Gradient Flow Sampler-based Distributionally Robust Optimization

Zusen Xu[1,2] **Jia-Jie Zhu**[1]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
jiajie@kth.se
[2]Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany
zusen.xu@wias-berlin.de

Abstract

We propose a mathematically principled PDE gradient flow framework for distributionally robust optimization (DRO). Exploiting the recent advances in the intersection of Monte Carlo sampling and statistical optimal transport, we show that our theoretical framework can be implemented as practical algorithms for sampling from worst-case distributions and, consequently, DRO. While numerous previous works have relied on dual reformulation techniques, we contribute a sound and complete gradient flow view based on SDEs or PDEs that can be used to construct new algorithms for general, potentially non-convex, losses. As an application, we solve the entropy-regularized Wasserstein DRO problem using the different gradient flows. Notably, we also show some simple reductions of our framework recover exactly previously proposed popular DRO methods, and provide new insights into their theoretical limits and optimization dynamics of DRO. Numerical studies based on stochastic gradient descent on machine learning tasks provide empirical backing for our theoretical findings.

1 Introduction

Distributionally robust optimization (DRO) (Delage and Ye, 2010; Kuhn et al., 2025) is a framework that aims to enhance the robustness of the solution to optimization problems. Since the original Wasserstein distributionally robust optimization (DRO) works (Mohajerin Esfahani and Kuhn, 2018; Zhao and Guan, 2018; Gao and Kleywegt, 2016), many subsequent works have presented variations of problems with various numerical solutions. In particular, Sinkhorn DRO (Wang et al., 2021), also referred to as entropy-regularized Wasserstein DRO (Dapogny et al., 2023), demonstrates significant advantages in handling tasks with class imbalance. Its inner maximization problem can be formulated as a penalized optimization problem:

$$\min_{\theta\in\Theta}\max_{\rho\in\mathcal{P}}\left\{\int \ell(\theta,y)\,\mathrm{d}\rho(y) - \tfrac{1}{2\tau}W_\epsilon^2(\rho,\widehat{\rho_N})\right\} \tag{1}$$

where $\tau > 0$ is the temperature parameter, $\widehat{\rho_N}$ is an empirical dataset, and $W_\epsilon^2(\mu,\nu) := \inf_{\gamma\in\Pi(\mu,\nu)}\{\mathbb{E}_\gamma[c] + \epsilon H(\gamma)\}$ is the entropy-regularized optimal transport (OT) divergence. Here, $\Pi(\mu,\nu)$ is the set of couplings with marginals μ and ν, $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a cost function, $H(\gamma) = \int \gamma\log\gamma$ is the negative entropy, and $\epsilon > 0$ is the entropy regularization parameter. This DRO problem was considered by Wang et al. (2021). If we set $\epsilon = 0$, we recover the penalized version of the Wasserstein DRO as considered by Sinha et al. (2017). Without loss of generality, we will temporarily focus on the interesting choice of W_ϵ^2, while other choices such as the KL divergences can also be straightforwardly adapted to our framework.

We observe that the inner maximization problem in (1) is equivalent to applying the entropy-regularized JKO operator (detailed in (3)) from the PDE literature(Jordan et al., 1998; Peyré, 2015), resulting the equivalent formulation to (1)

$$\min_{\theta\in\Theta} \mathbb{E}_{y\sim\pi_Y}\ell(\theta, y), \quad \pi_Y = \epsilon\text{-JKO}_{\tau V}(\widehat{\rho_N}) \tag{2}$$

where $V = -\ell$. Exploiting this connection, our key insight is solving the inner problem of DRO (1) can be achieved by sampling from the worst-case distribution $y \sim \pi_Y$ using a gradient flow. This simplification of the problem structure provides intuition for novel algorithmic design, which we use to effortlessly obtain novel algorithms such as Wasserstein-Fisher-Rao and Stein variational gradient-based DRO algorithms. We term our methodology *Gradient Flow Sampler-based Distributionally Robust Optimization* (GF-DRO). We illustrate the main idea in Figure 1.

To obtain theoretical guarantees, we connect the stochastic approximation of the gradient w.r.t. θ for the outer DRO problem $\widehat{g} = \nabla_\theta \ell(\theta, y), y \sim \pi_Y$, with that of the analysis of a gradient flow-based sampler. Subsequently, the gradient estimate is used for the outer DRO problem via SGD-like updates, e.g., $\theta^{s+1} \leftarrow \theta^s - \eta\widehat{g}$. The bound of $\widehat{g}$ deviating from the true gradient can be used in the downstream standard optimization error bound for DRO solution.

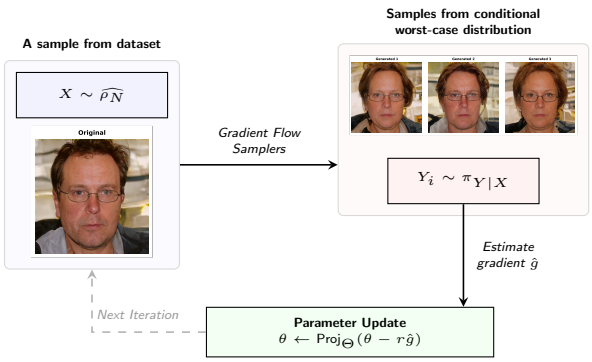


Figure 1: **GF-DRO** generates samples from the conditional worst-case distribution along the evolution of S/PDE gradient flows, which is fundamentally different from previous DRO methods. See text for detailed explanation and Section 4 for numerical experiments.

Contributions. Importantly, we focus on the algorithmic design for solving DRO, rather than proposing new ambiguity sets. Our core contributions are: 1. We propose GF-DRO, an algorithm framework bridging gradient-flow sampling and DRO—a connection previously missing in the literature. Unlike classical methods (Zhao and Guan, 2018; Mohajerin Esfahani and Kuhn, 2018; Gao and Kleywegt, 2016) relying on ad-hoc dual reformulations, GF-DRO directly solves the primal variational problem (3), enabling the treatment of general, potentially non-convex losses. 2. By framing the inner problem as a *distributional* stochastic gradient descent-ascent, we effortlessly integrate advanced samplers (e.g., Wasserstein-Fisher-Rao and Stein variational gradient flows). This advances the theoretical depth beyond previous ODE-based methods (Sinha et al., 2017) by utilizing S/PDEs and interacting particle systems.

2 Method

2.1 Schrödinger Problem Formulation of DRO

Our starting point is the following entropy-regularized JKO (Jordan-Kinderlehrer-Otto) operator from PDE(Jordan et al., 1998; Peyré, 2015):

$$\epsilon\text{-JKO}_{\tau V}(\rho_0) := \operatorname*{argmin}_{\rho\in\mathcal{P}} \left\{ \int V \, \mathrm{d}\rho + \tfrac{1}{2\tau} W_\epsilon^2(\rho, \rho_0) \right\} \tag{3}$$

This operation, hence the inner maximization problem of DRO, is a special case of the static Schrödinger problem with a free marginal, i.e., half bridge or one-sided bridge. We now provide a variational characterization enabling us to perform the sampling task in (2).

Proposition 1. *Solving for the entropy-regularized JKO operator* (3) *is equivalent to solving the Schrödinger half bridge problem*

$$\min_{\substack{\Pi\in\mathcal{M}^2 \\ \int \Pi \, \mathrm{d}y = \rho_0}} \left\{ \int V \, \mathrm{d}\Pi + \frac{1}{2\tau} \int c(x, y) \, \mathrm{d}\Pi + \frac{\epsilon}{2\tau} \int \log \Pi \, \mathrm{d}\Pi \right\} \tag{4}$$

The optimal marginal distribution of Y in (4) is given by a mixture distribution, for some normalization constants Z_x:

$$\pi_Y = \mathbb{E}_{x\sim\rho_0}\left[\frac{1}{Z_x}\exp\left[-\frac{2\tau V(y)+c(y,x)}{\epsilon}\right]\right]. \tag{5}$$

Consequently, (4) is equivalent to the minimization of the expected KL divergence with respect to the conditional distribution

$$\min_{\rho_{Y|X}} \mathbb{E}_{x\sim\rho_0}\,\mathrm{KL}\left(\rho_{Y|X=x}(y)\middle|\tfrac{1}{Z_x}\exp\left[-\tfrac{2\tau V(y)+c(y,x)}{\epsilon}\right]\right). \tag{6}$$

Letting $\pi_{Y|X} := \frac{1}{Z_x}\exp\left[-\frac{2\tau V(y)+c(y,x)}{\epsilon}\right]$, problem (6) provides the stochastic entropic transport map: $x \mapsto \pi_{Y|X=x}(y)$. This statement gives the overall variational structure of the DRO problem (1). We are now ready to introduce our proposed method GF-DRO in the next section.

2.2 Gradient Flow Sampler-based DRO

Using the variational problem (6), we define the energy functional as the KL divergence energy functional (without expectation):

$$F(\rho) := \tfrac{\epsilon}{2\tau}\,\mathrm{KL}\left(\rho\,\middle|\,\tfrac{1}{Z_x}\exp\left[-\tfrac{2\tau V(y)+c(y,x)}{\epsilon}\right]\right) = \int V\,\mathrm{d}\rho + \frac{1}{2\tau}W_c^2(\rho,\delta_x) + \frac{\epsilon}{2\tau}\int \rho\log\rho + \text{const.} \tag{7}$$

Note that the formulation on the right-hand side was also observed by Chen et al. (2022) in studying the proximal sampler.

With those ingredients, in this paper, we propose to achieve sampling from the conditional distribution $\pi_{Y|X}$ by simulating the *gradient system* $(\mathcal{P}, F, \mathbb{G})$. This results in the formal gradient flow equation:

$$\dot{\rho} = -\mathbb{G}^{-1}(\rho)\,DF(\rho). \tag{8}$$

where we can freely choose the dissipation geometry $\mathbb{G}$ for the gradient flow. Then, our central methodology is the following perspective of sampling from conditional distributions connected to the KL-minimization problem (6). Consider solving the inner maximization problem (1) using the gradient flow sampler, we now propose the following general-purpose gradient flow sampler-based DRO framework. Other straigforward variants, such as using sample average approximation (SAA)

Algorithm 1 Gradient Flow Sampler-based DRO

1: **Input:** Initial distribution ρ_0, e.g., empirical distribution $\rho_0 = \widehat{\rho_N}$, constraint set Θ, $\tau, \epsilon > 0$, stepsize r_s.
2: **for** iteration count $s = 0, \ldots$ **do**
3: Sample $x^s \sim \rho_0$
4: Sample $y^s \sim \pi_{Y|X}$ using a gradient flow (e.g. with energy functional given by (7))
5: DRO step: $\theta^{s+1} \leftarrow \mathrm{Proj}_\Theta(\theta^s - r_s\nabla_\theta\ell(\theta^s, y^s))$
6: **end for**

instead of stochastic approximation (SA) above, are possible.

2.3 Applications of GF-DRO

Wasserstein Gradient Flow Sampler As the first practical outcome of our theoretical insights, we instantiate a Wasserstein gradient flow (WGF) sampler. We consider the Wasserstein gradient system with the driving functional F and the Wasserstein metric as the dissipation geometry. In sampling algorithms, this gradient flow is typically implemented by discretizing the Langevin SDE, resulting in the forward Euler discretization known as the unadjusted Langevin algorithm (ULA):

$$y_{t+1} \leftarrow y_t - \eta\nabla V(y_t) + \sqrt{\eta\epsilon/\tau}\,\xi_t \tag{9}$$

where $\xi_t \sim \mathcal{N}(0, I)$ and the target distribution $\pi(x) \propto e^{-V(x)}$. For the penalized Wasserstein DRO problem ($\epsilon = 0$), this step specializes to the update rule that coincides with the inner SGD step used

in the WRM algorithm by Sinha et al. (2017). Hence, their WRM algorithm is a special case (with only ODE) of our GF-DRO framework. The detailed algorithm is deferred to Appendix C.1.

Wasserstein Fisher-Rao Flow Sampler Moreover, we go beyond the standard Wasserstein gradient flow and introduce the Wasserstein Fisher-Rao (WFR) gradient system. Specifically, we consider the HK/WFR gradient system associated with the reaction-diffusion PDE:

$$\frac{\partial \mu}{\partial t} = \alpha \operatorname{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}\right) - \beta\, \mu \left(\frac{\delta F}{\delta \mu} - \int \frac{\delta F}{\delta \mu}\, \mathrm{d}\mu\right) \tag{10}$$

The primary distinction of this sampler, compared to the WGF version, is the addition of a birth-death mechanism. Crucially, this mass reallocation mechanism enables the system to concentrate its search within low-energy regions and effectively escape from local optima; see details in Appendix C.2.

We also provide a Stein variational gradient (SVG) flow version of our framework in Appendix C.3.

3 Optimization Complexity Analysis

Letting $\Phi(\theta) := \max_{\rho\in\mathcal{P}} \left\{\int \ell(\theta, z)\, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})\right\}$, the DRO problem (1) can be written as $\min_{\theta\in\Theta} \Phi(\theta)$. We analyze the computational complexity of our proposed discrete-time algorithms, aiming to find an ϵ_{opt}-stationary point θ^S at the last iteration S, i.e. $\mathbb{E}[\|\nabla\Phi(\theta^S)\|^2] \le \epsilon_{\text{opt}}^2$. Our analysis uses a standard framework for non-convex stochastic optimization (Ghadimi and Lan, 2013). We rely on standard non-convex stochastic optimization assumptions (detailed in Appendix D.1), including L_Φ-smoothness of the objective, L_f-Lipschitz continuous gradients, and bounded variance for the gradient estimator (Assumption 8). Under these standard assumptions, we can bound the gradient bias and establish a general convergence result for the outer loop.

Theorem 2 (Outer Loop Convergence)**.** *Let Assumptions 8 hold. With a stepsize* $r = \sqrt{\frac{1}{SL_\Phi\sigma^2}}$*, the outer loop requires* $S = O(1/\epsilon_{opt}^4)$ *iterations to find an* ϵ_{opt}*-stationary point (i.e.* $\mathbb{E}[\|\nabla\Phi(\theta^S)\|^2] \le \epsilon_{opt}^2$*), provided the error* δ_{sample} *in Assumption 8 is controlled such that* $\delta_{sample} = O(\epsilon_{opt}/L_f)$*.*

This complexity aligns with standard non-convex stochastic algorithms. Next, we derive the complexity of the Algorithm 3 as an example. For the ULA specifically, we assume the conditional target distribution satisfies L_U-smoothness and the λ_U-log-Sobolev inequality (LSI) (Assumption 9). This relaxes the strict log-convexity assumption commonly adopted in previous works (Sinha et al., 2017; Wang et al., 2021). By combining the derived inner loop complexity analysis with the outer loop convergence rate from Theorem 2, we establish the total computational complexity.

Theorem 3 (Complexity of Algorithm 3)**.** *Under Assumptions 8 and 9, the total complexity for the ULA-based GF-DRO to find an* ϵ_{opt}*-stationary point is* $\tilde{O}\left(\frac{d}{\epsilon_{opt}^6}\right) = O\left(\frac{d}{\epsilon_{opt}^6} \cdot \log \frac{1}{\epsilon_{opt}}\right)$*.*

The detailed derivations for the gradient bias, the proofs of Theorem 2 and Theorem 3 are provided in Appendix D.

4 Experiments

We conduct numerical experiments on different tasks to validate our theoretical insights. Our goal is to compare the performance of four different DRO methods: Algorithm 4 (WFR-DRO), Algorithm 3 (WGF-DRO), the dual method for SDRO (Dual) (Wang et al., 2021), and WRM (Sinha et al., 2017). For methodological consistency across all experiments, the Dual method uses Randomized Truncation MLMC (Blanchet and Glynn, 2015) as suggested in (Wang et al., 2021), and the WRM solves its inner problem using gradient descent.

4.1 Robustness on Convolutional Neural Networks

We evaluate our methods on MNIST using the LeNet-5 architecture to compare their performance in high-dimensional, end-to-end training settings. We train the entire network using DRO algorithms. We assess robustness using PGD attacks under both L_2 and L_∞ norm constraints. We compare

our proposed WFR-DRO and WGF-DRO ($\tau = 1.0, \epsilon = 0.05, m = 8$) against the WRM baseline and the Dual Sinkhorn DRO approach. As shown in Tables 1 and 2, WFR-DRO consistently yields the lowest test error rates across all attack intensities Δ. While WGF-DRO consistently outperforms WRM, demonstrating that stochasticity and entropy regularization enhance robustness in non-convex landscapes, WFR's birth-death mechanism provides superior convergence for complex convolutional neural networks, outperforming the Dual baseline as well.

Table 1: PGD-L_2 Robustness (Test Error %)

Method	$\Delta = 0.00$	$\Delta = 0.05$	$\Delta = 0.10$	$\Delta = 0.15$
SDRO_Dual	0.98 ± 0.10	3.21 ± 0.29	9.54 ± 0.82	24.56 ± 2.35
WRM	1.00 ± 0.10	3.71 ± 0.34	12.28 ± 0.98	31.25 ± 2.12
WGF-DRO	0.99 ± 0.12	3.62 ± 0.37	11.45 ± 1.22	29.61 ± 2.69
WFR-DRO	$\mathbf{0.94 \pm 0.09}$	$\mathbf{2.99 \pm 0.33}$	$\mathbf{8.80 \pm 1.09}$	$\mathbf{22.71 \pm 2.77}$

Table 2: PGD-L_∞ Robustness (Test Error %)

Method	$\Delta = 0.00$	$\Delta = 0.05$	$\Delta = 0.10$	$\Delta = 0.15$
SDRO_Dual	0.98 ± 0.10	3.09 ± 0.37	8.61 ± 1.20	21.00 ± 3.50
WRM	1.00 ± 0.10	3.82 ± 0.41	12.51 ± 1.37	31.65 ± 3.40
WGF-DRO	0.99 ± 0.12	3.60 ± 0.43	11.11 ± 1.41	28.58 ± 3.68
WFR-DRO	$\mathbf{0.94 \pm 0.09}$	$\mathbf{2.78 \pm 0.36}$	$\mathbf{7.59 \pm 0.98}$	$\mathbf{18.47 \pm 2.56}$

4.2 Human Face Classification

In this section, we apply the WFR gradient flow sampler to the 512-dimensional latent space $\mathcal{W}$ of a pre-trained ALAE (Pidhorskyi et al., 2020) model to investigate how the sampler identifies the worst-case distribution for high-dimensional image data. We employ a linear logistic regression classifier trained on the FFHQ dataset to define gender attributes in the latent space; the gradient of this classifier then serves as the potential function. The WFR sampler is configured with parameters $T = 300$, $\eta = 0.05$, $\epsilon = 0.2$, and $\tau = 5$. As shown in Figure 2a, the sampler produces continuous semantic transitions while maintaining key identity details such as pose and expression. And ss illustrated in Figure 2b, the WFR sampler effectively generates multiple plausible perturbations simultaneously.

(a) Semantic transitions in the latent space

(b) Comparison of sample diversity

Figure 2: **Performance and diversity of the WFR sampler. (a)** Semantic transitions. Shown from left to right are the original images, intermediate samples at $T = 100$ and $T = 200$, and the final result at $T = 300$. **(b)** Sample diversity comparison. The first column displays the original images, while the following three columns show the three highest-weighted samples generated at $T = 300$.

5 Conclusion

In this work, we presented a novel framework that bridges distributionally robust optimization (DRO) and sampling via gradient flows. By formulating the inner maximization of DRO as a regularized variational sampling problem, our approach serves as a general-purpose distributional SGDA. This perspective offers significant unifying power, revealing that previous methods such as WRM (Sinha et al., 2017) are fundamentally rooted in ODE analysis, which can be seen as a special case of the S/PDE framework proposed here. Furthermore, by addressing the DRO problems with a sampling methodology, our methodology effortlessly integrates advanced, non-Wasserstein samplers such as Wasserstein-Fisher-Rao (WFR) and Stein Variational Gradient (SVG) flows. Our theory treats general DRO by simply changing the corresponding gradient flow energy function F in (7), opening

the door for future research into diverse DRO ambiguity notions and gradient flow dissipation geometries.

References

Jimmy Ba, Murat A Erdogdu, Marzyeh Ghassemi, Shengyang Sun, Taiji Suzuki, Denny Wu, and Tianzong Zhang. Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations*, 2021.

Jose H Blanchet and Peter W Glynn. Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization. In *2015 Winter Simulation Conference (WSC)*, pages 3656–3667. IEEE, 2015.

José A Carrillo, Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, and Dongyi Wei. Fisher-Rao gradient flow: geodesic convexity and functional inequalities. *arXiv preprint arXiv:2407.15693*, 2024.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

Charles Dapogny, Franck Iutzeler, Andrea Meda, and Boris Thibert. Entropy-regularized Wasserstein distributionally robust shape and topology optimization. *Structural and Multidisciplinary Optimization*, 66(3):42, 2023.

Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

Rui Gao and Anton J. Kleywegt. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Nicolas García Trillos and Daniel Sanz-Alonso. Continuum limits of posteriors in graph Bayesian inverse problems. *SIAM Journal on Mathematical Analysis*, 50(4):4020–4040, 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

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

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Daniel Kuhn, Soroosh Shafiee, and Wolfram Wiesemann. Distributionally robust optimization. *Acta Numerica*, 34:579–804, 2025.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Yulong Lu, Jianfeng Lu, and James Nolen. Accelerating Langevin sampling with birth-death. *arXiv preprint arXiv:1905.09863*, 2019.

Yulong Lu, Dejan Slepčev, and Lihan Wang. Birth-death dynamics for sampling: Global convergence, approximations and their asymptotics. *Nonlinearity*, 36(11):5731–5772, November 2023. ISSN 0951-7715, 1361-6544. doi: 10.1088/1361-6544/acf988.

Alexander Mielke. An introduction to the analysis of gradients systems. *arXiv preprint arXiv:2306.05026*, 2023.

Alexander Mielke and Jia-Jie Zhu. Hellinger-Kantorovich gradient flows: Global exponential decay of entropy functionals. *arXiv preprint arXiv:2501.17049*, 2025.

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.

Felix Otto. *Double degenerate diffusion equations as steepest descent.* Sonderforschungsbereich 256, 1996.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14104–14113, 2020.

Adil Salim, Lukang Sun, and Peter Richtarik. A convergence theory for SVGD in the population limit under Talagrand's inequality T1. In *International Conference on Machine Learning*, pages 19139–19152. PMLR, 2022.

Jiaxin Shi and Lester Mackey. A finite-particle convergence rate for Stein variational gradient descent. *Advances in Neural Information Processing Systems*, 36:26831–26844, 2023.

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

Santosh Vempala and Andre Wibisono. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems*, 32, 2019.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn distributionally robust optimization. *arXiv preprint arXiv:2109.11926*, 2021.

Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, March 2018. ISSN 01676377. doi: 10.1016/j.orl.2018.01.011.

A Further Background

A.1 Gradient Systems and Their Gradient Flow Equations

The Wasserstein gradient flow (WGF) framework of (Otto, 1996) was introduced into the sampling literature to provide a theoretical foundation; see (Chewi et al., 2024; García Trillos and Sanz-Alonso, 2018) for recent surveys. In that framework, one can write a flow equation formally as

$$\dot{\rho} = -\mathbb{G}_W(\rho)^{-1}(\rho)\frac{\delta F}{\delta \rho}[\rho] = \nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho}[\rho]\right) \tag{11}$$

using the inverse of the Wasserstein Riemannian metric tensor: $\mathbb{G}_W^{-1}(\rho) : T_\rho^*\mathcal{M} \to T_\rho\mathcal{M}, \xi \mapsto -\nabla \cdot (\rho \nabla \xi)$, where $\mathcal{M}$ is a space or a manifold, $T_\rho\mathcal{M}$ is the tangent space of $\mathcal{M}$ at ρ and $T_\rho^*\mathcal{M}$ is the cotangent space. With those ingredients, we can formally define the gradient systems that generate gradient flow equations such as the WGF equation (11). We refer to Mielke (2023) for more details.

Definition 4 (Gradient system)**.** *We refer to a tuple $(\mathcal{M}, F, \mathbb{G})$ as a gradient system. It has the gradient structure identified by:*

1. *a space or a manifold $\mathcal{M}$,*
2. *an energy functional F,*
3. *a dissipation geometry given by either: a distance metric defined on $\mathcal{M}$ or a Riemannian metric tensor $\mathbb{G}$.*

A.2 Gradient Flow Sampler

For sampling and inference, a common choice for the energy functional is the KL divergence, i.e., $F(\rho) = \mathrm{KL}(\rho|\pi)$. Through elementary calculation, we obtain from (11) the Fokker-Planck equation (FPE)

$$\partial_t \rho = \nabla \cdot \left(\rho \nabla \log \frac{\rho}{\pi}\right) = \Delta \rho - \nabla \cdot (\rho \nabla \log \pi)\,. \tag{12}$$

When we express the target as $\pi(x) = \frac{1}{Z}\exp(-V(x))$ where Z is a normalization constant (partition function), (12) is then $\partial_t \rho = \Delta\rho + \nabla \cdot (\rho \nabla V)$. Viewed as a dynamical system, the KL divergence energy functional dissipates along (12) in the steepest descent manner. Based on Definition 4, we say that PDE (12) has the *gradient structure* that entails the following key ingredients:

$$\begin{cases} \text{Space :} & \text{prob. space } \mathcal{P}, \\ \text{Energy functional :} & F(\cdot) := \mathrm{KL}(\cdot|\pi), \\ \text{Dissipation Geometry :} & \text{Wasserstein metric.} \end{cases}$$

A.3 Dual approach to Entropy-regularized Wasserstein DRO

The Sinkhorn DRO problem, as formulated by Wang et al. (2021), aims to find a decision that minimizes the worst-case expected loss over an ambiguity set defined by Sinkhorn divergence, which is identical to entropy-regularized Wasserstein distance in this problem. Given a loss function $\ell(\theta, z)$, a nominal distribution $\widehat{\rho_N}$, and a radius $r > 0$, the primal inner problem is:

$$\sup_{\rho \in \mathcal{B}_{r,\epsilon}(\widehat{\rho_N})} \mathbb{E}_{z\sim\rho}[\ell(\theta, z)], \tag{13}$$

where the ambiguity set is $\mathcal{B}_{r,\epsilon}(\widehat{\rho_N}) := \{\rho \in \mathcal{P}(\mathcal{Z}) : W_\epsilon^2(\widehat{\rho_N}, \rho) \le r\}$, and the entropy-regularized Wasserstein distance,

$$W_\epsilon^2(\rho_1, \rho_2) = \inf_{\gamma \in \Gamma(\rho_1, \rho_2)} \{\mathbb{E}_{(x,y)\sim\gamma}[c(x, y)] + H(\gamma)\}.$$

A dual approach, established by Wang et al. (2021), reformulates the inner maximization problem. Using the strong duality of this problem and substituting (5) into this problem, the inner problem can be expressed as a minimization over a dual variable τ:

$$\inf_{\tau \ge 0} \left\{ \frac{r}{2\tau} + \frac{\epsilon}{2\tau}\mathbb{E}_{x\sim\widehat{\rho_N}}\left[\log \mathbb{E}_{y\sim\rho_{x,\epsilon}}\left[\exp\left(\frac{2\tau\ell(\theta, y)}{\epsilon}\right)\right]\right]\right\}, \tag{14}$$

where $\rho_{x,\epsilon}(y) = \frac{\exp(\frac{-c(x,y)}{\epsilon})}{Z}$ is a kernel distribution centered at x, and Z is a normalization constant.

To approximate the value and the gradient of the nested expectation in (14), Wang et al. (2021) employ a two-level sampling procedure for a given τ: first, a sample x is drawn from the nominal distribution $\widehat{\rho_N}$; second, a set of samples are drawn from the kernel distribution $\rho_{x,\epsilon}$ to approximate the inner expectation.

However, a major limitation of this dual method is that the nested expectation structure in the objective leads to a biased subgradient estimator when using standard Monte Carlo sampling, especially for small values of ϵ. This bias can impede the convergence of the optimization procedure. This challenge motivates the exploration of alternative approaches, such as the gradient flow methods discussed in this paper, which can directly sample from the primal worst-case distribution to obtain a less biased estimate of the gradient.

Remark 5. *Consider the KL-DRO problem:*

$$\min_{\theta} \sup_{\rho \ll \rho_0} \left\{ \mathbb{E}_{y\sim\rho}[\ell(\theta, y)] - \frac{1}{2\tau} \operatorname{KL}(\rho \| \rho_0) \right\} \tag{15}$$

Let us define a reference distribution $\rho_0(y)$ as a kernel density estimator of the empirical distribution $\widehat{\rho_N} = \frac{1}{N}\Sigma_{i=1}^N \delta(x_i)$: $\rho_0(y) = \frac{1}{N}\Sigma_{i=1}^N K_\epsilon(y, x_i)$, where $K_\epsilon(y, x_i) = \alpha_\epsilon \exp(-c(y, x_i)/\epsilon)$ and $\alpha_\epsilon = \mathbb{E}[\exp(-c(y, x_i)/\epsilon)]^{-1}$. With this choice, and by setting $\tau = \tau'/\epsilon$, the solution to the inner maximization problem is given by Hu and Hong (2013):

$$\begin{aligned}\rho^{\mathrm{KL}}(y) &= \rho_0(y) \cdot \frac{\exp(2\tau'\ell(\theta, y)/\epsilon)}{\mathbb{E}_{y\sim\rho_0}[\exp(2\tau'\ell(\theta, y)/\epsilon)]} \\ &= \frac{1}{N}\Sigma_{i=1}^N \alpha_\epsilon \beta \exp((2\tau'\ell(\theta, y) - c(y, x_i))/\epsilon),\end{aligned}$$

where $\beta = \mathbb{E}_{y\sim\rho_0}[\exp(2\tau'\ell(\theta, y)/\epsilon)]^{-1}$. This resulting distribution bears a strong resemblance to the worst-case distribution for entropy-regularized DRO. However, they are not identical, as the normalization constants differ: $\alpha_\epsilon\beta \neq \alpha_x := \mathbb{E}[\exp((2\tau'\ell(\theta, y) - c(y, x))/\epsilon]^{-1}$. In fact, the term $\alpha_\epsilon\beta \exp((2\tau'\ell(\theta, y) - c(y, x))/\epsilon)$ does not integrate to one and thus does not define a valid conditional probability measure. Actually, $\rho^{\mathrm{KL}}(y)$ can be interpreted as a weighted expectation of conditional distributions:

$$\rho^{\mathrm{KL}} = \mathbb{E}_{x\sim\widehat{\rho_N}}[\frac{\alpha_\epsilon\beta}{\alpha_x}\rho_{Y|X=x}]$$

Consequently, the two-layer sampling procedure requires computing the corresponding weights, sampling from the resulting weighted empirical distribution, and subsequently sampling from the conditional distribution. However, the term α_x is intractable in practice. Therefore, unlike the entropy-regularized DRO case, the KL-DRO problem cannot be directly framed as a two-level sampling task.

A.4 Hellinger-Kantorovich a.k.a. Wasserstein-Fisher-Rao Gradient Flows

Consider the WFR gradient system of the energy functional F, i.e., the triple $(\mathcal{P}, F, \mathsf{WFR})$, where WFR is the Wasserstein-Fisher-Rao metric, a.k.a. the *Hellinger-Kantorovich* metric restricted to the probability space. Specifically, we consider the HK/WFR gradient system associated with reaction-diffusion PDE:

$$\partial_t \mu = \alpha \operatorname{div}\left(\mu \nabla \frac{\delta F}{\delta \mu}\right) - \beta\, \mu \left(\frac{\delta F}{\delta \mu} - \int \frac{\delta F}{\delta \mu} \,\mathrm{d}\mu\right). \tag{16}$$

The hope of this gradient flow is that the added reaction term on the right-hand side will help the gradient flow to converge more rapidly. This can be formally seen by checking the time derivative of the energy functional F along the gradient flow (16), we have

$$\partial_t F(\rho_t) \overset{(\mathrm{EDB})}{=} -\alpha \int \left|\nabla \log \frac{\rho}{\pi}(x)\right|^2 \mathrm{d}\rho(x) - \beta \int \left|\log \frac{\rho}{\pi}(x) - \operatorname{KL}(\rho|\pi)\right|^2 \mathrm{d}\rho(x)$$

where we used the Energy-Dissipation Balance (EDB, a.k.a. equality) of gradient flows. We observe the right-hand side is non-positive, hence its dissipation will be faster than the original pure Wasserstein gradient flow.

However, we must note that there is no global PL or strong convexity analogous to the case of the Wasserstein gradient flow, as discussed in the main text. A "warm-start" condition is necessary as it has been shown that global PL/gradient dominance cannot hold for the Fisher-Rao/(spherical)Hellinger gradient flows; see Mielke and Zhu (2025); Carrillo et al. (2024) for details. There exist fine-grained warm-start initialization conditions in the gradient flow literature; see Lu et al. (2023, 2019); Chen et al. (2023).

For example, Lu et al. (2023) showed that, after a warm-start initialization in the form of a density-ratio lower bound, it is possible to get an improved exponential convergence rate than the pure Wasserstein gradient flow as characterized in Lemma 11. However, it is important to note that the specific warm-start initialization condition cannot be verified in our DRO problems. Nonetheless, in practice, we observe significant speedups using the WFR gradient flow as compared to the pure Wasserstein gradient flow.

B Analysis of Gradient Flows

For algorithmic design, one paradigm is to first obtain intuition and insights from principled mathematical analysis before constructing practical numerical algorithms. While the development in machine learning research does not always follow this pattern, we argue in this paper that, by first understanding the theoretical limits of the gradient flow, we can design novel DRO algorithms that performs as expected theoretically. This is done without inventing new theoretical framework, but by leveraging the existing gradient flow and PDE analysis (Otto and Villani, 2000; Jordan et al., 1998) specialized to DRO.

In DRO works such as (Sinha et al., 2017) and subsequent variants, complexity analysis typically involves controlling the gradient estimation error. That is, to control the deviation from the true gradient used for DRO. For illustration, let us consider the DRO problem (1). Note that setting $\epsilon = 0$ recovers the problem of (Sinha et al., 2017). Our goal is to analyze the estimation of DRO stochastic gradient $\hat{g}$ using PDE tools. With our gradient flow sampler-based DRO algorithm, we aim to generate samples from the worst-case distribution π_Y, which is the entropic JKO solution

$$\pi_Y = \epsilon\text{-JKO}_{\tau V}(\widehat{\rho_N}) \tag{17}$$

Following the framework of Algorithm 2, for each $x_i \sim \rho_0$, we generate samples from the worst-case

Algorithm 2 Worst-case Distribution Sampler via Gradient Flows

1: **Input:** an initial distribution ρ_0 to sample from (e.g. empirical distribution $\widehat{\rho}_N$ in data-driven DRO)
2: Sample $X \sim \rho_0$
3: Sample $Y \sim \pi_{Y|X}$ using a gradient flow (e.g. with energy functional given by (7))
4: **Output:** sample Y from the worst-case distribution

distribution by sampling from the conditional distribution, $y_i \sim \rho^*_{Y|X=x_i}$. Then, for the outer DRO problem, we can use the sample-based gradient estimate $\widehat{g} := \nabla_\theta \ell(\theta, y)$ to update the parameter θ. The following result uses gradient flow analysis to bound the gradient estimate error. We note that the results are stated in terms of the ideal continuous-time gradient flow, which is not yet the mixing time of the discrete-time sampler; we will detail the discrete-time algorithmic analysis in a later section. The gradient flow analysis provides a key insight and guideline for the design of the gradient flow sampler-based DRO algorithm.

For notational simplicity, we abbreviate $\widetilde{V}_{x,\tau}$ as $\widetilde{V}$ in this section and, specifically, for the case of quadratic transport cost $c(y, x) = \|x - y\|^2$, we define the following. Recall the functional $F(\rho) := \frac{\epsilon}{2\tau}\,\mathrm{KL}\left(\rho \,\middle|\, \frac{1}{Z_x}\exp\left[-\frac{2\tau V(y)+c(y,x)}{\epsilon}\right]\right)$ as in (7).

Definition 6. *1. **L-smoothness:** A differentiable function f is L-smooth if its gradient ∇f is L-Lipschitz continuous, i.e., there exists a constant $L > 0$ such that*

$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|, \quad \forall x, x' \in \mathcal{X}. \tag{18}$$

*2. **Semi-convexity of the regularized potential $\widetilde{V}$:** There exists a constant $\lambda > 0$ such that*

$$\nabla^2 \widetilde{V} \geq \lambda I \iff \nabla^2 V + \frac{1}{2\tau} I \geq \lambda I \tag{19}$$

*3. **Gradient dominance/Polyak-Łojaciewicz (PL) inequality of $\widetilde{V}$:** There exists a constant $\lambda > 0$ such that*

$$\left|\nabla \widetilde{V}\right|^2 \geq \lambda \widetilde{V}, \quad \forall y \in \mathcal{X} \tag{20}$$

*4. **Functional PL inequality over measures along GF**: There exists a constant $\lambda > 0$ such that*

$$\left\| \frac{\mathrm{d}}{\mathrm{d}t} F(\rho_t) \right\|^2 \geq \lambda F(\rho_t), \tag{21}$$

*which is equivalent to the λ-**logarithmic Sobolev inequality** (LSI) for $\pi = \frac{1}{Z_x} \exp\left[-\frac{2\tau V(y) + c(y,x)}{\epsilon}\right]$,*

$$\int \left\| \nabla \log \frac{\rho}{\pi} \right\|^2 \mathrm{d}\rho \geq \lambda \, \mathrm{KL}(\rho|\pi) \tag{22}$$

along the solution ρ_t.

It is immediate that 1. $\implies$ 2. and 1. $\implies$ 3. We note that the first condition in Definition 6 does not require the function V itself to be convex. Hence, the DRO loss $\ell(\theta, x)$ does not necessarily need to be concave in variable x. In such cases, we can already obtain the geodesic convexity of F in the Wasserstein space, termed displacement convexity.

Letting the DRO objective $\Phi(\theta) := \max_{\rho \in \mathcal{P}} \int \ell(\theta, z) \, \mathrm{d}\rho(z) - \frac{1}{2\tau} W_\epsilon^2(\rho, \widehat{\rho_N})$. Using standard analysis of gradient flows, we obtain the following ideal gradient estimate.

Proposition 7. *Denote the initial sample $x^i \sim \widehat{\rho_N}$ after running the Wasserstein gradient flow sampler for time t as $y_t^i \sim \rho^*_{Y|X=x^i}$. For a fixed θ, suppose that the loss function $\ell(\theta, x)$ is L-smooth in x and either $\widetilde{V}_{x,\tau}(y) := -\ell(\theta, y) + \frac{1}{2\tau} c(y, x)$ is λ-convex with $\lambda > 0$ or the functional F satisfies the λ-LSI as in Definition 6. Then, in order to generate an ϵ-gradient estimate, i.e.,*

$$\left| \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \ell(\theta, y_t^i) - \nabla_\theta \Phi(\theta) \right| \leq \epsilon,$$

the Wasserstein gradient flow sampler needs to run for time at least $t \gtrsim \mathcal{O}\left(\frac{1}{\lambda} \log \frac{L}{\sqrt{\lambda}\epsilon}\right)$.

Results such as Proposition 7 are based on the displacement convexity or PL/LSI in the Wasserstein space. The value of our general gradient flow perspective is to let us freely choose the gradient flow geometry, not confined to Wasserstein. This is not just for theoretical considerations — the estimate such as (28) depends crucially on the (PL/convexity) constant λ, which has been a major limitation of the Langevin type samplers in the literature. Note that the speed-up is due to an advantage of the Hellinger or Fisher-Rao geometry when the warm-start initialization is given. It provides a simple insight for algorithmic design: the unbalanced WFR gradient flow can be used to speed up the sampling process for DRO. This insight is later validated in the numerical studies.

As noted earlier, Proposition 7 is aimed at providing the intuition for the behavior of the gradient flow sampler-based DRO algorithm; it is not the practical complexity estimate based on sampler's

mixing time. It serves as a guideline for the design of the gradient flow sampler-based DRO algorithm using existing tools from PDE/SDE.

C Applications of GF-DRO

C.1 Wasserstein Gradient Flow-based DRO

In this subsection, we present the detailed procedure for the Wasserstein Gradient Flow-based Distributionally Robust Optimization (DRO). The specific steps are summarized in Algorithm 3.

Algorithm 3 Entropy-regularized Wasserstein DRO via WGF

Input: Empirical distribution $\widehat{\rho_N}$, constraint set Θ, $\tau, \epsilon > 0$, stepsize sequence $\{r_s > 0\}_{s=0}^{S-1}$, inner stepsize η, inner iterations T, number of samples m.
for $s = 0, \ldots, S-1$ **do**
  Sample $x^s \sim \widehat{\rho_N}$
  Initialize $y_0^{i,s} \leftarrow x^s, i = 1, \ldots, m$
  for $t = 0, \ldots, T-1$ **do**
    Sample $\xi_t^{i,s} \sim \mathcal{N}(0, I)$
    $y_{t+1}^{i,s} \leftarrow y_t^{i,s} - \eta \nabla \widetilde{V}_{x^s,\tau}(y_t^{i,s}) + \sqrt{\eta\epsilon/\tau}\xi_t^{i,s}$
  end for
  $\theta^{s+1} \leftarrow \mathrm{Proj}_\Theta(\theta^s - r_s \sum_{i=1}^m \frac{1}{m} \nabla_\theta \ell(\theta^s, y_T^{i,s}))$
end for
return θ^S

C.2 Wasserstein Fisher-Rao Gradient Flow-based DRO

This subsection details the DRO method based on the Wasserstein Fisher-Rao gradient flow. Algorithm 4 provides a comprehensive summary of this procedure. Compared to the WGF-based method, this approach further refines the DRO framework by introducing birth-death mechanism.

Algorithm 4 Entropy-regularized Wasserstein DRO via WFR flow

Input: Empirical distribution $\widehat{\rho_N}$, constraint set Θ, $\tau, \epsilon > 0$, stepsize sequence $\{r_s > 0\}_{s=0}^{S-1}$, inner stepsize η, weight stepsize η_w, weight threshold $w_{\min}$, inner iterations T, number of samples m.
for $s = 0, \ldots, S-1$ **do**
  Sample $x^s \sim \widehat{\rho_N}$
  Initialize $y_0^{i,s} \leftarrow x^s, i = 1, \ldots, m$
  for $t = 0, \ldots, T-1$ **do**
    Sample $\xi_t^{i,s} \sim \mathcal{N}(0, I)$
    $y_{t+1}^{i,s} \leftarrow y_t^{i,s} - \eta \nabla \widetilde{V}_{x^s,\tau}(y_t^{i,s}) + \sqrt{\eta\epsilon/\tau}\xi_t^{i,s}$
    $w_{t+1}^{i,s} \leftarrow (w_t^{i,s})^{1-\epsilon\eta_w/(2\tau)} e^{-\eta_w \widetilde{V}_{x^t,\tau}(y_t^{i,s})}$
    Normalize weights: $w_{t+1}^{i,s} \leftarrow w_{t+1}^{i,s} / \sum_{j=1}^m w_{t+1}^{i,s}$
    Particle death/rejuvenation:
    if $w_{t+1}^{i,s} < w_{\min}$ **then**
      Uniformly select an index i' from $\{j : j \neq i\}$.
      $y_{t+1}^{i,s} \leftarrow y_{t+1}^{i',s}, \quad w_{t+1}^{i',s}, w_{t+1}^{i,s} \leftarrow (w_{t+1}^{i,s} + w_{t+1}^{i',s})/2.$
    end if
  end for
  $\theta^{s+1} \leftarrow \mathrm{Proj}_\Theta(\theta^s - r_s \sum_{i=1}^m w_T^{i,s} \nabla_\theta \ell(\theta^s, y_T^{i,s}))$
end for
return θ^S

Algorithm 5 Sinkhorn DRO via SVGD

Input: Empirical distribution $\widehat{\rho_N}$, constraint set Θ, $\tau, \epsilon > 0$, stepsize sequence $\{r_s > 0\}_{s=0}^{S-1}$, inner stepsize η, inner iterations T, initial deviation σ, number of samples m.
for $s = 0, \ldots, S-1$ **do**
  Sample $x^s \sim \widehat{\rho_N}$
  Sample $\xi^{s,i} \sim \mathcal{N}(0, \sigma I)$, $i = 1, \ldots, m$
  Initialize $y_0^{s,i} \leftarrow x^s + \xi^{s,i}$
  for $t = 0, \ldots, T-1$ **do**
    $\phi_t^{s,i} \leftarrow \frac{1}{m} \sum_{j=1}^{m} \left[-k(y_t^{s,i}, y_t^{s,j}) \cdot \frac{2\tau}{\epsilon} \nabla \widetilde{V}_{x^s,\tau}(y_t^{s,j}) + \nabla_2 k(y_t^{s,i}, y_t^{s,j}) \right]$
    $y_{t+1}^{s,i} \leftarrow y_t^{s,i} + \eta \phi_t^{s,i}$
  end for
  $\theta^{s+1} \leftarrow \mathrm{Proj}_\Theta(\theta^s - r_s \sum_{i=1}^{m} \frac{1}{m} \nabla_\theta \ell(\theta^s, y_T^{s,i}))$
end for
return θ^S

C.3 Stein Variational Gradient-based DRO

Stein Variational Gradient Descent (SVGD) is a variational inference algorithm that approximates a target probability density $p(x)$ by iteratively transporting a set of particles $\{x_i\}_{i=1}^n$ (Liu and Wang, 2016). The method is formulated as a functional gradient descent on the Kullback-Leibler (KL) divergence, $\mathrm{KL}(q||p)$, where q represents the empirical distribution of the particles. The particle updates are governed by a velocity field $\phi(x_i)$ that corresponds to the direction of steepest descent. For an empirical measure of n particles, this update is given by:

$$\boldsymbol{\phi}(x_i) \propto \frac{1}{n} \sum_{j=1}^{n} \left[k(x_j, x_i) \nabla_{x_j} \log p(x_j) + \nabla_{x_j} k(x_j, x_i) \right]$$

This update rule can be decomposed into two functional components: (i) a driving force, which is a kernel-smoothed average of the score function $\nabla \log p(x)$ that directs particles towards the modes of the target distribution, and (ii) a repulsive force, which arises from the kernel gradient $\nabla k(\cdot, \cdot)$ and ensures particle diversity to prevent mode collapse (Ba et al., 2021).

The iterative application of this update rule forms the basis of the SVGD algorithm. As detailed in (Liu and Wang, 2016), its computational complexity is dominated by the pairwise kernel computations, resulting in a cost of $O(m^2 d)$ per iteration for m particles in d dimensions. Regarding convergence, theoretical guarantees have been established under certain assumptions. As shown by (Salim et al., 2022), for target distributions that satisfy Talagrand's T1 inequality, a finite-iteration complexity bound is derived. A finite-particle convergence analysis is also provided in (Shi and Mackey, 2023).

By applying SVGD to sample from the worst-case distribution, we arrive at Algorithm 5.

D Proofs of Theoretical Results

D.1 Assumptions

Assumption 8.

1. ***Smoothness of the objective:*** *There exists a constant $L_\Phi > 0$ such that*
$$\|\nabla \Phi(\theta_1) - \nabla \Phi(\theta_2)\| \leq L_\Phi \|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \Theta$$

2. ***Lipschitz continuity of gradients:*** *There exists a constant $L_f > 0$ such that*
$$\|\nabla_\theta \ell(\theta_1, z) - \nabla_\theta \ell(\theta_2, z)\| \leq L_f \|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \Theta$$

3. ***Sampler error and bounded variance:*** *The inner-loop sampler (Algorithm 2) generates a distribution $\widehat{\rho}_{Y|X=x}$, and $\mathbb{E}_{x \sim \widehat{\rho_N}}[W_2(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x})] \leq \delta_{sample}$. The stochastic gradient estimator $\widehat{g}$ has bounded variance σ^2.*

Assumption 9. *For any fixed θ and x, the conditional distribution $\rho_{Y|X=x}$ is L_U-smooth and satisfies the λ_U-log-Sobolev inequality (LSI).*

D.2 Proof of Proposition 1

Proof of Proposition 1. We rewrite the problem using the definition of the entropy-regularized OT formulation:

$$\min_{\rho} \min_{\Pi} \left\{ \int V \,\mathrm{d}\rho + \frac{1}{2\tau} \int c(x,y) \,\mathrm{d}\Pi + \frac{\epsilon}{2\tau} \int \log \Pi \,\mathrm{d}\Pi \right\} \tag{23}$$

subject to the margianl constraints $\int \Pi \,\mathrm{d}x = \rho$ and $\int \Pi \,\mathrm{d}y = \rho_0$. Here, ρ_0 can be set to empirical distribution $\widehat{\rho_N}$ as in the data-driven DRO problem. Using the marginal constraints, we can then write the first linear term as $\int \int V(y) \,\mathrm{d}\Pi(x,y)$ and eliminate the variable ρ, i.e., one end is unconstrained – hence it's the half bridge problem:

$$\min_{\Pi} \left\{ \int V \,\mathrm{d}\Pi + \frac{1}{2\tau} \int c(x,y) \,\mathrm{d}\Pi + \frac{\epsilon}{2\tau} \int \log \Pi \,\mathrm{d}\Pi \;\middle|\; \int \Pi \,\mathrm{d}y = \rho_0 \right\}. \tag{24}$$

After straightforward linear optimization, we obtain the optimal solution to the variational problem (3) as

$$\rho^* \propto \mathbb{E}_{x \sim \rho_0} \left[\exp \left[-\frac{2\tau V(y) + c(y, x_k^i)}{\epsilon} \right] \right]. \tag{25}$$

Using elementary manipulations, the variational problem (3) can also be written down as the minimization of the expected KL divergence:

$$\min_{\rho_{Y|X}} \mathbb{E}_{x \sim \rho_0} \mathrm{KL} \left(\rho_{Y|X=x}(y) \middle| \frac{1}{Z_x} \exp \left[-\frac{2\tau V(y) + c(y,x)}{\epsilon} \right] \right). \tag{26}$$

□

D.3 Proof of Proposition 7

First, we recall some standard results.

Lemma 10. *Suppose the first condition in Definition 6 holds for some $\lambda > 0$. Then, F is λ-displacement convex.*

Lemma 11 (WGF gradient dominance/PL)**.** *Let ρ_t be a solution to the Wasserstein gradient flow of the KL divergence energy functional $\mathrm{KL}(\rho | \exp(-\widetilde{V}_{x,\tau}))$. Then,*

$$\widetilde{V} \text{ is } \lambda\text{-displacement convex or } \lambda\text{-PL} \implies \mathrm{KL}(\rho_t | \exp(-\widetilde{V}_{x,\tau})) \leq \mathrm{e}^{-2\lambda t} \mathrm{KL}(\rho_0 | \exp(-\widetilde{V}_{x,\tau})).$$

Lemma 12 (Talagrand's inequality)**.** *Under the same assumptions as in Lemma 11, we have*

$$W_2^2(\rho_t, \exp(-\widetilde{V}_{x,\tau})) \leq \sqrt{\frac{2}{\lambda} \mathrm{KL}(\rho_t | \exp(-\widetilde{V}_{x,\tau}))}$$

Recall also that $W_1(\rho_t, \pi_{Y|X=x}) \leq W_2(\rho_t, \pi_{Y|X=x})$.

From conditional to marginal bound We have the following relation between the marginal and conditional KL divergences. Suppose ρ_X is the marginal distribution of X. Recall the relation $\pi_Y = \int_x \rho^*_{Y|X=x} \,\mathrm{d}\rho_X(x)$. Then,

$$\mathrm{KL}(\rho_Y | \pi_Y) \leq \mathbb{E}_{X \sim \rho_X} \left[\mathrm{KL} \left(\rho_{t,Y|X} \middle| \rho^*_{Y|X} \right) \right].$$

Alternatively, the Wasserstein relation gives

$$W_p(\rho_Y, \pi_Y) \leq \mathbb{E}_{X \sim \rho_X} \left[W_p(\rho_{Y|X}(\cdot|X), \rho^*_{Y|X}(\cdot|X)) \right].$$

Marginal distribution error control Suppose we have the samples $y_t^i \sim \rho_t$, Our goal is to use the empirical measure $\widehat{\rho_Y} = \frac{1}{N}\sum_{i=1}^{N} \delta_{y_t^i}$ to approximate π_Y. An elementary manipulation shows that

$$W_1(\widehat{\rho_Y}, \pi_Y) \leq \mathbb{E}_{X\sim\rho_X}\left[W_1(\widehat{\rho}_{Y|X}, \rho^*_{Y|X})\right] \lesssim \frac{\mathrm{e}^{-\lambda t}}{\sqrt{\lambda}} \tag{27}$$

Therefore, to reach an ϵ-solution, we need to run the gradient flow (i.e. simulating the PDE/SDE) for time

$$t \gtrsim \frac{1}{\lambda}\log\frac{1}{\sqrt{\lambda}\epsilon}. \tag{28}$$

We will later show the discrete-time gradient descent version of this result.

We emphasize that the continuous-time gradient flow result is important not only for understanding the limit of discrete-time algorithms, but also for serving as a guide for designing new discrete-time algorithms. Let us get a quick insight from the above estimate: suppose we discretized algorithm with discretization stepsize η for in total T steps. Ignoring the discretization error, we have $t = T\eta$ and the (ideal) number of iterations needed to reach an ϵ-solution of the inner maximization problem (17) is $T \gtrsim \frac{1}{\eta\lambda}\log\frac{1}{\sqrt{\lambda}\epsilon}$.

DRO gradient oracle error control via gradient flow Let us now show how to convert the error estimate in distributions to the error estimate in the gradient oracle, needed for the optimization analysis later. If $W_1(\widehat{\rho_Y}, \pi_Y) \leq \delta$ for some $\delta > 0$ (as controlled by (27)) and g is L-Lipschitz (w.r.t. Euclidean norm). Then, taking expectations and using the optimal transport plan (coupling) γ of $(X, Y) \sim (\widehat{\rho_Y}, \pi_Y)$, we obtain:

$$\left|\int g\, d(\widehat{\rho_Y} - \pi_Y)\right| = \left|\int g(x) - g(y)\,\mathrm{d}\gamma(x,y)\right| \leq LW_1(\widehat{\rho_Y}, \pi_Y) \leq L\delta. \tag{29}$$

In the optimization analysis, we can later take g to be the gradient oracle of the DRO loss, $g := \nabla_\theta \ell(\theta, \cdot)$, to control the gradient oracle error. Recall that $\widehat{\rho_Y} = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_t^i}$ is the particle approximation obtained by running the gradient flow sampler for t steps, and π_Y is the target marginal distribution, which is the worst-case distribution of DRO inner maximization problem (17). Then, (29) results in the following bound on the gradient oracle error:

$$\begin{aligned}\Bigg|\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \ell(\theta, y_t^i) - \nabla_\theta \underbrace{\max_{\rho\in\mathcal{P}}\left(\int \ell(\theta, z)\,\mathrm{d}\rho(z) - \frac{1}{2\tau}W_\epsilon^2(\rho, \widehat{\rho_N})\right)}_{\text{DRO objective}}\Bigg| \\ = \left|\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta \ell(\theta, y_t^i) - \mathbb{E}_{y\sim\pi_Y}\nabla_\theta \ell(\theta, y)\right| \overset{(29)}{\leq} L\delta.\end{aligned} \tag{30}$$

Replacing $L\delta$ by ϵ, we obtain the desired bound on the gradient oracle error as in Proposition 7.

D.4 Proof of Theorem 2 (Outer Loop Convergence)

We provide a detailed proof for the convergence of the outer loop, which follows the standard analysis for non-convex Stochastic Gradient Descent with a persistent bias.

Lemma 13 (Bias of the Stochastic Gradient)**.** *Under Assumptions 8.**??**, **??**, the bias of the stochastic gradient estimator, defined as $B(\theta) := \mathbb{E}[\widehat{g}] - \nabla\Phi(\theta)$, is bounded as:*

$$\|B(\theta)\| \leq L_f \cdot \delta_{sample}.$$

Proof. By definition, the bias is

$$\begin{aligned} B(\theta) &= \mathbb{E}_{x\sim\widehat{\rho_N}}\left[\mathbb{E}_{y\sim\widehat{\rho}_{Y|X=x}}[\nabla_\theta \ell(\theta,y)]\right] - \mathbb{E}_{x\sim\widehat{\rho_N}}\left[\mathbb{E}_{y\sim\pi_{Y|X=x}}[\nabla_\theta \ell(\theta,y)]\right] \\ &= \mathbb{E}_{x\sim\widehat{\rho_N}}\left[\mathbb{E}_{y\sim\widehat{\rho}_{Y|X=x}}[\nabla_\theta \ell(\theta,y)] - \mathbb{E}_{y\sim\pi_{Y|X=x}}[\nabla_\theta \ell(\theta,y)]\right]. \end{aligned}$$

Taking norms and using Jensen's inequality:

$$\|B(\theta)\| \le \mathbb{E}_{x\sim\widehat{\rho_N}}\left[\|\mathbb{E}_{y\sim\widehat{\rho}_{Y|X=x}}[\nabla_\theta \ell(\theta,y)] - \mathbb{E}_{y\sim\pi_{Y|X=x}}[\nabla_\theta \ell(\theta,y)]\|\right].$$

For any two probability measures ρ_1, ρ_2 and a function h that is L_h-Lipschitz, we know that $\|\mathbb{E}_{\rho_1}[h(y)] - \mathbb{E}_{\rho_2}[h(y)]\| \le L_h W_1(\rho_1,\rho_2)$. Applying this with $h(y) = \nabla_\theta \ell(\theta,y)$, which is L_f-Lipschitz in y by Assumption 2, gives:

$$\begin{aligned} \|B(\theta)\| &\le \mathbb{E}_{x\sim\widehat{\rho_N}}\left[L_f \cdot W_1(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x})\right] \\ &\le L_f \cdot \mathbb{E}_{x\sim\widehat{\rho_N}}\left[W_2(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x})\right] \quad (\text{since } W_1 \le W_2) \\ &\le L_f \cdot \delta_{\text{sample}}. \end{aligned}$$

□

Proof of Theorem 2. From the L_Φ-smoothness of Φ (Assumption 8.**??**), we have the standard descent lemma. For simplicity, we assume $\Theta = \mathbb{R}^d$ and a constant stepsize $r_s = r$.

$$\begin{aligned} \Phi(\theta^{s+1}) &\le \Phi(\theta^s) + \langle \nabla\Phi(\theta^s), \theta^{s+1} - \theta^s\rangle + \frac{L_\Phi}{2}\|\theta^{s+1} - \theta^s\|^2 \\ &= \Phi(\theta^s) - r\langle \nabla\Phi(\theta^s), \widehat{g}^s\rangle + \frac{L_\Phi r^2}{2}\|\widehat{g}^s\|^2. \end{aligned}$$

Taking expectation conditioned on the filtration $\mathcal{F}_s$:

$$\begin{aligned} \mathbb{E}[\Phi(\theta^{s+1})|\mathcal{F}_s] &\le \Phi(\theta^s) - r\langle \nabla\Phi(\theta^s), \mathbb{E}[\widehat{g}^s|\mathcal{F}_s]\rangle + \frac{L_\Phi r^2}{2}\mathbb{E}[\|\widehat{g}^s\|^2|\mathcal{F}_s] \\ &= \Phi(\theta^s) - r\langle \nabla\Phi(\theta^s), \nabla\Phi(\theta^s) + B(\theta^s)\rangle + \frac{L_\Phi r^2}{2}\left(\|\mathbb{E}[\widehat{g}^s|\mathcal{F}_s]\|^2 + \mathrm{Var}(\widehat{g}^s|\mathcal{F}_s)\right). \end{aligned}$$

Using the variance bound and the bias definition:

$$\mathbb{E}[\Phi(\theta^{s+1})|\mathcal{F}_s] \le \Phi(\theta^s) - r\|\nabla\Phi(\theta^s)\|^2 - r\langle \nabla\Phi(\theta^s), B(\theta^s)\rangle + \frac{L_\Phi r^2}{2}\left(\|\nabla\Phi(\theta^s) + B(\theta^s)\|^2 + \sigma^2\right).$$

Applying Young's inequality to the inner product term:

$$-r\langle \nabla\Phi(\theta^s), B(\theta^s)\rangle \le \frac{r}{2}\|\nabla\Phi(\theta^s)\|^2 + \frac{r}{2}\|B(\theta^t)\|^2.$$

Substituting this in and simplifying with $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$:

$$\mathbb{E}[\Phi(\theta^{s+1})|\mathcal{F}_s] \le \Phi(\theta^s) - \frac{r}{2}\|\nabla\Phi(\theta^s)\|^2 + \frac{r}{2}\|B(\theta^s)\|^2 + L_\Phi r^2(\|\nabla\Phi(\theta^s)\|^2 + \|B(\theta^s)\|^2) + \frac{L_\Phi r^2\sigma^2}{2}.$$

Rearranging terms to isolate $\|\nabla\Phi(\theta^s)\|^2$:

$$r\left(\frac{1}{2} - L_\Phi r\right)\|\nabla\Phi(\theta^s)\|^2 \le \Phi(\theta^s) - \mathbb{E}[\Phi(\theta^{s+1})|\mathcal{F}_s] + \left(\frac{r}{2} + L_\Phi r^2\right)\|B(\theta^s)\|^2 + \frac{L_\Phi r^2\sigma^2}{2}.$$

Choosing a stepsize $r \le \frac{1}{4L_\Phi}$, we have $(\frac{1}{2} - L_\Phi r) \ge \frac{1}{4}$ and $(\frac{r}{2} + L_\Phi r^2) \le r$. This gives:

$$\frac{r}{4}\|\nabla\Phi(\theta^s)\|^2 \le \Phi(\theta^s) - \mathbb{E}[\Phi(\theta^{s+1})|\mathcal{F}_s] + r\|B(\theta^s)\|^2 + \frac{L_\Phi r^2\sigma^2}{2}.$$

Taking total expectation and using Lemma 1, $\|B(\theta^s)\|^2 \le (L_f\delta_{\text{sample}})^2$:

$$\frac{r}{4}\mathbb{E}[\|\nabla\Phi(\theta^s)\|^2] \le \mathbb{E}[\Phi(\theta^s)] - \mathbb{E}[\Phi(\theta^{s+1})] + r(L_f\delta_{\text{sample}})^2 + \frac{L_\Phi r^2\sigma^2}{2}.$$

Summing from $s = 0$ to $S-1$ yields a telescoping sum:

$$\frac{r}{4}\sum_{s=0}^{S-1}\mathbb{E}[\|\nabla\Phi(\theta^s)\|^2] \le \Phi(\theta^0) - \mathbb{E}[\Phi(\theta^S)] + Sr(L_f\delta_{\text{sample}})^2 + \frac{SL_\Phi r^2\sigma^2}{2}.$$

Dividing by $\frac{rS}{4}$ and using $\mathbb{E}[\Phi(\theta^S)] \geq \Phi_{\text{inf}}$:

$$\frac{1}{S}\sum_{s=0}^{S-1}\mathbb{E}[|\nabla\Phi(\theta^s)|^2] \leq \frac{4(\Phi(\theta^0)-\Phi_{\text{inf}})}{rS} + 4(L_f\delta_{\text{sample}})^2 + 2L_\Phi r\sigma^2.$$

Therefore, when δ_{sample} is controlled such that $\delta_{\text{sample}} = O(\epsilon_{\text{opt}}/L_f)$, and step size $r = \sqrt{\frac{1}{SL_\Phi\sigma^2}}$, and the iteration $S = O(1/\epsilon_{\text{opt}}^4)$, $\frac{1}{S}\sum_{s=0}^{S-1}\mathbb{E}[|\nabla\Phi(\theta^s)|^2] \leq O(\epsilon_{\text{opt}}^2)$.

D.5 Derivation of ULA Sampler Complexity (for Theorem 3)

The complexity of the ULA sampler depends on the geometric properties of the inner target distribution $\rho_{Y|X=x}$. Under Assumption 9, standard results on ULA convergence (Vempala and Wibisono, 2019) state that after T steps, the KL-divergence to the target is bounded. To achieve a final KL-divergence of δ_{KL}, the number of iterations required is $T = O\left(\frac{L_U^2 d}{\lambda_U^2\delta_{KL}}\log\frac{1}{\delta_{KL}}\right) = \tilde{O}\left(\frac{L_U^2 d}{\lambda_U^2\delta_{KL}}\right)$.

Our goal is to connect the required outer-loop sampling accuracy δ_{sample} (in W_2 distance) to the required inner-loop KL-divergence accuracy δ_{KL}. Under the LSI assumption, Talagrand's inequality gives the relationship:

$$W_2^2(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x}) \leq \frac{2}{\lambda_U}\,\text{KL}(\widehat{\rho}_{Y|X=x}||\pi_{Y|X=x})$$

From Theorem 2, we require $W_2(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x}) \leq \delta_{\text{sample}} = O(\epsilon_{\text{opt}}/L_f)$. This implies we need to achieve a KL-divergence of:

$$\delta_{KL} \leq \frac{\lambda_U}{2}W_2^2(\widehat{\rho}_{Y|X=x}, \pi_{Y|X=x}) = O\left(\lambda_U\frac{\epsilon_{\text{opt}}^2}{L_f^2}\right).$$

Substituting this required δ_{KL} into the ULA iteration complexity gives the number of inner loop steps:

$$T_{ULA} = \tilde{O}\left(\frac{L_U^2 d}{\lambda_U^2\cdot\lambda_U\frac{\epsilon_{\text{opt}}^2}{L_f^2}}\right) = \tilde{O}\left(\frac{L_U^2 L_f^2 d}{\lambda_U^3\epsilon_{\text{opt}}^2}\right).$$

The total complexity is the product of outer iterations $S = O(1/\epsilon_{\text{opt}}^4)$, inner iterations T_{ULA}, and the cost per inner gradient step $C_{\nabla_z} = O(d)$.

$$\begin{aligned}\text{Complexity} &= S\times T\times C_{\nabla_z}\\ &= O\left(\frac{1}{\epsilon_{\text{opt}}^4}\right)\times\tilde{O}\left(\frac{L_U^2L_f^2d}{\lambda_U^3\epsilon_{\text{opt}}^2}\right)\times O(d)\\ &= \tilde{O}\left(\frac{L_U^2L_f^2d^2}{\lambda_U^3\epsilon_{\text{opt}}^6}\right).\end{aligned}$$

Note that we absorb constants like L_Φ into the $\tilde{O}(\cdot)$ notation. This completes the derivation for Theorem 3.

E Additional Experiments

E.1 Ablation Study for Human Face Classification

In this section, we evaluate the sensitivity of the proposed WFR gradient flow sampler to its key hyperparameters. To ensure a controlled comparison, we vary one parameter at a time while keeping the others fixed at their baseline values. The baseline configuration is defined as follows: regularization $\tau = 5$, entropy regularization $\epsilon = 0.1$, learning rate $\eta = 0.01$, number of particles $m = 32$, and total iterations $T = 10,000$. All qualitative comparisons are conducted using a fixed representative sample from the FFHQ test set to ensure consistency.

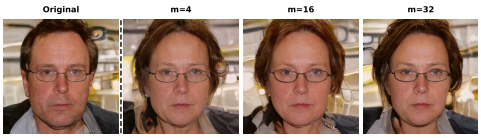
Figure 3 presents the qualitative results of varying the regularization strength, learning rate, and the number of particles. As shown in Figure 3a, a small regularization value ($\tau = 0.5$) overly constrains the latent code to the source distribution, causing the model to retain original features such as masculine traits. The baseline value of $\tau = 5$ achieves a successful attribute transition while effectively preserving the identity of the source image. However, when τ is excessively large, the sampler fails to maintain the underlying facial structure, leading to unrealistic results.

The learning rate η and the number of particles m also play critical roles in the sampling process. As observed in Figure 3b, a smaller learning rate can mitigate artifacts and distortions in the generated images, providing more stable updates. Regarding the number of particles, Figure 3c demonstrates that increasing m consistently improves the quality and stability of the generated samples by providing a better approximation of the gradient flow.

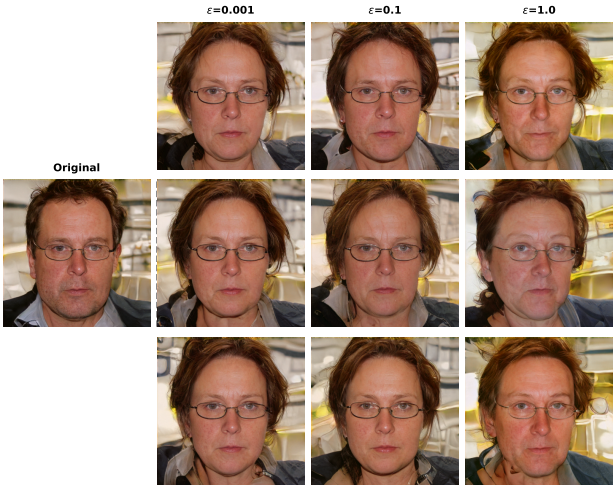
We further analyze the entropy regularization ϵ, which governs the stochasticity and diversity of the generated images. As illustrated in the multi-sample comparison in Figure 3d, low entropy leads to nearly deterministic trajectories with minimal variation between samples. At the intermediate baseline of $\epsilon = 0.1$, the sampler produces diverse yet realistic adversarial candidates. Conversely, increasing ϵ to 1.0 introduces excessive noise that dominates the drift term, resulting in significant image degradation and loss of semantic content.

(a) Effect of Regularization τ

(b) Effect of learning rate η



(c) Effect of the number of particles m



(d) Effect of Entropy Regularization ϵ on Diversity

Figure 3: Qualitative ablation study on the WFR sampler parameters. (a-c) show the sensitivity of single-sample generation to τ, η, and m. (d) visualizes the Top-3 candidates for each ϵ to demonstrate sample diversity. The leftmost column in each figure represents the original source image.

Finally, we visualize the temporal evolution of the latent code in Figure 4. The progression from iteration 250 to 1,000 demonstrates a stable and smooth transition toward the target attribute, confirming the ability of generating samples from the worst-case distribution of the proposed method.

Figure 4: Visualization of the convergence process. The sampler demonstrates a gradual and stable transition towards the target attribute.