly impairs its performance under strong domain shifts. Our GDRO outperforms all baselines and achieves much lower prediction errors and estimation errors in all settings, which shows our uncertainty set built on Geometric Wasserstein distance is much more practical and reasonable.

2. Selection Bias: Domain Shift via Selection Bias Mechanism

Secondly, a more complicated mechanism is designed via selection bias as:

$$S \sim \mathcal{N}(0, 2\mathbb{I}_{n_s}) \in \mathbb{R}^5, \;\; V \sim \mathcal{N}(0, 2\mathbb{I}_{n_v}) \in \mathbb{R}^5, \;\; Y = \beta^T S + 0.1 \cdot S_1 S_2 S_3 + \mathcal{N}(0, 0.5) \tag{15}$$

Similar to the toy example above, S are stable features while the relationships between V and Y are perturbed in different domains. Specifically, a data point is selected with probability $P(x_i, y_i) = |r|^{-5*|y_i - \text{sign}(r)\cdot v_i^b|}$, which induces the spurious correlation between a certain covariate $V^b \in V$ and Y. In training, we generate 10000 points, where the major group contains 95% data with $r = 1.9$ and the minor group contains 5% data with $r = -1.3$. In testing, we first report the performances of the two training groups, and then we further vary $r \in \{-1.5, -1.7, -1.9, -2.3, -2.7, -3.0\}$ to simulate more challenging domain shifts that cannot be obtained by interpolation between training groups. The average results over ten runs are shown in Table 2 (simulation 2). As for sub-population shifts, GDRO achieves good tail performance at a slight sacrifice of the performance of the major group. Further, for more challenging domain shifts, GDRO significantly outperforms all baselines in all testing distributions, even though the shifts are much stronger than the training.

Table 2: Results on the Selection Bias Experiments. We report the root mean square errors.

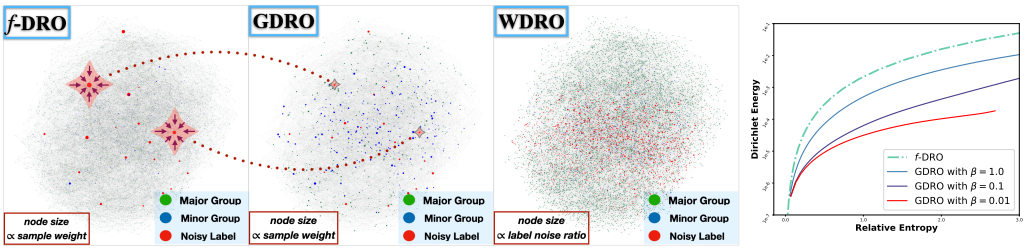
Simulation 2: Selection Bias Experiment without Label Noises									
	Train(major)	Train(minor)	Test						Parameter
Bias Ratio r	$r = 1.9$	$r = -1.3$	$r = -1.5$	$r = -1.7$	$r = -1.9$	$r = -2.3$	$r = -2.7$	$r = -3.0$	Est_ Error
ERM	**0.339**	0.876	0.892	0.884	0.864	0.880	0.843	0.888	0.423
WDRO	**0.339**	0.877	0.894	0.885	0.865	0.882	0.844	0.890	0.424
$\mathcal{X}_2$-DRO	0.411	0.744	0.757	0.741	0.733	0.742	0.714	0.755	0.367
KL-DRO	0.370	0.713	0.728	0.716	0.708	0.713	0.685	0.724	0.319
GDRO	0.493	**0.492**	**0.508**	**0.489**	**0.501**	**0.483**	**0.486**	**0.496**	**0.033**
Simulation 3: Selection Bias Experiment under Label Noises									
ERM	**0.335**	0.845	0.885	0.879	0.874	0.884	0.882	0.876	0.422
WDRO	**0.335**	0.896	0.887	0.880	0.875	0.886	0.884	0.877	0.423
$\mathcal{X}_2$-DRO	0.375	0.866	0.855	0.856	0.843	0.860	0.854	0.845	0.408
KL-DRO	0.393	0.879	0.868	0.866	0.856	0.876	0.866	0.861	0.391
GDRO	0.542	**0.537**	**0.553**	**0.549**	**0.534**	**0.539**	**0.555**	**0.550**	**0.058**

3. Label Noises: Add Label Noises

Since DRO methods are risk-aware, they are prone to be affected by label noises. Though the effect of label noises cannot be eliminated due to the nature of DRO, our proposed GDRO could significantly improve the resistance of DRO methods to a minor degree of label noises. Based on the selection bias experiment, we random sample 20 points and add label noises on them via $\tilde{Y} = Y + \text{Std}(Y)$ where $\text{std}(Y)$ denotes the standard derivation of the marginal distribution of Y. From results shown in Table

2(simulation 3), both f-DRO methods are significantly affected and perform similarly to ERM under such a minor degree of label noises. And our GDRO is only slightly affected by the label noises.

Further, to demonstrate the difference between f-DRO, WDRO, and GDRO, for the label noise experiment, we visualize the learned worst-case distribution of three methods. **(1)** In the first two figures in Figure 2(a), we draw the learned sample weights of KL-DRO and GDRO, where red points represent the noisy samples in training data, and the size of each point is proportional to its sample weight. We can see that KL-DRO puts heavy weights on the noisy points (the red nodes are much larger), while our GDRO only slightly increases them and the weights of data in the minor group are raised, which results in the difference between their performances in the label noise experiment. The reason for such phenomenon of KL-DRO is that it ignores the data geometry, allowing for some isolated nodes with much heavier weights than surrounding nodes, which corresponds with our Theorem 3.4. Since our GDRO intrinsically naturally constrains weight learning on the data manifold, the learned weights are smooth w.r.t the data manifold. **(2)** In the third figure in Figure 2(a), the red points represent the samples, to which WDRO introduces label noises with the ratio larger than 50%, and the size of each point is proportional to its label noise ratio. We can see that although WDRO extends the support, the created samples are quite noisy and greatly harm the learning process. **(3)** To quantify this property, we measure the smoothness via Dirichlet Energy and plot the Dirichlet Energy w.r.t the relative entropy $KL(p\|\hat{P}_{tr})$ between the learned distribution and training distribution in Figure 2(b), which shows that the learned weights of GDRO are much more smooth on the data manifold than that of KL-DRO.



(a) Visualization of learned weights of f-DRO and GDRO. (b) Weight smoothness.

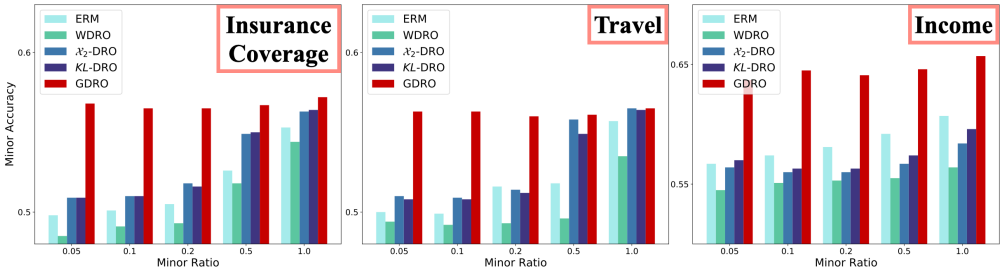
Figure 2: Explanatory studies for comparison between f-DRO ($f(x) = x \ln x$) and GDRO. **Figure (a)** visualizes the learned worst-case distribution of f-DRO, GDRO, and WDRO on kNN, and the size of each node is proportional to its sample weight or its label noise ratio. **Figure (b)** plots the Dirichlet Energy w.r.t the relative entropy, which measures the smoothness of learned weights given the same $D_{KL}(p\|\hat{P}_{tr})$.

4.2 Real-World Data

Finally, towards a comprehensive comparison with existing DRO methods, we evaluate our method on four real-world datasets with various kinds of distributional shifts, including sub-population shifts, domain shifts, class-wise shifts, and label noises. All experimental details can be found in Appendix.

1. Retiring Adults: Sub-population Shifts The Retiring Adults dataset [7] is derived from US Census surveys, where the sub-population shifts are natural since there are geographic variations across different states. Our experiments involve three prediction tasks defined in [7], including Income Prediction, Public Health Insurance Coverage Prediction, and Commuting Time Prediction. For each task, we randomly sample 2000 points from state A and $2000 \cdot r$ points from state B, where $r < 1$ and the second state is viewed as the minor group. In Figure 3, we report the prediction accuracy of the minor group for different minority ratios r in three tasks.

2. HIV-1: Sub-population Shifts & Label Imbalance HIV-1 Protease Cleavage Dataset[5] involves a task to predict whether an octamer would be cleaved by HIV-1 protease, given a 160-dimensional one-hot vector encoding the sequence of 8 amino acids composing the octamer. The dataset contains 4 splits, and following [15], we merge '746' and '1625' as subpopulations A and view 'Impens' as B. In training, subpopulations A and B are mixed at a ratio of $1 : r$, where $r \leq 1$ is the minor group ratio. Coupled with the sub-population shift, class labels are imbalanced with 33% positive in sub-population A and 16% positive in B, which is more challenging. In testing, we re-balance the two classes and plot the overall accuracy w.r.t. the minor group ratio in Figure 3.



(a) Health Insurance Coverage Data. **(b) Travel Time Data.** **(c) People Income Data.**

Figure 3: Results of Retiring Adults Dataset.

Table 3: Results of Colored MNIST data and Ionosphere data.

Method	Colored MNIST		Ionosphere		
	Train Acc	Test Acc	Easy Class Acc	Hard Class Acc	AUC Score
ERM	0.867	0.116	0.952	0.481	0.683
WDRO	**1.000**	0.335	0.944	0.630	0.774
$\mathcal{X}_2$-DRO	0.839	0.420	0.976	0.519	0.756
KLDRO	**1.000**	0.287	**0.984**	0.630	0.826
GDRO	0.717	**0.696**	0.962	**0.741**	**0.883**

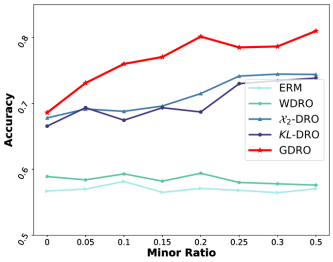


Figure 4: Results of the HIV-1 data.

3. Colored MNSIT: Domain Shifts & Label Noises Following Arjovsky *et al.*[1], we conduct a binary classification task on the MNIST dataset. Firstly, a binary label Y is assigned to each image according to its digit: $Y = 0$ for digit $0 \sim 4$ and $Y = 1$ for digit $5 \sim 9$. Secondly, we induce noisy labels $\tilde{Y}$ by randomly flipping the label Y with a probability of 0.2. Then we induce domain shifts by sampling the color id C spuriously correlated with $\tilde{Y}$. Specifically, we generate C by flipping $\tilde{Y}$ with probability r, which can be viewed as the indicator of different domains. In training, we randomly sample 5000 data points and set $r = 0.75$ and in testing, we set $r = 0$, which induces strong domain shifts between training and testing. Results are shown in Table 3.

4. Ionosphere Radar Classification: Class Difficulty Shifts Ionosphere Radar Dataset[6] consists of return signals from the ionosphere of a phased array radar system in Google Bay, Labrador. The electromagnetic signals were processed by an auto-correlation function to produce 34 continuous attributes. The target is to determine whether the return signal indicates specific physical structures in the ionosphere (good return) or not (bad return). Due to the disparity between classes, ERM was found to achieve a much lower accuracy on bad returns than good ones[18]. DRO methods are expected to achieve higher accuracy in the harder class. Results are shown in Table 3.

Analysis From the results on real-world data, we find that in most scenarios, WDRO only slightly outperforms ERM, and two f-DRO methods show significant promotions to ERM. Our proposed GDRO outperforms all baselines significantly in all scenarios. (1) The discrepancy between WDRO ($\approx$ ERM) and f-DRO ($>$ ERM) indicates that the extending the distribution support in practice is not as promising as might be expected, which we think is because creating new data points is nearly impossible in real scenarios. (2) The significant discrepancy between GDRO ($\gg$ ERM) and WDRO ($\approx$ ERM) shows the superiority of restricting the distribution support, which enables DRO to provide robustness in a much larger uncertainty set without worrying about generating unrealistic samples. (3) GDRO exhibits significant advantages under strong distributional shifts (small minor ratio in Figure 3 and Figure 4; strong shifts in Colored MNSIT in Table 3), which shows that by incorporating geometric properties, our uncertainty set mitigates the over-flexibility problem and our GDRO can resist stronger distributional shifts.

5 Conclusion

In this work, we take the first step to incorporate data geometry information to mitigate the over-flexibility problem in DRO. Theoretical analysis and empirical results validate the effectiveness of our proposed GDRO. We find there are potentials to incorporate more complicated models like GNNs for a better characterization of the data manifold, which we leave for the future work.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs are placed in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All details are placed in Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] The error bar is quite small in our experiments and therefore we omit it.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[3] Ruidi Chen and Ioannis C Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018.

[4] Shui-Nee Chow, Wuchen Li, and Haomin Zhou. Entropy dissipation of fokker-planck equations on graphs. *arXiv preprint arXiv:1701.04841*, 2017.

[5] HIV-1 Protease Cleavage Dataset. https://archive-beta.ics.uci.edu/ml/datasets/hiv+1+protease +cleavage. 2015.

[6] Ionosphere Radar Dataset. https://archive-beta.ics.uci.edu/ml/datasets/ionosphere.

[7] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Wei Dong, Moses Charikar, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 577–586. ACM, 2011.

[9] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[10] Antonio Esposito, Francesco S. Patacchini, André Schlichting, and Dejan Slepcev. Nonlocal-interaction equation on graphs: gradient flow structure and continuum limit. *CoRR*, abs/1912.09834, 2019.

[11] Charlie Frogner, Sebastian Claici, Edward Chien, and Justin Solomon. Incorporating unlabeled data into distributionally robust learning. *arXiv preprint arXiv:1912.07729*, 2019.

[12] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

[13] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018.

[14] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

[15] Thorsteinn S. Rögnvaldsson, Liwen You, and Daniel Garwicz. State of the art prediction of HIV-1 protease cleavage sites. *Bioinform.*, 31(8):1204–1210, 2015.

[16] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[17] Soroosh Shafieezadeh Abadeh, Peyman M Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. *Advances in Neural Information Processing Systems*, 28, 2015.

[18] Vincent G Sigillito, Simon P Wing, Larrie V Hutton, and Kile B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Technical Digest*, 10(3):262–266, 1989.

[19] Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[20] Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[21] Cédric Villani. Topics in optimal transportation. 58, 2021.