Distributionally Robust Optimization via Diffusion Ambiguity Modeling

Jiaqi Wen JWEN4@UH.EDU
Jianyi Yang JYANG66@UH.EDU
University of Houston, United State

Abstract

This paper studies Distributionally Robust Optimization (DRO), a fundamental framework for enhancing the robustness and generalization of statistical learning and optimization. An effective ambiguity set for DRO must involve distributions that remain consistent to the nominal distribution while being diverse enough to account for a variety of potential scenarios. Moreover, it should lead to tractable DRO solutions. To this end, we proposes a diffusion-based ambiguity set design that captures various adversarial distributions beyond the nominal support space while maintaining consistency with the nominal distribution. Building on this ambiguity modeling, we propose Diffusion-based DRO (`D-DRO`), a tractable DRO algorithm that solves the inner maximization over the parameterized diffusion model space. We formally establish the stationary convergence performance of `D-DRO` and empirically demonstrate its superior Out-of-Distribution (OOD) generalization performance in a ML prediction task.

Keywords: Distributionally Robust Optimization, Diffusion Models, OOD Generalization

1. Introduction

Distributionally Robust Optimization (DRO) is a fundamental framework for enhancing the robustness of statistical learning and optimization problems, particularly under Out-of-Distribution (OOD) scenarios [2, 19]. DRO formulates a minimax optimization problem, where the inner maximization identifies the worst-case distribution within an ambiguity set, and the outer minimization optimizes the decision variable against this worst-case scenario [3, 4, 16, 21, 29]. Unlike non-probabilistic robust optimization, DRO leverages probabilistic uncertainty modeling to enable improved generalization performance. This property has made DRO increasingly important in ML for addressing distribution shifts, noisy data, and adversarial conditions [18, 24–26, 37].

The performance of DRO algorithms critically depends on the design of the ambiguity set, which must contain meaningful distributional variations around the nominal distribution. A common approach is to model the ambiguity set using ϕ-divergences, such as the Kullback–Leibler (KL) divergence [11, 12, 14, 15, 21]. Although such ϕ-divergence-based formulations can sometimes yield closed-form solutions [11, 12], they require that any distribution P in the ambiguity set be *absolutely continuous* with respect to the nominal distribution P_0 (denoted $P \ll P_0$), meaning that for any measurable set $\mathcal{A}$, if $P_0(\mathcal{A}) = 0$, then $P(\mathcal{A}) = 0$. This implicit constraint limits robustness in scenarios with support shifts. In contrast, Wasserstein-DRO leverages the Wasserstein distance to define the ambiguity set, allowing for support shifts. However, solving Wasserstein-DRO over the infinite probability space is difficult. Some approaches [4, 9, 23] reformulate Wasserstein-DRO as a finite-dimensional optimization problem based on the convex assumptions which typically do

not hold in ML. Other methods approximate Wasserstein-DRO via adversarial optimization [9, 35], but such relaxations are overly conservative, limiting their ability to fully leverage the benefits of probabilistic uncertainty modeling.

Recent advances have aimed to address the challenges of ambiguity modeling in DRO. For example, [20] incorporates data geometric properties into the design of discrepancy metrics, thereby reducing the complexity of the ambiguity set. In addition, a recent work [41] studies DRO with ambiguity sets defined by a generalized Sinkhorn distance, which enables modeling uncertainty across distributions with different support space. Another work [26] constructs a Wasserstein-based ambiguity set in the latent space of generative models and subsequently applies Wasserstein-DRO methods to solve the problem. Additional related studies are discussed in Appendix A.

Different from these approaches, our work is the first to model the ambiguity set in the space of diffusion models, which offers several advantages: (1) Diffusion models have a strong capability to represent the underlying data distribution, ensuring that the distributions in the ambiguity set remain consistent with the nominal distribution. (2) Diffusion models are capable of producing diverse samples beyond the training support space, thereby enabling the discovery of worst-case distributions. (3) Diffusion models provide a finite, parameterized optimization space, avoiding the need to solve problems over an infinite probability space.

Our main contributions are summarized below: **(1)** We introduce a novel *Diffusion Ambiguity Set* for DRO, which encompasses diverse distributions while preserving consistency with the nominal distribution. **(2)** We design an inner maximization procedure for DRO with the proposed Diffusion Ambiguity Set, enabling tractable iterative optimization within a finite, parameterized space. **(3)** We propose `D-DRO` (Algorithm 1), which solves the resulting minimax optimization problem, and formally establish its stationary convergence in Theorems 1 and 2. **(4)** We demonstrate the superior performance of `D-DRO` on the challenging ML task of renewable energy prediction.

2. DRO and Ambiguity Modeling

Distributionally Robust Optimization. DRO optimizes for the worst-case performance given an ambiguity set constructed on a nominal distribution P_0 which can be an empirical distribution S_0. Consider an objective function $f(w, x)$ with the decision variable $w \in \mathcal{W}$ and the random parameter $x \in \mathcal{X}$. Given the nominal distribution $P_0(x)$ of the random parameter x, DRO solves the following minimax optimization problem.

$$w = \min_{w \in \mathcal{W}} \max_{P \in \mathcal{B}(P_0, \epsilon)} \mathbb{E}_{x \sim P}[f(w, x)], \tag{1}$$

where $\mathcal{B}(P_0, \epsilon)$ is the ambiguity set containing possible testing distributions which is typically modeled as a distribution ball $\mathcal{B}(P_0, \epsilon) = \{P \mid D(P, P_0) < \epsilon\}$ given a distribution discrepancy measure D and an adversary budget ϵ.

Ambiguity modeling in DRO. The choice of the ambiguity set in DRO has a significant impact on both generalization performance and solution tractability. We observe that a well-designed ambiguity set in DRO should satisfy the following properties: First, it should include diverse distributions beyond the support of the nominal distribution, enabling the identification of various worst-case distributions. Second, the distributions within the ambiguity set should remain realistic and consistent with the nominal distribution, balancing the average-case and worst-case performance. Finally, the ambiguity set should facilitate a tractable solution of the DRO problem despite the infinite probability space.

3. Method

3.1. Diffusion Ambiguity Modeling

Diffusion models. Diffusion models learn an underlying distribution P_0 from its finite dataset S_0 and can generate diverse samples from this distribution. The model training and inference are based on a forward process and a reverse process, detailed in Appendix B. The forward process begins with an initial sample $x_0 \in S_0$ and evolves according to a stochastic process to produce random variables $x_1, \ldots, x_T$ with marginal distributions P_t, $t \in [T]$. The reverse process starts with x_T drawn from a prior distribution π that approximates P_T and reconstructs $x_{T-1}, \ldots, x_0$ by following the reverse diffusion process, which depends on the score function (the gradient of the log-density of the underlying distribution $\nabla_x \log P_t(x)$).

The score-matching method employs a ML model $s_\theta(x, t)$ with parameters θ to approximate the distribution gradient $\nabla_x \log P_t(x)$ by minimizing an empirical score-matching loss $J(\theta, S_0)$ on the dataset S_0. Once the score model $s_\theta(x, t)$ is obtained, we can generate samples $x_0, \ldots, x_{T-1}$ according to the distribution $P_\theta(x_{0:T}) = \pi(x_T) \prod_{t=1}^{T} P_\theta(x_{t-1} \mid x_t)$, where $P_\theta(x_{t-1} \mid x_t) = \mathcal{N}\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big)$ where $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)\big)$ rely on the score model parameter θ.

Theorem 1 in [33] (restated in Lemma 3) shows that if the score-matching loss is bounded, i.e., $J(\theta, S_0) \leq \epsilon$, then the KL divergence $D_{\mathrm{KL}}(P_0 \,\|\, P_\theta)$ is bounded by ϵ plus additional approximation error terms. We note that the KL divergence $D_{\mathrm{KL}}(P_0 \,\|\, P_\theta)$ in Lemma 3 differs from the KL divergence $D_{\mathrm{KL}}(P_\theta \,\|\, P_0)$ commonly used in `KL-DRO`. Importantly, the former allows P_θ to have a broader support than P_0 (i.e., $P_0 \ll P_\theta$).

Diffusion ambiguity set. We can model the ambiguity set based on the score-matching loss of a diffusion model, due to its property to constrain the distributional discrepancy while allowing a broader support space. This leads to DRO with diffusion ambiguity sets:

$$\min_{w \in \mathcal{W}} \max_{\theta \in \Theta} \mathbb{E}_{x \sim P_\theta}[f(w, x)], \quad \text{s.t.} \quad J(\theta, S_0) \leq \epsilon, \tag{2}$$

where P_θ denotes the distribution of the diffusion reverse process, and ϵ is the adversarial budget.

The diffusion ambiguity set enhances DRO performance as follows. Due to the distribution modeling capability of diffusion models, score-matching constraint ensures that the diffusion-modeled distributions remain consistent with the nominal data, mitigating over-conservativeness issues in DRO. It also leverages diffusion models' ability to generate diverse samples beyond the nominal support, enabling the identification of worst-case distributions given any w. Furthermore, the inner maximization in (2) operates in a finite, parameterized space, ensuring tractable optimization.

3.2. Diffusion-Based Inner Maximization

To solve the inner maximization of (2), which is a constrained optimization over the diffusion parameter space, we adopt a dual learning approach: introducing a Lagrangian dual $\mu > 0$ to reformulate it as the unconstrained problem

$$\max_{\theta} \; \mathbb{E}_{x \sim P_\theta}[f(w, x)] - \mu J(\theta, S_0), \tag{3}$$

and updating μ via dual gradient descent. As is shown by `InnerMax` in Algorithm 1, we increase μ when J exceeds the budget ϵ and decreasing it otherwise.

We apply the policy optimization methods to transform the objective in (3) into a tractable form. Vanilla policy gradient [36] directly calculates the empirical gradient of the objective in (3)

Algorithm 1 Diffusion-based DRO (`D-DRO`)

Input: Training dataset S_0; Adversary budget $\epsilon > 0$; Step size $\eta > 0$, $\lambda > 0$.
Initialization: Initialize decision variable w, diffusion parameter θ and Lagrangian weight $\mu > 0$
for $j = 1, 2, \cdots, I$ **do**
  `// Diffusion-based inner maximization (InnerMax)`
  for $k = 1, 2, \cdots, K$ **do**
    Update diffusion parameter θ_k by solving (3) given μ
    Update Lagrangian parameter: $\mu \leftarrow \max\{0, \mu + \eta\big(J(\theta_k, S_0) - \epsilon\big)\}$
  end
  `// Outer minimization to update decision variable`
  Generate dataset S_j with diffusion model $P_{\theta^{(j)}}$ with $\theta^{(j)}$ uniformly selected from $[\theta_1, \cdots, \theta_K]$
  Update decision variable:w: $w_j = w_{j-1} - \lambda \cdot \nabla_w \mathbb{E}_{x \in S_j}[f(w_{j-1}, x)]$
end
return w uniformly selected from $[w_1, \cdots, w_I]$

as is detailed in C.1. Proximal Policy Optimization (PPO) [27] is believed to have more stable performance. It transforms the objective (3) into a differentiable form as is detailed in Section C.2.

3.3. `D-DRO` Algorithm

We design `D-DRO` in Algorithm 1 which solves the mini-max optimization following the framework of Gradient Descent with Max-Oracle (GDMO) in [13]. In each iteration, we first run `InnerMax` to search for the worst-case diffusion model $P_{\theta^{(j)}}$ that maximizes the expected loss of the current variable w. Next, we generate an adversarial dataset S_j based on $P_{\theta^{(j)}}$ and use it to update w. The convergence of Algorithm 1 in proved in Theorem 1 and Theorem 2.

4. Analysis

Theorem 1 (Convergence of Inner Maximization) *Let θ^* be the optimal diffusion parameter that solves the inner maximization (2) given a variable w. If the expected score-matching loss is bounded as $J(\theta) \leq \bar{J}$ and the step size is chosen as $\eta \sim \mathcal{O}(\frac{1}{\sqrt{K}})$, the inner maximization error holds that*

$$\Delta' := \mathbb{E}_{x \sim P_{\theta^*}}[f(w, x)] - \mathbb{E}_k \mathbb{E}_{x \sim P_{\theta_k}}[f(w, x)] \leq \frac{1}{\sqrt{K}} \max\{\epsilon, \bar{J}\} \|\mu^{(1)}\|, \tag{4}$$

where the outer expectation is taken over the randomness of output selection. In addition, the KL-divergence with respect to the nominal distribution is bounded as $\mathbb{E}_k[D_{\mathrm{KL}}(P_0 || P_{\theta_k})] \leq \epsilon + \frac{\max\{\epsilon, \bar{J}\}|\mu_C - \mu^{(1)}|}{\sqrt{K}(\mu_C - \mu^*)} + D_{\mathrm{KL}}(P_T || \pi) + C_1$, where $\mu_C > \mu^*$ with μ^* being the optimal dual variable and C_1 is a constant, P_T is the output distribution of the forward process, and π is the initial distribution of the reverse process.

Proofs of Theorem 1 are provided in Appendix D.2. The bound shows that when the inner iteration number K is sufficiently large, the inner maximization error will be small enough, so `D-DRO` can find a worst-case distribution in the Diffusion Ambiguity Set. Moreover, the reverse KL-divergence w.r.t. the nominal distribution is bounded by the budget ϵ, the KL-divergence between

P_T and π, and the constant gap C_1 due to the approximation of the score matching loss [32].This implies that we can adjust the adversarial budget ϵ to get a worst-case distribution that is consistent with the nominal distribution.

Theorem 2 (Convergence of D−DRO) *Assume that objective $f(w,x)$ is $\beta-$smooth and $L-$Lipschitz with respect to w and is upper bounded by $\bar{f}$. If each dataset S_j sampled from diffusion model contains n examples and the step size is chosen as $\lambda \sim \mathcal{O}(\sqrt{\frac{1}{\beta L^2 H}})$, then with probability $1-\delta, \delta \in (0,1)$, the average norm of Moreau envelope of $\phi(w) := \max_\theta \mathbb{E}_{P_\theta}[f(w,x)]$ satisfies*

$$\mathbb{E}_{j,k}\left[\|\nabla \phi_{\frac{1}{2\beta}}(w)\|^2\right] \leq 4\beta\Delta' + \frac{V_1}{\sqrt{n}} + \frac{V_2}{\sqrt{H}}, \tag{5}$$

where the expectation is taken over the randomness of output selection, Δ' is the error of inner maximization bounded in Theorem 1, $V_1 = 8\beta\bar{f}\sqrt{\log(2/\delta)}$ and $V_2 = 4L\sqrt{(\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w))\beta}$.

The proof of Theorem 2 is deferred to Appendix D.3. It shows that with sufficiently large iteration number H and sampling size n, the average gradient norm of the Moreau envelope for the optimal inner maximization function $\phi(w)$ is bounded by the error Δ' of the maximization oracle. Since Δ' decreases as the inner iteration number K becomes sufficiently large as proved by Theorem 1, the average gradient norm of the Moreau envelope can be small enough. As is detailed in Appendix D.3.1, this implies that the output w converges to an approximately stationary point of the optimal inner maximization function $\phi(w)$. Thus, D−DRO can find an approximated solution of (1) with enough diffusion-based maximization and minimization iterations.

5. Experiment

In this section, we present numerical studies on a ML for renewable prediction task based on the Electricity Maps[1] datasets (experiment setups in Appendix E). A part of results are given in Table 1 where D−DRO are compared with baselines in Appendix E.1.1 on various OOD testing datasets in Appendix E.1.2. We observe that all methods perform better on datasets with smaller Wasserstein distribution shifts. While DML and other DRO baselines outperform ML, D−DRO consistently outperforms them across all OOD datasets, owing to its diffusion-based ambiguity set that effectively captures worst-case yet realistic distributions. More evaluation results are provided in Appendix E.

Table 1: Test MSE on different datasets(Partial).

Datasets (Wasserstein Distance)	Algorithms				
	D−DRO	KL−DRO	W−DRO	DML	ML
BANC_22 (0.0240)	**0.0047**	0.0086	0.0073	0.0078	0.0183
BANC_21 (0.1213)	**0.0054**	0.0112	0.0121	0.0093	0.0238
QLD_22 (0.2782)	**0.0192**	0.0379	0.0557	0.0352	0.0667
QLD_21 (0.3054)	**0.0186**	0.0377	0.0574	0.0339	0.0696
GB_22 (0.1255)	**0.0105**	0.0197	0.0245	0.0172	0.0360
GB_21 (0.1359)	**0.0094**	0.0181	0.0229	0.0158	0.0340
Average	**0.0163**	0.0288	0.0342	0.0271	0.0450
Maximum	**0.0509**	0.0831	0.0879	0.0834	0.0946

6. Conclusion

In this paper, we propose D−DRO, which introduces a novel diffusion-based ambiguity modeling for DRO, and develops D−DRO to solve the DRO with diffusion ambiguity set. We prove the stationary convergence performance of D−DRO. The experiments demonstrate robust OOD generalization performance of D−DRO. Overall, this new DRO solution has the potential to enhance the robustness of critical statistical optimization and ML tasks under distribution shifts and imperfect data.

References

[1] Electricity maps datasets portal. https://portal.electricitymaps.com/datasets.

[2] Martin Arjovsky. *Out of distribution generalization in machine learning*. PhD thesis, New York University, 2020.

[3] Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13): 1–48, 2018.

[4] Ruidi Chen, Ioannis Ch Paschalidis, et al. Distributionally robust learning. *Foundations and Trends® in Optimization*, 4(1-2):1–243, 2020.

[5] Xuelong Dai, Yanjie Li, Mingxing Duan, and Bin Xiao. Diffusion models as strong adversaries. *IEEE Transactions on Image Processing*, 2024.

[6] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024.

[7] Bingqian Du, Zhiyi Huang, and Chuan Wu. Adversarial deep learning for online resource allocation. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, 6(4):1–25, 2022.

[8] Electricity Maps. The world's most comprehensive electricity data platform. https://www.electricitymaps.com, 2025.

[9] Rui Gao and Anton Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *Mathematics of Operations Research*, 48(2):603–655, 2023.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[11] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

[12] Hisham Husain, Vu Nguyen, and Anton van den Hengel. Distributionally robust bayesian optimization with ϕ-divergences. 2023.

[13] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pages 4880–4889. PMLR, 2020.

[14] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2174–2184. PMLR, 2020.

[15] Burak Kocuk. Conic reformulations for kullback-leibler divergence constrained distributionally robust optimization and applications. *arXiv preprint arXiv:2007.05966*, 2020.

[16] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[17] Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[18] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5051–5059, 2019.

[19] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[20] Jiashuo Liu, Jiayun Wu, Bo Li, and Peng Cui. Distributionally robust optimization with data geometry. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33689–33701. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/da535999561b932f56efdd559498282e-Paper-Conference.pdf.

[21] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.

[22] Xutao Ma, Chao Ning, and Wenli Du. Differentiable distributionally robust optimization layers. 2024. URL https://arxiv.org/abs/2406.16571.

[23] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

[24] Fabio Muratore, Fabio Ramos, Greg Turk, Wenhao Yu, Michael Gienger, and Jan Peters. Robot learning from randomized simulations: A review. *Frontiers in Robotics and AI*, 9: 799893, 2022.

[25] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022.

[26] Allen Z Ren and Anirudha Majumdar. Distributionally robust policy learning via adversarial environment generation. *IEEE Robotics and Automation Letters*, 7(2):1379–1386, 2022.

[27] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

[29] Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[32] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

[33] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021.

[34] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017.

[35] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017.

[36] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. 12, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

[37] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017.

[38] Prince Zizhuang Wang, Jinhao Liang, Shuyi Chen, Ferdinando Fioretto, and Shixiang Zhu. Gen-dfl: Decision-focused generative learning for robust decision making. 2025. URL https://arxiv.org/abs/2502.05468.

[39] Weicheng Xie, Zenghao Niu, Qinliang Lin, Siyang Song, and Linlin Shen. Generative imperceptible attack with feature learning bias reduction and multi-scale variance regularization. *IEEE Transactions on Information Forensics and Security*, 2024.

[40] Chen Xu, Jonghyeok Lee, Xiuyuan Cheng, and Yao Xie. Flow-based distributionally robust optimization. *IEEE Journal on Selected Areas in Information Theory*, 2024.

[41] Yufeng Yang, Yi Zhou, and Zhaosong Lu. Nested stochastic algorithm for generalized sinkhorn distance-regularized distributionally robust optimization. *arXiv preprint arXiv:2503.22923*, 2025.

[42] Linglingzhi Zhu and Yao Xie. Distributionally robust optimization via iterative algorithms in continuous probability spaces. *arXiv preprint arXiv:2412.20556*, 2024.

Appendix A. Related Work

A.1. Distributionally Robust Optimization

DRO algorithms are widely studied to improve the OOD generalization performance for various optimization and ML tasks [12, 21, 26]. ϕ-divergence-based DRO [11, 12, 14, 15] is one of the commonly-used DRO method. A closed-form solution to the inner maximization of (1) with KL-divergence-based ambiguity set is provided by [11]. While we can usually get tractable DRO solutions based on ϕ-divergence, the definition of ϕ-divergence requires any distribution P in the ambiguity set to be absolutely continuous with respect to the nominal distribution, which limits its application in statistical learning tasks. Alternatively, many studies adopt Wasserstein distance-based DRO [4, 9, 12, 17, 23]. Wasserstein measure has no restrictions on the support of the distribution, but it is difficult to get a tractable solution for W−DRO. Some methods [4, 9, 12, 17, 23] reformulate Wasserstein-constrained DRO into a tractable finite optimization based on the assumption of convex objectives which typically do not hold in deep learning. Other methods relax Wasserstein-constrained DRO into an adversarial optimization problem [9, 28, 34], but this relaxation can be overly conservative and cannot fully exploit the benefits of probabilistic ambiguity modeling.

A line of recent studies focuses on addressing the challenges in ambiguity modeling for DRO [20, 22, 26, 40–42]. Among them, [20] incorporates data geometric properties into the design of discrepancy metrics, reducing the size of the ambiguity set. [26] constructs a novel ambiguity set on the latent space of generative models such that the adversarial distribution is realistic and apply Wasserstein-based DRO solutions. Ma *et al.* [22] propose a differentiable parameterized Second-Order Cone (SOC) to characterize the ambiguity set and develop an end-to-end framework in which an ML model is trained to predict the ambiguity set for downstream DRO tasks. Moreover, a latest work [41] introduces a regularized nonconvex DRO method with generalized Sinkhorn distance, reformulating the problem as a contextual nested stochastic optimization and proving convergence without assuming strong convexity or large batches. Different from these methods, we utilize the strong distribution learning capability of diffusion models to build the ambiguity set, enabling the discovery of worst-case and realistic distributions. At the same time, the proposed algorithm D−DRO converts DRO into a finite tractable problem in the diffusion parameter space.

A.2. Generative Models for Robust Learning

Generative models have been widely studied to generate adversarial samples for robust training [5–7, 39]. The target of these works is to generate adversarial attacking examples which is fundamentally different from the worst-case distribution generation which is studied in this paper and aims to improve OOD generalization. A recent paper proposed DRAGEN that [26] models the adversarial distribution on the latent space of a generative model. However, it still lies in the Wasserstein-based framework. Wang *et al.* [38] introduced a Generate-then-Optimize framework, where a diffusion model is trained to generate data for downstream statistical optimization with a focus on the conditional value-at-risk (CVaR) objective. Although related to our work, their method mainly targets risk mitigation within in-distribution settings, while D−DRO is specifically designed to improve robustness under OOD scenarios.

Appendix B. Preliminaries of Diffusion Models

This paper exploits diffusion models to improve the performance of DRO, so we summarize the preliminaries about diffusion models in this section. We introduce a score-based diffusion modeling by Stochastic Differential Equations (SDEs) [32]. They rely on forward and backward stochastic processes introduced as follows.

Forward Process. The forward process incrementally injects noise into the data, generating a sequence of perturbed samples. It begins with an initial sample $x_0 \in \mathcal{R}^d$ drawn from the underlining distribution P_0, and evolves according to a stochastic process as:

$$\mathrm{d}x = b(x,t)\mathrm{d}t + r(t)\mathrm{d}w, \tag{6}$$

where $b(\cdot,t) : \mathcal{R}^d \to \mathcal{R}^d$ is a vector-valued function, $r(t) \in \mathcal{R}$, w is a standard Wiener process and $\mathrm{d}w$ is white Gaussian noise. By the forward process, we get a collection of random variables $\{x_t\}_{t\in[0,T]}$. We use P_t to represent the distribution of x_t and $P_{t|0}$ to denote the conditional distribution of x_t given $x_0 \sim P_0$. With a sufficiently long time T, the marginal distribution $P_T(x_T)$ approximates a tractable prior distribution $\pi(x)$ which is typically chosen as a standard Gaussian distribution.

Reverse Process. A reverse diffusion process is associated with the forward equation in (6) and is expressed as

$$\mathrm{d}x = \big(b(x,t) - r(t)^2\nabla_x \log P_t(x)\big)\,\mathrm{d}t + r(t)\mathrm{d}\bar{w}, \tag{7}$$

where $\bar{w}$ is a standard Wiener process in the reverse-time direction, $\nabla_x \log P_t(x)$ is the time-dependent score function.

Score Matching. In the reverse process, the score function $\nabla_x \log P_t(x)$ plays a critical role in directing the dynamics. To estimate the score function $\nabla_x \log P_t(x)$, we train a score-based model $s_\theta(x,t)$ based on samples generated from the forward diffusion process. The score-based model should minimize the following score-matching loss:

$$J_{\mathrm{SM}}(\theta) = \int_0^T \mathbb{E}_{P_t(x)}\left[\iota(t)\left\|\nabla_x \log P_t(x) - s_\theta(x,t)\right\|^2\right]\mathrm{d}t,$$

where $\iota(t) > 0$ is a positive weighting function. We usually approximate the score-matching loss by a tractable denoising score-matching loss up to a constant that does not rely on θ:

$$J(\theta) = \int_0^T \mathbb{E}_{P_0(x)P_{t|0}(x'|x)}\Big[\iota(t)\Big\|\nabla_{x'}\log P_{t|0}(x'|x) - s_\theta(x,t)\Big\|^2\Big]\mathrm{d}t, \tag{8}$$

Sampling. If we discretize the reverse process, initialize $x_T \sim \pi$ and replace $\nabla_x \log P_t(x)$ with the score-based model $s_\theta(x,t)$, we can generate samples with a Markov chain with T steps:

$$x_{t-1} = x_t + [b(x_t,t) - r^2(t)s_\theta(x_t,t)]\Delta t + r(t)\sqrt{|\Delta t|}z_t, \tag{9}$$

where Δ_t is a small enough constant and $z_t \sim \mathcal{N}(0,\mathbf{I})$. Most existing diffusion models generate samples following the Markov chain [10, 30, 31] and a common expression for the joint distribution of the reverse outputs is

$$P_\theta(x_{0:T}) = \pi(x_T)\prod_{t=1}^{T} P_\theta(x_{t-1} \mid x_t), \tag{10}$$

where $P_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$.

We have the following lemma which shows the reverse KL divergence between the diffusion model and the nominal distribution is bounded.

Lemma 3 *Given the above assumptions D.1.1, if the score-matching loss satisfy $J(\theta, S) \leq \epsilon$, the output distribution of the diffusion model P_θ satisfies:*

$$D_{\mathrm{KL}}(P_0 \,\|\, P_\theta) \leq \epsilon + D_{\mathrm{KL}}(P_T \,\|\, \pi) + C_1, \tag{11}$$

where P_T is the output distribution of the forward process and $P_T \approx \pi$ by the design of diffusion models, and C_1 is a constant from approximating the score-matching loss.

Note that the KL-divergence $D_{\mathrm{KL}}(P_0||P_\theta)$ in Lemma 3 is not the KL-divergence $D_{\mathrm{KL}}(P_\theta||P_0)$ commonly used in `KL-DRO`. The former KL-divergence allows P_θ to have broader support space than P_0 ($P_0 \ll P$). By Lemma 3, if we find an adversarial distribution P_θ by (2), P_θ also stays close enough to the training distribution P_0 through a KL-divergence depending on the budget ϵ. Therefore, the constraint in (2) can define a probabilistic ambiguity set for DRO.

Appendix C. Details of `InnerMax` in `D-DRO` Algorithm 1

C.1. Policy Gradient for Diffusion-based `InnerMax`

A vanilla policy gradient can transform (3) into

$$\max_\theta \hat{\mathbb{E}}_{P_\theta(x_{0:T})}[\ln P_\theta(x_{0:T}) \cdot l(h_w, x_0)] - \lambda \cdot J(\theta, S_0), \tag{12}$$

where $\hat{\mathbb{E}}_{P_\theta(x_{0:T})}$ is the empirical mean based on the $T-$step samples for the backward process of the diffusion model P_θ, and $\ln P_\theta(x_{0:T}) = -\sum_{t=1}^{T}[x_{t-1} - \mu_\theta(x_t, t)]^2 + C_2$ where C_2 is a constant. The derivation details are given below.

C.1.1. Objective Derivation by Policy Gradient

Let $x_{0:T}$ be the output vector of each step in the backward process of the diffusion model. Denote $P_{0:T,\theta}$ as the joint distribution of $x_{0:T}$. Since $x_0 \sim P_\theta$, we can express the first term of the Lagrangian-relaxed objective as

$$\mathbb{E}_{x\sim P_\theta}[f(w, x)] = \mathbb{E}_{x_{0:T}\sim P_{0:T,\theta}}[f(w, x_0)]. \tag{13}$$

Then, the gradient of Lagrangian-relaxed objective can be expressed as

$$\begin{aligned}
&\nabla_\theta \left(\mathbb{E}_{x\sim P_\theta}[f(w, x)] - \mu J(\theta, S_0)\right) \\
=&\nabla_\theta \left(\mathbb{E}_{x_{0:T}\sim P_{0:T,\theta}}[f(w, x_0)] - \mu J(\theta, S_0)\right) \\
=&\int_{x_{0:T}} f(w, x_0) \nabla_\theta P_{0:T,\theta}(x_{0:T}) \mathrm{d}x_{0:T} - \mu \nabla_\theta J(\theta, S_0) \\
=&\int_{x_{0:T}} P_{0:T,\theta}(x_{0:T}) f(w, x_0) \nabla_\theta \ln P_{0:T,\theta}(x_{0:T}) \mathrm{d}x_{0:T} - \mu \nabla_\theta J(\theta, S_0) \\
=&\mathbb{E}_{x_{0:T}\sim P_{0:T,\theta}}\left[f(w, x_0) \nabla_\theta \ln P_{0:T,\theta}(x_{0:T})\right] - \mu \nabla_\theta J(\theta, S_0)
\end{aligned} \tag{14}$$

The first term $\mathbb{E}_{x_{0:T}\sim P_{0:T,\theta}}$ can be calculated by empirical mean $\hat{\mathbb{E}}_{x_{0:T}\sim P_{0:T,\theta}}$ based on the examples sampled from $P_{0:T,\theta}$. Thus, we can equivalently implement the gradient ascent by optimizing the objective in (12):

$$\max_{\theta} \hat{\mathbb{E}}_{x_{0:T}\in P_{0:T,\theta}}[\ln P_{0:T,\theta}(x_{0:T}) \cdot f(w, x_0)] - \mu \cdot J(\theta, S_0), \tag{15}$$

Next, we derive the expression for $\ln P_{0:T,\theta}(x_{0:T})$. Consider a discrete-time diffusion backward process as

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t w_t, \tag{16}$$

where w_t is a standard multi-dimensional Gaussian variable. Given θ, the conditional probability at each step t is

$$P_{t-1,\theta}(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}, \mu_\theta(x_t, t), \sigma_t^2 \mathbf{I}). \tag{17}$$

We can explicitly express the joint distribution $P_{0:T,\theta}$ of $x_{0:T}$ for all T backward steps:

$$P_{0:T,\theta}(x_{0:T}) = P(x_T) \prod_{t=1}^{T} P_{t-1,\theta}(x_{t-1} \mid x_t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\|x_T\|^2}{2}} \cdot \frac{1}{\sqrt{2\pi}\sigma_t} e^{-\sum_{t=1}^{T} \frac{\|x_{t-1}-\mu_\theta(x_t,t)\|^2}{2\sigma_t^2}}. \tag{18}$$

Thus, we can get the expression of $\ln P_{0:T,\theta}(x_{0:T})$ as

$$\ln P_{0:T,\theta}(x_{0:T}) = -\sum_{t=1}^{T} \|x_{t-1} - \mu_\theta(x_t, t)\|^2/(2\sigma_t^2) + C_2, \tag{19}$$

where $C_2 = -\ln(2\pi\sigma_t) - \frac{\|x_T\|^2}{2}$.

C.2. PPO for Diffusion-based `InnerMax`

The policy gradient algorithm requires frequent resampling by the diffusion model P_θ, resulting in high complexity and unstable performance. A more popular policy optimization method is Proximal Policy Optimization (PPO) [27] which transforms (3) into

$$\max_{\theta} \hat{\mathbb{E}}_{P_{\theta_{\text{old}}}(x_{0:T})}[\min(r_\theta l(h_w, x_0)), \text{clip}(r_\theta(x_{0:T}), 1-\kappa, 1+\kappa) \cdot l(h_w, x_0))] - \lambda \cdot J(\theta, S_0), \tag{20}$$

where $P_{\theta_{\text{old}}}$ is a reference diffusion model used for sampling the backward sequence, the probability ratio is $r_\theta(x_{0:T}) = \frac{P_\theta(x_{0:T})}{P_{\theta_{\text{old}}}(x_{0:T})} = \exp\{-\sum_{t=1}^{T}(\frac{\|x_{t-1}-\mu_\theta(x_t,t)\|^2}{2\sigma_t^2} - \frac{\|x_{t-1}-\mu_{\theta_{\text{old}}}(x_t,t)\|^2}{2\sigma_t^2})\}$, and $\kappa \in (0, 1)$ is the clipping parameter to avoid overly-large policy updates. To reduce the training complexity, instead of optimizing all the T steps, we can only optimize the last T' steps of the backward process by choosing $r_\theta(x_{0:T'}) = \exp\{-\sum_{t=1}^{T'}(\frac{\|x_{t-1}-\mu_\theta(x_t,t)\|^2}{2\sigma_t^2} - \frac{\|x_{t-1}-\mu_{\theta_{\text{old}}}(x_t,t)\|^2}{2\sigma_t^2})\}$ and only keep the loss terms for corresponding steps in $J(\theta, S_0)$. This practice has been verified by our experiments in Appendix E. More derivation details are given below.

C.2.1. OBJECTIVE DERIVATION BY PROXIMAL POLICY OPTIMIZATION

In this PPO method, we convert the first term of (3) as a PPO-like objective given a reference diffusion model P_{θ_0}, i.e.

$$\begin{aligned}\mathbb{E}_{x\sim P_\theta}[f(w,x)] &= \mathbb{E}_{x_{0:T}\sim P_{0:T,\theta}}[f(w,x_0)] \\ &= \mathbb{E}_{x_{0:T}\sim P_{0:T,\theta_0}}\left[\frac{P_{0:T,\theta}(x_{0:T})}{P_{0:T,\theta_0}(x_{0:T})}f(w,x_0)\right],\end{aligned} \tag{21}$$

where the probability ratio is

$$r_\theta(x_{0:T}) = \frac{P_{0:T,\theta}(x_{0:T})}{P_{0:T,\theta_0}(x_{0:T})} = \exp\left\{-\sum_{t=1}^{T}\left(\frac{\|x_{t-1}-\mu_\theta(x_t,t)\|^2}{2\sigma_t^2} - \frac{\|x_{t-1}-\mu_{\theta_0}(x_t,t)\|^2}{2\sigma_t^2}\right)\right\}$$

given the joint probability expression. The expectation can be approximated by empirical mean based on the examples sampled by $P_{0:T,\theta_0}$. Like PPO, we apply clipping on the ratio to avoid overly-large updates, and we can get the objective in Eqn. (20):

$$\max_\theta \hat{\mathbb{E}}_{P_{0:T,\theta_0}(x_{0:T})}[\min(r_\theta(x_{0:T})f(w,x_0)), \mathrm{clip}(r_\theta(x_{0:T}),1-\kappa,1+\kappa)\cdot f(w,x_0))] - \mu\cdot J(\theta,S_0), \tag{22}$$

where $\kappa\in(0,1)$ is the clipping parameter.

Appendix D. Theorem Proofs

D.1. Proof of Lemma 3

D.1.1. BOUND OF THE SCORE-MATCHING LOSS BY KL DIVERGENCE

Lemma 3 is a result based on conclusion of Theorem 1 in [33]. For completeness, we give the full proof of Theorem 1 in [33] (Lemma 4), and then we derive the concrete expressions for the constants in Lemma 3.

Assumptions We make the following assumptions for all the lemmas and theorems in the paper.

1. $P_0(x)$ is a density function with continuous second-order derivatives and $\mathbb{E}_{x\sim P_0}\left[\|x\|_2^2\right] < \infty$.
2. The prior distribution $\pi(x)$ is a density function with continuous second-order derivatives and $\mathbb{E}_{x\sim\pi}\left[\|x\|_2^2\right] < \infty$.
3. $\forall t\in[0,T]$: $f(\cdot,t)$ is a function with continuous first order derivatives. $\exists C>0, \forall x\in\mathcal{R}^d, t\in[0,T]$: $\|b(x,t)\|_2 \le C(1+\|x\|_2)$
4. $\exists C>0, \forall x,y\in\mathcal{R}^d: \|b(x,t)-b(y,t)\|_2 \le C\|x-y\|_2$.
5. g is a continuous function and $\forall t\in[0,T], |r(t)|>0$.
6. For any open bounded set $\mathcal{O}$, $\int_{t=0}^{T}\int_{\mathcal{O}}\|P_t(x)\|_2^2 + dr(t)^2\|\nabla_x P_t(x)\|_2^2 \mathrm{d}x\mathrm{d}t$.
7. $\exists C>0, \forall x\in\mathcal{R}^d, t\in[0,T]: \|\nabla_x \log P_t(x)\|_2 \le C(1+\|x\|_2)$.
8. $\exists C>0, \forall x,y\in\mathcal{R}^d: \|\nabla_x \log P_t(x) - \nabla_x \log P_t(y)\|_2 \le C(\|x-y\|_2)$.

9. $\exists C > 0, \forall x \in \mathcal{R}^d, t \in [0,T] : \|s_\theta(x,t)\|_2 \leq C(1 + \|x\|_2)$.
10. $\exists C > 0, \forall x, y \in \mathcal{R}^d : \|s_\theta(x,t) - s_\theta(y,t)\|_2 \leq C(\|x - y\|_2)$.
11. Novikov's condition: $\mathbb{E}\left[\exp(\frac{1}{2}\int_{t=0}^{T} \|\nabla_x \log P_t(x) - s_\theta(x,t)\|_2^2 \mathrm{d}t)\right] < \infty$.

Lemma 4 *Let $P_0(x)$ be the underlining data distribution, $\pi(x)$ be a known prior distribution, and P_θ be marginal distribution of $\hat{x}_\theta(0)$, the output of reverse-time SDE defined as below.*

$$\mathrm{d}\hat{x} = [b(\hat{x},t) - r(t)^2 s_\theta(\hat{x},t)]\mathrm{d}t + r(t)\mathrm{d}\bar{w},\ \hat{x}_\theta(T) \sim \pi. \tag{23}$$

With the assumptions in Section D.1.1, we have

$$D_{\mathrm{KL}}(P_0 \| P_\theta) \leq J_{SM}(\theta, r(\cdot)^2) + D_{\mathrm{KL}}(P_T \| \pi), \tag{24}$$

where $J_{SM}(\theta, r(\cdot)^2) = \frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{p_t(x)}\left[r(t)\|\nabla_x \log P_t(x) - s_\theta(x,t)\|_2^2\right] \mathrm{d}t$.

Proof We denote the path measure of the forward outputs $\{x_t\}_{t\in[0,T]}$ as p and the path measure of the backward outputs $\{\hat{x}_{\theta,t}\}_{t\in[0,T]}$ as q. By assumptions 1- 5, 9, 10, both p and q are uniquely given by the forward and backward SDEs, respectively. Consider a Markov kernel $M(\{z_t\}_{t\in[0,T]}, y) := \delta(z_0 = y)$ given any Markov chain $\{z_t\}_{t\in[0,T]}$. Since $x_0 \sim P_0$ and $\hat{x}_{\theta,0} \sim P_\theta$, we have

$$\int M(\{x_t\}_{t\in[0,T]}, x)\mathrm{d}p(\{x_t\}_{t\in[0,T]}) = P_0(x) \tag{25}$$

$$\int M(\{\hat{x}_{\theta,t}\}_{t\in[0,T]}, x)\mathrm{d}q(\hat{x}_{\theta,t}\}_{t\in[0,T]}) = P_\theta(x) \tag{26}$$

Here the Markov kernel M essentially performs marginalization of path measures to obtain distributions at $t = 0$. We can use the data processing inequality with this Markov kernel to obtain

$$\begin{aligned} &D_{\mathrm{KL}}(P_0 \| P_\theta) \\ =&D_{\mathrm{KL}}\left(\int M(\{x_t\}_{t\in[0,T]}, x)\mathrm{d}p(\{x_t\}_{t\in[0,T]}) \| \int M(\{\hat{x}_{\theta,t}\}_{t\in[0,T]}, x)\mathrm{d}q(\hat{x}_{\theta,t}\}_{t\in[0,T]})\right) \\ \leq&D_{\mathrm{KL}}(p,q). \end{aligned} \tag{27}$$

Since $x_T \sim P_T$ and $\hat{x}_{\theta,T} \sim \pi$. Leveraging the chain rule of KL divergence, we have

$$\begin{aligned} D_{\mathrm{KL}}(p,q) &= \mathbb{E}_p\left[\log(\frac{p(x_{1:T} \mid x_T)P_T(x_T)}{q((\hat{x}_{\theta,1:T} \mid \hat{x}_{\theta,T})\pi(\hat{x}_{\theta,T})})\right] \\ &= D_{\mathrm{KL}}(P_T \| \pi) + \mathbb{E}_{z\sim P_T}\left[D_{\mathrm{KL}}(p(\cdot \mid x_T = z) \| q(\cdot \mid \hat{x}_{\theta,T} = z))\right]. \end{aligned} \tag{28}$$

Under assumptions 1,3-8, the SDE in Eqn. (6) has a corresponding reverse-time SDE as

$$\mathrm{d}x = [b(x,t) - r(t)^2 \nabla_x \log P_t(x)] + r(t)\mathrm{d}\bar{w}. \tag{29}$$

Since Eqn. (29) is the time reversal of Eqn. (6), they share the same path measure p. Thus, $\mathbb{E}_{z\sim P_T}\left[D_{\mathrm{KL}}(p(\cdot \mid x_T = z) \| q(\cdot \mid \hat{x}_{\theta,T} = z))\right]$ can be viewed as the KL divergence between the path

measures induced by the two SDEs in Eqn. (6) and Eqn. (23) with the same starting points $x_T = \hat{x}_{\theta,T} = z$.

The KL divergence between two SDEs with shared diffusion coefficients and starting points exists under assumptions 7,-11, and can be bounded by the Girsanov theorem

$$\begin{aligned}
&D_{\mathrm{KL}}(p(\cdot \mid x_T = z) \| q(\cdot \mid \hat{x}_{\theta,T} = z)) = \mathbb{E}_p\left[\log \frac{\mathrm{d}p}{\mathrm{d}q}\right] \\
=&\mathbb{E}_p\left[\int_{t=0}^{T} r(t)(\nabla_x \log P_t(x) - s_\theta(x,t))\mathrm{d}\bar{w}_t + \frac{1}{2}\int_{t=0}^{T} r(t)^2 \|\nabla_x \log P_t(x) - s_\theta(x,t)\|_2^2 \mathrm{d}t\right] \\
=&\mathbb{E}_p\left[\frac{1}{2}\int_{t=0}^{T} r(t)^2 \|\nabla_x \log P_t(x) - s_\theta(x,t)\|_2^2 \mathrm{d}t\right] \\
=&\frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{P_t(x)}\left[r(t)^2 \|\nabla_x \log P_t(x) - s_\theta(x,t)\|_2^2\right] \mathrm{d}t = J_{SM}(\theta, r(\cdot)^2),
\end{aligned} \tag{30}$$

where the second equality holds by Girsanov Theorem II, and the third equality holds because $Y_s = \int_{t=0}^{s} r(t)(\nabla_x \log P_t(x) - s_\theta(x,t))\mathrm{d}\bar{w}_t$ is a continuous-time Martingale process ($\mathbb{E}[Y_s \mid Y_{\tau,\tau\le s'}] = Y_{s'}, \forall s' \le s$) and we have $\mathbb{E}[Y_s - Y_{s'}] = 0, \forall s' < s$. ■

D.1.2. PROOF OF LEMMA 3

Lemma 3. *Given the assumptions in Appendix D.1.1, if the score-matching loss satisfies $J(\theta, S) \le \epsilon$, the output distribution of the diffusion model P_θ satisfies*

$$D_{\mathrm{KL}}(P_0||P_\theta) \le \epsilon + D_{\mathrm{KL}}(P_T||\pi) + C_1,$$

where P_T is the output distribution of the forward process, π is a prior distribution of the diffusion model and $P_T \approx \pi$ by the design of diffusion models, and C_1 is a constant that does not rely on θ.

The score matching loss $J(\theta, S_0)$ in (8) with $\iota(t) = r(t)^2$ is actually the denoising score matching loss $J_{DSM}(\theta, r(\cdot)^2)$, i.e.

$$J(\theta, S_0) = J_{DSM}(\theta, r(\cdot)^2) = \frac{1}{2}\int_0^T \mathbb{E}_{x_0\sim P_0}\mathbb{E}_{x_t\sim P_{t|0}}\left[r(t)^2 \left\|\nabla_{x_t} \log P_{t|0}(x_t \mid x_0) - s_\theta(x_t,t)\right\|^2\right] \mathrm{d}t, \tag{31}$$

which is used to compute the original score matching loss $J_{SM}(\theta, r(\cdot)^2)$ given a dataset S_0. The gap between $J_{DSM}(\theta, r(\cdot)^2)$ and $J_{SM}(\theta, r(\cdot)^2)$ is a constant C_1 that does not depend on θ, which is shown as below.

The difference is expressed as

$$\begin{aligned}
&J_{SM}(\theta, r(\cdot)^2) - J_{DSM}(\theta, r(\cdot)^2) \\
=&\frac{1}{2}\int_{t=0}^{T} \mathbb{E}_{P_{0,t}(x_0,x_t)}\left[r(t)^2\left(\|\nabla_{x_t} \log P_t(x_t) - s_\theta(x_t,t)\|_2^2 - \|\nabla_{x_t} \log P_{t|0}(x_t \mid x_0) - s_\theta(x_t,t)\|_2^2\right)\right] \mathrm{d}t \\
=&\int_{t=0}^{T} r(t)^2 \left(\underbrace{\mathbb{E}_{P_{0,t}(x_0,x_t)}\left[-\left\langle s_\theta(x_t,t), \nabla_{x_t} \log P_t(x_t) + \nabla_{x_t} \log P_{t|0}(x_t \mid x_0)\right\rangle\right]}_{(1)} + C_1'(x_0,x_t)\right) \mathrm{d}t,
\end{aligned} \tag{32}$$

where $P_{0,t}$ is the joint distribution of x_0 and x_t, and $C_1'(x_0, x_t) = \mathbb{E}_{P_{0,t}(x_0,x_t)}\left[\frac{1}{2}\|\nabla_{x_t}\log P_t(x_t)\|_2^2 - \frac{1}{2}\|\nabla_{x_t}\log P_{t|0}(x_t \mid x_0)\|_2^2\right]$.

The first term (1) is zero because

$$\begin{aligned}
&\mathbb{E}_{P_{0,t}(x_0,x_t)}\left[\langle s_\theta(x_t,t), \nabla_{x_t}\log P_t(x_t)\rangle\right] = \mathbb{E}_{P_t(x_t)}\left[\langle s_\theta(x_t,t), \nabla_{x_t}\log P_t(x_t)\rangle\right] \\
=&\int_{x_t}\left\langle s_\theta(x_t,t), \frac{1}{P_t(x_t)}\nabla_{x_t}P_t(x_t)\right\rangle P_t(x_t)\mathrm{d}x_t \\
=&\int_{x_t}\left\langle s_\theta(x_t,t), \nabla_{x_t}\int_{x_0}P_{t|0}(x_t \mid x_0)P_0(x_0)\mathrm{d}x_0\right\rangle \mathrm{d}x_t \\
=&\int_{x_t}\left\langle s_\theta(x_t,t), \int_{x_0}P_{t|0}(x_t \mid x_0)P_0(x_0)\nabla_{x_t}\log(P_t(x_t \mid x_0))\mathrm{d}x_0\right\rangle \mathrm{d}x_t \\
=&\int_{x_0,x_t}P_{0,t}(x_0,x_t)\left\langle s_\theta(x_t,t), \nabla_{x_t}\log(P_{t|0}(x_t \mid x_0))\right\rangle \mathrm{d}x_0\mathrm{d}x_t \\
=&\mathbb{E}_{P_{0,t}(x_0,x_t)}\left[\left\langle s_\theta(x_t,t), \nabla_{x_t}\log P_{t|0}(x_t \mid x_0)\right\rangle\right]
\end{aligned} \tag{33}$$

Thus, we can bound the gap between $J_{DSM}(\theta, r(\cdot)^2)$ and $J_{SM}(\theta, r(\cdot)^2)$ as

$$C_1 = J_{SM}(\theta, r(\cdot)^2) - J_{DSM}(\theta, r(\cdot)^2) = \int_{t=0}^{T} r(t)^2 C_1'(x_0, x_t)\mathrm{d}t, \tag{34}$$

where $C_1'(x_0, x_t) = \mathbb{E}_{P_{0,t}(x_0,x_t)}\left[\frac{1}{2}\|\nabla_{x_t}\log P_t(x_t)\|_2^2 - \frac{1}{2}\|\nabla_{x_t}\log P_{t|0}(x_t \mid x_0)\|_2^2\right]$.

Therefore, if $J(\theta, S_0) = J_{DSM}(\theta, r(\cdot)^2) \le \epsilon$, by Lemma 4, we have

$$\begin{aligned}
D_{\mathrm{KL}}(P_0\|P_\theta) \le& J_{SM}(\theta, r(\cdot)^2) + D_{\mathrm{KL}}(P_T\|\pi) \\
=& J_{DSM}(\theta, r(\cdot)^2) + D_{\mathrm{KL}}(P_T\|\pi) + C_1 \\
\le& \epsilon + D_{\mathrm{KL}}(P_T\|\pi) + C_1,
\end{aligned} \tag{35}$$

which completes the proof.

D.2. Proof of Theorem 1

Theorem 1. *Let θ^* be the optimal diffusion parameter that solves the inner maximization (2) given a variable w. If the expected score-matching loss is bounded as $J(\theta) \le \bar{J}$ and the step size is chosen as $\eta \sim \mathcal{O}(\frac{1}{\sqrt{K}})$, the inner maximization error holds that*

$$\Delta' := \mathbb{E}_{x\sim P_{\theta^*}}[f(w,x)] - \mathbb{E}_k\mathbb{E}_{x\sim P_{\theta_k}}[f(w,x)] \le \frac{1}{\sqrt{K}}\max\{\epsilon, \bar{J}\}\|\mu^{(1)}\|, \tag{36}$$

where the outer expectation is taken over the randomness of output selection. In addition, the KL-divergence with respect to the nominal distribution is bounded as $\mathbb{E}_k[D_{\mathrm{KL}}(P_0||P_{\theta_k})] \le \epsilon + \frac{\max\{\epsilon,\bar{J}\}|\mu_C - \mu^{(1)}|}{\sqrt{K}(\mu_C - \mu^*)} + D_{\mathrm{KL}}(P_T||\pi) + C_1$, where $\mu_C > \mu^*$ and C_1 are constants, P_T is the output distribution of the forward process, and π is the initial distribution of the reverse process.

D.2.1. CONVERGENCE OF DUAL GRADIENT DESCENT

To prove Theorem 1, we need the convergence analysis of a general dual gradient descent in the following lemma.

Lemma 5 *Assume that the dual variable is updated following*

$$\mu_{k+1} = \max\{\mu_k - \eta \cdot b_k, 0\},$$

where $\eta > 0$ and b_k has the same dimension as μ and $\max\{\cdot, 0\}$ is an element-wise non-negativity operation. With η as the step size in dual gradient descent, given any $\mu > 0$, we have

$$\frac{1}{K}\sum_{k=1}^{K}\langle \mu_k - \mu, b_k\rangle \leq \eta \frac{1}{K}\sum_{k=1}^{K}\|b_k\|^2 + \frac{1}{2K\eta}\|\mu - \mu^{(1)}\|^2 \tag{37}$$

Proof For any dimension j such that $\mu_{k,j} \geq \eta b_{k,j}$, we have $\mu_{k+1,j} = \mu_{k,j} - \eta \cdot b_{k,j}$, so it holds for any $\mu > 0$ that

$$\left(b_{k,j} + \frac{1}{\eta}(\mu_{k+1,j} - \mu_{k,j})\right)(\mu_j - \mu_{k+1,j}) = 0. \tag{38}$$

For any dimension j such that $\mu_{k,j} < \eta b_{k,j}$, we have $\mu_{k+1,j} = 0$, and it holds for any $\mu > 0$ that

$$\left(b_{k,j} + \frac{1}{\eta}(\mu_{k+1,j} - \mu_{k,j})\right)(\mu_j - \mu_{k+1,j}) = (b_{k,j} - \frac{\mu_{k,j}}{\eta})\mu_j \geq (b_{k,j} - \frac{\eta b_{k,j}}{\eta})\mu_j = 0. \tag{39}$$

Combing (38) and (39) and we have for any $\mu > 0$ that

$$\left(b_k + \frac{1}{\eta}(\mu_{k+1} - \mu_k)\right)^\top (\mu - \mu_{k+1}) \geq 0. \tag{40}$$

Therefore, it holds for any $k \in [K]$ and $\mu > 0$ that

$$\begin{aligned}
(\mu_k - \mu)^\top b_k =& (\mu_k - \mu_{k+1})^\top b_k + (\mu_{k+1} - \mu)^\top b_k \\
\leq& (\mu_k - \mu_{k+1})^\top b_k + \frac{1}{\eta}(\mu_{k+1} - \mu_k)^\top (\mu - \mu_{k+1}) \\
=& (\mu_k - \mu_{k+1})^\top b_k + \frac{1}{2\eta}\left(\|\mu - \mu_k\|^2 - \|\mu - \mu_{k+1}\|^2 - \|\mu_{k+1} - \mu_k\|^2\right) \\
\leq& \eta\|b_k\|^2 + \frac{1}{2\eta}\left(\|\mu - \mu_k\|^2 - \|\mu - \mu_{k+1}\|^2\right),
\end{aligned} \tag{41}$$

where the first inequality holds by (40), the second equality holds by three-point property, and the last inequality holds because $\|a\|^2 + \|b\|^2 \geq 2a^\top b$.

Taking the sum over $k \in [K]$, we have

$$\sum_{k=1}^{K}\langle \mu_k - \mu, b_k\rangle \leq \eta\sum_{k=1}^{K}\|b_k\|^2 + \frac{1}{2\eta}\left(\|\mu - \mu_1\|^2\right), \tag{42}$$

which proves Lemma 5. ■

D.2.2. BOUND OF EXPECTED OBJECTIVE IN THEOREM 1

Proof Denote $F(\theta) = \mathbb{E}_{x\sim P_\theta}[f(w,x)]$ and the optimal parameter to solve the inner maximization of (2) is θ^*. Define the dual problem of the inner maximization of (2) as

$$Q(\mu) = \max_\theta F(\theta) + \mu(\epsilon - J(\theta, S_0)) \tag{43}$$

Given any $\mu > 0$, it holds be weak duality that

$$F(\theta^*) \le F(\theta^*) + \mu(\epsilon - J(\theta^*, S_0)) \le Q(\mu), \tag{44}$$

where the second inequality holds because $Q(\mu)$ maximizes $F(\theta) + \mu(\epsilon - J(\theta, S_0))$ over θ. Thus, the average gap between the expected loss of θ^* and θ_k is

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^K (F(\theta^*) - F(\theta_k)) &\le \frac{1}{T}\sum_{t=1}^T (Q(\mu_k) - F(\theta_k)) \\
&= \frac{1}{K}\sum_{k=1}^K \mu_k(\epsilon - J(\theta_k, S_0)) \\
&\le \eta \frac{1}{K}\sum_{k=1}^K (\epsilon - J(\theta_k, S_0))^2 + \frac{1}{2\eta}\frac{1}{K}\|\mu^{(1)}\|^2 \\
&\le \eta(\max\{\epsilon, \bar{J}\})^2 + \frac{1}{\eta}\frac{1}{K}\|\mu^{(1)}\|^2
\end{aligned} \tag{45}$$

where the equality holds because $\theta_k = \arg\max_\theta F(\theta) + \mu_k(\epsilon - J(\theta, S_0))$ by (3) and so $Q(\mu_k) = F(\theta_k) + \mu_k(\epsilon - J(\theta_k, S_0))$, the second inequality holds by Lemma 5 with the choice of $\mu = 0$, and the last inequality holds by the assumption $J(\theta_k, S_0) \le \bar{J}$.

Choosing $\eta = \frac{\mu^{(1)}}{\sqrt{K}\max\{\epsilon, \bar{J}\}}$, it holds that

$$\frac{1}{K}\sum_{k=1}^K (F(\theta^*) - F(\theta_k)) \le \frac{1}{\sqrt{K}}\max\{\epsilon, \bar{J}\}\mu^{(1)}, \tag{46}$$

which proves the bound of the expected loss given the uniformly selected $k \in [K]$.

D.2.3. BOUND OF KL DIVERGENCE IN THEOREM 1

For iteration k, denote $b(\theta) = \epsilon - J(\theta, S_0)$ and $b_k = \epsilon - J(\theta_k, S_0)$. Denote the constraint violation on the score matching loss at round k as $v_k = J(\theta_k, S_0) - \epsilon$. Denote the optimal dual variable as $\mu^* = \arg\min_\mu Q(\mu)$. Choose a dual variable $\mu_C > \mu^*$, we have the following decomposition.

$$\sum_{k=1}^K (\mu_k - \mu_C) b_k = \sum_{k=1}^K \mu_k \cdot b_k + \sum_{k=1}^K (-\mu_C) \cdot b_k \tag{47}$$

For the first term, we have

$$\begin{aligned}\sum_{k=1}^{K}\mu_k\cdot b_k &= \sum_{k=1}^{K}F(\theta_k)+\mu_k\cdot b_k-F(\theta_k)=\sum_{k=1}^{K}Q(\mu_k)-F(\theta_k)\\ &\geq KQ(\mu^*)-\sum_{k=1}^{K}F(\theta_k)\\ &\geq KQ(\mu^*)-\sum_{k=1}^{K}\max_{\theta\in\{\theta|J(\theta,S_0)\leq\epsilon+v_k\}}F(\theta)\\ &\geq KQ(\mu^*)-\sum_{k=1}^{K}\max_{\theta}(F(\theta)+\mu^*(\epsilon+v_k-J(\theta,S_0)))\\ &= KQ(\mu^*)-KQ(\mu^*)-\mu^*\sum_{k=1}^{K}v_k=-\mu^*\sum_{k=1}^{K}v_k,\end{aligned} \tag{48}$$

where the first inequality is because μ^* minimizes $Q(\mu_k)$, the second inequality holds because $\theta_k\in\{\theta\mid J(\theta,S_0)\leq\epsilon+v_k\}$, the third inequality holds by weak duality for μ^*. Continuing with (47), we have

$$\sum_{k=1}^{K}(\mu_k-\mu_C)b_k\geq\sum_{k=1}^{K}(-\mu^*v_k+\mu_C(J(\theta_k,S_0)-\epsilon))=(-\mu^*+\mu_C)\sum_{k=1}^{K}v_k, \tag{49}$$

where the equality holds because $v_k=J(\theta_k,S_0)-\epsilon$.

By Lemma 5 with the choice of $\mu=\mu_C$, we have

$$\frac{1}{K}\sum_{k=1}^{K}(\mu_k-\mu_C)b_k\leq\eta\frac{1}{K}\sum_{k=1}^{K}\|b_k\|^2+\frac{1}{2K\eta}(\mu_C-\mu^{(1)})^2\leq\eta(\max\{\epsilon,\bar{J}\})^2+\frac{1}{2K\eta}(\mu_C-\mu^{(1)})^2 \tag{50}$$

If the step size is chosen as $\eta=\frac{\mu^{(1)}}{\sqrt{K}\max\{\epsilon,\bar{J}\}}$, it holds that

$$\frac{1}{K}\sum_{k=1}^{K}(\mu_k-\mu_C)b_k\leq\frac{1}{\sqrt{K}}\max\{\epsilon,\bar{J}\}\left(\mu^{(1)}+\frac{|\mu_C-\mu^{(1)}|^2}{2\mu^{(1)}}\right). \tag{51}$$

Since μ_C is larger than μ^*, by (49), we have

$$\frac{1}{K}\sum_{k=1}^{K}v_k\leq\frac{1}{K(\mu_C-\mu^*)}\sum_{k=1}^{K}(\mu_k-\mu_C)b_k\leq\frac{\max\{\epsilon,\bar{J}\}\left(\mu^{(1)}+\frac{|\mu_C-\mu^{(1)}|^2}{2\mu^{(1)}}\right)}{\sqrt{K}(\mu_C-\mu^*)}=\frac{C_4}{\sqrt{K}}, \tag{52}$$

which means $\frac{1}{K}\sum_{k=1}^{K}J(\theta_k,S_0)\leq\epsilon+\frac{C_4}{\sqrt{K}}$.

Since $J(\theta_k,S_0)$ is the denoising score matching loss $J_{DSM}(\theta,r(\cdot)^2)$, by Lemma 3, we complete the proof by

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}D_{\mathrm{KL}}(P_0,P_{\theta_k})&\leq\frac{1}{K}\sum_{k=1}^{K}J(\theta_k,S_0)+D_{\mathrm{KL}}(P_T\|\pi)+C_1\\ &\leq\epsilon+\frac{C_4}{\sqrt{K}}+D_{\mathrm{KL}}(P_T||\pi)+C_1,\end{aligned} \tag{53}$$

where $C_1 = \int_{t=0}^{T} r(t)^2 \left(\mathbb{E}_{P_{0,t}(x_0,x_t)}\left[\frac{1}{2}\|\nabla_{x_t}\log P_t(x_t)\|_2^2 - \frac{1}{2}\|\nabla_{x_t}\log P_{t|0}(x_t \mid x_0)\|_2^2\right]\right)\mathrm{d}t$, $C_4 = \frac{\max\{\epsilon,\bar{J}\}\left(\mu^{(1)}+\frac{|\mu_C-\mu^{(1)}|^2}{2\mu^{(1)}}\right)}{(\mu_C-\mu^*)}$ with $\mu_C > \mu^*$. ■

D.3. Proof of Theorem 2

Theorem 2. *Assume that the DRO objective $f(w,x)$ is $\beta-$smooth and $L-$Lipschitz with respect to w and is upper bounded by $\bar{f}$. If each sampled dataset S_j from diffusion model has n samples and the step size for minimization is chosen as $\lambda \sim \mathcal{O}(\sqrt{\frac{1}{\beta L^2 H}})$, then with probability $1-\delta, \delta \in (0,1)$, the average Moreau envelope of the optimal inner maximization function $\phi(w) := \max_\theta \mathbb{E}_{P_\theta}[f(w,x)]$ satisfies*

$$\mathbb{E}\left[\|\nabla\phi_{\frac{1}{2\beta}}(w)\|^2\right] \leq 4\beta\Delta' + \frac{V_1}{\sqrt{n}} + \frac{V_2}{\sqrt{H}}, \tag{54}$$

where Δ' is the error of inner maximization bounded in Theorem 1, $V_1 = 8\beta\bar{f}\sqrt{\log(2/\delta)}$ and $V_2 = 4L\sqrt{(\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w))\beta}$.

Proof The proof of Theorem 2 follows the techniques of [13] with the difference that the maximization oracle is an empirical approximation. The optimization variable w is updated as $w_j = w_{j-1} - \lambda \cdot \nabla_w \mathbb{E}_{x\in S_j}[f(w_{j-1},x)]$. Define $\phi(w) := \max_\theta \mathbb{E}_{P_\theta}[f(w,x)]$ as the inner maximization function and define $\phi_\rho(w) := \min_{w'} \phi(w') + \frac{1}{2\rho}\|w-w'\|^2$ as the Moreau envelop of ϕ.

Denote $\hat{w}_j = \arg\min_w \phi(w) + \beta\|w - w_j\|^2$ as the proximal point of the Moreau envelop $\phi_{\frac{1}{2\beta}}(w)$. Since f is $\beta-$smooth, we have

$$f(\hat{w}_j,x) \geq f(w_j,x) + \langle\nabla_w f(w_j,x), \hat{w}_j - w_j\rangle - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2. \tag{55}$$

Thus, it holds with probability at least $1-\delta, \delta \in (0,1)$ that

$$\begin{aligned}
\phi(\hat{w}_j) &\geq \mathbb{E}_{P_{\theta_j}}[f(\hat{w}_j,x)] \geq \mathbb{E}_{S_j}[f(\hat{w}_j,x)] - \frac{\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} \\
&\geq \mathbb{E}_{S_j}[f(w_j,x)] + \langle g(w_j,S_j), \hat{w}_j - w_j\rangle - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 - \frac{\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} \\
&\geq \mathbb{E}_{P_{\theta_j}}[f(w_j,x)] + \langle g(w_j,S_j), \hat{w}_j - w_j\rangle - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 - \frac{2\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} \\
&\geq \phi(w_j) - \Delta_j' + \langle g(w_j,S_j), \hat{w}_j - w_j\rangle - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 - \frac{2\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}},
\end{aligned} \tag{56}$$

where the second inequality holds by applying McDiarmid's inequality on $\mathbb{E}_{P_{\theta_j}}[f(\hat{w}_j,x)]$ and $\bar{f}$ is the upper bound of f, the third inequality holds by (55) and $g(w_j,S_j) = \mathbb{E}_{S_j}[\nabla_w f(w_j,x)]$, the forth inequality holds by applying McDiarmid's inequality on $\mathbb{E}_{S_j}[f(w_j,x)]$, and the last inequality holds by Theorem 1.

Now, it holds that

$$\begin{aligned}
&\phi_{\frac{1}{2\beta}}(w_{j+1}) \\
=&\min_{w'} \phi(w') + \beta\|w_{j+1} - w'\|^2 \\
\leq&\phi(\hat{w}_j) + \beta\|w_j - \lambda g(w_j, S_j) - \hat{w}_j\|^2 \\
=&\phi(\hat{w}_j) + \beta\|w_j - \hat{w}_j\|^2 + 2\beta\lambda < g(w_j, S_j), \hat{w}_j - w_j > +\lambda^2\beta\|g(w_j, S_j)\|^2 \\
\leq&\phi_{\frac{1}{2\beta}}(w_j) + 2\beta\lambda < g(w_j, S_j), \hat{w}_j - w_j > +\lambda^2\beta\|g(w_j, S_j)\|^2 \\
\leq&\phi_{\frac{1}{2\beta}}(w_j) + 2\beta\lambda \left(\phi(\hat{w}_j) - \phi(w_j) + \Delta' + \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 + \frac{2\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}}\right) + \lambda^2\beta L^2,
\end{aligned} \tag{57}$$

where the last inequality holds by (56).

Summing up inequality (57) from $j = 1$ to $j = H$, we have

$$\begin{aligned}
&\phi_{\frac{1}{2\beta}}(w_{j+1}) \\
\leq&\phi_{\frac{1}{2\beta}}(w_1) + 2\beta\lambda \sum_{j=1}^{H} \left(\phi(\hat{w}_j) - \phi(w_j) + \Delta'_j + \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 + \frac{2\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}}\right) + \lambda^2\beta L^2 H,
\end{aligned} \tag{58}$$

and we further have

$$\begin{aligned}
&\frac{1}{H}\sum_{j=1}^{H} \left(\phi(w_j) - \phi(\hat{w}_j) - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2\right) \\
\leq&\Delta'_j + \frac{2\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} + \frac{\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w)}{2\beta\lambda H} + \frac{\lambda L^2}{2},
\end{aligned} \tag{59}$$

Also, it holds that

$$\begin{aligned}
&\phi(w_j) - \phi(\hat{w}_j) - \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 \\
=&\phi(w_j) + \beta\|w_j - w_j\|^2 - \phi(\hat{w}_j) - \beta\|\hat{w}_j - w_j\|^2 + \frac{\beta}{2}\|\hat{w}_j - w_j\|^2 \\
\geq&\frac{\beta}{2}\|\hat{w}_j - w_j\|^2 = \frac{1}{4\beta}\|\nabla\phi_{\frac{1}{2\beta}}(w_j)\|^2,
\end{aligned} \tag{60}$$

where the inequality holds by the definition of $\hat{w}_j$, and the last equality holds by the property of Moreau envelope such that $\frac{1}{2\beta}\nabla\phi_{\frac{1}{2\beta}}(w) = w - \hat{w}_j$ for any $w \in \mathcal{W}$. Therefore, we have

$$\begin{aligned}
&\frac{1}{H}\sum_{j=1}^{H} \|\nabla\phi_{\frac{1}{2\beta}}(w_j)\|^2 \\
\leq&4\beta\frac{1}{H}\sum_{j=1}^{H} \Delta'_j + \frac{8\beta\bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} + \frac{2(\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w))}{\lambda H} + 2\beta\lambda L^2.
\end{aligned} \tag{61}$$

By optimally choosing $\lambda = \sqrt{\frac{\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w)}{\beta L^2 H}}$, we have

$$\begin{aligned} &\mathbb{E}\left[\|\nabla \phi_{\frac{1}{2\beta}}(w_j)\|^2\right] \\ \leq & 4\beta\Delta' + \frac{8\beta \bar{f}\sqrt{\log(2/\delta)}}{\sqrt{n}} + 4L\sqrt{\frac{(\phi_{\frac{1}{2\beta}}(w_1) - \min_w \phi(w))\beta}{H}}. \end{aligned} \tag{62}$$

■

D.3.1. EXPLANATION OF CONVERGENCE BY MOREAU ENVELOP

The gradient bound of Moreau envelop indicates that the algorithm converges to an approximately stationary point, which is explained as below. The Moreau envelop $\nabla \phi_{\frac{1}{2\beta}}(w_j)$ satisfies $\nabla \phi_{\frac{1}{2\beta}}(w_j) = 2\beta \cdot (w_j - \hat{w}_j)$. Since the proximal point is $\hat{w}_j = \arg\min_w \phi(w) + \beta\|w - w_j\|^2$, we have $\nabla\phi(\hat{w}_j) + 2\beta(\hat{w}_j - w_j) = 0$. Thus, it holds that $\nabla \phi_{\frac{1}{2\beta}}(w_j) = 2\beta \cdot (w_j - \hat{w}_j) = \nabla\phi(\hat{w}_j)$ and $\|\hat{w}_j - w_j\| = \|\frac{1}{2\beta}\nabla\phi(\hat{w}_j)\| = \frac{1}{2\beta}\|\nabla \phi_{\frac{1}{2\beta}}(w_j)\|$. Therefore, if the gradient of Moreau envelop is bounded for the decision variable w, i.e. $\|\nabla \phi_{\frac{1}{2\beta}}(w)\| \leq \Delta$, its proximal point $\hat{w}$ is an approximately stationary point for the optimal inner maximization function ϕ (bounded gradient of the inner maximization function $\|\nabla\phi(\hat{w})\| \leq \Delta$), and the distance between w and $\hat{w}$ is close enough: $\|\hat{w}_j - w_j\| = \frac{1}{2\beta}\|\nabla \phi_{\frac{1}{2\beta}}(w_j)\| \leq \frac{\Delta}{2\beta}$. Thus, Algorithm 1 approximately converges.

Appendix E. Details of Experiments

E.1. Experiment Setups

This section reports numerical experiments on a representative ML task—time series forecasting based on the Electricity Maps [1] datasets —to evaluate the effectiveness of the proposed algorithms.

E.1.1. BASELINES

The baselines which are compared with our algorithms in our experiments are introduced as below.

Standard ML (ML): This method trains the ML to minimize the time series forecasting error without DRO.

Diffusion-based ML (DML): This is an ML model fine-tuned with diffusion-generated augmented datasets. Compared to D-DRO, DML performs standard training based on the augmented datasets rather than a distributionally robust training.

Wasserstein-based DRO (W-DRO): In this DRO framework, the ambiguity set is characterized by the Wasserstein metric. For our experiments, we employ the FWDRO algorithm proposed in [35], which transforms the inner maximization of W-DRO into an adversarial optimization with a mixed norm ball and then alternatively solve the adversarial examples and the ML weights.

KL-divergence-based DRO (KL-DRO): In this DRO framework, the ambiguity set is characterized by the KL- divergence. We employ the standard KL-DRO solution derived in [11], which is commonly adopted in practice.

E.1.2. DATASETS

The experiments are conducted based on Electricity Maps [1], a widely utilized global platform that provides high-resolution spatio-temporal data on electricity system operations, including carbon intensity (gCO_2eq/kWh) and energy mix, and is actively employed for carbon-aware scheduling and carbon footprint estimation in real systems such as data centers [8].

We utilize datasets from Electricity Maps that record hourly electricity carbon intensity over the period 2021∼2024 across four representative regions: California, United States (**BANC_21∼24**), Texas, United States (**ERCO_21∼24**), Queensland, Australia (**QLD_21∼24**), and the United Kingdom (**GB_21∼24**). These datasets capture fine-grained temporal variations in carbon intensity (measured in gCO_2eq/kWh) arising from different energy mixes in diverse geographical and regulatory contexts, thereby providing a comprehensive benchmark for evaluating carbon-aware forecasting models. For model training, we construct a dataset by merging **BANC_23** and **BANC_24**, resulting in 438 sequence samples. Model evaluation is then performed on multiple independent test sets, each consisting of 312 sequence samples drawn from other years and regions to ensure heterogeneous and challenging testing scenarios. To quantify the degree of distributional shift between the training and test sets, we compute the Wasserstein distance, which provides a principled measure of discrepancy between probability distributions. The calculated distances are reported alongside the dataset names in Table 2.

E.1.3. TRAINING SETUPS

The experimental setup is divided into the following parts:

Predictor: The predictors in `D-DRO` and all the baselines share the same two-layer stacked LSTM architecture with 128 and 64 hidden neurons.

Diffusion Model: The diffusion model in `D-DRO` is DDPM [10] which has $T = 500$ steps in a forward or a backward process.

Training: For `D-DRO`, we adopt the PPO-based reformulation in (20) for inner maximization. We train the reference DDPM θ_0 in (20) based on the original training dataset **BANC_2324** (**BANC_23** & **BANC_24**) and use it to generate an initial dataset z_0 to calculate r_θ in (20). The sampling variance of DDPM is chosen from a range $[0.1, 0.5]$. To improve training efficiency, only the last $T' = 15$ backward steps of the DDPM model are fine-tuned by (20). We choose a slightly higher clipping parameter $\kappa = 0.4$ in (20) to encourage the maximization while maintaining stability. We choose $\epsilon = 0.015$ as `InnerMax`'s adversarial budget which gives the best average performance over all validation datasets. We choose $\eta = 0.01$ as the rate to update the Lagrangian parameter α in Algorithm 1. We use the Adam optimizer with a learning rate of 10^{-5} for both the diffusion training in the maximization and the predictor update in minimization. The diffusion model is trained for 10 inner epochs with a batch size of 64. The predictor is trained for 15 epochs with a batch size of 64.

For the baseline methods, we choose the same neural network architecture as `D-DRO`. We carefully tuned the hyperparameters of the baseline algorithms to achieve optimal average performance over all validation datasets. For `W-DRO`, we consider the Wasserstein distance with respect to l_2-norm and set the adversarial budget as $\epsilon = 0.3$. For `KL-DRO`, we choose the adversarial budget $\epsilon = 4$. The predictors in all baseline methods are trained by Adam optimizer with a learning rate of 1×10^{-5}.

In terms of training, all baselines and D−DRO are initialized from the same ML model pretrained for 100 epochs, using a batch size of 64 in both the pretraining and subsequent training phases. For our method, the outer minimization in DRO is performed for 15 iterations, and each outer iteration contains 10 inner maximization steps, during which the diffusion model is fine-tuned. After these 10 inner maximization steps, the current diffusion model generates a dataset z_θ of the same size as z_0, and the pretrained ML model is then further trained for two epochs on z_θ. Consequently, our algorithm effectively fine-tunes the ML model for 30 rounds in total, while utilizing 15 different augmented datasets z_θ. To ensure a fully fair ablation setup, DML is also fine-tuned for 30 rounds, but always with the same dataset z_0. In contrast, since W−DRO and KL−DRO are not part of the ablation study, they are trained for 100 epochs to achieve their best performance.

E.2. Main results

We tune the hyperparameters of all methods based on validation datasets and present their best performance in Table 2 where all the datasets are OOD testing datasets with different discrepancies.

Table 2: Test MSE on different datasets.

Datasets (Wasserstein Distance)	Algorithms				
	D−DRO	**KL−DRO**	**W−DRO**	**DML**	**ML**
BANC_22 (0.0240)	**0.0047**	0.0086	0.0073	0.0078	0.0183
BANC_21 (0.1213)	**0.0054**	0.0112	0.0121	0.0093	0.0238
QLD_24 (0.2171)	**0.0450**	0.0754	0.0823	0.0766	0.0887
QLD_23 (0.2033)	**0.0509**	0.0831	0.0879	0.0834	0.0946
QLD_22 (0.2782)	**0.0192**	0.0379	0.0557	0.0352	0.0667
QLD_21 (0.3054)	**0.0186**	0.0377	0.0574	0.0339	0.0696
GB_24 (0.0419)	**0.0119**	0.0178	0.0176	0.0172	0.0285
GB_23 (0.0666)	**0.0100**	0.0178	0.0200	0.0164	0.0311
GB_22 (0.1255)	**0.0105**	0.0197	0.0245	0.0172	0.0360
GB_21 (0.1359)	**0.0094**	0.0181	0.0229	0.0158	0.0340
ERCO_24 (0.1206)	**0.0158**	0.0241	0.0266	0.0224	0.0379
ERCO_23 (0.1207)	**0.0106**	0.0179	0.0196	0.0162	0.0319
ERCO_22 (0.1581)	**0.0076**	0.0146	0.0187	0.0123	0.0312
ERCO_21 (0.1417)	**0.0093**	0.0189	0.0263	0.0160	0.0382
Average	**0.0163**	0.0288	0.0342	0.0271	0.0450
Maximum	**0.0509**	0.0831	0.0879	0.0834	0.0946

We can find that all DRO methods improve the testing performance for OOD datasets comparing to ML by optimizing the worst-case expected loss while DML improves upon ML by leveraging diffusion-generated augmented datasets. Using the performance of ML as the reference, D−DRO achieves the largest performance gain of 63.7%, followed by DML with a 39.7% improvement. In contrast, KL−DRO and W−DRO yield improvements of 36.1% and 24.0%, respectively.

Notably, DML surpasses KL−DRO by 3.6% only after augmentation training with diffusion-generated datasets, highlighting the significant positive impact of diffusion model. Moreover, DML differs from D−DRO only in that it disables the DRO component, while their training procedures remain identical (see Appendix E for details). Thus, D−DRO, DML, and ML together form two ablation studies: the performance gap between D−DRO and DML confirms that DRO contributes a 39.7% performance gain to D−DRO, while the gap between DML and ML verifies that the diffusion model also provides a 39.7% performance gain.

Although both DML and the other two DRO deliver noticeable performance gains, D−DRO still outperforms them by an average margin of 30.4%. The underlying reason may lie in the fact that D−DRO leverages the distribution learning capability of the diffusion model when constructing the ambiguity set, enabling the generation of adversarial distributions that are both strong and realistic,

thereby achieving superior OOD generalization. In contrast, the KL-divergence used in KL−DRO requires all distributions to be absolutely continuous with respect to the training distribution P_0, which constrains the support space of the ambiguity set and limits the search for strong adversaries. Among the baselines, W−DRO performs the worst, likely due to the fact that solving Wasserstein-based DRO usually involves optimal transport problems or their relaxations, which are computationally more demanding and often rely on approximate methods such as dual reformulation or adversarial training, resulting in suboptimal solutions. Moreover, since W−DRO allows adversarial distributions with supports different from the training distribution, it is theoretically more flexible and closer to real distribution shifts, but this flexibility can produce overly extreme adversaries and thus lead to overly conservative model training.

E.3. More Out-Of-Distribution Tests

E.3.1. EFFECT OF NOISY TYPES

In this test, we add a certain amount of different types of noise into each test set to compare the noise robustness of different algorithms. The noise types include Gaussian Noise, Perlin Noise, and Cutout Noise.

Gaussian Noise: In the Gaussian Noise test, we add Gaussian Noise with $\sigma = 0.1$ to each test set. As shown in Table 3, all algorithms exhibit performance degradation compared to the noise-free setting. Nevertheless, D−DRO still significantly outperforms the others: taking ML as the reference, our method achieves a 48.3% improvement, followed by DML and KL−DRO with gains of 31.7% and 29.15%, respectively. The weakest performer is W−DRO, which surpasses ML by only 20%. In fact, W−DRO is relatively adept at handling Gaussian Noise compared to other noise types, and thus maintains a noticeable advantage even under $\sigma = 0.1$ Gaussian Noise, a trend further confirmed in subsequent experiments. On the other hand, both DML and D−DRO are trained with diffusion-generated augmented datasets, which inherently possess noise characteristics due to the Gaussian-based diffusion process. As a result, their performance remains robust under Gaussian Noise perturbations.

Table 3: Gaussian-Corrupted Test.

Dataset	Algorithms				
	D−DRO	KL−DRO	W−DRO	DML	ML
BANC_22	**0.0170**	0.0197	0.0183	0.0189	0.0291
BANC_21	**0.0177**	0.0214	0.0216	0.0199	0.0337
QLD_24	**0.0560**	0.0839	0.0911	0.0847	0.0990
QLD_23	**0.0615**	0.0925	0.0960	0.0934	0.1034
QLD_22	**0.0307**	0.0476	0.0639	0.0445	0.0766
QLD_21	**0.0297**	0.0475	0.0649	0.0431	0.0772
GB_24	**0.0238**	0.0280	0.0276	0.0279	0.0381
GB_23	**0.0221**	0.0260	0.0305	0.0269	0.0414
GB_22	**0.0227**	0.0302	0.0343	0.0279	0.0458
GB_21	**0.0211**	0.0279	0.0334	0.0256	0.0438
ERCO_24	**0.0281**	0.0350	0.0369	0.0340	0.0471
ERCO_23	**0.0228**	0.0281	0.0294	0.0267	0.0418
ERCO_22	**0.0206**	0.0256	0.0293	0.0230	0.0416
ERCO_21	**0.0217**	0.0295	0.0357	0.0264	0.0475
Average	**0.0282**	0.0388	0.0438	0.0373	0.0547
Maximum	**0.0615**	0.0925	0.0960	0.0934	0.1034

Perlin Noise: Perlin Noise is a smooth pseudo-random gradient noise commonly used to simulate natural textures such as clouds, terrains, and wood grains. By combining multiple Perlin Noise components with different frequencies and amplitudes (known as octaves), more complex fractal noise can be produced. In this experiment, we superimpose 8 layers of Perlin Noise, with the noise amplitude normalized to the range [-1, 1].

As shown in Table 4, taking ML as the reference, our method outperforms ML by 76.4%, while DML achieves a 58.0% improvement, and both perform better than in the Gaussian Noise test. Although KL−DRO also shows a larger gain relative to ML, its MSE remains roughly the same as in the Gaussian Noise test; the apparent improvement is primarily due to the substantial performance drop of ML in this experiment. In contrast, W−DRO outperforms ML by only 9.6%, a significant decline compared to its performance under Gaussian Noise, with the MSE reduced by as much as 50%. This indicates that W−DRO is not well-suited for handling Perlin Noise, likely because the Wasserstein distance measures global distributional transport cost and is more effective in capturing smooth, small perturbations (e.g., Gaussian Noise), but fails to adequately model the long-range correlated patterns of Perlin Noise within the Wasserstein ball.

Cutout Noise: Cutout Noise is a commonly used perturbation method that randomly selects a region of the input data and sets its values to a constant, thereby simulating partial information loss. In our experiment, we randomly mask 30% of the sequence and set the masked values to a constant of 1. As shown in Table 5, the performance of all algorithms is very close to their performance under Perlin Noise. We attribute this to the fact that, although Perlin and Cutout Noises differ in form, both represent structured local perturbations that disrupt the continuity of the input patterns, thereby posing similar challenges to all algorithms and resulting in comparable performance under these two types of noise at a given perturbation level. This is further confirmed in the subsequent gradient-perturbation tests.

Table 4: Perlin-Corrupted Test.

Dataset	Algorithms				
	D−DRO	KL−DRO	W−DRO	DML	ML
BANC_22	**0.0117**	0.0296	0.0620	0.0227	0.0663
BANC_21	**0.0110**	0.0288	0.0591	0.0211	0.0662
QLD_24	**0.0355**	0.0636	0.0862	0.0588	0.0928
QLD_23	**0.0406**	0.0685	0.0787	0.0669	0.0860
QLD_22	**0.0186**	0.0476	0.0852	0.0341	0.0904
QLD_21	**0.0192**	0.0475	0.0865	0.0355	0.0940
GB_24	**0.0147**	0.0272	0.0499	0.0252	0.0563
GB_23	**0.0133**	0.0295	0.0560	0.0239	0.0628
GB_22	**0.0132**	0.0311	0.0611	0.0247	0.0666
GB_21	**0.0114**	0.0311	0.0586	0.0224	0.0677
ERCO_24	**0.0154**	0.0267	0.0436	0.0238	0.0525
ERCO_23	**0.0142**	0.0382	0.0767	0.0262	0.0829
ERCO_22	**0.0109**	0.0324	0.0655	0.0223	0.0732
ERCO_21	**0.0104**	0.0283	0.0506	0.0194	0.0605
Average	**0.0171**	0.0379	0.0657	0.0305	0.0727
Maximum	**0.0406**	0.0685	0.0865	0.0669	0.0940

Table 5: Cutout-Corrupted Test.

Dataset	Algorithms				
	D−DRO	KL−DRO	W−DRO	DML	ML
BANC_22	**0.0063**	0.0181	0.0297	0.0125	0.0385
BANC_21	**0.0095**	0.0302	0.0547	0.0192	0.0637
QLD_24	**0.0404**	0.0798	0.1096	0.0697	0.1148
QLD_23	**0.0426**	0.0802	0.1028	0.0743	0.1092
QLD_22	**0.0196**	0.0507	0.0881	0.0384	0.0971
QLD_21	**0.0208**	0.0534	0.0916	0.0400	0.1005
GB_24	**0.0145**	0.0351	0.0598	0.0252	0.0688
GB_23	**0.0122**	0.0343	0.0603	0.0230	0.0681
GB_22	**0.0129**	0.0361	0.0639	0.0245	0.0720
GB_21	**0.0127**	0.0356	0.0651	0.0245	0.0740
ERCO_24	**0.0157**	0.0375	0.0643	0.0275	0.0732
ERCO_23	**0.0134**	0.0349	0.0599	0.0241	0.0678
ERCO_22	**0.0134**	0.0349	0.0599	0.0241	0.0678
ERCO_21	**0.0116**	0.0354	0.0671	0.0225	0.0758
Average	**0.0174**	0.0425	0.0699	0.0319	0.0782
Maximum	**0.0426**	0.0802	0.1096	0.0743	0.1148

E.3.2. EFFECT OF NOISY LEVELS

In this experiment, we progressively increased the intensity of three types of noise. For Gaussian Noise, the perturbation range is set to $\sigma \in [0.05, 0.2]$; for Perlin Noise, the amplitude is controlled within the range $[0.05, 1]$; and for Cutout Noise, the Cutout Mask Ratio is adjusted between

$[10\%, 40\%]$. As shown in Figures 1, 2, 3, D−DRO consistently outperforms all baseline methods across different noise types and intensity levels. With increasing noise strength, we observe that variations in Gaussian Noise have a more pronounced impact on all methods compared to Perlin and Cutout Noise. In contrast, the impact of stronger Perlin and Cutout Noise on D−DRO remains limited, and even at higher noise levels, D−DRO maintains stable and superior performance, highlighting its strong robustness. For W−DRO, however, the performance degradation under Perlin and Cutout Noise is much greater, with trends almost identical to ML, indicating that W−DRO is not effective in handling Perlin and Cutout Noise but is relatively better at coping with Gaussian Noise. Interestingly, when the Cutout Mask Ratio is 30%, the performance of all algorithms is nearly identical to their performance under Perlin Noise with amplitude 1, suggesting that at specific noise levels, Perlin and Cutout—though different in form—both represent structured local perturbations that disrupt input continuity to a similar degree, thereby producing comparable impacts on the algorithms.

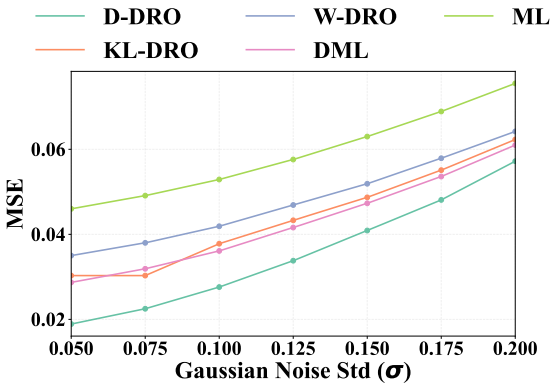

Figure 1: Gaussian perturbation strength

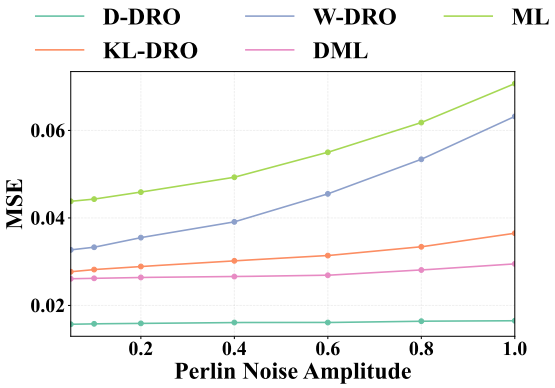

Figure 2: Perlin perturbation strength

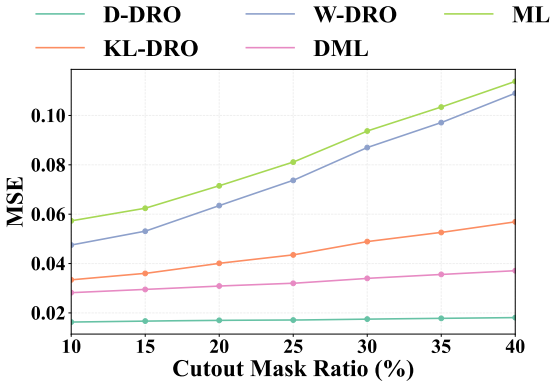

Figure 3: Cutout perturbation strength

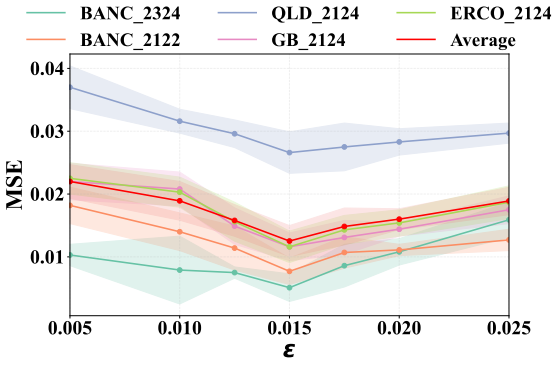

Figure 4: Effect of budget ϵ in D−DRO

E.4. Effects of DRO Budget

Finally, we examine the impact of the budget parameter ϵ in (2) on the performance of D−DRO. As illustrated in Fig. 4, the loss–ϵ curves across all datasets display a concave trend, with the best average performance achieved around $\epsilon = 0.015$. When ϵ is smaller than this threshold, the diffusion-modeled distributions are overly restricted to the training data, thereby hindering the ability of D−DRO to generalize to OOD datasets. In contrast, when ϵ becomes excessively large, the enlarged ambiguity set causes D−DRO to conservatively optimize against irrelevant distributions, which degrades its performance on real OOD datasets. Hence, selecting an appropriate value of ϵ is essential for constructing effective adversarial distributions, ensuring a proper balance between average-case and worst-case performance.