Geometry-Calibrated DRO: Combating Over-Pessimism with Free Energy Implications

Jiashuo Liu[1,†], **Jiayun Wu**[1], **Tianyu Wang**[2], **Hao Zou**[3], **Peng Cui**[1,‡]
[1]Department of Computer Science and Technology, Tsinghua University
[2]Department of Industrial Engineering and Operations Research, Columbia University
[3]Zhongguancun Lab
[†]liujiashuo77@gmail.com, [‡]cuip@tsinghua.edu.cn

Abstract

Distributionally Robust Optimization (DRO) optimizes the worst-case risk within an uncertainty set to resist distribution shifts. However, DRO suffers from over-pessimism, leading to low-confidence predictions, poor parameter estimations as well as poor generalization in practice. In this work, we uncover one probable root cause of over-pessimism: excessive focus on noisy samples. To alleviate the impact of noise, we incorporate data geometry into calibration terms in DRO, resulting in our novel Geometry-Calibrated DRO (GCDRO) *for regression*. We establish that our risk objective aligns with the Helmholtz free energy in statistical physics, which could extend to standard DRO methods. Leveraging gradient flow in Wasserstein space, we develop an approximate minimax optimization algorithm with a bounded error ratio and elucidate how our approach mitigates noisy sample effects. A full version of this paper can be found at https://arxiv.org/pdf/2311.05054.pdf.

1 Introduction

Machine learning algorithms with empirical risk minimization (ERM) have been shown to perform poorly under distributional shifts, especially sub-population shifts where substantial data subsets are underrepresented in the average risk due to their small sample sizes. As an alternative, Distributionally Robust Optimization (DRO) [28, 5, 4, 13, 39, 25, 18, 17] aims to optimize against the worst-case risk distribution within a predefined uncertainty set. However, DRO methods have been found to experience the over-pessimism problem in practice [20, 39] (*i.e.*, low-confidence predictions, poor parameter estimations, and generalization), recent studies have sought to address this issue. Recently, Słowik et al. [35] observed that DRO may overly focus on sub-populations with higher noise levels, leading to sub-optimal generalization. Consequently, they suggest incorporating calibration terms to mitigate this issue. Nevertheless, applicable calibration terms either require expert knowledge or are computationally intensive, and few practical algorithms have been proposed.

To devise a practical calibration term for DRO, we first attribute the probable root cause to the excessive focus on noisy samples that frequently exhibit higher prediction errors. We theoretically demonstrate that typical DRO algorithms tend to put higher densities on noisy samples, which, based on a simple yet insightful linear example (in Appendix A), we prove will greatly amplify the variance of estimated parameters, in line with the empirical findings reported in [39]. Furthermore, (in Appendix B) we demonstrate that existing outlier-robust regression methods are not directly applicable for mitigating noisy samples in DRO scenarios where both noisy samples and distribution shifts coexist, highlighting the non-trivial nature of this problem.

In this work, inspired by the ideas in [35, 1], we design calibration terms, $i.e.$, total variation and entropy regularization, to prevent DRO from excessively focusing on random noisy samples. In

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

conjunction with the Geometric Wasserstein uncertainty set [26] utilized in our methods, these calibration terms effectively incorporate information from the data manifold. We prove that these terms could effectively de-emphasizes noisy samples, leading to improved regulation of the worst-case distribution in DRO. Furthermore, from a statistical physics perspective, we relate our risk objective to the Helmholtz free energy, comprising three components: interaction energy, potential energy, and entropy. This physical interpretation provides a novel perspective for understanding different DRO methods by drawing parallels between the worst-case distribution and the steady state in statistical physics, offering valuable insights for designing new DRO algorithms.

2 Preliminaries: Noisy Samples Bring Over-Pessimism in DRO

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 $\theta \in \Theta$. $\hat{P}_N$ denotes the empirical counterpart of distribution $P(X, Y)$ with N samples, and $\mathbf{p} = (p_1, \ldots, p_N)^T \in \mathbb{R}_+^N$ is the probability vector. $[N] = \{1, 2, \ldots, N\}$ denotes the set of integers from 1 to N. The random variable of data points is denoted by $Z = (X, Y) \in \mathcal{Z}$. The random vector of n dimension is denoted by $\vec{h}_n = (h_1, \ldots, h_n)^T$. $G_N = (V, E, W)$ denotes a finite weighted graph with N nodes, where $V = [N]$ is the vertex set, E is the edge set and $W = \{w_{ij}\}_{(i,j)\in E}$ is the weight matrix of the graph. And $(x)_+ = \max(x, 0)$.

Distributionally Robust Optimization (DRO) is formulated as:

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

where ℓ is the loss function (typically mean square error) and $\mathcal{P}(P) = \{Q : \text{Dist}(Q, P) \leq \rho\}$ denotes the ρ-radius uncertainty ball around the distribution P. Although designed to resist distribution shifts, they have been observed to have poor generalization performances [20, 15, 35] in practice, which is referred to as over-pessimism. In this section, we identify one of the root causes of the over-pessimism of DRO: the *excessive focus on noisy samples with typically high prediction errors*.

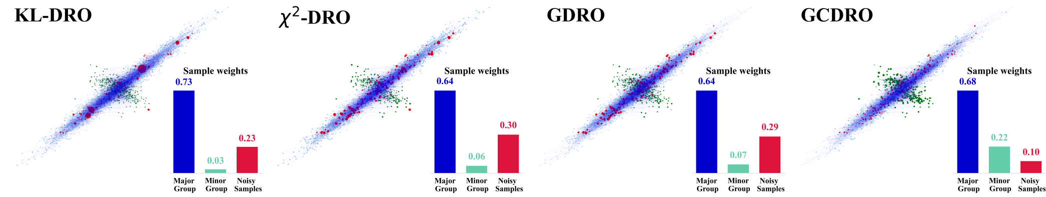

Figure 1: Visualizing the Worst-Case Distribution for Different DRO Methods: We show the data manifold and sample weights for each point, where blue points represent the major group, green ones represent the minor group, and red ones are noisy samples. The bars display the total sample weights of different groups, and the *original* group ratio is major (**93.1%**), minor (**4.9%**), (noisy **2%**).

In this section, we identify one of the root causes of the over-pessimism of DRO: the *excessive focus on noisy samples with typically high prediction errors*.

- We showcase DRO methods' excessive focus on noisy samples in practice and reveal their probability densities are linked to high prediction errors in worst-case distributions.
- Through a simple yet insightful regression example, we prove that such a phenomenon leads to high estimation variances and subsequently poor generalization performance.
- We demonstrate that existing outlier-robust regression methods are not directly applicable for mitigating noisy samples in DRO scenarios, emphasizing the non-trivial nature of this problem.

Problem Setting Given the *underlying* clean distribution $P_{clean} = (1-\alpha)P_{major} + \alpha P_{minor}, 0 < \alpha < \frac{1}{2}$, the **goal of DRO can be viewed as achieving good performance across all possible sub-populations** P_{minor}. Denote the observed contaminated training distribution by P_{train}. Based on Huber's ϵ-contamination model [21], we formulate P_{train} as:

$$P_{train} = (1-\epsilon)P_{clean} + \epsilon\tilde{Q} = \underbrace{(1-\epsilon)(1-\alpha)P_{major}}_{\text{major sub-population}} + \underbrace{(1-\epsilon)\alpha P_{minor}}_{\text{minor sub-population}} + \underbrace{\epsilon\tilde{Q}}_{\text{noisy sub-population}}, \tag{2}$$

where $\tilde{Q}$ is an arbitrary *noisy* distribution (typically with larger noise scale), $0 < \epsilon < \frac{1}{2}$ is the noise level. The *minor sub-population could represent any distribution with a proportion of α in*

P. However, we explicitly specify it here to emphasize the distinction between our setting and the traditional Huber's ϵ-contaminated setting, as the latter does *not* take sub-population shifts into account.

Empirical Observations. Following a typical regression setting [13, 26], we demonstrate the worst-case distribution of KL-DRO, χ^2-DRO, and GDRO in Figure 3, where the size of each point is proportional to its density. In this scenario, the underlying distribution P comprises a known major sub-population (95%, blue points) and a minor sub-population (5%, green points). And the noise level ϵ in P_{train} is 2%. DRO methods are expected to upweigh samples from minor sub-population to learn a model with uniform performances w.r.t. sub-populations. However, from Figure 3, we could observe that KL-DRO, χ^2-DRO and GDRO excessively focus on noisy samples, resulting in a noise level 10 to 15 times larger than the original.

We first analyze the worst distribution of KL-DRO, χ^2-DRO and GDRO [26].

Proposition 2.1 (Worst-case Distribution). *Let $\hat{Q}_N^* = (q_1^*, q_2^*, \ldots, q_N^*)^T \in \mathbb{R}_+^N$ denotes the worst-case distribution, and $\ell(f_\theta(x_i), y_i)$* (abbr. *ℓ_i) denotes the prediction error of sample $i \in [N]$. For different choices of Dist$(\cdot,\cdot)$ in $\mathcal{P}(P) = \{Q : \textit{Dist}(Q, P) \le \rho\}$, we have:*

- *KL-DRO: $q_i^*/q_j^* \propto \exp(\ell_i - \ell_j)$;*
- *GDRO's final state (gradient flow step $T \to \infty$): $q_i^*/q_j^* \propto \exp(\ell_i - \ell_j)$;*
- *χ^2-DRO: $q_i^*/q_j^* = (\ell_i - \lambda)_+/(\ell_j - \lambda)_+$, and $\lambda \ge 0$ is the dual parameter independent of i.*

Proposition 2.1 demonstrates that for KL-DRO, χ^2-DRO, and GDRO (large gradient flow step), the *relative density* between samples is solely determined by their prediction errors, indicating that a larger prediction error results in a higher density. However, in our problem setting, the presence of noisy samples in $\tilde{Q}$ significantly interferes with this objective and hurts model learning.

Due to space limits, in Appendix A, we use a simple example with the weighted least square model to demonstrate how this excessive focus on noisy samples can lead to high estimation variance, ultimately causing over-pessimism. Based on the analysis above, we stress the importance of integrating more data-derived information. In pursuit of this, we propose to leverage the unique geometric properties that distinguish noisy samples from minor sub-populations to address this issue.

Relationship with Conventional Outlier-robust Regression. We would like to explain why conventional outlier-robust regression methods cannot be directly applied to our problem. The main challenge stems from the *coexistence* of noisy samples and minor sub-populations, both of which typically exhibit high prediction errors, leading to a misleading worst-case distribution in DRO. Conventional outlier-robust regression methods [9, 23, 10] primarily focus on mitigating the effects of outliers without considering sub-population shifts. For instance, the L_2-estimation-error of outlier-robust linear regression is $\mathcal{O}(\epsilon \log(1/\epsilon))$ [9], where ϵ represents the noise level in Equation 1. However, as analyzed in Proposition 2.1 and demonstrated in Figure 3, during the optimization of DRO, the noise level ϵ significantly increases, rendering even outlier-robust estimation quite inaccurate. Moreover, [23] propose finding a pseudo distribution with minimal prediction errors to avoid outliers (see Algorithm 5.2 in [23]). Nevertheless, this approach might inadvertently exclude minor sub-populations, which should be the focus under sub-population shifts, due to the main challenge: the *coexistence* of noisy samples and minor sub-populations. Zhai et al. [39] incorporate this idea into DRO. Still, their method requires an implicit assumption that the prediction errors of noisy samples are higher than those of minor sub-populations, which does not always hold in practice. And Bennouna et al. [3] build the uncertainty set via two measures, KL-divergence and Wasserstein distance, leading to a combined approach of KL-DRO and ridge regression. Despite this, as we discussed earlier, DRO tends to increase the noise level in data, making it difficult to fix using ridge regression.

3 Proposed Method

In this work, with a focus on *regression*, we introduce our Geometry-Calibrated DRO (GCDRO). The fundamental idea is to utilize data geometry to distinguish between random noisy samples and minor sub-populations. It is motivated by the fact that prediction errors for minor sub-populations typically exhibit local smoothness along the data manifold, a property that is *not* shared by noisy samples.

Formulation Given training dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^N$ and a finite weighted graph $G_N = (V, E, W)$ representing the inherent structure of sample covariates (V is the node set and E denotes the edge set). Denote the empirical marginal distribution as $\hat{P}_X$, the formulation of GCDRO is:

$$\min_{\theta\in\Theta} \underbrace{\sup_{\mathbf{q}:\mathcal{GW}^2_{G_N}(\hat{P}_X,\mathbf{q})\le\rho}}_{\text{Geometric Wasserstein set}} \left\{\mathcal{R}_N(\theta,\mathbf{q}) := \sum_{i=1}^N q_i\ell(f_\theta(x_i),y_i) - \underbrace{\frac{\alpha}{2}\cdot\sum_{(i,j)\in E} w_{ij}q_iq_j(\ell_i-\ell_j)^2}_{\text{Calibration Term I}} - \underbrace{\beta\cdot\sum_{i=1}^N q_i\log q_i}_{\text{Calibration Term II}}\right\}, \tag{3}$$

where ρ is the pre-defined radius of the uncertainty set, ℓ_i is the loss on the i-th sample, $w_{ij} \in W$ denotes the edge weight between sample i and j, $\mathcal{GW}^2_{G_N}(\cdot,\cdot)$ is the Discrete Geometric Wasserstein Distance [26] (see Appendix C for a detailed review). α and β are hyper-parameters.

In our formulation, for any distribution $\mathbf{q}$ within the uncertainty set,
Calibration term I ($\sum_{(i,j)\in E} w_{ij}q_iq_j(\ell_i-\ell_j)^2$) calculates the *graph total variation* of prediction errors along the data manifold that is characterized by G_N. Intuitively, when *selecting the worst-case distribution*, this term imposes a penalty on distributions that allocate high densities to random noisy samples, as this allocation significantly amplifies the overall variation in prediction errors. Conversely, this term does not penalize distributions that allocate high densities to minor sub-populations, as their errors are smooth and have a relatively small impact on the total variation along the manifold. Further, *during the optimization of model parameter θ*, this term acts like a variance term, resulting in a quantile-like risk objective, which helps to mitigate the effects of outliers.
Calibration term II ($\sum_{i=1}^N q_i \log q_i$) represents the negative entropy of distribution $\mathbf{q}$. As discussed in Section E, during optimization, this term transforms into a non-linear *graph Laplacian operator* that encourages sample weights to be smooth along the manifold, avoiding extreme sample weights in the worst-case distribution.

3.1 Free Energy Implications on Worst-case Distribution

We introduce the free energy implications of our risk objective $\mathcal{R}_N(\theta,\mathbf{q})$. Intuitively, the change of sample weights across N samples (the inner maximization problem of $\mathcal{R}_N(\theta,\mathbf{q})$) can be analogously related to the dynamics of particles in a system, wherein the concentration of densities coincides with the aggregation of particle masses at N distinct locations (in the case of infinite samples, these locations converge to the data manifold). Building on this analogy, we can dive deeper into the physics of particle interactions. When particles exist within a potential energy field, they are subject to external forces. Simultaneously, there are interactions among the particles themselves, leading to a constant state of motion within the system. In statistical physics, a key point of interest is identifying when a system reaches a steady state. In a standard process like the reversible isothermal process, it is established that spontaneous reactions consistently move in the direction of decreasing *Helmholtz free energy* [16, 32, 14], which consists of interaction energy, potential energy and the negative entropy:

$$\mathcal{E}(\mathbf{q}) = \underbrace{\mathbf{q}^\top K\mathbf{q}}_{\text{Interaction Energy}} + \underbrace{\mathbf{q}^\top V}_{\text{Potential Energy}} \underbrace{-\beta\sum_{i=1}^N(-q_i\log q_i)}_{\text{Temperature}\times\text{Entropy}} = -\mathcal{R}_N(\theta,\mathbf{q}). \tag{4}$$

By taking $V = -\vec{\ell}$ and $K_{ij} = \frac{\alpha}{2}w_{ij}(\ell_i-\ell_j)^2$ for $(i,j)\in E$, our risk objective is a special case of Helmholtz free energy, where the potential energy of sample i is $-\ell_i q_i$ and the interaction energy between sample i and j is $\frac{\alpha}{2}w_{ij}(\ell_i-\ell_j)^2q_iq_j$. Specifically, such mutual interactions can manifest as *repulsive forces between adjacent particles*, thereby preventing the concentration of mass in locations where local prediction errors are significantly high. And this explains from a physical perspective why our calibration term **I** could mitigate random noisy samples.

In Appendix D, we derive Proposition D.1 to offer physical interpretations to comprehend the worst-case distribution of various DRO methods including KL-DRO, χ^2-DRO, MMD-DRO, Marginal DRO and GDRO. In Appendix E, we utilize gradient flow in Wasserstein space to derive an approximate minimax optimization algorithm with a bounded error ratio.

3.2 Mitigate the Effects of Random Noisy Samples

Finally, we prove that our GCDRO method effectively de-emphasizes 'noisy samples' with locally non-smooth prediction errors. Due to the challenge of assessing intermediate states in gradient flow,

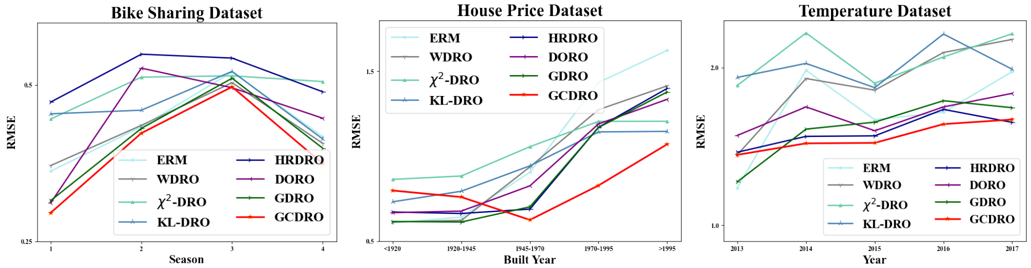

Figure 2: Results of real-world datasets with natural shifts. We do not manually add label noises here, since real-world datasets intrinsically contain noises.

we focus on its final state (as $T_{in} \to \infty$). Notably, in Proposition E.1, the convergence rate of gradient flow is $\mathcal{O}(e^{-CT_{in}})$, implying that an efficient approximation of the final state is feasible.

For the worst-case distribution q^*, we denote the density ratio between samples as $\gamma(i,j) := q_i^*/q_j^*$. In sensitivity analysis, when *only* sample i is perturbed with label noises, we denote the density ratio in the new worst-case distribution $\tilde{q^*}$ as $\gamma^{\text{noisy}}(i,j) := \tilde{q_i^*}/\tilde{q_j^*}$. The sample weight sensitivity $\xi(i,j)$ is defined as $\xi(i,j) = \log \gamma^{\text{noisy}}(i,j) - \log \gamma(i,j)$, which measures how much density ratio changes under perturbations on one sample. Larger $\xi(i,j)$ indicates larger sensitivity to noisy samples.

Proposition 3.1. *Assume $\ell_i^{noisy} - \ell_i \geq 2(\frac{\sum_{k\in N(i)} q_k^* w_{ik}\ell_k}{\sum_{k\in N(i)} q_k^* w_{ik}} - \ell_i)$ which is locally non-smooth. For any $\alpha > 0$ (in Equation 3), we have $\xi_{GCDRO} < \xi_{GDRO}$. Furthermore, there exists $M > 0$ such that for any $\alpha > M$, we have $\xi_{GCDRO}(i,j) < 0 < \min\{\xi_{\chi^2-DRO}(i,j), \xi_{GDRO}(i,j)(= \xi_{KL\text{-}DRO}(i,j))\}$, indicating that GCDRO is not sensitive to locally non-smooth noisy samples.*

In practice, we do a grid search over $\alpha \in [0.1, 10]$ on an independent held-out validation dataset to select the best α. The complexity of gradient flow scales *linearly* with sample size.

4 Experiments

In this section, we test the empirical performances of our proposed GCDRO on real-world *regression* datasets with natural distributional shifts. We compare with empirical risk minimization (ERM), WDRO, two typical f-DRO methods, including KL-DRO, χ^2-DRO [13], GDRO [26], HRDRO [3] and DORO [39], where HRDRO and DORO are designed to mitigate label noises. We use three real-world regression datasets, including bike-sharing prediction, house price, and temperature prediction, and introduce natural distribution shifts like spatial-temporal shifts. For all these experiments, we use a two-layer *MLP* model with mean square error (MSE). We use the Adam optimizer [22] with the default learning rate $1e-3$. And all methods are trained for $5e3$ epochs.

Due to space limits, we leave the results of simulation data in Appendix F, and experimental details could be found in Appendix J.

Analysis (1) From the results in Figure 2, we could see that the performances of ERM drop a lot under distributional shifts, and DRO methods have better performance as well as robustness. (2) Our proposed GCDRO outperforms all baselines under strong shifts, with the most stable performances under natural distributional shifts. (3) As for the kNN graph's fitting accuracy of the data manifold, we visualize the learned manifold in Appendix I and we could see that the learned kNN graph fits the data manifold well. Besides, we show in Figure 5 that the performances of our GCDRO are relatively stable across different choices of k. Also, our GCDRO only needs the input graph G_N to represent the data structure and *any manifold learning or graph learning* methods could be plugged in to give a better estimation of G_N.

5 Future Directions

Our work deals with the over-pessimism in DRO via geometric calibration terms and provides free energy implications. The high-level idea could inspire future research on (1) relating free energy with DRO; (2) designing more reasonable calibration terms in DRO; (3) incorporating data geometry in general risk minimization algorithms.

Acknowledgement

Peng Cui was supported in part by National Key R&D Program of China (No. 2018AAA0102004), National Natural Science Foundation of China (No. U1936219, 62141607).

References

[1] Alekh Agarwal and Tong Zhang. Minimax regret optimization for robust machine learning under distribution shift. *arXiv preprint arXiv:2202.05436*, 2022.

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

[3] Amine Bennouna and Bart Van Parys. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022.

[4] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

[5] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

[6] Bradley C. A. Brown, Anthony L. Caterini, Brendan Leigh Ross, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. The union of manifolds hypothesis and its implications for deep generative modelling. *CoRR*, abs/2207.02862, 2022.

[7] Theophilos Cacoullos. Estimation of a multivariate density. Technical report, University of Minnesota, 1964.

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

[9] Ilias Diakonikolas and Daniel M Kane. Algorithmic high-dimensional robust statistics. *Webpage http://www. iliasdiakonikolas. org/simons-tutorial-robust. html*, 2018.

[10] Ilias Diakonikolas, Daniel M Kane, Ankit Pensia, and Thanasis Pittas. Streaming algorithms for high-dimensional robust statistics. In *International Conference on Machine Learning*, pages 5061–5117. PMLR, 2022.

[11] Dheeru Dua and Casey Graff. Uci machine learning repository, 2017.

[12] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally robust losses for latent covariate mixtures. *Operations Research*, 2022.

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

[14] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

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

[16] XC Fu, WX Shen, TY Yao, and WH Hou. Physical chemistry. *Higher Education, Beijing*, 1990.

[17] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

[18] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 2022.

[19] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[20] 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.

[21] Peter J Huber. Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, pages 492–518, 1992.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

[24] Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 777–784, 2004.

[25] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, and Bo Li. Distributionally robust learning with stable adversarial training. *IEEE TKDE*, 2022.

[26] Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. Distributionally robust optimization with data geometry. In *Advances in Neural Information Processing Systems*, 2022.

[27] Dalton Lunga, Saurabh Prasad, Melba M Crawford, and Okan Ersoy. Manifold-learning-based feature extraction for classification of hyperspectral data: A review of advances in manifold learning. *IEEE Signal Processing Magazine*, 31(1):55–66, 2013.

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

[29] Hariharan Narayanan and Sanjoy K. Mitter. Sample complexity of testing the manifold hypothesis. In John D. Lafferty, Christopher K. I. Williams, John Shawe-Taylor, Richard S. Zemel, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada*, pages 1786–1794. Curran Associates, Inc., 2010.

[30] Arkadas Ozakin and Alexander G. Gray. Submanifold density estimation. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1375–1382. Curran Associates, Inc., 2009.

[31] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *ICLR*, 2021.

[32] Linda E Reichl. A modern course in statistical physics, 1999.

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

[34] Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[35] Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *International Conference on Artificial Intelligence and Statistics*, pages 1283–1297. PMLR, 2022.

[36] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[38] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J. Smola, and Zheng Zhang. Deep graph library: Towards efficient and scalable deep learning on graphs. *CoRR*, abs/1909.01315, 2019.

[39] Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Doro: Distributional and outlier robust optimization. In *International Conference on Machine Learning*, pages 12345–12355. PMLR, 2021.

A Example of Weighted Least Square

Example (Weighted Least Square)**.** *Consider the data generation process as $Y = kX + \xi$, where $X, Y \in \mathbb{R}$ and random noise ξ satisfies $\xi \perp X$, $\mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2]$ (*abbr. *σ^2) is finite. Assume that the training dataset X_D consists of clean samples $\{x_c^{(i)}, y_c^{(i)}\}_{i\in[N_c]}$ and noisy samples $\{x_o^{(i)}, y_o^{(i)}\}_{i\in[N_o]}$ with $\sigma_c^2 < \sigma_o^2$. Consider the weighted least-square model $f(X) = \theta X$. Denote the sample weight of a clean sample $(x_c^{(i)}, y_c^{(i)})$ as $w_c^{(i)} \in \mathbb{R}_+, i \in [N_c]$, and the sample weight of a noisy sample $(x_o^{(i)}, y_o^{(i)})$ as $w_o^{(i)} \in \mathbb{R}_+, i \in [N_o]$ with $\sum_{i\in[N_c]} w_c^{(i)} + \sum_{i\in[N_o]} w_o^{(i)} = 1$. The variance of the estimator $\hat{\theta}$ is given by:*

$$\mathit{Var}[\hat{\theta}|X_D] = \frac{\sum_{i=1}^{N_c}(w_c^{(i)})^2(x_c^{(i)})^2\sigma_c^2 + \sum_{i=1}^{N_o}(w_o^{(i)})^2(x_o^{(i)})^2\sigma_o^2}{\left[\sum_{i=1}^{N_c} w_c^{(i)}(x_c^{(i)})^2 + \sum_{i=1}^{N_o} w_o^{(i)}(x_o^{(i)})^2\right]^2}, \tag{5}$$

where $X_D = \{x_c^{(i)}\}_1^{N_c} \cup \{x_o^{(i)}\}_1^{N_o}$ are the sampled covariates in the dataset. Besides, the minimum variance is achieved if and only if $\forall 1 \le i \le N_c, 1 \le j \le N_o, w_o^{(j)}/w_c^{(i)} = \sigma_c^2/\sigma_o^2 < 1$.

From the results, we make the following remarks:
- If noisy samples have higher weights than clean samples (e.g., $w_o/w_c > 1$), the variance of the estimated parameter $\hat{\theta}$ will be larger, suggesting that the learned $\hat{\theta}$ could be significantly unstable.
- In conjunction with Proposition 2.1, DRO methods tend to assign high weights to noisy samples, which can lead to unstable parameter estimation.

B Relationship with Conventional Outlier-robust Regression

We would like to explain why conventional outlier-robust regression methods cannot be directly applied to our problem. The main challenge stems from the *coexistence* of noisy samples and minor sub-populations, both of which typically exhibit high prediction errors, leading to a misleading worst-case distribution in DRO. Conventional outlier-robust regression methods [9, 23, 10] primarily focus on mitigating the effects of outliers without considering sub-population shifts. For instance, the L_2-estimation-error of outlier-robust linear regression is $\mathcal{O}(\epsilon \log(1/\epsilon))$ [9], where ϵ represents the noise level in Equation 1. However, as analyzed in Proposition 2.1 and demonstrated in Figure 3, during the optimization of DRO, the noise level ϵ significantly increases, rendering even outlier-robust estimation quite inaccurate. Moreover, [23] propose finding a pseudo distribution with minimal prediction errors to avoid outliers (see Algorithm 5.2 in [23]). Nevertheless, this approach might inadvertently exclude minor sub-populations, which should be the focus under sub-population shifts, due to the main challenge: the *coexistence* of noisy samples and minor sub-populations. **(author?)** [39] incorporate this idea into DRO. Still, their method requires an implicit assumption that the prediction errors of noisy samples are higher than those of minor sub-populations, which does not always hold in practice. And **(author?)** [3] build the uncertainty set via two measures, KL-divergence and Wasserstein distance, leading to a combined approach of KL-DRO and ridge regression. Despite this, as we discussed earlier, DRO tends to increase the noise level in data, making it difficult to fix using ridge regression.

C Definition of Discrete Geometric Wasserstein Distance

We briefly revisit the definition of the discrete geometric Wasserstein distance. Given a weighted finite graph $G_N = (V, E, W)$, the probability set $\mathscr{P}(G_N)$ supported on the vertex set V is defined as $\mathscr{P}(G_N) = \{\mathbf{p} \in \mathbb{R}^N | \sum_{i=1}^N p_i = 1, p_i \ge 0, \text{for } i \in V\}$, and its interior is denoted as $\mathscr{P}_o(G_N)$. A velocity field $\mathbf{v} = (v_{ij})_{i,j\in V} \in \mathbb{R}^{N\times N}$ on G_N is defined on the edge set E satisfying that $v_{ij} = -v_{ji}$ if $(i,j) \in E$. $\xi_{ij}(\mathbf{p})$ is a function interpolated with the associated nodes' densities p_i, p_j. The flux function $\mathbf{pv} \in \mathbb{R}^{N\times N}$ on G_N is defined as $\mathbf{pv} := (v_{ij}\xi_{ij}(\mathbf{p}))_{(i,j)\in E}$ and its divergence is defined as $\text{div}_{G_N}(\mathbf{pv}) := -(\sum_{j\in V:(i,j)\in E} \sqrt{w_{ij}} v_{ij}\xi_{ij}(\mathbf{p}))_{i=1}^N \in \mathbb{R}^N$. Then for distributions

$\mathbf{p}_0, \mathbf{p}_1 \in \mathscr{P}_o(G_N)$, the discrete geometric Wasserstein distance [8, 26] is defined as:

$$\mathcal{GW}^2_{G_N}(\mathbf{p}_0, \mathbf{p}_1) := \inf_v \left\{ \int_0^1 \frac{1}{2} \sum_{(i,j)\in E} \xi_{ij}(\mathbf{p}(t)) v_{ij}^2 dt \quad \text{s.t.} \frac{d\mathbf{p}}{dt} + \text{div}_{G_N}(\mathbf{p}\mathbf{v}) = 0, \mathbf{p}(0) = \mathbf{p}_0, \mathbf{p}(1) = \mathbf{p}_1 \right\}. \tag{6}$$

Equation 6 computes the shortest (geodesic) length among all potential plans, integrating the total kinetic energy of the velocity field throughout the transportation process. A key distinction from the Wasserstein distance is that it only permits density to appear at the graph nodes.

D Free Energy Implications

Proposition D.1 offers physical interpretations to comprehend the worst-case distribution of various DRO methods. We make some remarks: (1) current DRO methodologies, except MMD-DRO, do not explicitly formulate the interaction term between samples in their design considerations (χ^2-DRO does not involve interaction between samples), despite the corresponding interaction energy between particles being a common phenomenon in physics; (2) MMD-DRO simply uses kernel gram matrix for interaction and lacks efficient optimization algorithms; (3) by *considering this interaction energy*, our proposed GCDRO is capable of mitigating the impacts of random noisy samples.

Proposition D.1 (Free Energy Implications)**.** *The dual reformulations of some typical DRO methods are equivalent to the free-energy-based minimax problem* $\min_{\theta\in\Theta,\lambda\geq 0} \max_{\mathbf{q}\in\mathscr{P}} \left\{ \lambda\rho - \mathcal{E}(\mathbf{q}, \theta, \lambda) \right\}$ *with different choices of* $\mathscr{P}$, ρ *and* $K, V, H[q]$ *in the free energy* $\mathcal{E}$*. Details are shown in Table 1.*

Table 1: Free energy implications of some DRO methods. Δ_N denotes the N-dimensional simplex, η in marginal DRO is the dual parameter.

Method	Energy Type			Specific Formulation			
	Interaction	Potential	Entropy	K	V	$H[\mathbf{q}]$	$\mathscr{P}$
KL-DRO	✗	✓	✓	-	$-\vec{\ell}$	$H[\mathbf{q}]$	Δ_N
χ^2-DRO	✓	✓	✗	λI	$-\vec{\ell}$	-	Δ_N
MMD-DRO	✓	✓	✗	Kernel Gram Matrix K	$-\vec{\ell} - \frac{2\lambda}{N} K^\top \mathbf{1}$	-	Δ_N
Marginal χ^2-DRO	✗	✓	✗	-	$-(\vec{\ell} - \eta)_+$	-	Δ_N with Hölder continuity
GDRO	✗	✓	✓	-	$-\vec{\ell}$	$H[\mathbf{q}]$	Geometric Wasserstein Set
GCDRO	✓	✓	✓	Interaction Matrix K	$-\vec{\ell}$	$H[\mathbf{q}]$	Geometric Wasserstein Set

Through free energy, we could understand the type of energy or steady state that DRO methods strive to achieve, and design better interaction energy terms in DRO. Moreover, our optimization, as outlined in Section E, could accommodate multiple quadratic forms of interaction energy.

E Optimization

Then we derive an approximate minimax optimization for our GCDRO. For the *inner maximization* problem, we approximately deal with it via the gradient flow of $-\mathcal{R}_N(\theta, Q)$ w.r.t. Q in the geometric Wasserstein space $(\mathscr{P}_o(G_N), \mathcal{GW}_{G_N})$. We show that the error rate is $\mathcal{O}(e^{-CT_{in}})$ after T_{in} iterations inner loop, which gives a nice approximation. For the *outer minimization* w.r.t. model parameters θ, we analyze the convergence rate of $\mathcal{O}(1/\sqrt{T_{out}})$ after T_{out} iterations outer loop when the risk function satisfies Lipschitzian smoothness conditions.

Inner Maximization. We denote the *Continuous gradient flow* as $\mathbf{q} : [0, T] \to \mathscr{P}_o(G_N)$, the probability density of sample i at time t is abbreviated as $q_i(t)$, and the *Time-discretized gradient flow* with time step τ as $\hat{\mathbf{q}}_\tau$. For inner maximization, we utilize the τ-time-discretized gradient flow [37] for $-\mathcal{R}_N(\theta, \mathbf{q})$ in the geometric Wasserstein space $(\mathscr{P}_o(G_N), \mathcal{GW}^2_{G_N})$ as:

$$\hat{\mathbf{q}}_\tau(t+\tau) = \underset{\mathbf{q}\in\mathscr{P}_o(G_N)}{\text{argmax}} \; \mathcal{R}_N(\theta, \mathbf{q}) - \frac{1}{2\tau}\mathcal{GW}^2_{G_N}(\hat{\mathbf{q}}_\tau(t), \mathbf{q}). \tag{7}$$

The gradient of $\mathbf{q}$ in Equation 7 is given as (when $\tau \to 0$):

$$\frac{dq_i}{dt} = \sum_{(i,j)\in E} w_{ij}\xi_{ij}\bigg(\mathbf{q},\; \ell_i - \ell_j + \beta(\log q_j - \log q_i) + \alpha\big(\sum_{h\in N(j)} (\ell_h - \ell_j)^2 w_{jh} q_h - \sum_{h\in N(i)} (\ell_h - \ell_i)^2 w_{ih} q_h\big)\bigg), \tag{8}$$

where E is the edge set of G_N, w_{ij} is the edge weight between node i and j, $N(i)$ denotes the set of neighbors of node i, ℓ_i denotes the loss of sample i, and $\xi_{ij}(\cdot,\cdot) : \mathscr{P}(G_N) \times \mathbb{R} \to \mathbb{R}$ is:

$$\xi_{ij}(\mathbf{q}, v) := v \cdot \big(\mathbb{I}(v > 0)q_j + \mathbb{I}(v \le 0)q_i\big), v \in \mathbb{R}, \tag{9}$$

which is the *upwind interpolation* commonly used in statistical physics and guarantees that the probability vector $\mathbf{q}$ keeps positive. From the gradient, we could see that the entropy regularization acts as a non-linear graph Laplacian operator to make the sample weights smooth along the manifold. In our algorithm, we fix the steps of the gradient flow to be T_{in} and prove that the error ratio is $e^{-CT_{in}}$ compared with the *ground-truth* worst-case risk $\mathcal{R}_N(\theta, \mathbf{q}^*)$ constrained in an $\rho(\theta, T_{in})$-radius ball.

Proposition E.1 (Approximation Error Ratio)**.** *Given the model parameter θ, denote the distribution after time T_{in} as $\mathbf{q}^{T_{in}}(\theta)$, and the distance to training distribution $\hat{P}_X$ as $\rho(\theta, T_{in}) := \mathcal{GW}^2_{G_N}(\hat{P}_X, \mathbf{q}^{T_{in}}(\theta))$ (*abbr. *$\rho(\theta)$). Assume $\mathcal{R}_N(\theta, \mathbf{q})$ is convex w.r.t $\mathbf{q}$. Then define the ground-truth worst-case distribution $q^*(\theta)$ within the $\rho(\theta)$-radius ball as:*

$$\mathbf{q}^*(\theta) := \arg \sup_{\mathbf{q}: \mathcal{GW}^2_{G_N}(\hat{P}_X, \mathbf{q}) \le \rho(\theta)} \mathcal{R}_N(\theta, \mathbf{q}). \tag{10}$$

The upper bound of the error rate of the objective function $\mathcal{R}_N(\theta, \mathbf{q}^{T_{in}})$ satisfies:

$$(\mathcal{R}_N(\theta, \mathbf{q}^*) - \mathcal{R}_N(\theta, \mathbf{q}^{T_{in}}))/\Big(\mathcal{R}_N(\theta, \mathbf{q}^*) - \mathcal{R}_N(\theta, \hat{P}_X)\Big) < e^{-CT_{in}}, \tag{11}$$

$$C = 2m\lambda_{sec}(\hat{L})\lambda_{min}(\nabla^2 \mathcal{R}_N)\frac{1}{(r+1)^2} > 0, \tag{12}$$

where $\hat{L}$ is the Laplacian matrix of G_N. $\lambda_{sec}, \lambda_{min}$ are the second smallest and smallest eigenvalue, m, r are constants depending on $\mathcal{R}_N, G_N, \beta$.

We make some remarks:

- For the assumption that $\mathcal{R}_N$ is convex w.r.t. $\mathbf{q}$, the Hessian is given by $\nabla^2 \mathcal{R}_N = \beta \text{diag}(1/q_1, ..., 1/q_N) + 2K$. Since K is a sparse matrix whose nonzero elements in each row is far smaller than N, it is easily satisfied in empirical settings that the Hessian matrix $\nabla^2 \mathcal{R}$ is diagonally dominant and thus positive definite, making the inner maximization concave w.r.t $\mathbf{q}$.
- During the optimization, our algorithm finds an approximate worst-case distribution that is close to the ground-truth one within a $\rho(\theta)$-radius uncertainty set. Our robustness guarantee is similar to [34] (see Equation 12 in [34]).
- The error ratio is $e^{-CT_{in}}$, enabling to find a nice approximation efficiently with finite T_{in} steps.

Outer Minimization. The convergence property relies on the risk objective $\mathcal{R}_N(\theta, \mathbf{q})$. When $\mathcal{R}_N(\theta, \mathbf{q})$ is *smooth* w.r.t. θ, the following proposition guarantees convergence to a stationary point of problem 3 at a standard rate of $\mathcal{O}(1/\sqrt{T})$.

F Empirical Results on Simulation Data

Table 2: Results on the simulation data. We report the root mean square errors.

	Weak Label Noise (noise level 0.5%)					Strong Label Noise (noise level 5%)				
	Train (major)	Train (minor)	Test Mean	Test Std	Parameter Est Error	Train (major)	Train (minor)	Test Mean	Test Std	Parameter Est Error
ERM	0.337	0.850	0.598	0.264	0.423	0.368	0.855	0.599	0.243	0.431
WDRO	0.337	0.851	0.589	0.292	0.424	0.368	0.857	0.600	0.268	0.432
χ^2-DRO	0.596	0.765	0.680	0.088	0.447	1.072	0.708	0.875	0.193	0.443
KL-DRO	0.379	1.616	0.974	0.660	0.886	0.468	1.683	1.037	0.621	0.913
HRDRO	0.325	1.298	0.794	0.516	0.693	0.330	1.343	0.801	0.522	0.694
DORO	0.347	0.793	0.565	0.230	0.384	0.334	0.919	0.611	0.295	0.449
GDRO	0.692	0.516	0.605	0.094	0.198	0.618	0.752	0.677	0.063	0.421
GCDRO	0.411	0.554	**0.482**	**0.070**	**0.190**	0.494	0.591	**0.540**	**0.044**	**0.268**

Data Generation. We design simulation settings with both sub-population shifts and noisy samples. The input covariates $X = [S, U, V]^T \in \mathbb{R}^{10}$ consist of stable covariates $S \in \mathbb{R}^5$, irrelevant ones $U \in \mathbb{R}^4$ and the unstable covariate $V \in \mathbb{R}$:

$$[S, U] \sim \mathcal{N}(0, 2\mathbb{I}_9), Y = \theta_S^T S + 0.1 S_1 S_2 S_3 + \mathcal{N}(0, 0.5), V \sim \text{Laplace}(\text{sign}(r) \cdot Y, 1/5 \ln |r|), \tag{13}$$

where $\theta_S \in \mathbb{R}^5$ is the coefficients of the true model, $|r| > 1$ is the adjustment factor for each sub-population, and $\text{Laplace}(\cdot, \cdot)$ denotes the Laplace distribution. From the data generation, the

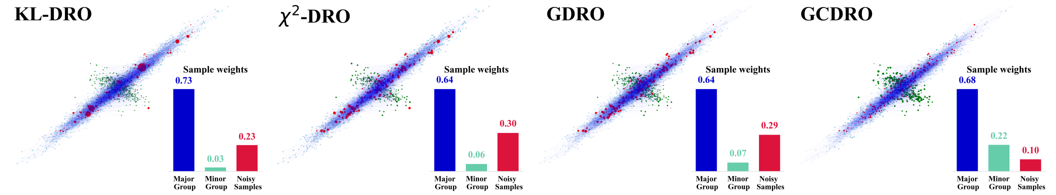

Figure 3: Visualizing the Worst-Case Distribution for Different DRO Methods: We show the data manifold and sample weights for each point, where blue points represent the major group, green ones represent the minor group, and red ones are noisy samples. The bars display the total sample weights of different groups, and the *original* group ratio is major (**93.1%**), minor (**4.9%**), (noisy **2%**).

relationship between S and Y stays invariant under different r, $U \perp Y$, while the relationship between V and Y is controlled by r, which *varies across sub-populations*. Intuitively, sign(r) controls whether the spurious correlation V-Y is positive or negative. And $|r|$ controls the strength of the spurious correlation: the larger $|r|$ is, the stronger the spurious correlation is. Furthermore, in order to conform to real data which are naturally assembled with label noises [39], we introduce label noises by an ϵ proportion of labels as $Y' \sim \mathcal{N}(0, \text{Std}(Y))$. ϵ controls the noise level.

Settings. In training, we generate 9,500 points with $r = 1.9$ (*majority*, strong positive spurious correlation V-Y) and 500 points with $r = -1.3$ (*minority*, weak negative spurious correlation V-Y). In testing, we vary $r \in \{3.0, 2.3, -1.9, -2.7\}$ to simulate different spurious correlations V-Y. We use *linear model* with mean square error (MSE) and report the prediction root-mean-square errors (RMSE) for each sub-population, the mean and standard deviation of prediction errors among all testing sub-populations. Also, we report the parameter estimation errors $\|\hat{\theta} - \theta^*\|_2$ of all methods ($\theta^* = (\theta_S^T, 0, \ldots, 0)^T$). The results over 10 runs are shown in Table 2.

Analysis. From Table 2, (1) compared with ERM, all typical DRO methods, especially χ^2-DRO and KL-DRO, are strongly affected by label noises. (2) Although DORO is designed to mitigate outliers, it does not perform well under strong noises ($\kappa = 5\%$), because it relies on the assumption that noisy points have the largest prediction errors, which does not always hold. (3) Our proposed GCDRO outperforms all baselines under different strengths of label noises, which demonstrates its effectiveness. (4) Compared with GDRO, we could see that our *calibration terms* in Equation 3 is effective to mitigate label noises. From Figure 3, the worst-case distribution of our GCDRO *significantly upweighs on the minority* (green points) and does not put much density on the noisy data (red points), while the others put much higher weights on the noisy samples and perform poorly.

G Implementation

For our GCDRO, G_N is constructed as a k-nearest neighbor (kNN) graph from training data *once and for all* ***only at the initialization step***. For large-scale datasets, we use NN-Descent to estimate the kNN graph with an almost linear complexity of $\mathcal{O}(\mathbf{N^{1.14}})$. Since the sample weights are transferred along the edges of the graph, the simulation of gradient flow can be implemented similarly to message propagation with DGL package [38], which ***scales linearly with sample size*** and enjoys *parallelization by GPU*. The implementation above ensures the adaptability to large-scale data.

H Improvements of our work.

In Section 2, we have introduced the typical DRO methods in detail and demonstrated the over-pessimism problem. Here we compare our work with several DRO works and clarify their differences.
(1) With MMD-DRO: MMD-DRO [36] also has a quadratic term in its dual reformulation, while [36] focuses on the equivalence between MMD-DRO and Hilbert norms and there is no efficient or applicable algorithm yet. Further, it remains the risk objective unchanged (the quadratic term is from MMD distance) and just uses the Gaussian RBF kernel. Our work firstly incorporates the data geometry into the design of the calibration term and demonstrates its relationship with Helmholtz free energy, and we propose an applicable algorithm that could be used under deep models.
(2) With GDRO: GDRO [26] uses the discrete geometric Wasserstein distance to build the uncertainty set, and intuitively demonstrates its superiority. Our work theoretically analyzes the over-pessimism

problem and attributes the cause of over-pessimism to the excessive focus on noisy samples in DRO. And for the risk objective function, our work further introduces the graph total variation term to mitigate the effects of noisy samples, which is theoretically justified and empirically verified. From our results, GDRO is heavily affected by noisy samples, while our GCDRO has a much better performance. Further, this work relates the newly-proposed risk objective to the Helmholtz free energy and unifies some typical DRO methods into it, which is a new perspective to view DRO methods and could inspire future research.
(3) With DORO: DORO [39] proposes to dismiss data samples with the top losses and then performs DRO, and we compare with it in our experiments. Theoretically, this method relies on the implicit assumption that noisy samples must have larger prediction errors than hard clean samples. However, this assumption does not always hold, and as shown in our experiments, it has some effects but does not work very well.

I Why uses kNN graph?

Manifold Assumption. The data manifold hypothesis indicates that high-dimensional data often lies in an unknown lower-dimensional manifold embedded in ambient space [33, 2, 24, 27, 6] and is supported by strong evidence. From a theoretical perspective, **(author?)** [30, 29] prove that when such hypothesis holds, manifold learning and density estimation scale exponentially with the *low intrinsic* dimension, but otherwise scale exponentially with the *high ambient* dimension [7]. Therefore, as **(author?)** [6] point out, one most plausible explanation for the success of machine learning methods on real-world data is the existence of such lower intrinsic dimension, which enables learning on datasets of fairly reasonable size, which is empirically verified by **(author?)** [31]. Also, for two of the real-world tabular datasets used in this work, we visualize their 3-dimensional manifolds and calculate their intrinsic dimensions in Figure 4.

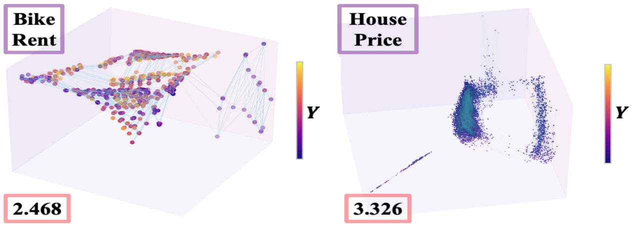

Figure 4: Visualization of the 3-dimensional manifold of the tabular datasets, and the numbers in the lower left represent the intrinsic dimension according to [24]

Our GCDRO algorithm uses an input-weighted graph G_N to approximate the data manifold. The kNN graph is a fundamental and basic way to represent the data structure, and manifold learning is an area with intensive research. We have to clarify that manifold learning is not the focus of this paper, which takes the data structure G_N as input to design a DRO objective and optimization algorithm that incorporates data geometric information for more reasonable worst-case distribution. Notably, our GCDRO achieves significant performance in the experiments even with the simple kNN representation of data structure. It proves that this direction for geometric-aware DROs is promising, and our proposed method could efficiently leverage the geometric properties encoded in the input graph to mitigate the effects of harmful data points (note that no target information is leaked into G_N). Actually, our GCDRO is compatible with any manifold learning or graph learning method. We do believe that a more accurate estimated data structure with advanced manifold learning algorithms will further boost the performance of GCDRO, and we leave this to future work.

Not Sensitive to k. For the house pricing dataset, we plot the results of our GCDRO with varying ks in Figure 5. We could see that the performance of our algorithm is not affected much.

J Experimental Details

Model & Loss function. For simulation data, we use linear models for all methods. For real-world data, we use two-layer MLPs for all methods.

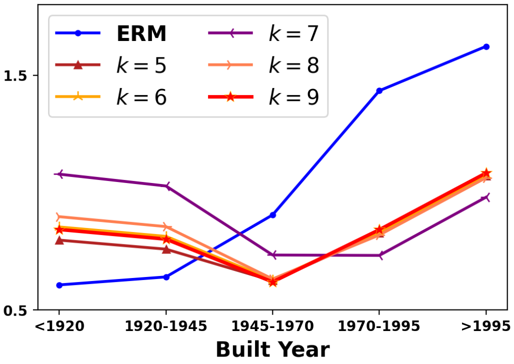

Figure 5: Results with varying k.

Optimizer. For all experiments, we use Adam with a learning rate of $1e-3$ in PyTorch for all methods.

Hyper-parameters. For KLDRO, WDRO and χ^2-DRO, we grid search the radius of the uncertainty set within the range of $[1e-3, 2e2]$, and we select the best hyper-parameters according to their testing performances. For GDRO, we grid search the number of gradient flow steps within the range of $[1e2, 2e3]$, the parameter $\beta \in [1, 20]$ and we select the best hyper-parameters according to its testing performances. For DORO, we set the noisy ratio to the ground-truth value for the simulation data, and we grid search the ratio of noisy points within the range of $[1e-2, 5e-1]$ for the real-world data. For HRDRO, we use L_1 loss as proposed in [3] and grid search $\epsilon \in [1e-3, 1]$. For GCDRO, we grid search the number of gradient flow steps within the range of $[1e2, 2e3]$, $\beta \in [1, 20]$ and $\alpha \in [1e-1, 1e1]$. We select the best hyper-parameters according to their testing performances.

Note that in our experiments, we found that model selection without domain information in the validation set is very hard, which is also verified by [39, 19]. And we believe this is still an open problem and is fairly non-trivial.

Real-World Datasets
(1) **Bike-sharing** dataset [11] contains the daily count of rental bikes in the Capital bike-sharing system with the corresponding 11 weather and seasonal covariates. The task is to predict the count of rental bikes of *casual users*. Note that the count of casual users is likely to be more *random and noisy*, which is suitable to verify the effectiveness of our method. We split the dataset according to the season for natural shifts. In the training data, the ratio of four seasons' data is $9:7:5:3$. We test on the rest of the data and report the prediction error of each season.
(2) **House Price** dataset[1] contains house sales prices from King County, USA. The task is to predict the transaction price of the house via 17 predictive covariates such as the number of bedrooms, square footage of the house, etc. We divide the data into 5 sub-populations according to the built year of each house with each sub-population covering a span of 25 years. In training, we use data from the first group (built year < 1920) and report the prediction error for each testing group.
(3) **Temperature** dataset [11] is largely composed of the LDAPS model's next day's forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables in South Korea from 2013 to 2017. The task is to predict the next-day's maximum air temperatures based on the 22 covariates. We divide the data into 5 groups corresponding with 5 years. In the training data, the ratio of five years' data is $9:7:5:3:1$. We test on the rest of the data and report the prediction error of each year. More details could be found in Appendix.

K Examples on Label Noise

Theorem K.1. *Assume that the training data is a mixture of n_c clean samples $\{x_c^{(i)}, y_c^{(i)}\}$ drawn from distribution $P_c(X,Y)$ and n_o noisy samples $\{x_o^{(i)}, y_o^{(i)}\}$ drawn from distribution $P_o(X,Y)$. Consider a linear data generation process, i.e. $Y = kX + \xi$ and $\xi \perp X, \mathbb{E}[\xi] = 0$ and $\mathbb{E}[\xi^2]$ is finite*

[1] https://www.kaggle.com/c/house-prices-advanced-regression- techniques/data

(abbr. σ^2). The regression model is parameterized as $f(x) = \theta \cdot x$ and trained with Weighted Least Square estimation:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n_c} w_c^{(i)} \|(y_c^{(i)} - \theta \cdot x_c^{(i)})\|^2 + \sum_{i=1}^{n_o} w_o^{(i)} \|(y_o^{(i)} - \theta \cdot x_o^{(i)})\|^2. \tag{14}$$

$$s.t. \quad \sum_{i=1}^{n_c} w_c^{(i)} + \sum_{i=1}^{n_o} w_o^{(i)} = 1, \tag{15}$$

where $w_c^{(i)}, w_o^{(i)} \geq 0$ are weights on clean and noisy samples respectively, and $\sigma_c^2 < \sigma_o^2$. Then the variance of the least square estimate $\hat{\theta}$ is given by:

$$\mathit{Var}[\hat{\theta}|X_D] = \frac{\sum_{i=1}^{n_c}(w_c^{(i)})^2 (x_c^{(i)})^2 \sigma_c^2 + \sum_{i=1}^{n_o}(w_o^{(i)})^2 (x_o^{(i)})^2 \sigma_o^2}{\left[\sum_{i=1}^{n_c} w_c^{(i)} (x_c^{(i)})^2 + \sum_{i=1}^{n_o} w_o^{(i)} (x_o^{(i)})^2\right]^2}, \tag{16}$$

where $X_D = \{x_c^{(i)}\} \cup \{x_o^{(i)}\}$ is the sampled covariates in the dataset. Further, the variance of the estimator $\hat{\theta}$ achieves the minimum if and only if:

$$\forall 1 \leq i \leq n_c, 1 \leq j \leq n_o, \;\; \gamma(i,j) = w_o^{(j)} / w_c^{(i)} = \sigma_c^2 / \sigma_o^2, \tag{17}$$

where $\gamma(i,j)$ denotes the sample weight ratio between i and j.

The theorem is a direct corollary of the following lemma.

Lemma K.1. *Assume that the training data contains n samples $\{x^{(i)}, y^{(i)}\}$. Consider a linear data generation process with heterogeneous noise, i.e. $y^{(i)} = kx^{(i)} + \xi_i$ with $\xi_i \perp X, \mathbb{E}[\xi_i] = 0$, and $\mathbb{E}[\xi_i^2]$ is finite. The regression model is parameterized as $f(x) = \theta \cdot x$ and trained with Weighted Least Square estimation:*

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} w^{(i)} \|(y^{(i)} - \theta \cdot x^{(i)})\|^2. \tag{18}$$

$$s.t. \quad \sum_{i=1}^{n} w^{(i)} = 1, \tag{19}$$

where $w^{(i)} \geq 0$ are sample weights. Then the variance of the least square estimate $\hat{\theta}$ is given by:

$$\mathit{Var}[\hat{\theta}|X_D] = \frac{\sum_{i=1}^{n}(w^{(i)})^2 (x^{(i)})^2 \sigma_i^2}{\left[\sum_{i=1}^{n} w^{(i)} (x^{(i)})^2\right]^2}, \tag{20}$$

where $X_D = \{x^{(i)}\}$ is the sampled covariates in the dataset. Further, the variance of the estimator $\hat{\theta}$ achieves the minimum if and only if:

$$\forall 1 \leq i \leq n, 1 \leq j \leq n, \;\; w^{(i)} \sigma_i^2 = w^{(j)} \sigma_j^2. \tag{21}$$

Proof. According to the heterogeneous noise distribution, let $y^{(i)} = x^{(i)} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. The least square estimation of $\hat{\theta}$ is given by:

$$\hat{\theta} = k + \frac{\sum_{i=1}^{n} w^{(i)} x^{(i)} \epsilon_i}{\sum_{i=1}^{n} w^{(i)} (x^{(i)})^2}. \tag{22}$$

Since $\mathbb{E}[\hat{\theta}|X_D] = k$, we have

$$\begin{aligned} \mathrm{Var}[\hat{\theta}|X_D] &= \mathbb{E}\left|\frac{\sum_{i=1}^{n} w^{(i)} x^{(i)} \epsilon_i}{\sum_{i=1}^{n} w^{(i)} (x^{(i)})^2}\right|^2 & (23) \\ &= \frac{\sum_{i=1}^{n}(w^{(i)})^2 (x^{(i)})^2 \sigma_i^2}{\left[\sum_{i=1}^{n} w^{(i)} (x^{(i)})^2\right]^2}. & (24) \end{aligned}$$

Next, we solve the minimum of Eq.20 w.r.t. sample weights $w^{(i)}$. Let $\alpha_i = w^{(i)}(x^{(i)})^2$. We could formulate the variance in Eq.20 as a function of $\alpha = (\alpha_1, ..., \alpha_n)$:

$$V(\alpha) = \frac{\sum_{i=1}^n \alpha_i^2 \sigma_i^2/(x^{(i)})^2}{\left(\sum_{i=1}^n \alpha_i\right)^2}. \tag{25}$$

Since $V(\lambda\alpha) = V(\alpha)$ for any $\lambda > 0$, we could assume $\sum_{i=1}^n \alpha_i = 1$ without loss of generality. Then the minimization of $V(\alpha)$ is equivalent to:

$$\min_\alpha V(\alpha) = \sum_{i=1}^n \alpha_i^2 \sigma_i^2/(x^{(i)})^2. \tag{26}$$

$$s.t. \ \ \sum_{i=1}^n \alpha_i = 1. \tag{27}$$

The first-order KKT condition gives:

$$\exists C, \forall 1 \le i \le n, \ \ \alpha_i^* = C(x^{(i)})^2/\sigma_i^2, \tag{28}$$

from which we can solve:

$$\alpha_i^* = \frac{(x^{(i)})^2/\sigma_i^2}{\sum_{j=1}^n (x^{(j)})^2/\sigma_j^2}. \tag{29}$$

Since $\nabla_\alpha^2 V(\alpha) = diag\left[2\sigma_1^2/(x^{(1)})^2, ..., 2\sigma_n^2/(x^{(n)})^2\right]$ is always positive definite, Eq.29 minimizes $V(\alpha)$. Correspondingly $w^{(i)} \propto 1/\sigma_i^2$, which finishes the proof. □

L Proofs

L.1 Proof of Proposition 2.1

Proof. (1) For KL-divergence as the distance function, we have the following optimization problem under finite samples.

$$\min_{\theta\in\Theta,\lambda\ge 0} \sup_{\mathbf{p}\in\Delta_n} \left\{ \sum_{i=1}^n p_i \ell(f_\theta(x_i), y_i) - \lambda \sum_{i=1}^n p_i \log p_i + \lambda(\epsilon - \log n) \right\}, \tag{30}$$

Solve the inner supremum problem, and the worst-case distribution is like:

$$p_i = \exp\left(\frac{\ell_i - \eta}{\lambda} - 1\right), \ \eta(\ell) = \lambda \log \lambda + \lambda \log \left(\sum_{i=1}^n \exp(\frac{\ell_i}{\lambda} - 1)\right), \tag{31}$$

and the objective function becomes:

$$\min_{\theta\in\Theta,\lambda\ge 0} \lambda \log \left(\sum_{i=1}^n \exp(\frac{\ell(f_\theta(x_i), y_i))}{\lambda})\right) + \lambda(\epsilon + \log\lambda - \log n). \tag{32}$$

And we could compare the sample weights of different samples as:

$$\frac{p_i}{p_j} = \exp(\frac{\ell_i - \ell_j}{\lambda}). \tag{33}$$

(2) For χ^2-divergence which is defined as $f(x) = (x-1)^2$, we have the following optimization problem.

$$\min_{\theta\in\Theta,\lambda\ge 0} \sup_{\mathbf{p}\in\Delta_n} \left\{ \sum_{i=1}^n p_i \ell(f_\theta(x_i), y_i) + \lambda\epsilon - \frac{\lambda}{n}\sum_{i=1}^n (np_i - 1)^2 \right\}. \tag{34}$$

Solve the inner supremum problem, and we have the worst-case distribution like:

$$p_i = \frac{1}{\lambda n}(\ell_i + \lambda - \eta)_+, \tag{35}$$

and the objective function becomes:

$$\min_{\theta\in\Theta,\lambda\geq 0,\eta\in\mathbb{R}} \sum_{i=1}^{n} \frac{1}{2\lambda}(\ell_i+\lambda-\eta)_+^2 + \lambda\epsilon + \eta - \frac{\lambda}{2}. \tag{36}$$

And we could compare the sample weights of different samples as:

$$\frac{p_i}{p_j} = \frac{(\ell_i+\lambda-\eta)_+}{(\ell_j+\lambda-\eta)_+}, \tag{37}$$

if $p_j > 0$.

(3) For Maximal Mean Discrepancy (MMD) distance, we have the following optimization problem:

$$\sup_{\mathbf{p}} \left\{ \sum_{i=1}^{n} p_i\ell_i + \lambda\epsilon - \lambda(\mathbf{p}-\frac{\mathbf{1}}{\mathbf{n}})^{\mathbf{T}}\mathbf{K}(\mathbf{p}-\frac{\mathbf{1}}{\mathbf{n}}) \right\} \tag{38}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i = 1 \tag{39}$$

$$p_i \geq 0, \text{for } i = 1, \ldots, n \tag{40}$$

Solve the inner supremum problem, and we have the worst-case distribution like:

$$p^* = \frac{1}{2\lambda}K^{-1}(\ell - \eta + \frac{2\lambda}{n}K\mathbf{1})_+, \tag{41}$$

and the objective function becomes:

$$\min_{\theta\in\Theta,\lambda\geq 0,\eta\in\mathbb{R}} \frac{1}{4\lambda}(\ell + \frac{2\lambda}{n}K\mathbf{1} - \eta)_+ K^{-1}(\ell + \frac{2\lambda}{n}K\mathbf{1} - \eta)_+ + \lambda\epsilon + \eta - \frac{\lambda}{n^2}\mathbf{1}^T K\mathbf{1}. \tag{42}$$

□

L.2 Proof of Proposition 3.1

Proof. It is easy to prove that the final state of $\mathcal{R}_N(\theta, q)$ w.r.t. q is given as

$$q_i^\infty = \frac{1}{Z}\exp(\frac{\ell_i - \alpha(\sum_{j\in N(i)} q_j^\infty w_{ij}(\ell_i-\ell_j)^2)}{\beta}), \tag{43}$$

where

$$Z = \sum_{i=1}^{N} \exp(\frac{\ell_i - \alpha(\sum_{j\in N(i)} q_j^\infty w_{ij}(\ell_i-\ell_j)^2)}{\beta}). \tag{44}$$

(1) When $\beta \to \infty$, $q_i^\infty \to \frac{1}{N}$. When $\beta \ll \infty$, the gradient flow is like:

$$\frac{dq_i}{dt} = \sum_{(i,j)\in E} w_{ij}\xi_{ij}\Big(\ell_i - \ell_j + \beta(\log q_j - \log q_i) + \alpha\big(\sum_{h\in N(j)}(\ell_h-\ell_j)^2 w_{hj}q_h - \sum_{h\in N(i)}(\ell_h-\ell_i)^2 w_{hi}q_h\big)\Big), \tag{45}$$

and

$$\xi_{ij}(v) := v\cdot\big(\mathbb{I}(v>0)q_j + \mathbb{I}(v\leq 0)q_i\big). \tag{46}$$

Therefore, when $q_i > q_j$ and $\ell_i > \ell_j$, we have $\log q_j - \log q_i < 0$, which decreases the gradient of q_i. Thus, the entropy term prompts the sample weights to be smooth between neighbors. When the sample weight of sample i is larger than its neighbors, this term will decrease the gradient of q_i to prevent it from gaining too much weights.

(3) Under the assumptions, we have

$$(\sum_{j\in N(i)} q_j^\infty w_{ij}(\ell_i+\Delta_i-\ell_j)^2 - \sum_{j\in N(i)} q_j^\infty w_{ij}(\ell_i-\ell_j)^2) \tag{47}$$

$$= \sum_{j\in N(i)} q_j^\infty w_{ij}(2\ell_i - 2\ell_j + \Delta_i)\Delta_i \tag{48}$$

$$\geq \Delta_i\big(\sum_{j\in N(i)} q_j^\infty w_{ij}(\Delta_i - 2L_x\|x_i - x_j\|_2)\big) \tag{49}$$

$$> \Delta_i. \tag{50}$$

Therefore, define $\delta_i = \ell_i^{\text{noisy}} - \ell_i$, it is easy to prove that for $\alpha > 0$

$$\xi_{\text{GCDRO}}(i, j) < \xi_{\text{GDRO}}(i, j), \tag{51}$$

and when $\alpha > \frac{1}{\sum_{k \in N(i)} q_k w_{ik} (2\ell_i - 2\ell_k + \delta_i)}$, we have $\xi_{\text{GCDRO}}(i, j) < 0$. □

L.3 Proof of Proposition D.1

Proof. Please refer to the proof of Proposition 2.1 for the proof of KL-DRO, χ^2-DRO and MMD DRO. For marginal DRO, it is easy to prove following [12]. For GDRO, it is easy to prove following [26]. □

L.4 Proof of Proposition E.1

Proof. The proof is based on the Theorem 5 in [8]. From [8], we have

$$\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q(t)) \leq e^{-Ct}(\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q^0)). \tag{52}$$

Furthermore,

$$C := 2m\lambda_{\text{sec}}(\hat{L})\lambda_{\text{min}}(\nabla^2\mathcal{R}_N)\frac{1}{(r+1)^2} > 0, \tag{53}$$

and

$$r = \sqrt{2}k \max_{(i,j)\in E} w_{ij} \frac{\|\text{Hess}\mathcal{R}_N\|_1}{\lambda_{\text{min}}(\text{Hess}\mathcal{R}_N)^{1.5}} \frac{1-m}{m^2} \frac{\lambda_{\text{max}}(\hat{L})}{\lambda_{\text{sec}}(\hat{L})^2} \sqrt{\mathcal{R}_N(q^0) - \mathcal{R}_N(q^\infty)}, \tag{54}$$

where k denotes the number of neighbors in the kNN graph, $\hat{L}$ is the graph Laplacian matrix, $\lambda_{\text{sec}}, \lambda_{\text{min}}$ are the second smallest and smallest eigenvalue, and

$$\|\text{Hess}\mathcal{R}_N\|_1 = \sup_{q \in \mathscr{P}(G_N)} \|\text{Hess}\mathcal{R}_N(q)\|_1, \quad \lambda_{\text{min}}(\text{Hess}\mathcal{R}_N) = \min_{q \in \mathscr{P}(G_N)} \lambda_{\text{min}}(\text{Hess}\mathcal{R}_N(q)), \tag{55}$$

and

$$m = \frac{1}{2}(\frac{1}{(1+2M)^{\frac{1}{\beta}}})^{N-2} \min\{\frac{1}{(1+2M)^{\frac{1}{\beta}}}), \frac{1}{N}\}. \tag{56}$$

Then denote the real worst-case distribution within the $\epsilon(\theta)$-radius discrete Geometric Wasserstein-ball as q^*, that is,

$$q^* = \arg \sup_{q:\mathcal{GW}^2_{G_N}(\hat{P}_{tr}, q) \leq \epsilon(\theta)} \mathcal{R}_N(\theta, q), \tag{57}$$

and we have

$$\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q^*) + \mathcal{R}_N(q^*) - \mathcal{R}_N(q(t)) \leq e^{-Ct}(\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q^*) + \mathcal{R}_N(q^*) - \mathcal{R}_N(q^0)). \tag{58}$$

Therefore, we have

$$\mathcal{R}_N(q^*) - \mathcal{R}_N(q(t)) \leq e^{-Ct}(\mathcal{R}_N(q^*) - \mathcal{R}_N(q^0)) - (1 - e^{-Ct})(\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q^*)), \tag{59}$$

and

$$\frac{\mathcal{R}_N(q^*) - \mathcal{R}_N(q(t))}{\mathcal{R}_N(q^*) - \mathcal{R}_N(q^0)} \leq e^{-Ct} - (1 - e^{-Ct})\frac{\mathcal{R}_N(q^\infty) - \mathcal{R}_N(q^*)}{\mathcal{R}_N(q^*) - \mathcal{R}_N(q^0)} < e^{-Ct}. \tag{60}$$

□