Information and Inference: A Journal of the IMA (2025) **14**, iaaf028
https://doi.org/10.1093/imaiai/iaaf028

Robust generative learning with Lipschitz-regularized α-divergences allows minimal assumptions on target distributions

Ziyu Chen*, Hyemin Gu, Markos A. Katsoulakis, Luc Rey-Bellet
Department of Mathematics and Statistics, University of Massachusetts Amherst, 710 N Pleasant Street, Amherst, MA 01002, USA
*Corresponding author. Email: ziyuchen@unc.edu

AND

Wei Zhu
School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332, USA

[Received on 23 November 2024; revised on 31 May 2025; accepted on 4 September 2025]

This paper demonstrates the robustness of Lipschitz-regularized α-divergences as objective functionals in generative modeling, showing they enable stable learning across a wide range of target distributions with minimal assumptions. We establish that these divergences remain finite under a mild condition—that the source distribution has a finite first moment—regardless of the properties of the target distribution, making them adaptable to the structure of target distributions. Furthermore, we prove the existence and finiteness of their variational derivatives, which are essential for stable training of generative models such as generative adversarial networks and gradient flows. For heavy-tailed targets, we derive necessary and sufficient conditions that connect data dimension, α and tail behaviour to divergence finiteness, that also provide insights into the selection of suitable α's. We also provide the first sample complexity bounds for empirical estimations of these divergences on unbounded domains. As a byproduct, we obtain the first sample complexity bounds for empirical estimations of these divergences and the Wasserstein-1 metric with group symmetry on unbounded domains. Numerical experiments confirm that generative models leveraging Lipschitz-regularized α-divergences can stably learn distributions in various challenging scenarios, including those with heavy tails or complex, low-dimensional, or fractal support, all without any prior knowledge of the structure of target distributions.

Keywords: probability divergences; Lipschitz regularization; generative modeling; heavy tails; manifolds; attractors.

1. Introduction

In generative modeling, the goal is to create new samples that resemble those from an unknown data distribution by designing algorithms that minimize a probability divergence or metric between the generated distribution and the target distribution. However, the diverse characteristics of real-world data distributions—such as heavy tails, low-dimensional structures, manifold constraints, or fractal-like supports—introduce significant challenges in the training of generative models. These challenges are manifested as instabilities, reduced robustness and a need for specialized architectures, as standard generative frameworks struggle to adapt to complex data structures. Addressing these issues is essential for developing models that are not only accurate but also robust across a wide range of scenarios for the target distribution.

Features such as heavy-tailed distributions arise in various fields, including extreme events in ocean waves [16], floods [35], social sciences [27, 43], human activities [29, 55], biology [30] and computer science [46]. Learning to generate heavy-tailed distributions has been explored with generative adversarial networks (GANs). However, GANs based on integral probability metrics (IPMs), such as the Wasserstein-1 metric, may struggle to learn these distributions without additional tail estimation strategies [1, 17, 23]. This limitation arises because the Wasserstein-1 metric between two distributions becomes infinite when one lacks a finite first moment, and accurately estimating tail behaviour often requires extensive data from that tail, which may be difficult to obtain. Consequently, capturing discrepancies between distributions with a metric that remains finite, is stable to compute, and is less sensitive to the need for extensive tail data is essential for stable and effective learning.

On the other hand, many empirical results suggest that real-world data, such as images, exhibit low-dimensional structures [44]. While there are theoretical guarantees for GANs to learn distributions with low-dimensional support [22, 28], recent works on flow-based models, such as continuous normalizing flows (CNFs), neural ordinary differential equations (ODEs) and score-based diffusion models, often rely on density assumptions [9, 32]. These models can struggle to learn low-dimensional structures without additional regularization or specific architectures, such as autoencoders (see Section 7). This limitation arises because their performance is typically evaluated using the Kullback–Leibler (KL) or f-divergences, which require absolute continuity between probability measures. Thus, it is crucial to select a divergence that remains flexible and inherently compatible with the structure of the data distribution.

In this work, we demonstrate that the Lipschitz-regularized α-divergence, as proposed in [4, 15], is a suitable objective functional for generative modeling with minimal assumptions on the target distribution, denoted by Q from now on. First, we revisit the definition of the Lipschitz-regularized α-divergence between two distributions P and Q defined as:

$$D_\alpha^L(P\|Q) := \sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)} \left\{\mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right\}, \tag{1.1}$$

where $\mathrm{Lip}_L(\mathbb{R}^d)$ is the class of L-Lipschitz functions on $\mathbb{R}^d$; see more details in Section 3. In particular, we show that the Lipschitz-regularized α-divergences are suitable for stably learning a broad range of distributions from three perspectives:

- **Finiteness.** The objective of generative modeling using (1.1) can be formulated as $\min_\theta D_\alpha^L(P_\theta\|Q)$, where P_θ is the generated distribution parametrized by θ and Q is the target distribution. Thus, the divergence needs to be finite. On the contrary, an infinite or large divergence value can be an indicator of the divergence of an algorithm (see Table G1 in Appendix G). We prove that these divergences remain finite whenever the generated distribution has a finite first moment, with no assumptions necessary on the target distribution Q. When both distributions have power-law-decay densities, we provide sufficient and necessary conditions for the divergences to be finite. Notably, the Lipschitz-regularized KL divergences require minimal assumptions on both the tails of the generated and the target distributions.
- **Existence of variational derivatives.** To find the optimal parameter θ in the optimization $\min_\theta D_\alpha^L(P_\theta\|Q)$, one often uses gradient-based algorithms. Formally, the gradient of $D_\alpha^L(P_\theta\|Q)$ in terms of θ can be evaluated as

$$\nabla_\theta D_\alpha^L(P_\theta\|Q) = \int \frac{\delta D_\alpha^L(P\|Q)}{\delta P}(P_\theta(x))\cdot\nabla_\theta P_\theta(x)\,\mathrm{d}x, \tag{1.2}$$

therefore it is essential that the variational derivative $\frac{\delta D_\alpha^L(P\|Q)}{\delta P}$ is well-defined. We prove that these divergences have well-defined variational derivatives for any target distribution Q, given P has a finite first moment. This is a crucial property for stable optimizations in generative learning and the associated gradient flows, and it illustrates that algorithms using this class of divergences can stably learn distributions without extensive prior knowledge of the tail behaviour or density formulation of the target. In contrast, those using divergences without Lipschitz regularization generally can fail to learn (see Section 7).

- **Convergence of empirical estimations.** As distributions are only accessible through their finite samples, it is important to know how fast the divergence between their empirical measures converges to the true value of the divergence. We prove the first result of empirical estimations of this class of divergences on $\mathbb{R}^d$, and as a byproduct of the proof, we offer the first sample complexity bounds for empirical estimations of the Lipschitz-regularized α-divergences and the Wasserstein-1 metric with group invariance on $\mathbb{R}^d$ with sub-Weibull assumptions. The key to these results is the Lipschitz regularization, without which we cannot prove such bounds.

The rest of the paper is organized as follows. We review and discuss some related work in Section 2. Section 3 provides background and motivation for the proposed divergences. Finiteness results, including the variational derivatives and their gradient flow for the Lipschitz-regularized α-divergences are presented in Section 4. Section 5 provides the first convergence rate for finite-sample estimations of these divergences in $\mathbb{R}^d$. Based on the results and proofs from Section 5, in Section 6, we provide the first sample complexity bounds for empirical estimations of the Lipschitz-regularized α-divergences and the Wasserstein-1 metric with group symmetry in $\mathbb{R}^d$. Numerical experiments are detailed in Section 7 including synthetic heavy-tailed distributions, distributions on a low-dimensional manifold, real keystroke data and trajectories from the attractor of the Lorentz system, which is known to exhibit fractal properties. Finally, we conclude this paper in Section 8.

2. Related work

Generative models for heavy-tailed distributions. Although heavy-tailed distributions are common, there are few results to date in their generative modeling, primarily using GANs. For example, [53] generates heavy-tailed financial time series data by logarithmically transforming the data and then exponentiating the output, which produces distributions whose tails follow lognormal asymptotic rather than distributions with power-law tails considered in our paper. In a different approach, GANs are used for cosmological analysis [17], sharing a similarity with Pareto GANs [23] in their use of a heavy-tailed latent variable. However, both papers require accurate estimations of the tail decay rate for each marginal distribution. Exterme-Value (EV)-GANs [1] use neural network approximations of the quantile function to encode the tail decay rate in an asymptotic sense, which is essentially also a tail estimation approach. We note that the focus of our work is to devise appropriate divergences as objective functionals for comparing and learning heavy-tailed distributions stably, *without* prior knowledge of the tail behaviour.

Generative models for distributions with low-dimensional structures. In [22, 28] it is rigorously shown that IPM-GANs are able to learn distributions with low-dimensional support. There are some other generative models that learn high-dimensional distributions from the low-dimensional latent space provided by auto-encoders [33, 51], such as Bidirectional GANs [14], Variational Auto-Encoders [26] and Generalized Denoising Auto-Encoders [3]. However, it is not clear if the low-dimensional latent space matches the low-dimensional structure of the data distribution and no convergence guarantees have been provided, and these results are largely empirical.

Empirical estimations of divergences. [37, 45, 48] estimate f-divergences using various assumptions and estimators, and [13] considers in particular the α-divergences. However, these studies either make additional structural assumptions or consider light tails or without establishing a convergence rate of the estimation. Recently, [31, 34] studied the convergence rate of entropic optimal transport (OT) and OT with smooth costs. While our proof of the convergence rate of the empirical estimations of the Lipschitz-regularized α-divergences is inspired by these works, the structure inherited from the α-divergences in our study requires different, non-trivial treatment due to the nonlinear and asymmetric variational form, particularly as we consider even heavier tails. When the distributions are invariant to some group actions, [10] shows that empirical estimations of the Lipschitz-regularized α-divergences and the Wasserstein-1 metric enjoy a faster convergence using symmetry-informed estimators on bounded domains of $\mathbb{R}^d$, and later [49] extends the result to closed Riemannian manifolds with group symmetry only for Sobolev-IPMs that are symmetric.

Lipschitz-regularized divergences. The class of Lipschitz-regularized f-divergences was first proposed in [15] in the context of Lipschitz-regularized KL-divergences with its first variation formula, under the assumptions that both the source and the target distributions have finite first moments. Later, [4] generalized it to the class of Lipschitz-regularized f-divergences and observed that GANs optimizing Lipschitz-regularized f-divergences outperform those optimizing either the Wasserstein-1 metric or the f-divergences in learning heavy-tailed distributions. In [20], under the assumption that Q has a finite first moment, the gradient flows of the Lipschitz-regularized α-divergences were introduced, using the variational derivatives to define a corresponding generative particle algorithm (GPA), outperforming other generative models in scarce and high-dimensional data regimes. In this paper, we provide the first theoretical explanations, not only for learning heavy-tailed distributions but also for learning distributions with manifold or fractal support, essentially making the generative modeling agnostic to the target data assumptions.

3. Background

Let $\mathscr{P}(\mathbb{R}^d)$ be the space of probability measures on $\mathbb{R}^d$. A map $D : \mathscr{P}(\mathbb{R}^d) \times \mathscr{P}(\mathbb{R}^d) \to [0, \infty]$ is called a *divergence* on $\mathscr{P}(\mathbb{R}^d)$ if

$$D(P, Q) = 0 \iff P = Q \in \mathscr{P}(\mathbb{R}^d), \tag{3.1}$$

hence providing a notion of 'distance' between probability measures. In particular, the class of α-divergences [2, 21], denoted by D_α, which is a sub-class of f-divergences [11], is defined as

$$D_\alpha(P\|Q) := \int_{\mathbb{R}^d} f_\alpha\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q, \quad \text{if } P \ll Q, \tag{3.2}$$

where $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$, with $\alpha > 0$ and $\alpha \neq 1$, and $P \ll Q$ means P is absolutely continuous with respect to Q. When P is not absolutely continuous with respect to Q, we write $D_\alpha(P\|Q) = \infty$.

Remark 1. Note that the α-divergences can be equivalently defined as $D_\alpha(P\|Q) = \int \tilde{f}_\alpha(\frac{\mathrm{d}P}{\mathrm{d}Q})\,\mathrm{d}Q$, where $\tilde{f}_\alpha(x) = \frac{x^\alpha - x}{\alpha(\alpha-1)}$ by noticing that $\int (f_\alpha - \tilde{f}_\alpha)(\frac{\mathrm{d}P}{\mathrm{d}Q})\,\mathrm{d}Q = 0$ for any $P \ll Q$. In the limiting case for $\tilde{f}_\alpha(x)$ when $\alpha \to 1$, we have $\lim_{\alpha\to 1} \frac{x^\alpha - x}{\alpha(\alpha-1)} = x \ln x$, recovering the KL divergence. In this paper, we use $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$, and simply mean to replace f_α in (3.2) by $f(x) = x \ln x$ whenever we refer to $\alpha = 1$.

The α-divergence can be equivalently formulated in its dual form [4, 39] as

$$D_\alpha(P\|Q) = \sup_{\gamma \in \mathscr{M}_b(\mathbb{R}^d)} \left\{ \mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)] \right\}, \tag{3.3}$$

where $\mathscr{M}_b(\mathbb{R}^d)$ is the set of bounded measurable functions and f_α^* is the convex conjugate (Legendre transform) of f_α,

$$f_\alpha^*(y) = \begin{cases} \alpha^{-1}(\alpha-1)^{\frac{\alpha}{\alpha-1}} y^{\frac{\alpha}{\alpha-1}} \mathbf{1}_{y>0} + \frac{1}{\alpha(\alpha-1)}, & \alpha > 1, \\ \infty \mathbf{1}_{y\geq 0} + \left(\alpha^{-1}(1-\alpha)^{-\frac{\alpha}{1-\alpha}} |y|^{-\frac{\alpha}{1-\alpha}} - \frac{1}{\alpha(1-\alpha)} \right) \mathbf{1}_{y<0}, & \alpha \in (0,1). \end{cases} \tag{3.4}$$

Compared to (3.3), the formulation of the Lipschitz-regularized α-divergences in (1.1) can be viewed as imposing Lipschitz regularization on the space of test functions in the variational form of α-divergences. In our work, we focus on the case when $\alpha > 1$ or $\alpha = 1$ (corresponding to the KL divergence). It has been proved in [4] that the Lipschitz-regularized α-divergence defined in (1.1) has an equivalent primal formulation

$$D_\alpha^L(P\|Q) = \inf_{\eta \in \mathscr{P}(\mathbb{R}^d)} \{ D_\alpha(\eta\|Q) + L \cdot W_1(P,\eta) \}, \tag{3.5}$$

where W_1 is the Wasserstein-1 metric. One can easily verify that D_α^L satisfies the conditions for being a divergence using (3.5). Equation (3.5) can be viewed as the infimal convolution between the α-divergence and the Wasserstein-1 metric. Though (1.1) is more often used in generative modeling as training objectives, its primal formulation is also theoretically very important. For example, we have from (3.5) that

$$D_\alpha^L(P\|Q) \leq \min\{ D_\alpha(P\|Q), L \cdot W_1(P,Q) \}. \tag{3.6}$$

In practical tasks, such as in generative modeling, we estimate the divergence from finite samples of P and Q, where the absolute continuity assumption in (3.2) typically no longer holds. Meanwhile, $D_\alpha^L(P\|Q)$ is always finite if P and Q are discrete measures of finitely many points with possibly different support since $D_\alpha^L(P\|Q) \leq L \cdot W_1(P,Q) < \infty$ by (3.6).

The following example shows that we can have a strict inequality in (3.6).

EXAMPLE 1. Let P and Q be distributions on $\mathbb{R}$ such that

$$p(x) = (1+\delta) x^{-(2+\delta)} \mathbf{1}_{x\geq 1}, \quad q(x) = \frac{1}{2}\mathbf{1}_{0\leq x<1} + \frac{1}{x^2}\mathbf{1}_{x\geq 2}.$$

Then neither $D_\alpha(P\|Q)$ nor $W_1(P,Q)$ is finite for any $\alpha > 1, \delta > 0$, while $D_\alpha^L(P\|Q) < \infty$.

Proof. Since P is not absolutely continuous with respect to Q, we have $D_\alpha(P\|Q) = \infty$; applying the cumulative distribution function formula for the one-dimensional Wasserstein-1 distance, it is straightforward to see $W_1(P,Q) = \infty$ as Q does not have a finite first moment. Consider the formula

(3.5) and in particular, we design the intermediate probability measure as

$$\mathrm{d}\eta = (1+\delta)2^{1+\delta}x^{-(2+\delta)}\mathbf{1}_{x\geq 2}.$$

Then we have

$$D_\alpha(\eta\|Q) = \int_2^\infty \frac{(1+\delta)^\alpha 2^{\alpha(1+\delta)}x^{-\alpha\delta}-1}{\alpha(\alpha-1)}\cdot\frac{1}{x^2}\,\mathrm{d}x < \infty,$$

and

$$\begin{aligned} W_1(P,\eta) &= \int_1^2\int_1^y (1+\delta)x^{-(2+\delta)}\,\mathrm{d}x\mathrm{d}y \\ &\quad + \int_2^\infty \left|\int_1^y (1+\delta)x^{-(2+\delta)}\,\mathrm{d}x - \int_2^y (1+\delta)2^{1+\delta}x^{-(2+\delta)}\,\mathrm{d}x\right|\mathrm{d}y \\ &= \int_1^2 1-y^{-(1+\delta)}\mathrm{d}y + \int_2^\infty \left|(1-y^{-(1+\delta)}) - (1-2^{1+\delta}y^{-(1+\delta)})\right|\mathrm{d}y \\ &= \int_1^2 1-y^{-(1+\delta)}\mathrm{d}y + \int_2^\infty (2^{1+\delta}-1)y^{-(1+\delta)}\mathrm{d}y < \infty. \end{aligned}$$

Therefore, $D_\alpha^L(P\|Q) \leq D_\alpha(\eta\|Q) + L\cdot W_1(P,\eta) < \infty$. □

Example 1 is not a special example when $D_\alpha^L(P\|Q)$ is finite but neither $D_\alpha(P\|Q)$ nor $W_1(P,Q)$ is finite. In fact, D_α^L can be applied to much wider situations. As we will see in Theorem 2 and its proof, the Lipschitz regularization plays a key role.

For the rest of the paper, we denote by $\mathscr{P}_k(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ that have a finite kth moment, $k \geq 1$ and we assume that k can be a non-integer; we also denote by $\mathscr{P}_{<k}(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ that have a finite sth moment for any $s < k$.

4. Finiteness and variational derivatives of D_α^L

In generative modeling, the goal is to approximate a target data distribution Q by a generated distribution P_{g_θ}, where g_θ is typically a neural net parametrization. A specific divergence between the target and the generated distributions is often chosen as the loss function. We want to build the best approximation $P_{g_{\theta^*}}$ of Q using the optimization of a probability divergence or metric:

$$g_{\theta^*} = \arg\min_{g_\theta\in\mathscr{G}} D\left(P_{g_\theta}, Q\right) \approx Q, \tag{4.1}$$

where $\mathscr{G}$ is a family of neural nets with certain constraints on the parameters θ. To optimize or minimize this loss, it is essential to ensure that the loss function or divergence is *finite*. In Section 4.1, we first demonstrate that when P has a finite first moment, $D_\alpha^L(P\|Q)$ remains finite without requiring any assumptions on Q. In Section 4.2, assuming P and Q have densities and tails, we provide necessary and sufficient conditions for $D_\alpha^L(P\|Q)$ to be finite.

4.1 *Minimal assumptions on the target Q*

We make the following assumption on P and Q for this subsection.

Assumption 1. Let P and Q be arbitrary probability measures on $\mathbb{R}^d$. In addition, we assume that P has a finite first moment, i.e. $P \in \mathscr{P}_1(\mathbb{R}^d)$.

We show in Theorem 2 that $D_\alpha^L(P\|Q)$ is finite whenever $P \in \mathscr{P}_1(\mathbb{R}^d)$ without any assumption on Q. This includes cases when Q has heavy tails, even without a finite first moment, and when Q is supported on a low-dimensional manifold and does not have a density. Before stating and proving the theorem, we need the following lemma for measures that are not necessarily probability measures that generalizes Lemma A.12 in [10], and the proofs are the same in essence.

Lemma 1. For $\alpha > 1$ and any non-negative measures P and Q defined on some bounded $\Omega \subset \mathbb{R}^d$ with non-zero integrals, $\Gamma = \mathrm{Lip}_L(\Omega)$, we have

$$\sup_{\gamma\in\Gamma}\left\{\int_\Omega \gamma(x)\,\mathrm{d}P - \int_\Omega f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\} = \sup_{\gamma\in\mathscr{F}}\left\{\int_\Omega \gamma(x)\,\mathrm{d}P - \int_\Omega f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\}, \tag{4.2}$$

where

$$\mathscr{F} = \left\{\gamma \in \mathrm{Lip}_L(\Omega) : \|\gamma\|_\infty \le (\alpha-1)^{-1}\left(\frac{\int_\Omega \mathrm{d}P}{\int_\Omega \mathrm{d}Q}\right)^{\alpha-1} + L\cdot\mathrm{diam}(\Omega)\right\}.$$

Proof of Lemma 1. For any fixed $\gamma \in \Gamma$, define

$$h(\nu) = \int_\Omega (\gamma(x)+\nu)\,\mathrm{d}P - \int_\Omega f_\alpha^*[\gamma(x)+\nu]\,\mathrm{d}Q.$$

Since $\sup_{x\in\Omega}\gamma(x) - \inf_{x\in\Omega}\gamma(x) \le L\cdot\mathrm{diam}(\Omega)$, interchanging the integration with differentiation is allowed by the dominated convergence theorem:

$$h'(\nu) = \int_\Omega \mathrm{d}P - \int_\Omega f_\alpha^{*\prime}(\gamma+\nu)\,\mathrm{d}Q,$$

where

$$f_\alpha^{*\prime}(y) = (\alpha-1)^{\frac{1}{\alpha-1}} y^{\frac{1}{\alpha-1}}\mathbf{1}_{y>0}. \tag{4.3}$$

If $\inf_{x\in\Omega}\gamma(x) > (\alpha-1)^{-1}\left(\frac{\int \mathrm{d}P}{\int \mathrm{d}Q}\right)^{\alpha-1}$, then $h'(0) < 0$. So there exists some $\nu_0 < 0$ such that $h(\nu_0) > h(0)$. This indicates the supremum on the left side of (4.2) is attained only if $\sup_{x\in\Omega}\gamma(x) \le (\alpha-1)^{-1}\left(\frac{\int \mathrm{d}P}{\int \mathrm{d}Q}\right)^{\alpha-1} + L\cdot\mathrm{diam}(\Omega)$. On the other hand, if $\sup_{x\in\Omega}\gamma(x) < 0$, then there exists $\nu_0 > 0$ that

satisfies $\sup_{x\in\Omega}\gamma(x)+\nu_0<0$ such that

$$\begin{aligned}\int_\Omega \big(\gamma(x)+\nu_0\big)\,\mathrm{d}P-\int_\Omega f_\alpha^*[\gamma(x)+\nu_0]\,\mathrm{d}Q&=\int_\Omega \big(\gamma(x)+\nu_0\big)\,\mathrm{d}P\\&>\int_\Omega \gamma(x)\,\mathrm{d}P\\&=\int_\Omega \gamma(x)\,\mathrm{d}P-\int_\Omega f_\alpha^*[\gamma(x)]\,\mathrm{d}Q.\end{aligned}$$

This indicates that the supremum on the left side of (4.2) is attained only if $\inf_{x\in\Omega}\gamma(x)\geq -L\cdot\mathrm{diam}(\Omega)$. Therefore, we have that the supremum on the left side of (4.2) is attained only if $\|\gamma\|_\infty\leq(\alpha-1)^{-1}\left(\frac{\int \mathrm{d}P}{\int \mathrm{d}Q}\right)^{\alpha-1}+L\cdot\mathrm{diam}(\Omega)$. □

THEOREM 2. Suppose $\alpha\geq 1$ ($\alpha=1$ refers to the KL) and P,Q satisfy Assumption 1, namely $P\in\mathscr{P}_1(\mathbb{R}^d)$, then $D_\alpha^L(P\|Q)<\infty$.

The key is the Lipschitz regularization, without which the result will not be true; see the proof below.

Proof. We first prove the case when $\alpha>1$. Let $\Gamma=\mathrm{Lip}_L(\mathbb{R}^d)$, and we have

$$\begin{aligned}D_\alpha^L(P\|Q)&=\sup_{\gamma\in\Gamma}\left\{\int\gamma(x)\,\mathrm{d}P-\int f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\}\\&\leq\sup_{\gamma\in\mathrm{Lip}_L(\|x\|<R)}\left\{\int_{\|x\|<R}\gamma(x)\,\mathrm{d}P-\int_{\|x\|<R}f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\}\\&\quad+\sup_{\gamma\in\mathrm{Lip}_L(\|x\|\geq R)}\left\{\int_{\|x\|\geq R}\gamma(x)\,\mathrm{d}P-\int_{\|x\|\geq R}f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\}\\&:=I_1+I_2.\end{aligned}$$

For I_1, by Lemma 1, we have

$$I_1\leq C\int_{\|x\|<R}\mathrm{d}P+\left(\alpha^{-1}(\alpha-1)^{\frac{\alpha}{\alpha-1}}C^{\frac{\alpha}{\alpha-1}}+\alpha^{-1}(\alpha-1)^{-1}\right)\int_{\|x\|<R}\mathrm{d}Q<\infty,$$

where $C=(\alpha-1)^{-1}\left(\frac{\int_{\|x\|<R}\mathrm{d}P}{\int_{\|x\|<R}\mathrm{d}Q}\right)^{\alpha-1}+2LR$.

Now we prove that $I_2<+\infty$. Let $M(\gamma)=\sup_{\|x\|=R}|\gamma(x)|$, where $\gamma\in\mathrm{Lip}_L(\|x\|\geq R)$. We show that there exists some $\overline{M}>0$ such that

$$I_2=\sup_{\gamma\in\mathscr{G}}\left\{\int_{\|x\|\geq R}\gamma(x)\,\mathrm{d}P-\int_{\|x\|\geq R}f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\},\tag{4.4}$$

where

$$\mathscr{G} = \left\{\gamma \in \mathrm{Lip}_L(\|x\| \geq R) : M(\gamma) \leq \overline{M}\right\}. \tag{4.5}$$

Indeed, we have for any $\gamma \in \mathrm{Lip}_L(\|x\| \geq R)$,

$$\begin{aligned}
&\int_{\|x\|\geq R} \gamma(x)\,\mathrm{d}P - \int_{\|x\|\geq R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q \\
&\quad= \int_{R\leq\|x\|<2R} \gamma(x)\,\mathrm{d}P - \int_{R\leq\|x\|<2R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q \\
&\qquad+ \int_{\|x\|\geq 2R} \gamma(x)\,\mathrm{d}P - \int_{\|x\|\geq 2R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q \\
&\quad\leq \int_{R\leq\|x\|<2R} \gamma(x)\,\mathrm{d}P - \int_{R\leq\|x\|<2R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q + \int_{\|x\|\geq 2R} \gamma(x)\,\mathrm{d}P \\
&\quad\leq (M(\gamma)+LR)\int_{R\leq\|x\|<2R} \mathrm{d}P - \int_{R\leq\|x\|<2R} f_\alpha^*(M(\gamma)-3LR)\,\mathrm{d}Q \\
&\qquad+ \int_{\|x\|\geq 2R} (M(\gamma)+LR+L\|x\|)\,\mathrm{d}P \\
&\quad= LR\int_{\|x\|\geq R} \mathrm{d}P + L\int_{\|x\|\geq 2R} \|x\|\,\mathrm{d}P + M(\gamma)\int_{\|x\|\geq R} \mathrm{d}P \\
&\qquad- f_\alpha^*(M(\gamma)-3LR)\int_{R\leq\|x\|<2R} \mathrm{d}Q,
\end{aligned}$$

where the last inequality is due to the fact that $\gamma(x)$ is L-Lipschitz and that for any $x : \|x\| \geq R$, we have $|\gamma(x) - M(\gamma)| \leq L(R + \|x\|)$. The first two terms are finite and are independent of γ since $P \in \mathscr{P}_1(\mathbb{R}^d)$. For the difference between the last two terms, we have

$$\lim_{M(\gamma)\to+\infty} M(\gamma)\int_{\|x\|\geq R} \mathrm{d}P - f_\alpha^*(M(\gamma)-3LR)\int_{R\leq\|x\|<2R} \mathrm{d}Q = -\infty,$$

since the exponent of x in $f_\alpha^*(x)$ is $\frac{\alpha}{\alpha-1} > 1$. This indicates that the supremum in I_2 should be taken over γ such that $M(\gamma) \leq \overline{M}$ for some $\overline{M} > 0$. Therefore,

$$\begin{aligned}
I_2 &= \sup_{\gamma\in\mathscr{G}} \left\{\int_{\|x\|\geq R} \gamma(x)\,\mathrm{d}P - \int_{\|x\|\geq R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q\right\} \\
&\leq \sup_{\gamma\in\mathscr{G}} \int_{\|x\|\geq R} \gamma(x)\,\mathrm{d}P \\
&\leq \sup_{\gamma\in\mathscr{G}} \int_{\|x\|\geq R} (LR + L\|x\| + M(\gamma))\,\mathrm{d}P \\
&\leq \int_{\|x\|\geq R} \left(LR + L\|x\| + \overline{M}\right)\mathrm{d}P < \infty.
\end{aligned}$$

For $\alpha = 1$, we bound I_1 using a similar Lemma B1 in Appendix B, and the bound for I_2 can be derived exactly in the same way as for $\alpha > 1$ by replacing f_α^* by f_{KL}^*. □

REMARK 2. Lemma 1 and Theorem 2 indeed work for any Lipschitz-regularized f-divergences, if f^*, the convex conjugate of f, is bounded below and superlinear, i.e. $\lim_{x\to\infty} \frac{f^*(x)}{x} = \infty$.

REMARK 3. Theorem 2 has important implications in generative modeling that one can learn a data distribution Q, without any prior knowledge of whether Q has heavy tails (even without a finite first moment) or lies on a low-dimensional manifold such that Q does not have a density, whenever P has a finite first moment, which is a very weak assumption; e.g. P can start with the Gaussian which is very easy to sample from. In this sense, the generative learning task can be agnostic to the structure of the data distribution using Lipschitz-regularized α-divergences as the objective functionals.

In what follows, we discuss the applicability of two generative models based on Theorem 2. Their numerical implementations can be seen in several numerical examples in Section 7.

Lip-α-GANs GANs based on the Lipschitz-regularized α-divergences, abbreviated as Lip-α-GANs, can be formulated as

$$\inf_{g\in\mathscr{G}} D_\alpha^L(g_\sharp P\|Q) = \inf_{g\in\mathscr{G}} \sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)} \left\{\mathbb{E}_{g_\sharp P}[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right\}, \tag{4.6}$$

where P is the initial source distribution, typically chosen as a Gaussian, and Q is the target data distribution, and $\mathscr{G}$ is the class of generators, and $g_\sharp P$ is the push-forward measure of P by the map g. Theorem 2 informs us that we can learn any probability measure Q if $g_\sharp P \in \mathscr{P}_1(\mathbb{R}^d)$; e.g. the generator can be realized using a ReLU network with a Gaussian source distribution as P. Key to obtaining the optimal generator is calculating the gradient of the loss relative to generator parameters, shown by the chain rule (Regarding the chain rule calculation (4.7), we also refer to a related formal calculation in Sec. 3.3 of [38]):

$$\nabla_\theta D_\alpha^L(P_{g_\theta}\|Q) = \int \frac{\delta D_\alpha^L(P\|Q)}{\delta P}(P_{g_\theta}(x)) \cdot \nabla_\theta P_{g_\theta}(x)\,\mathrm{d}x, \tag{4.7}$$

where $\frac{\delta D_\alpha^L(P\|Q)}{\delta P}$ is the variational derivative or the first variation of $D_\alpha^L(P\|Q)$, formally defined in Theorem 3. Therefore, even with a well-designed neural network architecture for the generator g_θ, a robust and well-defined variational derivative $\frac{\delta D_\alpha^L(P\|Q)}{\delta P}(P_{g_\theta}(x))$ is crucial for stable and effective optimization in the parameter θ because it directly impacts the parameter gradient $\nabla_\theta D_\alpha^L(P_{g_\theta}\|Q)$ via (4.7), otherwise computing $\nabla_\theta D_\alpha^L(P_{g_\theta}, Q)$ could become unstable, leading to erratic parameter updates that hinder convergence. While GANs use discriminators rather than explicit variational derivatives, Theorem 3 shows that the finiteness of a variational derivative can provide mathematical insight into GAN training. On the other hand, it is worth noting that, in light of (3.5), D_α^L offers advantages over both D_α and W_1:

- The variational derivative does not exist in general for the Wasserstein-1 metric alone (as is used in WGANs). For example, let $P = \delta_{x_1}$ and $Q = \delta_{x_2}$ be two Dirac delta distributions centered at points x_1 and x_2 in $\mathbb{R}$ with the usual distance function. Then the variational derivative in the sense of

Theorem 3 has a discontinuity:

$$\frac{\partial}{\partial\epsilon}\left|x_1 - x_2 + \epsilon v\right|\Big|_{\epsilon=0} = \begin{cases} v, & \text{if} x_1 - x_2 > 0, \\ -v, & \text{if} x_1 - x_2 < 0. \end{cases}$$

- Unregularized f-divergences (such as the KL-divergence) may yield large variational derivatives when P_{g_θ} and Q do not overlap significantly, potentially causing gradient spikes. This instability can lead to large, uncontrolled updates in θ, which might result in mode collapse or oscillations in GAN training. In contrast, the Lipschitz-regularized α-divergences always have well-defined variational derivatives by Theorem 3. For example, let $P = \mathcal{N}(\mu_1, \sigma)$ and $Q = \mathcal{N}(\mu_2, \sigma)$ be two univariate Gaussians with different means but the same variance. Then through a direct calculation, we have $D_{\mathrm{KL}}(P\|Q) = \frac{-(\mu_1-\mu_2)^2}{2\sigma^2}$, so that $\frac{\mathrm{d}D_{\mathrm{KL}}(P\|Q)}{\mathrm{d}\mu_1} = \frac{-(\mu_1-\mu_2)}{\sigma^2}$. We can think of P and Q do not overlap significantly if $\mu_1 - \mu_2$ has a large magnitude and σ is small, so that both $D_{\mathrm{KL}}(P\|Q)$ and its derivative in μ_1 will have a large magnitude.

Gradient flows of D_α^L. To further illustrate the significance of Theorem 2, we provide perspectives from the Wasserstein gradient flows of D_α^L for a feasible distribution learning task. As a particular case of the Lipschitz-regularized gradient flows proposed in [20], the Lipschitz-regularized α-divergences can be used to construct gradient flows of the form

$$\partial_t P_t = \operatorname{div}\left(P_t \nabla \frac{\delta D_\alpha^L(P_t\|Q)}{\delta P_t}\right), \tag{4.8}$$

for an initial source probability measure P_0 and a target measure Q, where $\frac{\delta D_\alpha^L(P\|Q)}{\delta P}$ is the first variation of $D_\alpha^L(P\|Q)$, defined in Theorem 3. This type of gradient flows was inspired by the gradient flows in the 2-Wasserstein space of probability measures in [24, 42]. In [20], the first variation form of $D_\alpha^L(P\|Q)$ is proved under the assumption that both $P, Q \in \mathscr{P}_1(\mathbb{R}^d)$. In Theorem 3, we extend it to the case when we only require $P \in \mathscr{P}_1(\mathbb{R}^d)$ but impose no assumptions on Q. This corresponds to the condition in Theorem 2. The key to the extension is our Lemma E3 and the proof can be found in Appendix C.

THEOREM 3. Under Assumption 1, namely $P \in \mathscr{P}_1(\mathbb{R}^d)$ and Q can be any probability measure, we define

$$\gamma^\star := \underset{\gamma\in\mathrm{Lip}_L(\mathbb{R}^d)}{\arg\max}\left\{\mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right\}, \tag{4.9}$$

where the optimizer $\gamma^\star \in \mathrm{Lip}_L(\mathbb{R}^d)$ exists, and is defined on $\mathrm{supp}(P) \cup \mathrm{supp}(Q)$, and is unique. Subsequently, we can extend $\gamma^\star$ to all of $\mathbb{R}^d$ as $\hat{\gamma}$ with the same Lipschitz constant. Let ρ be a signed measure of total mass 0 and let $\rho = \rho_+ - \rho_-$, where both $\rho_\pm \in \mathscr{P}_1(\mathbb{R}^d)$ are non-negative and mutually singular. If $P + \epsilon\rho \in \mathscr{P}_1(\mathbb{R}^d)$ for sufficiently small $\epsilon > 0$, then

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(D_\alpha^L(P+\epsilon\rho\|Q) - D_\alpha^L(P\|Q)\right) = \int \hat{\gamma}\,\mathrm{d}\rho, \tag{4.10}$$

and we write

$$\frac{\delta D_\alpha^L(P\|Q)}{\delta P}(P) = \hat{\gamma}. \tag{4.11}$$

As a result, Theorem 3 provides a reformulation of (4.8) as in [20]:

$$\begin{aligned} &\partial_t P_t + \operatorname{div}(P_t v_t^L) = 0, \quad P_0 = P \in \mathscr{P}_1(\mathbb{R}^d), \\ &v_t^L = -\nabla\gamma_t^\star, \quad \gamma_t^\star = \underset{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}{\arg\max} \left\{\mathbb{E}_{P_t}[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right\}. \end{aligned} \tag{4.12}$$

Moreover, Theorem 2 in [20] tells us that if P_t is sufficiently smooth, then we have

$$\frac{\mathrm{d}}{\mathrm{d}t} D_\alpha^L(P_t\|Q) = -I_\alpha(P_t\|Q) \le 0, \tag{4.13}$$

where $I_\alpha(P_t\|Q)$ is the Lipschitz-regularized Fisher Information:

$$I_\alpha(P_t\|Q) := \mathbb{E}_{P_t}[\left|\nabla\gamma_t^\star\right|^2].$$

Then for any $T \ge 0$, we have

$$D_\alpha^L(P_T\|Q) = D_\alpha^L(P_0\|Q) - \int_0^T I_\alpha(P_s\|Q)\,\mathrm{d}s \le D_\alpha^L(P_0\|Q). \tag{4.14}$$

Therefore, both the finiteness and the variational derivative of $D_\alpha^L(P_0\|Q)$ are crucial for the divergence to dissipate from the gradient flow perspective. While the convergence of the gradient flow is also important, we do not address its partial differential equations (PDE) theory in this work, but rather its feasibility to learn any distribution Q.

4.2 *When P and Q have densities and heavy tails*

In this subsection, we show that D_α^L is applicable to comparing heavy-tailed distributions, by providing necessary and sufficient conditions that relate the tail behaviours of P and Q with α. This also provides insights into the selection of suitable α's. For this purpose, including cases when $P \notin \mathscr{P}_1(\mathbb{R}^d)$ –compare to Theorem 2–we make the following assumptions on P and Q.

Assumption 2. Let P and Q be distributions on $\mathbb{R}^d$ whose densities $p(x)$ and $q(x)$ are absolutely continuous with respect to the Lebesgue measure. However, P and Q are not necessarily absolutely continuous with respect to each other on some bounded subset.

Definition 1. For a pair of distributions (P, Q) on $\mathbb{R}^d$, we say they are of heavy-tail (β_1, β_2), $\beta_1, \beta_2 > d$, if there exists some $R > 0$, such that

$$p(x) \asymp \|x\|^{-\beta_1}, \quad q(x) \asymp \|x\|^{-\beta_2},$$

for $\|x\| \geq R$. That is, there exist constants $0 < c_{p,1} \leq c_{p,2}$ and $0 < c_{q,1} \leq c_{q,2}$ such that

$$c_{p,1} \|x\|^{-\beta_1} \leq p(x) \leq c_{p,2} \|x\|^{-\beta_1}, \quad c_{q,1} \|x\|^{-\beta_2} \leq q(x) \leq c_{q,2} \|x\|^{-\beta_2},$$

for $\|x\| \geq R$.

Then we prove the following necessary and sufficient conditions on the tail behaviours of (P, Q) for $D_\alpha^L(P\|Q)$ to be finite. The proof makes extensive use of the variational formula (1.1) and Lipschitz regularization and is provided in Appendix D.

THEOREM 4 (Necessary and sufficient conditions for $D_\alpha^L < \infty, \alpha > 1$). Suppose $\alpha > 1$, and (P, Q) are distributions on $\mathbb{R}^d$ of heavy-tail (β_1, β_2). Then $D_\alpha^L(P\|Q) < \infty$ if and only if one of the following two conditions holds:

(i) $d < \beta_1 \leq d + 1$ and $\beta_2 - \beta_1 < \frac{\beta_1 - d}{\alpha - 1}$;

(ii) $\beta_1 > d + 1$.

REMARK 4. We can relax the assumption in Definition 1 to allow different tail behaviour in different directions as follows. Let Ω_k be a finite partition of the spherical coordinates $[0, \pi]^{d-2} \times [0, 2\pi)$, where each Ω_k has non-zero Lebesgue measure of $[0, \pi]^{d-2} \times [0, 2\pi)$. We can assume that $p(x) \asymp \|x\|^{-\beta_{1,k}}$ and $q(x) \asymp \|x\|^{-\beta_{2,k}}$ on each Ω_k. Then the $D_\alpha^L(P\|Q) < \infty$ if and only if $\beta_{1,k}$ and $\beta_{2,k}$ satisfy one of the conditions of Theorem 4 on each Ω_k. The proof is the same as that of Theorem 4 constrained on each Ω_k. This relaxation can be adopted in the same way for Theorem 5 and Corollary 6.

For the Lipschitz-regularized KL-divergence, we have the following result whose proof can be found in Appendix D.

THEOREM 5 (Necessary and sufficient conditions for $D_{\mathrm{KL}}^L < \infty$). Suppose $\alpha = 1$ (the KL case), and (P, Q) are distributions on $\mathbb{R}^d$ of heavy-tail (β_1, β_2), then $D_{\mathrm{KL}}^L(P\|Q) < \infty$ for any $\beta_1, \beta_2 > d$.

REMARK 5. Since $\beta_1, \beta_2 > d$ are the minimal assumptions for P and Q to be probability distributions, Theorem 5 suggests that using the Lipschitz-regularized KL-divergence is the most robust choice, as it can be agnostic to both the tails of P and Q, compared to the conditions in Theorem 4.

In cases where both P and Q lie on a low-dimensional submanifold, we have the following corollary. The proof can be found in Appendix D.

COROLLARY 6 (Necessary and sufficient conditions on embedded submanifolds). Let $\mathscr{M}$ be a d^*-dimensional smooth embedded submanifold of $\mathbb{R}^d$ via an L^*-Lipschitz embedding $\varphi : \mathbb{R}^{d^*} \to \mathbb{R}^d$ with $\mathscr{M} = \varphi(\mathbb{R}^{d^*})$ for $d^* < d$. Suppose (P, Q) are of heavy-tail (β_1, β_2) on $\mathbb{R}^{d^*}$, and let $p_{\mathscr{M}}$ and $q_{\mathscr{M}}$ be their push-forward distributions on $\mathscr{M}$, i.e. $p_{\mathscr{M}} = p \circ \varphi^{-1}$ and $q_{\mathscr{M}} = q \circ \varphi^{-1}$. Then the Lipschitz-regularized α-divergence between $p_{\mathscr{M}}$ and $q_{\mathscr{M}}$, defined as

$$D_\alpha^L(p_{\mathscr{M}} \| q_{\mathscr{M}}) = \sup_{\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)} \left\{ \mathbb{E}_{p_{\mathscr{M}}}[\gamma] - \mathbb{E}_{q_{\mathscr{M}}}[f_\alpha^*(\gamma)] \right\},$$

is finite if and only if one of the following two conditions holds for $\alpha > 1$:

(1) $d^* < \beta_1 \le d^* + 1$ and $\beta_2 - \beta_1 < \frac{\beta_1 - d^*}{\alpha - 1}$;

(2) $\beta_1 > d^* + 1$;

and $D_\alpha^L(p_{\mathscr{M}} \| q_{\mathscr{M}}) < \infty$ for any $\beta_1, \beta_2 > d^*$ if $\alpha = 1$.

REMARK 6. The Lipschitz condition on the embedding φ is necessary to guarantee that the tails of $p_{\mathscr{M}}$ and $q_{\mathscr{M}}$ do not become heavier than those of p and q.

5. Lipschitz regularization implies finite-sample estimation of D_α^L on $\mathbb{R}^d$

In practice, we only have finite i.i.d. samples drawn from P and Q. We denote by $X = \{x_1, \dots, x_m\}$ and $Y = \{y_1, \dots, y_n\}$ the i.i.d. samples from P and Q, with empirical distributions $P_m = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$ and $Q_n = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$, respectively. Thus it is essential to provide guarantees for how fast $D_\alpha^L(P_m \| Q_n)$ converges to $D_\alpha^L(P\|Q)$ in average. This type of convergence rate for the Lipschitz-regularized α-divergences has been proved in [10] on bounded domains of $\mathbb{R}^d$. Here, we derive the first result of the convergence of the finite-sample estimations on the unbounded domain $\mathbb{R}^d$, under certain tail conditions. The result for $d \ge 3$ is stated below, with its proof deferred to Appendix E. The results for $d = 1, 2$ can be found as and in Appendix E.

THEOREM 7 (Finite sample estimation of D_α^L on $\mathbb{R}^d$). Assume $d \ge 3$. For $\alpha > 1$, let P and Q be probability measures on $\mathbb{R}^d$ such that $P \in \mathscr{P}_{<\beta_1 - d}(\mathbb{R}^d)$ and $Q \in \mathscr{P}_{<\beta_2 - d}(\mathbb{R}^d)$, where $\beta_1 > 3d$ and $\beta_2 > 5d$. Suppose α satisfies $\frac{2d\alpha}{\alpha-1} < \beta_1 - d$ and $\frac{2\alpha}{\alpha-1} < \frac{\beta_2}{d} - 3$. Then we have

$$\mathbb{E}_{X,Y}\left| D_\alpha^L(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \le \frac{C_1}{m^{1/d}} + \frac{C_2}{n^{1/d}}, \tag{5.1}$$

where C_1 depends on $M_{\frac{d}{d-1}}(P)$ and C_2 depends on $M_{\frac{2d\alpha}{\alpha-1}}(P)$, $M_{\frac{2d\alpha}{\alpha-1}}(Q)$, and $M_{dr_2}(Q)$ for any $2 + \frac{2\alpha}{\alpha-1} < r_2 < \frac{\beta_2}{d} - 1$. Here, we use $M_r(P)$ to denote the rth moment of P. Both C_1 and C_2 are independent of m, n, but they depend on L such that $C_1, C_2 \to \infty$ when $L \to \infty$.

REMARK 7. The key to proving Theorem 7 is to leverage the Lipschitz condition of the test functions in the variational form (1.1).

6. Finite-sample estimations of D_α^L and W_1 with group symmetry on $\mathbb{R}^d$

Based on Theorem 7 and its proof, we are able to consider one special situation when the distributions are invariant with respect to some group symmetry and to provide convergence results for the empirical estimations of D_α^L with group symmetry in $\mathbb{R}^d$. Empirical estimations of divergences with group symmetry have been studied in [10, 49] on bounded domains of $\mathbb{R}^d$ or on closed Riemannian manifolds. Here we provide the first sample complexity bound with group symmetry on unbounded domains, in particular, for D_α^L and later for W_1 in this section. Before presenting the theorems, we first briefly review the related concepts of group symmetry. Readers of interest can refer to [5, 10, 49] for more details. We leave all the proofs for this section in Appendix F.

A *group* is a set G equipped with a group product satisfying the axioms of associativity, identity and invertibility. Given a group G and a set $\mathscr{X} \subset \mathbb{R}^d$, a map $\theta : G \times \mathscr{X} \to \mathscr{X}$ is called a *group action on* $\mathscr{X}$ if $\theta_g := \theta(g, \cdot) : \mathscr{X} \to \mathscr{X}$ is an automorphism on $\mathscr{X}$ for all $g \in G$, and $\theta_{g_2} \circ \theta_{g_1} = \theta_{g_2 \cdot g_1}, \forall g_1, g_2 \in G$. By convention, we will abbreviate $\theta(g, x)$ as gx. We make the following assumptions on G.

ASSUMPTION 3. For any $g \in G$ and $x \in \mathbb{R}^d$, $\theta_g(x) = A_g \cdot x$, for some unitary matrix $A_g \in \mathbb{R}^{d \times d}$.

A function $\gamma : \mathscr{X} \to \mathbb{R}$ is called *G-invariant* if $\gamma \circ \theta_g = \gamma, \forall g \in G$. Let Γ be a set of measurable functions $\gamma : \mathscr{X} \to \mathbb{R}$; its subset, Γ_G, of G-invariant functions is defined as

$$\Gamma_G := \{\gamma \in \Gamma : \gamma \circ \theta_g = \gamma, \forall g \in G\}. \tag{6.1}$$

On the other hand, a probability measure $P \in \mathscr{P}(\mathscr{X})$ is called *G-invariant* if $P = (\theta_g)_\sharp P, \forall g \in G$, where $(\theta_g)_\sharp P := P \circ (\theta_g)^{-1}$ is the push-forward measure of P under θ_g. We denote the set of all G-invariant distributions on $\mathscr{X}$ as $\mathscr{P}_G(\mathscr{X}) := \{P \in \mathscr{P}(\mathscr{X}) : P \text{ is } G\text{-invariant}\}$. For $P, Q \in \mathscr{P}_G(\mathscr{X})$, [10] proposes the following symmetry-informed estimator

$$D_\alpha^{L,G}(P_m \| Q_n) := \sup_{\gamma \in \mathrm{Lip}_L^G(\mathbb{R}^d)} \{\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)]\} \tag{6.2}$$

for $D_\alpha^L(P\|Q)$, where $\mathrm{Lip}_L^G(\mathbb{R}^d) \subset \mathrm{Lip}_L(\mathbb{R}^d)$ that consists of G-invariant L-Lipschitz functions. It is shown in Theorem 4.6 in [5] that when P_m, Q_n are replaced by $P, Q \in \mathscr{P}_G(\mathscr{X})$ in (6.2), we have $D_\alpha^{L,G}(P\|Q) = D_\alpha^L(P\|Q)$; i.e. the divergence value between P and Q does not change if the supremum is taken over $\mathrm{Lip}_L^G(\mathbb{R}^d) \subset \mathrm{Lip}_L(\mathbb{R}^d)$ when both P and Q are G-invariant.

In particular, we consider the case when both P and Q are sub-Weibull, defined as follows.

DEFINITION 2 (sub-Weibull distributions). We call a distribution $P \in \mathscr{P}(\mathbb{R}^d)$ sub-Weibull, if

$$\Pr(x \sim P : \|x\| \geq r) \leq a \exp(-br^{1/\theta}) \text{ for all} r > 0, \text{ for some} a, b, \theta > 0. \tag{6.3}$$

REMARK 8. Sub-Gaussian and sub-exponential distributions are special examples of sub-Weibull distributions.

The following definition of intrinsic dimension is adopted from the capacity dimension from [25].

DEFINITION 3. The intrinsic dimension of a bounded $\mathscr{X} \subset \mathbb{R}^D$, denoted by $dim(\mathscr{X})$, is defined as

$$\dim(\mathscr{X}) := -\lim_{\epsilon \to 0^+} \frac{\ln \mathscr{N}(\mathscr{X}, \epsilon)}{\log \epsilon}, \tag{6.4}$$

where $\mathscr{N}(\mathscr{X}, \epsilon)$ is the covering number of $\mathscr{X}$ with ϵ-balls in the standard Euclidean metric of $\mathbb{R}^d$.

For example, if $\mathscr{X} \subset \mathbb{R}^D$ has non-empty interior, then $\dim(\mathscr{X}) = D$; if $\mathscr{X}$ is a d-dimensional submanifold of $\mathbb{R}^D$, then $\dim(\mathscr{X}) = d$.

We have the following theorem for the empirical estimation of D_α^L with group symmetry on unbounded domains.

THEOREM 8 (Finite sample estimation of D_α^L with finite group symmetry). For $\alpha > 1$, let $P, Q \in \mathscr{P}_G(\mathscr{X})$ for some $\mathscr{X} \subset \mathbb{R}^d$, where G satisfies Assumption 3. Suppose the quotient space $\mathscr{X}/G$ is connected, and for any bounded $\mathscr{X}_0 \subset \mathscr{X}/G$ with non-empty interior with respect to the subspace topology ($\mathscr{X}/G \hookrightarrow \mathbb{R}^d$). Let $|G| < \infty$ be the cardinality of G, and we further assume that both P and Q are sub-Weibull on $\mathbb{R}^d$. Then

- If $\dim(\mathscr{X}_0) = d^* \geq 3$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1}{(|G|\,m)^{1/d^*}} + \frac{C_2}{(|G|\,n)^{1/d^*}}; \tag{6.5}$$

- If $\dim(\mathscr{X}_0) = d^* = 2$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1 \ln m}{(|G|\,m)^{1/2}} + \frac{C_2 \ln n}{(|G|\,n)^{1/2}}; \tag{6.6}$$

- If $\dim(\mathscr{X}_0) = d^* = 1$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1}{(|G|\,m)^{1/2}} + \frac{C_2}{(|G|\,n)^{1/2}}, \tag{6.7}$$

where C_1 and C_2 depend on $M_d(P), M_d(Q)$. Both C_1 and C_2 are independent of m, n and G.

When G is a continuous group, we have the following theorem.

THEOREM 9 (Finite sample estimation of D_α^L with infinite group symmetry). For $\alpha > 1$, let $P, Q \in \mathscr{P}_G(\mathscr{X})$ for some $\mathscr{X} \subset \mathbb{R}^d$, where G satisfies Assumption 3. Suppose the quotient space $\mathscr{X}/G$ is connected, and for any bounded $\mathscr{X}_0 \subset \mathscr{X}/G$ with non-empty interior with respect to the subspace topology ($\mathscr{X}/G \hookrightarrow \mathbb{R}^d$). Assume that both P and Q are sub-Weibull on $\mathbb{R}^d$. Then

- If $\dim(\mathscr{X}_0) = d^{**} \geq 3$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1}{m^{1/d^{**}}} + \frac{C_2}{n^{1/d^{**}}}; \tag{6.8}$$

- If $\dim(\mathscr{X}_0) = d^{**} = 2$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1 \ln m}{m^{1/2}} + \frac{C_2 \ln n}{n^{1/2}}; \tag{6.9}$$

- If $\dim(\mathscr{X}_0) = d^{**} = 1$, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \leq \frac{C_1}{m^{1/2}} + \frac{C_2}{n^{1/2}}, \tag{6.10}$$

where C_1 and C_2 depend on $M_d(P), M_d(Q)$. Both C_1 and C_2 are independent of m, n.

REMARK 9. If $\mathscr{X}$ is a d^*-dimensional connected submanifold of $\mathbb{R}^d$, and G is a compact Lie group acting locally smoothly on $\mathscr{X}$, then $d^{**} = d - \dim(G)$, where $\dim(G)$ is the dimension of a principal orbit (i.e. the maximal dimension among all orbits) by Theorem IV 3.8 in [6].

The proofs of Theorems 8 and 9 also imply the convergence bound for the Wasserstein-1 distance with group symmetry on unbounded domains, since the variational form is shift-invariant with respect to the test function. We consider the symmetry-informed estimator for $P, Q \in \mathscr{P}_G(\mathscr{X})$, proposed in [10, 49], defined as

$$W_1^G(P_m, Q_n) := \sup_{\gamma \in \mathrm{Lip}_L^G(\mathbb{R}^d)} \{\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[\gamma]\} \tag{6.11}$$

for $W_1(P, Q)$.

THEOREM 10 (Finite sample estimation of W_1 with finite group symmetry). Let $P, Q \in \mathscr{P}_G(\mathscr{X})$ for some $\mathscr{X} \subset \mathbb{R}^d$, where G satisfies Assumption 3. Suppose the quotient space $\mathscr{X}/G$ is connected, and for any bounded $\mathscr{X}_0 \subset \mathscr{X}/G$ with non-empty interior with respect to the subspace topology $(\mathscr{X}/G \hookrightarrow \mathbb{R}^d)$. Let $|G| < \infty$ be the cardinality of G, and we further assume that both P and Q are sub-Weibull on $\mathbb{R}^d$. Then

- If $\dim(\mathscr{X}_0) = d^* \geq 3$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1}{(|G|\, m)^{1/d^*}} + \frac{C_2}{(|G|\, n)^{1/d^*}}; \tag{6.12}$$

- If $\dim(\mathscr{X}_0) = d^* = 2$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1 \ln m}{(|G|\, m)^{1/2}} + \frac{C_2 \ln n}{(|G|\, n)^{1/2}}; \tag{6.13}$$

- If $\dim(\mathscr{X}_0) = d^* = 1$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1}{(|G|\, m)^{1/2}} + \frac{C_2}{(|G|\, n)^{1/2}}, \tag{6.14}$$

where C_1 and C_2 depends on $M_d(P)$, $M_d(Q)$. Both C_1 and C_2 are independent of m, n and G.

When G is a continuous group, we have the following theorem.

THEOREM 11 (Finite sample estimation of W_1 with infinite group symmetry). Let $P, Q \in \mathscr{P}_G(\mathscr{X})$ for some $\mathscr{X} \subset \mathbb{R}^d$, where G satisfies Assumption 3. Suppose the quotient space $\mathscr{X}/G$ is connected, and for any bounded $\mathscr{X}_0 \subset \mathscr{X}/G$ with non-empty interior with respect to the subspace topology $(\mathscr{X}/G \hookrightarrow \mathbb{R}^d)$. Assume that both P and Q are sub-Weibull on $\mathbb{R}^d$. Then we have

- If $\dim(\mathscr{X}_0) = d^{**} \geq 3$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1}{m^{1/d^{**}}} + \frac{C_2}{n^{1/d^{**}}}; \tag{6.15}$$

- If $\dim(\mathscr{X}_0) = d^{**} = 2$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1 \ln m}{m^{1/2}} + \frac{C_2 \ln n}{n^{1/2}}; \tag{6.16}$$

- If $\dim(\mathscr{X}_0) = d^{**} = 1$, we have

$$\mathbb{E}_{X,Y}\left|W_1^G(P_m, Q_n) - W_1(P, Q)\right| \leq \frac{C_1}{m^{1/2}} + \frac{C_2}{n^{1/2}}, \tag{6.17}$$

where C_1 and C_2 depend on $M_d(P), M_d(Q)$. Both C_1 and C_2 are independent of m, n.

Remark 10. Although the multiplicative constants in Theorems 10 and 11 are not optimal, but the rate is optimal compared to Theorem 1 in [18] for W_1, when d^* or d^{**} are greater than or equal to three, or equal to one.

7. Numerical experiments

In this section, we demonstrate how using the Lipschitz-regularized α-divergences as objective functionals enables stable learning of heavy-tailed distributions and distributions with low-dimensional manifolds or fractal structures with various generative models. Note that the Lipschitz-regularized α-divergences have an equivalent primal formulation in (3.5), which can be viewed as α-divergences with W_1-proximal regularization. One may consider replacing the W_1-proximal regularization with a W_2-proximal regularization, where W_2 is the Wasserstein-2 distance, as the W_2 distance and proximal regularization is widely used in generative modeling; e.g. see [41, 52]. The α-divergences with W_2-proximal regularization are defined as

$$D^{\lambda}_{\alpha,2}(P\|Q) := \inf_{\eta \in \mathscr{P}(\mathbb{R}^d)} \{D_\alpha(\eta\|Q) + \lambda \cdot W_2^2(P, \eta)\}. \tag{7.1}$$

In Section 7.1, we introduce the generative models used and explain how their learning objectives relate to α-divergences with W_1 or W_2 proximals. We illustrate our points with four examples. In Section 7.2, we compare the effects of incorporating W_1 or W_2 proximals in the learning objectives by training on a two-dimensional Student-t distribution and on a real-world keystroke dataset. In Section 7.3, we show the importance of Lipschitz-regularized α-divergences when learning distributions with low-dimensional structures with an example of learning a strange attractor from the Lorenz 63 model. In Section 7.4, we present the task of learning an anisotropic heavy-tailed distribution embedded in a high-dimensional space and the results highlight that the Lipschitz-regularized α-divergences make generative learning agnostic to heavy-tailed and manifold assumptions. We use Gaussian priors for all our experiments, and the implementation details including the network architectures can be found in the Supplementary Material.

7.1 *Generative models with different learning objectives*

W_1 and W_2 proximals can be found, sometimes implicitly, in the learning objectives of several existing generative models. Below, we list various models based on α-divergences used in our experiments and explain why some of them are (either implicitly or explicitly) regularized by Wasserstein proximal.

(1) **Generative models without proximal regularization:**

- **α-GAN**: GANs [19, 40] based on the variational representation of the α-divergence (3.3);
- **α-GPA**: GPA based on the α-divergence [20];
- **CNF**: CNFs by [8], where the loss function is based on the KL divergence, a special case of the α-divergence when $\alpha = 1$.

(2) **Generative models with W_1-proximal regularization:**

- **Lip-α-GAN** [4]: GANs using the Lipschitz-regularized α-divergence (1.1) as the objective function, with the Lipschitz constant set to $L = 1$ in our experiment;
- **Lip-α-GPA** [20]: GPAs using the Lipschitz-regularized α-divergence (1.1) as the objective function, with the Lipschitz constant set to $L = 1$ in our experiment. This is the implementation of the gradient flow formulation (4.12).

(3) **Generative models with W_2-proximal regularization:** We consider the following class of flow-based models, which minimize α-divergences with W_2 proximal (7.1) written as (7.2) via the Benamou-Brenier formula,

$$\inf_{v,\rho} \mathscr{F}(\rho(\cdot, T)) + C \int_0^T \frac{1}{2} |v(x,t)|^2 \rho(x,t)\, \mathrm{d}x\, \mathrm{d}t. \tag{7.2}$$

Here, $\rho : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ is the evolution of the probability measure via the (trainable) velocity field $v : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$, satisfying the Fokker–Planck equation:

$$\rho_t + \nabla \cdot (\rho v) = \frac{\sigma^2}{2} \Delta \rho, \quad \rho(\cdot, 0) = \rho_0 \text{ is a tractable prior distribution, e.g. Gaussian.} \tag{7.3}$$

- **OT flow** [41]: OT normalizing flows, which are equivalent to the W_2-proximal of CNFs, with $\mathscr{F}(\rho(\cdot, T)) = D_{\mathrm{KL}}(Q \| \rho(\cdot, T))$ and $\sigma = 0$ in (7.2);
- **Variance-exploding (VE)-Score-based generative model (SGM)** [47]: SGMs with VE forward stochastic differential equations (SDEs) [47]. According to the mean-field game formulation by [54], it is equivalent to (7.2) with stochastic dynamics ($\sigma > 0$) and a cross-entropy terminal cost $\mathscr{F}(\rho(\cdot, T)) = -\mathbb{E}_{\rho(\cdot,T))}[\log Q]$, essentially also a W_2-proximal of CNFs.

We refer to Fig. 1 for a visual illustration of the relationships among the models being compared.

7.2 *Learning heavy-tailed distributions*

7.2.1 *Two-dimensional Student-t example* We compare various generative models for learning a heavy-tailed two-dimensional isotropic Student-t distribution with ν degrees of freedom, $q(x) \propto (1 + \frac{|x|^2}{\nu})^{\frac{\nu+2}{2}}$. This synthetic example allows us to adjust the tail decay rate $\beta = \nu + 2$ by selecting different degrees of freedom ν. In the main text, we present a heavy-tailed example with $\beta = 3$ that does not have a finite first moment, while the relatively easier case of $\beta = 5$ is deferred to the Supplementary Material. We use 10,000 samples to train the models.

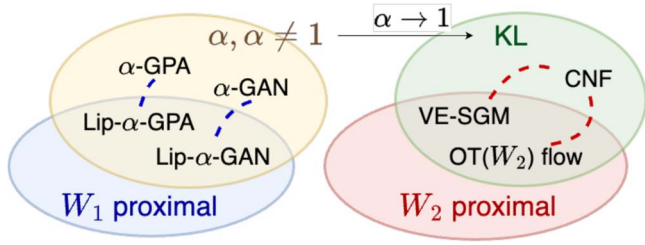
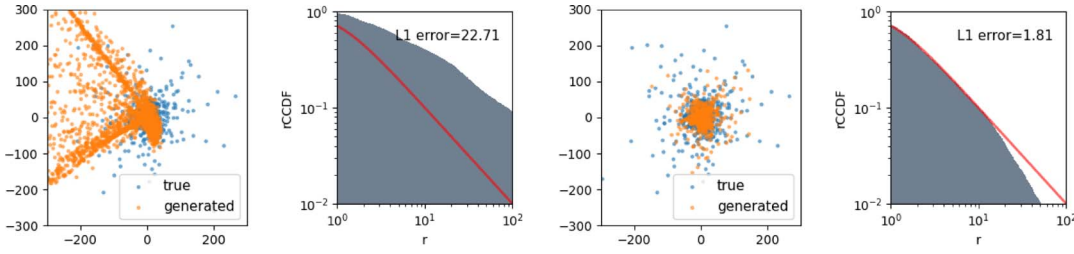
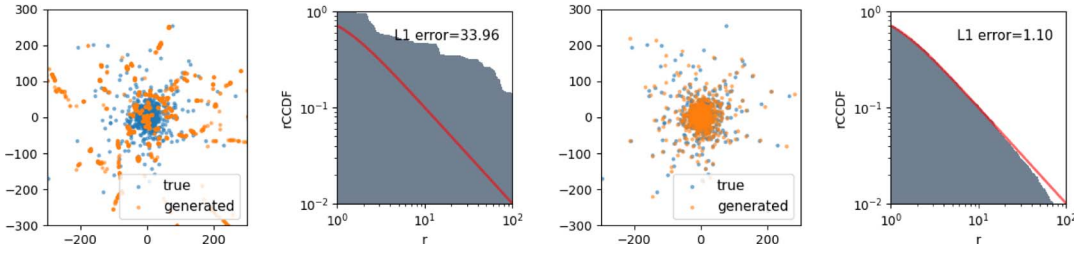


FIG. 1. Generative models in the experiment and their relationship with the α-divergences with W_1 or W_2 proximal regularization. See Section 7.1 for detailed explanations of the models and notations.



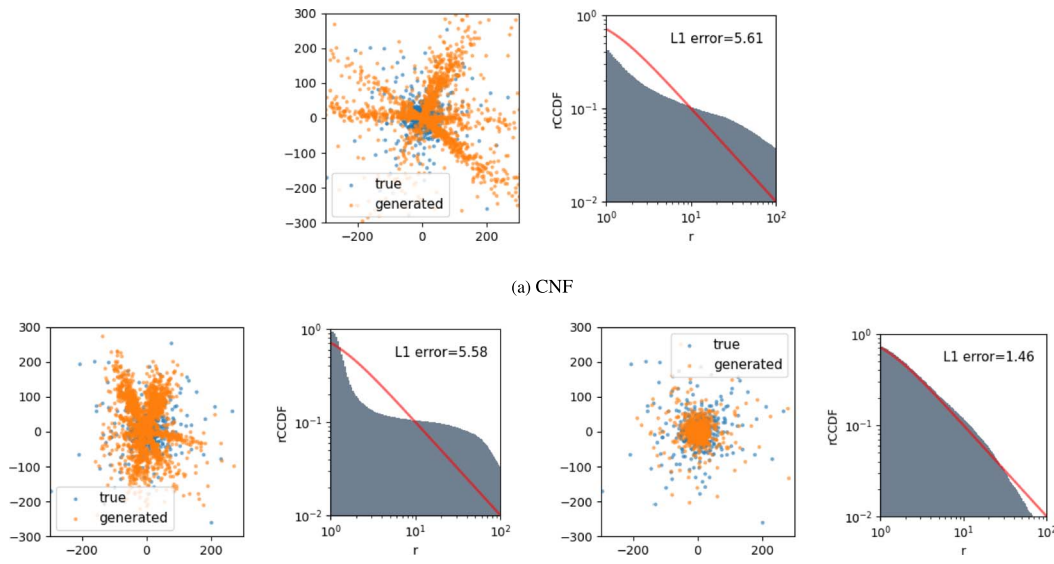
(a) α-GAN (left) and its counterpart with W_1-proximal regularization, Lip-α-GAN (right)



(b) α-GPA (left) and its counterpart with W_1-proximal regularization, Lip-α-GPA (right)

FIG. 2. Learning a two-dimensional isotropic Student-t with degree of freedom $\nu = 1$ (tail index $\beta = 3.0$) using generative models based on α-divergences with $\alpha = 2$ with or without Lipschitz regularization. Models with Lipschitz regularization (right) learn the heavy-tailed distribution significantly better than those without (left). See Section 7.1 for detailed explanations of the models.

Figures 2 and 3 present the performance of various generative models. Each model is evaluated in two plots. First, a two-dimensional scatter plot displays the generated samples (orange) and the true samples (blue), providing a visual assessment of the sample quality. Next, the tail behaviour is assessed by plotting the ground truth Radial Complementary Cumulative Distribution Function (rCCDF) (red curve) and the histogram of the radii of generated samples (gray). The rCCDF is defined as $\text{rCCDF}(r) = 1 - \text{CDF}(r)$, where $\text{CDF}(r)$ is the cumulative distribution function of the radius. We then calculate the L_1 error between the ground truth rCCDF and the generated sample histogram. Generative models with Lipschitz regularization (W_1-proximal) significantly outperform the others in learning heavy-tailed distributions, corroborating our theoretical results in Section 4.

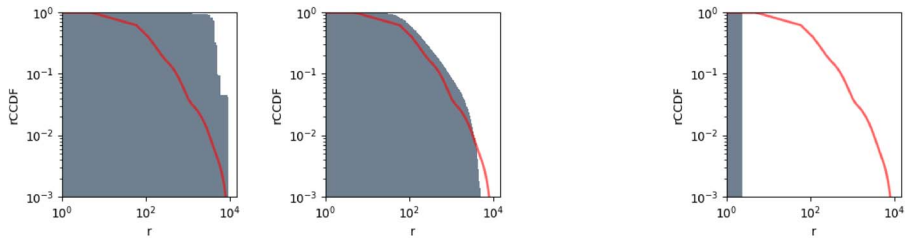


(a) CNF

(b) OT flow

(c) VE-SGM

FIG. 3. Learning a two-dimensional isotropic Student-t with degree of freedom $\nu = 1$ (tail index $\beta = 3.0$) using generative models based on α-divergences with or without W_2-proximal regularization and $\alpha = 2$. See Section 7.1 for detailed explanations of the models.



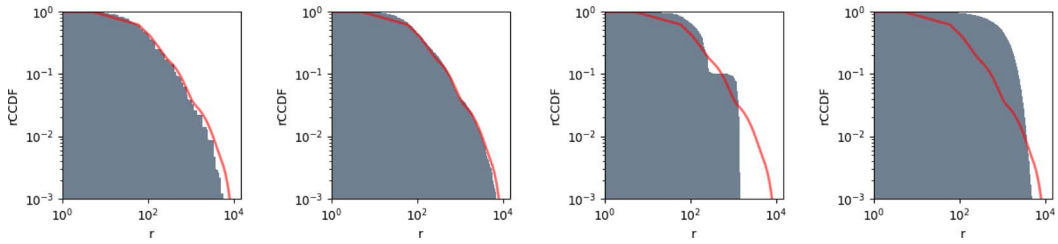
(a) α-GPA (left), α-GAN (right)

(b) CNF

FIG. 4. Sample generation of inter-arrival time between keystrokes. Generative models based on the α-divergences with $\alpha = 2$ (a), and the KL divergence (b).

7.2.2 *Keystroke example* For a real-world heavy-tailed example, we consider learning the inter-arrival time between keystrokes from multiple users typing sentences [12]. The target dataset consists of 7,160 scalar samples, and we generated 10,000 samples using generative models with W_1 or W_2 proximal regularization.

We display the tail behaviour by plotting the ground truth CCDF (red curve) and the corresponding histogram of the generated samples (gray) in Fig. 4 and Fig. 5. Unlike the previous synthetic example, the ground truth CCDF here is obtained by interpolating the heights of the histogram bins of the true samples. In Fig. 5, generative models with W_1-proximal regularization (Lip-α-GPA and Lip-α-GAN) outperform those regularized with W_2-proximals (OT flow and VE-SGM) in capturing the tails. This observation suggests that W_1-proximal algorithms can potentially handle heavier tails more effectively



(a) Lip-α GPA (left), Lip-α GAN (right)

(b) OT flow (left), VE SGM (right)

FIG. 5. Sample generation of inter-arrival time between keystrokes. Generative models with W_1-proximal regularization, panel (a), outperform those with W_2-proximal regularization, panel (b), in capturing the tails. This observation suggests that W_1-proximal algorithms can potentially handle heavier tails more effectively than W_2-proximal methods.

than W_2-proximal methods. In other words, algorithms based on the Lipschitz-regularized α-divergences are more agnostic to heavy-tailed assumptions.

7.3 *Learning attractors of chaotic dynamical systems*

Strange attractor from Lorenz 63 example The Lorenz 63 model is renowned for its strange attractor, which exhibits a complex fractal structure characterized by a non-integer Hausdorff dimension. In this example, we use various generative models to learn the geometric shape of the attractor, without accounting for its underlying dynamics. The target dataset $\mathscr{T}$ for the generative models consists of $N = 5,000$ positions, defined as: $\mathscr{T} = \{\mathbf{x}(t_i) = (x_1(t_i), x_2(t_i), x_3(t_i)) : t_i \sim \text{Unif}([9,900, 10,000])\}_{i=1}^N$ where $(x_1(t_i), x_2(t_i), x_3(t_i))$ is a numerically computed solution trajectory of the Lorenz 63 model with the standard parameter values $a = 10, b = 28, c = 8.3$. The generated samples are represented as $\mathscr{G} = \{\mathbf{y}_i = (y_{1i}, y_{2i}, y_{3i})\}_{i=1}^M$, where M is the number of generated points which does not necessarily match N. We use $M = 10,000$ generated samples across various generative models for this example.

Because the generated samples lack time labels, the dynamics cannot be directly observed. Instead, we consider two standards: (1) measurement of how close the generated particles land on the attractor and (2) characteristic of the fractal structure. These standards are measured by corresponding metrics:

(a) **Mean square sum of the errors (MSE)** between generated samples $\mathbf{y}_i$ and their closest validation sample $\mathbf{v}_i^* = \text{argmin}_{\mathbf{v}_j \in \mathscr{V}} |\mathbf{y}_i - \mathbf{v}_j|$ where the validation dataset is given as $\mathscr{V} = \{\mathbf{v}_j = (v_1(t_j), v_2(t_j), v_3(t_j)) : t_j = 9,900 + 0.01 \cdot j\}_{j=1}^{10,000}$

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^{M} |\mathbf{y}_i - \mathbf{v}_i^*|^2, \tag{7.4}$$

which measures the deviation of generated samples from the attractor trajectory.

(b) Adapted **Correlation dimension** for measuring dimensionality of the space occupied by point clouds of generated samples $\{\mathbf{y}_i\}_{i=1}^M$ without time information. Original correlation dimension is a characteristic measure to distinguish between deterministic chaos and random noise, to detect potential faults [7]. Real correlation dimension for the attractor of Lorenz 63 should be 2.05. We

TABLE 1 *Performance metrics: (i) MSE (7.4) between generated samples and the validation dataset $\mathscr{V}$ that measures how close the generated particles land on the attractor, and (ii) Correlation dimension for $M = 10,000$ generated samples from different generative models*

Model	MSE	Correlation dimension	Computation time (sec)
Lip-$\alpha = 2$ GAN	**0.1240**	**2.00**	491.851
Lip-KL GAN	**0.1226**	**2.01**	505.330
$\alpha = 2$ GAN	0.945	1.99	336.272
KL GAN	0.1612	1.99	486.941
Lip-$\alpha = 2$ GPA	0.2984	1.60	410.385
Lip-KL GPA	0.1369	1.91	398.344
$\alpha = 2$ GPA	-	-	-
KL GPA	-	-	-
OT(W_2) flow	0.6231	2.29	$\geq 60,000$
CNF	1.2674	2.31	$\geq 60,000$
VE SGM	**0.0791**	2.31	2,382.733

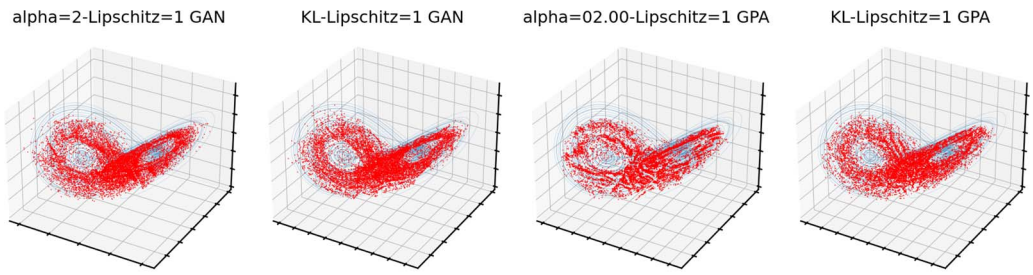
The ground truth correlation dimension measured on the validation dataset $\mathscr{V}$ is 2.04. A higher correlation dimension implies that noise dominates the shape of the attractor. A lower correlation dimension implies that the point clouds are more sparsely populated on the attractor; see, for instance, Lip-$\alpha = 2$ GPA compared to Lip-$\alpha = 2$ GAN in Fig. 6. We do not report the MSE and correlation dimension for $\alpha = 2$ GPA and KL GPA (no Lipschitz regularization) since generated particles diverged in the early stage of training. Although SGM has the smallest MSE, it takes a significant longer time to train, requiring much deeper network architecture (otherwise it does not converge), and it still significantly over-estimates the fractal dimension. See also Fig. 6 for visualizations.

obtained a reference value 2.04 by applying to our validation dataset $\mathscr{V}$ from a selection of the algorithm's parameter radius $r \in [0.7, 1.1]$.

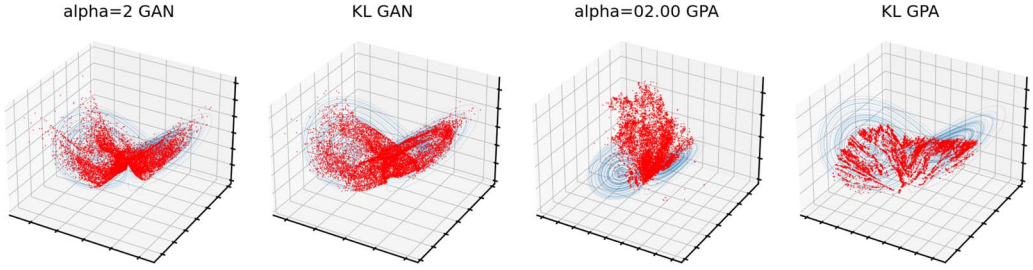
The results can be found in Table 1. The results illustrate that (1) Lipschitz-regularized methods in general capture the attractor and its structure while those without Lipschitz regularization fail; (2) other methods such as OT flows, CNFs and SGMs fail to accurately capture the attractor even they are trained for a longer time with more complicated network architecture. We additionally visualize generated samples in Fig. 6. Similar results when $N = 1,000$ and $M = 2,000$ can be found in the Supplementary Material.

7.4 *Learning distributions supported on low dimensional manifolds*

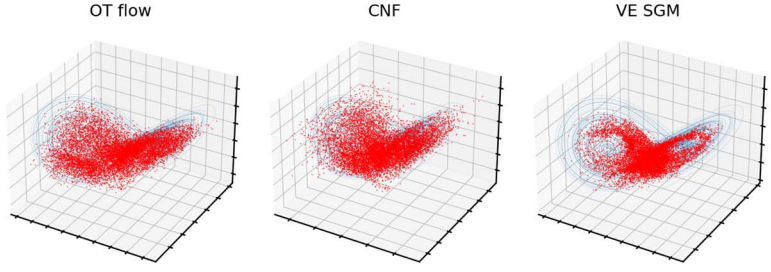
7.4.1 *10D heavy-tailed manifold embedded in 110D* We provide a high-dimensional example adapted from [23]. In this example, a 10D heavy-tailed distribution is embedded in $\mathbb{R}^{110}$. Each of the first 10 axes is drawn from the standard Cauchy distribution $w_i \sim Cauchy$, then powered by a random exponent $t_i \sim \text{Unif}([0.5, 2])$, i.e. $x_i = sign(w_i)|w_i|^{t_i}$ for $i = 1, \ldots, 10$. Values of the remaining axes are set to zero: $x_i = 0$ for $i = 11, \cdots, 110$. In our experiment, we fix the exponents $t_i, i = 1, \cdots 10$, to $(1.31, 0.91, 1.13, 1.76, 0.50, 0.68, 1.50, 1.73, 0.70, 1.36)$. We present two metrics similar to those used in the multivariate distributions example in [23] to demonstrate (a) whether the algorithm can capture the heavy tails in the first 10 dimensions and (b) whether the generated distribution correctly lies on the 10-dimensional plane. For (a), we calculate the averaged L_1 error over the first 10 dimensions between the empirical rCCDF F_v built from a validation dataset consisting of 100K target samples and the empirical



(a) Generated samples from generative models with Lipschitz-regularized α-divergences as learning objectives. $\alpha = 2$-Lipschitz-1 GAN (first), KL-Lipschitz-1 GAN (second), $\alpha = 2$-Lipschitz-1 GPA (third), KL-Lipschitz-1 GPA (fourth)



(b) Generated samples from generative models with un-regularized α-divergences as learning objectives. $\alpha = 2$ GAN (first), KL GAN (second), $\alpha = 2$ GPA (third), KL GPA (fourth). Snapshots from $\alpha = 2$ and KL GPAs are transient and eventually blew up.



(c) Generated samples from generative models with different learning objectives. OT-flow: W_2-reverse KL divergence (left), CNF: reverse KL divergence (center), VE-SGM: W_2-proximal regularized cross-entropy (right)

FIG. 6. Generated samples ($M = 10,000$) of the Lorenz 63 strange attractor from $N = 5,000$ target samples. Lipschitz-regularized methods in general capture the attractor and its dimension while those without Lipschitz regularization fail. Other methods such as OT flows, CNFs and SGMs cannot accurately capture the fractal structure. See Table 1 for error metrics.

rCCDF F_g built from generated samples:

$$L_1(F_v, F_g) = \sum_{i=1}^{20,000} |F_v(z_i) - F_g(z_i)|(z_{i+1} - z_i), \tag{7.5}$$

where z_i are sampled in equi-distance from the interval $[1, 5 \times 10^6]$. For (b), we calculate the Euclidean distance of the generated samples to their projections on the first 10-dimensional subspace which is written as $\sum_{i=11}^{110} \mathbb{E}_{y_i}[\|y_i\|]$ where the orthogonal subspace is represented as zero $[0, \cdots, 0] \in \mathbb{R}^{100}$.

The results in Table 2 verify that models with the Lipschitz-regularized α-divergences as objectives are more agnostic to both heavy-tailed and manifold assumptions.

TABLE 2 *Learning 10D heavy-tailed data embedded in $\mathbb{R}^{110}$ using 10K target samples*

Model	**Heavy-tailed subspace avg L_1 error**	**Orthogonal subspace avg Euclidean distance**
Lip-α GPA	$\mathbf{3.1155e+02}$	$3.4179e+00$
α GPA	$4.9993e+06$	$1.7150e+15$
Lip-α GAN	$\mathbf{3.4645e+02}$	$\mathbf{1.0990e-01}$
α GAN	$4.4994e+06$	$\mathbf{2.4480e-03}$
OT(W_2) flow	$4.9993e+06$	inf
CNF	$4.9993e+06$	inf
VE SGM	$3.6031e+02$	$1.4441e+03$

We report the L_1 error defined in (7.5) averaging over the first 10 dimensions. Generative models without Lipschitz-regularized learning objectives, such as unregularized models or those using W_2-proximal regularization, either fail to capture the heavy tails or fail to capture the manifold. In contrast, Lipschitz-regularized α-divergence enables generative models to learn heavy-tailed distributions even when the tails exhibit different power-law behaviours, i.e. $Q(x_i) \sim |x_i|^{-\beta_i}$ for $i = 1, \cdots, 10$. In addition, the Lipschitz-regularized α-divergence encourages generated samples to lie near the data manifold. The unconstrained discriminator in α-GAN produces large values outside the manifold, forcing the generator to map the source onto the 10D plane. However, the unconstrained α-GAN fails to learn the tails. For further comparison of training objective function values for GANs and GPAs, see Table G1 in Appendix G.

8. Conclusions and discussions

In this paper, we prove that Lipschitz-regularized α-divergences, introduced in previous works, enable robust and stable learning for target distributions with minimal assumptions. In particular, we prove that these divergences are always finite and have a well-defined variational derivative when the first input distribution has a finite first moment. We also prove the sufficient and necessary conditions for the divergence to be finite when both distributions have power-law-decay tails. A first convergence rate of the finite-sample estimations of these divergences on $\mathbb{R}^d$ is proved. As a result, we derive the first sample complexity bounds for the empirical estimations of D_α^L and W_1 with group symmetry on $\mathbb{R}^d$. Numerical simulations further confirm the robustness of these divergences, showing that they significantly improve the learning process across a range of challenging scenarios, such as heavy-tailed distributions or distributions supported on low-dimensional manifolds or fractals.

Some future directions are unexplored in this work. First, it is not clear if there is an optimal α or if the α should be chosen adaptively to make the learning more efficient. Second, the PDE theory of the Lipschitz-regularized gradient flow is not established, and the convergence of the gradient flow is an important topic and may require some new functional inequalities. Lastly, Theorem 7 is not sharp, and a sharp convergence bound will help better understand this class of divergences and further derive better generalization bounds for algorithms based on this class of divergences.

Acknowledgement

The authors would like to thank the anonymous reviewers for their careful reading and constructive feedback, which helped improve the manuscript.

Supplementary data

Supplementary data is available at *Information and Inference: a Journal of the IMA* online.

Funding

National Science Foundation [DMS-2307115 to M.K. and L.R.-B., TRIPODS CISE-1934846 to H.G. and M.K., DMS-2052525 to Z.C. and W.Z., DMS-2140982 to Z.C. and W.Z. and DMS-2244976 to Z.C. and W.Z.]; and Air Force Office of Scientific Research [FA9550-21-1-0354 to Z.C., H.G., M.K. and L.R.-B.].

Data availability statement

All codes in Section 7 can be found at https://github.com/HyeminGu/Proximal_generative_models. The implementation details can be found in the Supplementary Material. The keystroke data is available in [Observations on Typing from 136 Million Keystrokes], at http://userinterfaces.aalto.fi/136Mkeystrokes. All the other datasets can be generated on a local computer.

References

1. Allouche, M., Girard, S. & Gobet, E. (2022) EV-GAN: simulation of extreme events with ReLU neural networks. *J. Mach. Learn.*, **23**, 1–39.
2. Amari, S.-I. & Nagaoka, H. (2000) *Methods of Information Geometry*, vol. **191**. USA: American Mathematical Soc.
3. Bengio, Y., Yao, L., Alain, G. & Vincent, P. (2013) Generalized denoising auto-encoders as generative models. *Advances in Neural Information Processing Systems*, **26**.
4. Birrell, J., Dupuis, P., Katsoulakis, M. A., Pantazis, Y. & Rey-Bellet, L. (2022) (f, Γ)-divergences: interpolating between f-divergences and integral probability metrics. *J. Mach. Learn.*, **23**, 1–70.
5. Birrell, J., Katsoulakis, M., Rey-Bellet, L. & Zhu, W. (2022) Structure-preserving GANs. In *International Conference on Machine Learning*. PMLR, pp. 1982–2020.
6. Bredon, G. E. (1972) *Introduction to Compact Transformation Groups*. New York and London: Academic Press.
7. Caesarendra, W., Kosasih, B., Tieu, K. & Moodie, C. A. (2013) An application of nonlinear feature extraction - a case study for low speed slewing bearing condition monitoring and prognosis. In *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 1713–1718.
8. Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. (2018) Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, **31**.
9. Chen, S., Chewi, S., Li, J., Li, Y., Salim, A. & Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*.
10. Chen, Z., Katsoulakis, M., Rey-Bellet, L. & Zhu, W. (2023) Sample complexity of probability divergences under group symmetry. In *International Conference on Machine Learning*. PMLR, pp. 4713–4734.
11. Csiszar, I. (1963) Eine information's theoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoschen ketten, magyar tud. *Akad. Mat*. **8**, 85–108.
12. Dhakal, V., Feit, A. M., Kristensson, P. O. & Oulasvirta, A. Observations on typing from 136 million keystrokes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18 ,* page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.
13. Ding, R. & Mullhaupt, A. (2023) Empirical squared Hellinger distance estimator and generalizations to a family of α-divergence estimators. *Entropy*, **25**, 612.
14. Donahue, J., Krähenbühl, P. & Darrell, T. *Adversarial feature learning*. In *International Conference on Learning Representations*, 2022.
15. Dupuis, P. & Mao, Y. (2022) Formulation and properties of a divergence used to compare probability measures without absolute continuity. *ESAIM Control Optim. Calc. Var.*, **28**, 10.
16. Farazmand, M. & Sapsis, T. P. (2019) Closed-loop adaptive control of extreme events in a turbulent flow. *Phys. Rev. E (3)*, **100**, 033110.

17. Feder, R. M., Berger, P. & Stein, G. (2020) Nonlinear 3D cosmic web simulation with heavy-tailed generative adversarial networks. *Phys. Rev. D*, **102**, 103504, 1–18.
18. Fournier, N. & Guillin, A. (2015) On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields*, **162**, 707–738.
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014) Generative adversarial nets. In *Advances in Neural Information Processing Systems*, **27**.
20. Gu, H., Birmpa, P., Pantazis, Y., Rey-Bellet, L. & Katsoulakis, M. A. (2024) Lipschitz-regularized gradient flows and generative particle algorithms for high-dimensional scarce data. *SIAM J. Math. Data Sci.* **6**, 1205–1235.
21. Havrda, J. & Charvát, F. (1967) Quantification method of classification processes. Concept of structural a-entropy. *Kybernetika*, **3**, 30–35.
22. Huang, J., Jiao, Y., Li, Z., Liu, S., Wang, Y. & Yang, Y. (2022) An error analysis of generative adversarial networks for learning distributions. *J. Mach. Learn.*, **23**, 1–43.
23. Huster, T., Cohen, J., Lin, Z., Chan, K., Kamhoua, C., Leslie, N. O., Chiang, C.-Y. J. & Sekar, V. (2021) Pareto GAN: Extending the representational power of gans to heavy-tailed distributions. In *International Conference on Machine Learning*, pages 4523–4532. PMLR.
24. Jordan, R., Kinderlehrer, D. & Otto, F. (1998) The variational formulation of the fokker–planck equation. *SIAM J. Math. Anal.*, **29**, 1–17.
25. Kégl, B. (2002) Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems*, **15**.
26. Kingma, D. P. & Welling, M. (2013) Auto-encoding variational bayes arXiv preprint arXiv:1312.6114.
27. Klebanov, L. B. & Kuvaeva, Y. V. (2023) Heavy-tailed probability distributions in social sciences arXiv preprint arXiv:2301.09393.
28. Liang, T. (2021) How well generative adversarial networks learn distributions. *J. Mach. Learn.*, **22**, 1–41.
29. Lotka, A. J. (1926) The frequency distribution of scientific productivity. *J. Wash. Aca. Sci.*, **16**, 317–323.
30. Lynn, C. W., Holmes, C. M. & Palmer, S. E. (2024) Heavy-tailed neuronal connectivity arises from Hebbian self-organization. *Nat. Phys.*, **20**, 1–8.
31. Manole, T. & Niles-Weed, J. (2024) Sharp convergence rates for empirical optimal transport with smooth costs. *Ann. Appl. Probab.*, **34**, 1108–1135.
32. Marzouk, Y., Ren, Z. R., Wang, S. & Zech, J. (2024) Distribution learning via neural differential equations: a nonparametric statistical perspective. *J. Mach. Learn.*, **25**, 1–61.
33. McClelland, J. L., Rumelhart, D. E., and PDP Research Group (1987) *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models, Volume 2*. Cambridge, MA, USA: MIT Press.
34. Mena, G. & Niles-Weed, J. (2019) Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, **32**.
35. Merz, B., Basso, S., Fischer, S., Lun, D., Blöschl, G., Merz, R., Guse, B., Viglione, A., Vorogushyn, S., Macdonald, E., Wietzke, L. and Schumann, A. (2022) Understanding heavy tails of flood peak distributions. *Water Resour. Res.*, **58**, e2021WR030506.
36. Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2018) *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press.
37. Moon, K. & Hero, A. (2014) Multivariate f-divergence estimation with confidence. *Advances in Neural Information Processing Systems*, **27**.
38. Mroueh, Y., Sercu, T. & Raj, A. (April 2019) Sobolev descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2976–2985 ISSN: 2640-3498.
39. Nguyen, X., Wainwright, M. J. & Jordan, M. I. (2010) Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inform. Theory*, **56**, 5847–5861.
40. Nowozin, S., Cseke, B. & R. (2016) Tomioka. f-GAN: training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, **29**.
41. Onken, D., Fung, S. W., Li, X. & Ruthotto, L. (2021) Ot-flow: fast and accurate continuous normalizing flows via optimal transport. *In Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 9223–9232.

42. Otto, F. (2001) The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Comm. Partial Differential Equations*, **26**, 101–174. Taylor & Francis.
43. Pareto, V., *Cours d'économie politique*, vol. **1**. Lausanne: Editor: F. Rouge, Librairie de l'Universite (1896). - 430p. Rue Haldimand.
44. Pope, P., Zhu, C., Abdelkader, A., Goldblum, M. & Goldstein, T. (2020) The intrinsic dimension of images and its impact on learning. *In International Conference on Learning Representations*.
45. Rubenstein, P., Bousquet, O., Djolonga, J., Riquelme, C. & Tolstikhin, I. O. (2019) Practical and consistent estimation of f-divergences. In *Advances in Neural Information Processing Systems*, **32**.
46. Sasaki, H., Su, F.-H., Tanimoto, T. & Sethumadhavan, S. (2017) Why do programs have heavy tails? In*2017 IEEE International Symposium on Workload Characterization (IISWC)*. IEEE, pp. 135–145.
47. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. (2021) Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. .
48. Sreekumar, S., Zhang, Z. & Goldfeld, Z. (2021) Non-asymptotic performance guarantees for neural estimation of f-divergences. In *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 3322–3330.
49. Tahmasebi, B. & Jegelka, S. (2024) Sample complexity bounds for estimating probability divergences under invariances. *In Forty-first International Conference on Machine Learning*, PMLR **235**:47396–47417.
50. van der Vaart, A. W. & Wellner, J. A. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.
51. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. (2008) Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*, pages 1096–1103.
52. Wang, Z. O., Baptista, R., Marzouk, Y., Ruthotto, L. & Verma, D. (2023) Efficient neural network approaches for conditional optimal transport with applications in bayesian inference. *SIAM J. Sci. Comput.*, **47**, C979–C1005.
53. Wiese, M., Knobloch, R., Korn, R. & Kretschmer, P. (2020) Quant GANs: deep generation of financial time series. *Quant. Finance*, **20**, 1419–1440.
54. Zhang, B. J. & Katsoulakis, M. A. (2023) A mean-field games laboratory for generative modeling. arXiv preprint arXiv:2304.13534 (2023).
55. Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA, USA: Addison-Wesley Press, Inc.

A. Notation for the proofs

We denote by $A \lesssim B$ if there are some $c, d > 0$, such that $A \leq cB + d$; and $A \asymp B$ if both $A \lesssim B$ and $B \lesssim A$ hold. For a bounded set $\Omega \subset \mathbb{R}^d$, $\operatorname{diam}(\Omega) = \sup_{x,y\in\Omega} \|x - y\|_2$, where $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^d$. Moreover, given a probability density $p(x)$, we use $M_r(p)$ to denote the rth moment of $p(x)$. For convenience, we will abuse notation and use symbols p, q and P, Q, to represent probability distributions as well as the density functions associated with them. Whether a character refers to a probability distribution or a density should be clear from the context.

B. Additional lemma of Theorem 2

For the Lipschitz-regularized KL-divergence, we have the following lemma similar to Lemma 1.

Lemma B1. For the KL case, i.e. $f^*_{\mathrm{KL}}(y) = e^{y-1}$ and any non-negative measures P and Q defined on some bounded $\Omega \subset \mathbb{R}^d$ with non-zero integrals, $\Gamma = \mathrm{Lip}_L(\Omega)$, we have

$$\sup_{\gamma\in\Gamma} \left\{ \int_\Omega \gamma(x)\,\mathrm{d}P - \int_\Omega f^*_{\mathrm{KL}}[\gamma(x)]\,\mathrm{d}Q \right\} = \sup_{\gamma\in\mathscr{F}} \left\{ \int_\Omega \gamma(x)\,\mathrm{d}P - \int_\Omega f^*_{\mathrm{KL}}[\gamma(x)]\,\mathrm{d}Q \right\}, \tag{B1}$$

where

$$\mathscr{F} = \left\{ \gamma \in \mathrm{Lip}_L(\Omega) : \ln \frac{\int_\Omega \mathrm{d}P}{\int_\Omega \mathrm{d}Q} + 1 - L \cdot \mathrm{diam}(\Omega) \le \gamma \le \ln \frac{\int_\Omega \mathrm{d}P}{\int_\Omega \mathrm{d}Q} + 1 + L \cdot \mathrm{diam}(\Omega) \right\}.$$

Proof. For any fixed $\gamma \in \Gamma$, define

$$h(v) = \int_\Omega (\gamma(x) + v)\, \mathrm{d}P - \int_\Omega f_{\mathrm{KL}}^*[\gamma(x) + v]\, \mathrm{d}Q.$$

Since $\sup_{x\in\Omega} \gamma(x) - \inf_{x\in\Omega} \gamma(x) \le L \cdot \mathrm{diam}(\Omega)$, interchanging the integration with differentiation is allowed by the dominated convergence theorem:

$$h'(v) = \int_\Omega \mathrm{d}P - \int_\Omega f_{\mathrm{KL}}^{*\prime}(\gamma + v)\, \mathrm{d}Q.$$

If $\inf_{x\in\Omega} \gamma(x) > \ln \frac{\int \mathrm{d}P}{\int \mathrm{d}Q} + 1$, then $h'(0) < 0$. So there exists some $v_0 < 0$ such that $h(v_0) > h(0)$. This indicates the supremum on the left side of (B1) is attained only if $\sup_{x\in\Omega} \gamma(x) \le \ln \frac{\int \mathrm{d}P}{\int \mathrm{d}Q} + 1 + L \cdot \mathrm{diam}(\Omega)$. On the other hand, if $\sup_{x\in\Omega} \gamma(x) < \ln \frac{\int \mathrm{d}P}{\int \mathrm{d}Q} + 1$, then $h'(0) > 0$. So there exists some $v_0 > 0$ such that $h(v_0) > h(0)$. This indicates that the supremum on the left side of (B1) is attained only if $\inf_{x\in\Omega} \gamma(x) \ge \ln \frac{\int \mathrm{d}P}{\int \mathrm{d}Q} + 1 - L \cdot \mathrm{diam}(\Omega)$. □

C. Proof of Theorem 3

Proof of Theorem 3. The existence and uniqueness of $\gamma^\star$ follow from Theorem 4.9 in [15] and Theorem 25 in [4]. We extend $\gamma^\star$ from $\mathrm{supp}(P) \cup \mathrm{supp}(Q)$ to all of $\mathbb{R}^d$ by

$$\hat{\gamma}(y) = \sup_{x \in \mathrm{supp}(P)\cup\mathrm{supp}(Q)} \{\gamma^\star(x) + L\,|x - y|\}. \tag{C1}$$

And it is a well-known result (e.g. see the proof of Lemma 2.3 in [20]) that $\hat{\gamma}$ is L-Lipschitz continuous on $\mathbb{R}^d$ and

$$\hat{\gamma} = \sup_h \{h(x) : h \in \mathrm{Lip}_L(\mathbb{R}^d), h(y) = \gamma^\star(y), \forall y \in \mathrm{supp}(P) \cup \mathrm{supp}(Q)\}. \tag{C2}$$

We need to show that

$$\liminf_{\epsilon\to 0+} \frac{1}{\epsilon} \left(D_\alpha^L(P + \epsilon\rho \| Q) - D_\alpha^L(P\|Q) \right) \ge \int \hat{\gamma}\, \mathrm{d}\rho, \tag{C3}$$

and

$$\limsup_{\epsilon\to 0+} \frac{1}{\epsilon} \left(D_\alpha^L(P + \epsilon\rho \| Q) - D_\alpha^L(P\|Q) \right) \le \int \hat{\gamma}\, \mathrm{d}\rho. \tag{C4}$$

If $P + \epsilon\rho \in \mathscr{P}_1(\mathbb{R}^d)$, then by Theorem 2, $D_\alpha^L(P + \epsilon\rho\|Q) < \infty$ and thus we have

$$\begin{aligned} D_\alpha^L(P + \epsilon\rho\|Q) &= \sup_{\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)} \left(\mathbb{E}_{P+\epsilon\rho}[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right) \\ &\geq \mathbb{E}_{P+\epsilon\rho}[\hat{\gamma}] - \mathbb{E}_Q[f_\alpha^*(\hat{\gamma})] \\ &= \epsilon \int_{\mathbb{R}^d} \hat{\gamma}\, \mathrm{d}\rho + \mathbb{E}_P[\hat{\gamma}] - \mathbb{E}_Q[f_\alpha^*(\hat{\gamma})] \\ &= \epsilon \int_{\mathbb{R}^d} \hat{\gamma}\, \mathrm{d}\rho + D_\alpha^L(P\|Q). \end{aligned}$$

Thus, we have

$$\liminf_{\epsilon \to 0+} \frac{1}{\epsilon} \left(D_\alpha^L(P + \epsilon\rho\|Q) - D_\alpha^L(P\|Q)\right) \geq \int \hat{\gamma}\, \mathrm{d}\rho. \tag{C5}$$

To prove the other direction, we define $F(\epsilon) = D_\alpha^L(P + \epsilon\rho\|Q)$. Then by Theorem 18 in [4], $F(\epsilon)$ is convex, lower semi-continuous and finite on $[0, \epsilon_0]$ for some $\epsilon_0 > 0$. Due to the convexity of F, it is differentiable on $(0, \epsilon_0)$ except for a countable number of points. If $\hat{\gamma}_\epsilon$ is the optimizer for $D_\alpha^L(P+\epsilon\rho\|Q)$, similar to (C5), we have for $\delta > 0$ sufficiently small

$$D_\alpha^L(P + (\epsilon + \delta)\rho\|Q) - D_\alpha^L(P + \epsilon\rho\|Q) \geq \delta \int \hat{\gamma}_\epsilon\, \mathrm{d}\rho, \tag{C6}$$

and

$$D_\alpha^L(P + (\epsilon - \delta)\rho\|Q) - D_\alpha^L(P + \epsilon\rho\|Q) \geq -\delta \int \hat{\gamma}_\epsilon\, \mathrm{d}\rho. \tag{C7}$$

If F is differentiable at ϵ, this implies that

$$\begin{aligned} \int \hat{\gamma}_\epsilon\, \mathrm{d}\rho &\leq \lim_{\delta \to 0} \frac{1}{\delta} \left(D_\alpha^L(P + (\epsilon + \delta)\rho\|Q) - D_\alpha^L(P + \epsilon\rho\|Q)\right) \\ &= F'(\epsilon) \\ &= \lim_{\delta \to 0} \frac{1}{\delta} \left(D_\alpha^L(P + \epsilon\rho\|Q) - D_\alpha^L(P + (\epsilon - \delta)\rho\|Q)\right) \\ &\leq \int \hat{\gamma}_\epsilon\, \mathrm{d}\rho. \end{aligned}$$

Consequently,

$$F'(\epsilon) = \int \hat{\gamma}_\epsilon\, \mathrm{d}\rho. \tag{C8}$$

Let $F'_+(0)$ be the right derivative at $\epsilon = 0$, i.e. $F'_+(0) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon}(F(\epsilon) - F(0))$. By convexity, for any sequence ϵ_n such that F is differentiable at ϵ_n and $\epsilon_n \searrow 0$, we have

$$F'_+(0) = \lim_{n \to \infty} F'(\epsilon_n) = \lim_{n \to \infty} \int \hat{\gamma}_{\epsilon_n}\, \mathrm{d}\rho. \tag{C9}$$

We write $\mathbb{R}^d = \cup_{m\in\mathbb{N}} K_m$ with $K_m \subset \mathbb{R}^d$ being a compact set and $K_m \subset K_{m+1}$. The optimizers $\hat{\gamma}_{\epsilon_n}$ are unique. Moreover, by Lemma E3, they satisfy $\left|\hat{\gamma}_{\epsilon_n}(x)\right| \le L(|x|+R) + M_n$, where

$$M_n = \inf_M \left\{ (M + LR) + L\int_{\mathbb{R}^d} |x|\,\mathrm{d}P + \epsilon_n L \int_{\mathbb{R}^d} |x|\,\mathrm{d}\rho < f_\alpha^*(M - 3LR)\int_{|x|<2R} \mathrm{d}Q \right\} \tag{C10}$$

where $R > 0$ is fixed for all n such that $\int_{|x|<2R} \mathrm{d}Q > 0$. Thus, by the linear dependence on ϵ_n on the left side inside the infimum in (C10), we have $M_n \le \overline{M}$ for all sufficiently large n. Therefore, the sequence $\{\hat{\gamma}_{\epsilon_n}\}$ is equibounded and equicontinuous on K_m. By the Arzelà-Ascoli theorem, there exists a subsequence of $\hat{\gamma}_{\epsilon_n}$ that converges uniformly in K_m. Using diagonal argument, by taking subsequences sequentially along $\{K_m\}_{m\in\mathbb{N}}$ we conclude there exists a subsequence such that $\hat{\gamma}_{\epsilon_{n_k}}$ converges uniformly in any K_m and thus $\hat{\gamma}_{\epsilon_{n_k}}$ converges pointwise in $\mathbb{R}^d$. Let $\hat{\gamma}_0$ be the limit, then $\hat{\gamma}_0$ is L-Lipschitz due to the uniform convergence of L-Lipschitz functions. For simplicity, we also denote by $\hat{\gamma}_{\epsilon_n}$ the convergent subsequence. Thus, given $\rho_\pm \in \mathscr{P}_1(\mathbb{R}^d)$, we have by the dominated convergence theorem,

$$F'_+(0) = \lim_{n\to\infty} \int_{\mathbb{R}^d} \hat{\gamma}_{\epsilon_n}\,\mathrm{d}\rho = \int_{\mathbb{R}^d} \hat{\gamma}_0\,\mathrm{d}\rho. \tag{C11}$$

By the lower semi-continuity of $D_\alpha^L(\cdot\|Q)$, we have

$$\begin{aligned}
D_\alpha^L(P\|Q) &\le \liminf_{n\to\infty} D_\alpha^L(P + \epsilon_n\rho\|Q) \\
&= \liminf_{n\to\infty} \left\{ \mathbb{E}_{P+\epsilon_n\rho}[\hat{\gamma}_{\epsilon_n}] - \mathbb{E}_Q[f_\alpha^*(\hat{\gamma}_{\epsilon_n})] \right\} \\
&= \lim_{n\to\infty} \mathbb{E}_{P+\epsilon_n\rho}[\hat{\gamma}_{\epsilon_n}] - \limsup_{n\to\infty} \mathbb{E}_Q[f_\alpha^*(\hat{\gamma}_{\epsilon_n})] \\
&= \mathbb{E}_P[\hat{\gamma}_0] - \limsup_{n\to\infty} \mathbb{E}_Q[f_\alpha^*(\hat{\gamma}_{\epsilon_n})] \\
&\le \mathbb{E}_P[\hat{\gamma}_0] - \mathbb{E}_Q[f_\alpha^*(\hat{\gamma}_0)] \\
&\le D_\alpha^L(P\|Q),
\end{aligned}$$

where in the third equality we use the dominated convergence theorem, and in the second-to-last inequality we apply the Fatou's lemma. Thus, we have $\hat{\gamma}_0 = \hat{\gamma}$ P, Q– a.s., and $\hat{\gamma}_0 \le \hat{\gamma}$ for all $x \in \mathbb{R}^d$. The latter is true since $\hat{\gamma}$ is the Lipschitz extension of $\gamma_0^\star$ by (C1), and (C2) guarantees that $\hat{\gamma}$ is the supremum of all the L-Lipschitz functions whose restriction on $\mathrm{supp}(P) \cup \mathrm{supp}(Q)$ is equal to $\hat{\gamma}_0$. It can be shown (as in the beginning of the proof of Theorem 1 in [20]) that ρ_- is absolutely continuous with respect to P, then we have

$$F'_+(0) = \int \gamma_0^\star\,\mathrm{d}\rho = \int \gamma_0^\star\,\mathrm{d}\rho_+ - \int \gamma_0^\star\,\mathrm{d}\rho_- = \int \gamma_0^\star\,\mathrm{d}\rho_+ - \int \hat{\gamma}\,\mathrm{d}\rho_- \le \int \hat{\gamma}\,\mathrm{d}\rho. \tag{C12}$$

Thus, (C4) is proved. □

D. Proofs of Theorems 4 and 5

Proof of Theorem 4. **1. Sufficiency**. Let $\Gamma = \mathrm{Lip}_L(\mathbb{R}^d)$, and we have

$$\begin{aligned} D_\alpha^L(P\|Q) &= \sup_{\gamma\in\Gamma}\left\{\int \gamma(x)p(x)\,\mathrm{d}x - \int f_\alpha^*[\gamma(x)]q(x)\,\mathrm{d}x\right\} \\ &\le \sup_{\gamma\in\mathrm{Lip}_L(\|x\|<R)}\left\{\int_{\|x\|<R}\gamma(x)p(x)\,\mathrm{d}x - \int_{\|x\|<R} f_\alpha^*[\gamma(x)]q(x)\,\mathrm{d}x\right\} \\ &\quad + \sup_{\gamma\in\mathrm{Lip}_L(\|x\|\ge R)}\left\{\int_{\|x\|\ge R}\gamma(x)p(x)\,\mathrm{d}x - \int_{\|x\|\ge R} f_\alpha^*[\gamma(x)]q(x)\,\mathrm{d}x\right\} \\ &:= I_1 + I_2. \end{aligned}$$

For I_1, by Lemma 1, we have

$$I_1 \le C\int_{\|x\|<R} p(x)\,\mathrm{d}x + \left(\alpha^{-1}(\alpha-1)^{\frac{\alpha}{\alpha-1}}C^{\frac{\alpha}{\alpha-1}} + \alpha^{-1}(\alpha-1)^{-1}\right)\int_{\|x\|<R} q(x)\,\mathrm{d}x < \infty,$$

where $C = (\alpha-1)^{-1}\left(\frac{\int_{\|x\|<R}p(x)\,\mathrm{d}x}{\int_{\|x\|<R}q(x)\,\mathrm{d}x}\right)^{\alpha-1} + 2LR$.

For I_2, we have

$$\int_{\|x\|\ge R}\gamma(x)p(x)\,\mathrm{d}x - \int_{\|x\|\ge R} f_\alpha^*[\gamma(x)]q(x)\,\mathrm{d}x = \int_{\|x\|\ge R} p(x)\left(\gamma(x) - f_\alpha^*[\gamma(x)]\frac{q(x)}{p(x)}\right)\mathrm{d}x.$$

(1) If $d < \beta_1 \le d+1$ and $\beta_2 - \beta_1 < \frac{\beta_1 - d}{\alpha-1}$:

Note that the set of bounded L-Lipschitz functions on $\{x : \|x\| \ge R\}$ is a subset of $\mathscr{M}_b(x : \|x\| \ge R)$, and the supremum over all the L-Lipschitz functions can be bounded by taking the supremum over all the measurable functions. Moreover, we can solve for the optimal $\hat{\gamma}(x)$ that maximizes $\gamma(x) - f_\alpha^*[\gamma(x)]\frac{q(x)}{p(x)}$ within the class of measurable functions: the stationary point of $\gamma(x) - f_\alpha^*[\gamma(x)]\frac{q(x)}{p(x)}$ in γ for every x provides $\hat{\gamma}(x) = \frac{1}{\alpha-1}\left(\frac{p(x)}{q(x)}\right)^{\alpha-1}$. Therefore, we have

$$\begin{aligned} &\sup_{\gamma\in\mathrm{Lip}_L(x:\|x\|\ge R)}\int_{\|x\|\ge R} p(x)\left(\gamma(x) - f_\alpha^*[\gamma(x)]\frac{q(x)}{p(x)}\right)\mathrm{d}x \\ &\le \int_{\|x\|\ge R} p(x)\left(\hat{\gamma}(x) - f_\alpha^*[\hat{\gamma}(x)]\frac{q(x)}{p(x)}\right)\mathrm{d}x \\ &= \int_{\|x\|\ge R}\frac{1}{\alpha(\alpha-1)}\left(\left[\frac{p(x)}{q(x)}\right]^\alpha - 1\right)q(x)\,\mathrm{d}x \\ &\asymp \int_{\|x\|\ge R}\|x\|^{\alpha(\beta_2-\beta_1)-\beta_2}\,\mathrm{d}x < \infty, \end{aligned}$$

since $\alpha(\beta_2-\beta_1)-\beta_2 = (\alpha-1)(\beta_2-\beta_1) - \beta_1 < -d$.

(2) If $\beta_1 > d+1$: the proof follows that of Theorem 2.

2. Necessity.
Suppose $\beta_1 \le d+1$ and $\beta_2 - \beta_1 \ge \frac{\beta_1 - d}{\alpha - 1}$. We split $\beta_2 - \beta_1 \ge \frac{\beta_1 - d}{\alpha - 1}$ into two cases.

(1) If $\beta_2 - \beta_1 \ge \frac{1}{\alpha - 1}$:

Let $\widehat{\gamma}(x) = \tau \|x\|$, where $\tau \in (0, L]$ is to be determined. Then we have $\widehat{\gamma} \in \mathrm{Lip}_L(\mathbb{R}^d)$. Using this $\widehat{\gamma}$, we have

$$D_\alpha^L(P\|Q) \ge \int \widehat{\gamma}(x)p(x)\,\mathrm{d}x - \int f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x$$
$$= \int_{\|x\|<R} \widehat{\gamma}(x)p(x) - f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x + \int_{\|x\|\ge R} \widehat{\gamma}(x)p(x) - f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x.$$

It is straightforward that the first integral over $\|x\| < R$ is finite. For the latter one, we have

$$\int_{\|x\|\ge R} \widehat{\gamma}(x)p(x)\,\mathrm{d}x - \int_{\|x\|\ge R} f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x \gtrsim \int_{\|x\|\ge R} \left(\tau\,\|x\|^{1-\beta_1} - \tau^{\frac{\alpha}{\alpha-1}}\,\|x\|^{\frac{\alpha}{\alpha-1}-\beta_2}\right)\mathrm{d}x.$$

We need to show the right-hand side is infinite. First, since $\frac{\alpha}{\alpha-1} > 1$, we can choose τ sufficiently small such that $\tau > \tau^{\frac{\alpha}{\alpha-1}}$. Moreover, by the assumption, we have $1-\beta_1 \ge -d$ and $\frac{\alpha}{\alpha-1} - \beta_2 \le 1 - \beta_1$, so that we have

$$\int_{\|x\|\ge R} \left(\tau\,\|x\|^{1-\beta_1} - \tau^{\frac{\alpha}{\alpha-1}}\,\|x\|^{\frac{\alpha}{\alpha-1}-\beta_2}\right)\mathrm{d}x = \infty,$$

and thus $D_\alpha^L(P\|Q) = \infty$.

(2) If $\frac{\beta_1 - d}{\alpha - 1} \le \beta_2 - \beta_1 < \frac{1}{\alpha - 1}$:

Define

$$\widehat{\gamma}(x) = \begin{cases} \tau R^{(\alpha-1)(\beta_2-\beta_1)}, & \text{if } \|x\| < R; \\ \tau\,\|x\|^{(\alpha-1)(\beta_2-\beta_1)}, & \text{if } \|x\| \ge R, \end{cases}$$

where $\tau \in (0, L]$ is to be determined. Since in this case we have $(\beta_2 - \beta_1)(\alpha - 1) < 1$, we have $\widehat{\gamma}(x) \in \mathrm{Lip}_L(\mathbb{R}^d)$ if we pick R sufficiently large which is independent of $\tau \le L$. Using this $\widehat{\gamma}(x)$, we have

$$D_\alpha^L(P\|Q) \ge \int \widehat{\gamma}(x)p(x)\,\mathrm{d}x - \int f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x$$
$$= \int_{\|x\|<R} \widehat{\gamma}(x)p(x) - f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x + \int_{\|x\|\ge R} \widehat{\gamma}(x)p(x) - f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x.$$

By the definition of $\widehat{\gamma}$, we know that the first integral over $\|x\| < R$ is finite. For the latter one, we have in this case

$$\int_{\|x\|\ge R} \widehat{\gamma}(x)p(x)\,\mathrm{d}x - \int_{\|x\|\ge R} f_\alpha^*[\widehat{\gamma}(x)]q(x)\,\mathrm{d}x$$
$$\gtrsim \int_{\|x\|\ge R} \left(\tau\,\|x\|^{(\alpha-1)(\beta_2-\beta_1)-\beta_1} - \tau^{\frac{\alpha}{\alpha-1}}\,\|x\|^{(\alpha-1)(\beta_2-\beta_1)-\beta_1}\right)\mathrm{d}x.$$

We show the right-hand side is infinite. Again, we can choose τ sufficiently small such that $\tau > \tau^{\frac{\alpha}{\alpha-1}}$. On the other hand, by the assumption in this case, we have $(\alpha-1)(\beta_2-\beta_1)-\beta_1 \geq -d$, so that we have

$$\int_{\|x\|\geq R} \left(\tau \|x\|^{(\alpha-1)(\beta_2-\beta_1)-\beta_1} - \tau^{\frac{\alpha}{\alpha-1}} \|x\|^{(\alpha-1)(\beta_2-\beta_1)-\beta_1}\right) \mathrm{d}x = \infty,$$

hence $D_\alpha^L(P\|Q) = \infty$. □

Proof of Theorem 5. Same as in the beginning of the proof of Theorem 4, we can split $D_{\mathrm{KL}}^L(P\|Q)$ into I_1 and I_2, where I_1 is bounded by Lemma B1 with appropriate R.

For I_2, we have

$$\begin{aligned}
&\sup_{\gamma\in \mathrm{Lip}_L(x:\|x\|\geq R)} \int_{\|x\|\geq R} \gamma(x)p(x)\,\mathrm{d}x - \int_{\|x\|\geq R} f_{\mathrm{KL}}^*[\gamma(x)]q(x)\,\mathrm{d}x \\
&\leq \sup_{\gamma\in \mathscr{M}_b(x:\|x\|\geq R)} \int_{\|x\|\geq R} \gamma(x)p(x)\,\mathrm{d}x - \int_{\|x\|\geq R} f_{\mathrm{KL}}^*[\gamma(x)]q(x)\,\mathrm{d}x \\
&= \int_{\|x\|\geq R} \ln\frac{p(x)}{q(x)} p(x)\,\mathrm{d}x \\
&\asymp \int_{\|x\|\geq R} \|x\|^{-\beta_1} \ln\|x\|\,\mathrm{d}x < \infty,
\end{aligned}$$

since $\beta_1 > d$ and the equality is due to the dual formula of KL divergence. □

Proof of Corollary 6. Note the change-of-variable formula

$$\int_{\mathbb{R}^d} \gamma(y)\,\mathrm{d}p_{\mathscr{M}}(y) = \int_{\mathbb{R}^{d^*}} (\gamma\circ\varphi)(x)\cdot p(x)\,\mathrm{d}x, \text{ (similarly for } q_{\mathscr{M}} \text{ and } q)$$

and $\gamma\circ\varphi$ is an LL^*-Lipschitz function on $\mathbb{R}^{d^*}$ for any $\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)$. Then the proof of Theorem 4 can be followed. □

E. Proofs of results in Section 5

To prove Theorem 7, we need a few lemmas. Let $x_1, x_2, \ldots, x_m \in \mathbb{R}^d$ be i.i.d. samples of distribution P, and P_m be the corresponding empirical distributions. We define $L_2(P_m)$ the metric between any functions f, g as $L_2(P_m)(f,g) = \sqrt{\frac{1}{m}\sum_{i=1}^m |f(x_i)-g(x_i)|^2}$.

Lemma E1 (Metric entropy with empirical measures). Let $\mathscr{F}$ be a class of real-valued functions on $\mathbb{R}^d$ and $0 \in \mathscr{F}$. Let $\xi = \{\xi_1, \xi_2, \ldots, \xi_m\}$ be a set of independent random variables that take values on $\{-1, 1\}$ with equal probabilities (also known as Rademacher variables). Suppose $X = \{x_1, x_2, \ldots, x_m\} \subset \mathbb{R}^d$ are i.i.d. samples of distribution P, then we have

$$\mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i f(x_i)\right| \leq \inf_{0<\theta<M_X} \left(4\theta + \frac{12}{\sqrt{m}} \int_\theta^{M_X} \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta, L_2(P_m))}\,\mathrm{d}\delta\right),$$

where $M_X = \sup_{f\in\mathscr{F}} \sqrt{\frac{1}{m}\sum_{i=1}^m |f(x_i)|^2}$.

Proof. Let $N \in \mathbb{N}$ be an arbitrary positive integer and $\delta_k = M_X \cdot 2^{-(k-1)}$, $k = 1, \ldots, N$, with $M_X = \sup_{f\in\mathscr{F}} \sqrt{\frac{1}{m}\sum_{i=1}^m \left|f(x_i)\right|^2}$. Let V_k be the cover achieving $\mathscr{N}(\mathscr{F}, \delta_k, L_2(P_m))$, and denote $\left|V_k\right| = \mathscr{N}(\mathscr{F}, \delta_k, L_2(P_m))$. For any $f \in \mathscr{F}$, let $\pi_k(f) \in V_k$, such that

$$\sqrt{\frac{1}{m}\sum_{i=1}^m \left|f(x_i) - \pi_k(f)(x_i)\right|^2} \le \delta_k. \tag{E1}$$

We have

$$\begin{aligned}
&\mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i f(x_i)\right| \\
&\le \mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \left(f(x_i) - \pi_N(f)(x_i)\right)\right| + \sum_{j=1}^{N-1} \mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \left(\pi_{j+1}(f)(x_i) - \pi_j(f)(x_i)\right)\right| \\
&\quad + \mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \pi_1(f)(x_i)\right|.
\end{aligned}$$

For the third term, observe that it suffices to take $V_1 = \{0\}$ so that $\pi_1(f)$ is the zero function and the third term vanishes. The first term can be bounded using Cauchy–Schwartz inequality as

$$\begin{aligned}
\mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \left(f(x_i) - \pi_N(f)(x_i)\right)\right| &\le \frac{1}{m}\sqrt{\sum_{i=1}^m \mathbb{E}_\xi(\xi_i)^2}\sqrt{\sup_{f\in\mathscr{F}} \sum_{i=1}^m \left(f(x_i) - \pi_N(f)(x_i)\right)^2} \\
&\le \delta_N.
\end{aligned}$$

To handle the middle term, for each j, let $W_j = \{\pi_{j+1}(f) - \pi_j(f) : f \in \mathscr{F}\}$. We have $\left|W_j\right| \le \left|V_{j+1}\right|\left|V_j\right| \le \left|V_{j+1}\right|^2$, then

$$\sum_{j=1}^{N-1} \mathbb{E}_\xi \sup_{f\in\mathscr{F}} \left|\frac{1}{m}\sum_{i=1}^m \xi_i \left(\pi_{j+1}(f)(x_i) - \pi_j(f)(x_i)\right)\right| = \sum_{j=1}^{N-1} \mathbb{E}_\xi \sup_{w\in W_j} \left|\frac{1}{m}\sum_{i=1}^m \xi_i w(x_i)\right|.$$

Moreover, we have

$$\begin{aligned}
&\sup_{w\in W_j} \sqrt{\sum_{i=1}^m w(x_i)^2} \\
&= \sup_{f\in\mathscr{F}} \sqrt{\sum_{i=1}^m \left(\pi_{j+1}(f)(x_i) - \pi_j(f)(x_i)\right)^2} \\
&\le \sup_{f\in\mathscr{F}} \sqrt{\sum_{i=1}^m \left(\pi_{j+1}(f)(x_i) - f(x_i)\right)^2} + \sup_{f\in\mathscr{F}} \sqrt{\sum_{i=1}^m \left(f(x_i) - \pi_j(f)(x_i)\right)^2} \\
&\le \sqrt{m}\delta_{j+1} + \sqrt{m}\cdot\delta_j \\
&= 3\sqrt{m}\delta_{j+1}.
\end{aligned}$$

By the Massart finite class lemma (see, e.g. [36]), we have

$$\mathbb{E}_{\xi} \sup_{w \in W_j} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i w(x_i) \right| \leq \frac{3\sqrt{m}\delta_{j+1}\sqrt{2 \ln \left| W_j \right|}}{m} \leq \frac{6\delta_{j+1}\sqrt{\ln \left| V_{j+1} \right|}}{\sqrt{m}}.$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\xi} \sup_{f \in \mathscr{F}} \left| \frac{1}{m} \sum_{i=1}^{m} \xi_i f(x_i) \right| &\leq \delta_N + \frac{6}{\sqrt{m}} \sum_{j=1}^{N-1} \delta_{j+1} \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta_{j+1}, L_2(P_m))} \\
&\leq \delta_N + \frac{12}{\sqrt{m}} \sum_{j=1}^{N} (\delta_j - \delta_{j+1}) \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta_j, L_2(P_m))} \\
&\leq \delta_N + \frac{12}{\sqrt{m}} \int_{\delta_{N+1}}^{M_X} \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta, L_2(P_m))}\, \mathrm{d}\delta.
\end{aligned}$$

Finally, select any $\theta \in (0, M_X)$ and let N be the largest integer with $\delta_{N+1} > \theta$, (implying $\delta_{N+2} \leq \theta$ and $\delta_N = 4\delta_{N+2} \leq 4\theta$), so that

$$\delta_N + \frac{12}{\sqrt{m}} \int_{\delta_{N+1}}^{M_X} \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta, L_2(P_m))}\, \mathrm{d}\delta \leq 4\theta + \frac{12}{\sqrt{m}} \int_{\theta}^{M_X} \sqrt{\ln \mathscr{N}(\mathscr{F}, \delta, L_2(P_m))}\, \mathrm{d}\delta.$$

□

LEMMA E2. Suppose P_m is the empirical distribution of $P \in \mathscr{P}_1(\mathbb{R}^d)$, and $\Lambda = \frac{1}{m} \sum_{i=1}^{m} \|x\|^{\hat{\beta}}$ with $1 \leq \hat{\beta} < \beta - d$, then for $1 \leq z \leq \hat{\beta}$, we have

$$\mathbb{E}_{P_m} \|x\|^z \leq \Lambda + 1.$$

Proof. Note that $\|x\|^z \leq \max\{1, \|x\|^{\hat{\beta}}\} \leq 1 + \|x\|^{\hat{\beta}}$, so we have the bound. □

We provide the following lemma that sets up a landmark for the magnitude of the Lipschitz functions under the supremum.

LEMMA E3. Suppose $\alpha > 1$, and $P \in \mathscr{P}_1(\mathbb{R}^d)$. Let $M(\gamma) = \sup_{\|x\|=R} |\gamma(x)|$, then there exists $\overline{M}$ that depends on P, Q, L and R, such that

$$D_{\alpha}^{L}(P\|Q) = \sup_{\substack{\gamma \in \mathrm{Lip}_L(\mathbb{R}^d) \\ M(\gamma) \leq \overline{M}}} \left\{ \mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_{\alpha}^{*}(\gamma)] \right\},$$

where

$$\overline{M} = \inf \left\{ \hat{M} : (M(\gamma) + LR) \int \mathrm{d}P + L \int \|x\|\, \mathrm{d}P - f_{\alpha}^{*}(M(\gamma) - 3LR) \int_{\|x\| < 2R} \mathrm{d}Q < 0, \forall M(\gamma) > \hat{M} \right\}.$$

Proof. For any $\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)$, let

$$J_1 := \int_{\|x\|<R} \gamma(x)\,\mathrm{d}P - \int_{\|x\|<R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q, \quad J_2 := \int_{\|x\|\geq R} \gamma(x)\,\mathrm{d}P - \int_{\|x\|\geq R} f_\alpha^*[\gamma(x)]\,\mathrm{d}Q,$$

then

$$\int \gamma(x)\,\mathrm{d}P - \int f_\alpha^*[\gamma(x)]\,\mathrm{d}Q = J_1 + J_2.$$

We have for any $\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)$,

$$\begin{aligned} J_1 &\leq \int_{\|x\|<R} (M(\gamma) + LR)\,\mathrm{d}P - \int_{\|x\|<R} f_\alpha^*(M(\gamma) - 3LR)\,\mathrm{d}Q \\ &= (M(\gamma) + LR) \cdot \int_{\|x\|<R} \mathrm{d}P - f_\alpha^*(M(\gamma) - 3LR) \cdot \int_{\|x\|<R} \mathrm{d}Q. \end{aligned}$$

On the other hand, by the same argument in the proof of Theorem 2 (for proving $I_2 < \infty$ therein), we have

$$\begin{aligned} J_2 \leq\ & LR \int_{\|x\|\geq R} \mathrm{d}P + L \int_{\|x\|\geq 2R} \|x\|\,\mathrm{d}P + M(\gamma) \int_{\|x\|\geq R} \mathrm{d}P \\ & - f_\alpha^*(M(\gamma) - 3LR) \int_{R\leq\|x\|<2R} \mathrm{d}Q, \end{aligned}$$

Both the upper bounds for J_1 and J_2 tend to $-\infty$ as $M(\gamma) \to \infty$. Thus, there exists such $\overline{M}$ as claimed. Moreover, we have

$$\begin{aligned} J_1 + J_2 \leq\ & (M(\gamma) + LR) \int \mathrm{d}P + L \int \|x\|\,\mathrm{d}P \\ & - f_\alpha^*(M(\gamma) - 3LR) \int_{\|x\|<2R} \mathrm{d}Q. \end{aligned}$$

Therefore, we can pick $\overline{M} > 0$ as

$$\inf\left\{\hat{M} : (M(\gamma) + LR) \int \mathrm{d}P + L \int \|x\|\,\mathrm{d}P - f_\alpha^*(M(\gamma) - 3LR) \int_{\|x\|<2R} \mathrm{d}Q < 0, \forall M(\gamma) > \hat{M}\right\},$$

and it is obvious that $\overline{M} > 0$ only depends on P, Q and R. □

Let $\overline{M}_{m,n}$ be the quantity in Lemma E3 where (P, Q) are replaced by their empirical counterparts (P_m, Q_n), then $\overline{M}_{m,n}$ is a random variable. We have the following lemma to estimate the expectation of the rth moment ($r \geq 1$) of $\overline{M}_{m,n}$. The proof is different from that for Lemma E3.

Lemma E4. Suppose $\alpha > 1$, and (P, Q) are distributions on $\mathbb{R}^d$ of heavy-tail (β_1, β_2) with $\beta_1, \beta_2 > d + r$ for some $r \geq 1$. Let $M(\gamma) = \sup_{\|x\|=R} |\gamma(x)|$, then there exists $\overline{M}_{m,n}$ that depends on P_m, Q_n and R, such that

$$D_\alpha^L(P_m \| Q_n) = \sup_{\substack{\gamma\in\mathrm{Lip}_L(\mathbb{R}^d) \\ M(\gamma)\leq\overline{M}_{m,n}}} \left\{\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)]\right\},$$

Moreover, we have

$$\mathbb{E}_{X,Y}\left[\overline{M}_{m,n}^{r}\right] \leq M_{p,q,r},$$

where $M_{p,q,r}$ depends on $\alpha, L, R, M_r(p)$ and $M_r(q)$, and is independent of m, n.

Proof. We have

$$\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)] \leq \sum_{i=1}^{m} \frac{M(\gamma) + L\left|\|x_i\| - R\right|}{m} - \sum_{j=1}^{n} \frac{f_\alpha^*\left(M(\gamma) - 2LR - L\left|\left\|y_j\right\| - R\right|\right)}{n}.$$

Hence $\overline{M}_{m,n}$ can be taken as

$$\overline{M}_{m,n} = \inf\left\{z : \sum_{i=1}^{m} \frac{s + L\left|\|x_i\| - R\right|}{m} < \sum_{j=1}^{n} \frac{f_\alpha^*\left(s - 2LR - L\left|\left\|y_j\right\| - R\right|\right)}{n}, \forall s > z\right\}.$$

Moreover, by Jensen's inequality, we have

$$\sum_{j=1}^{n} \frac{f_\alpha^*\left(s - 2LR - L\left|\left\|y_j\right\| - R\right|\right)}{n} \geq f_\alpha^*\left(s - 2LR - L\sum_{j=1}^{n} \frac{\left|\left\|y_j\right\| - R\right|}{n}\right),$$

since the convex conjugate f_α^* is convex, and so that

$$\begin{aligned}\overline{M}_{m,n} &\leq \inf\left\{z : \sum_{i=1}^{m} \frac{s + L\left|\|x_i\| - R\right|}{m} < f_\alpha^*\left(s - 2LR - L\sum_{j=1}^{n} \frac{\left|\left\|y_j\right\| - R\right|}{n}\right), \forall s > z\right\} \\ &:= \widetilde{M}_{m,n}.\end{aligned}$$

It is obvious that $\widetilde{M}_{m,n}$ solves the following equation in variable z:

$$f_\alpha^*(z - c_1) = z + c_2, \tag{E2}$$

where

$$c_1 = 2LR + L\sum_{j=1}^{n} \frac{\left|\left\|y_j\right\| - R\right|}{n},$$

$$c_2 = \sum_{i=1}^{m} \frac{L\left|\|x_i\| - R\right|}{m}.$$

Equation (E2) can be reformulated as to find y^* that solves:

$$f_\alpha^*(y) - y = c_1 + c_2, \tag{E3}$$

where $z - c_1 = y$. We derive an upper bound for y^* as follows. Let $g(y) = f_\alpha^*(y) - y$, then

$$g'(y) = (\alpha - 1)^{\frac{1}{\alpha-1}} y^{\frac{1}{\alpha-1}} \mathbf{1}_{y>0} - 1,$$

such that $g'(y) \geq 1$ for $y > 2^{\alpha-1}(\alpha-1)^{-1}$. Given that $g\left(2^{\alpha-1}(\alpha-1)^{-1}\right) = \frac{2^\alpha}{\alpha} + \frac{1}{\alpha(\alpha-1)} - \frac{2^{\alpha-1}}{\alpha-1}$, we can take $y^* \leq 2^{\alpha-1}(\alpha-1)^{-1} + c_1 + c_2 + \left|\frac{2^\alpha}{\alpha} + \frac{1}{\alpha(\alpha-1)} - \frac{2^{\alpha-1}}{\alpha-1}\right|$. Therefore, we have

$$\overline{M}_{m,n} \leq \widetilde{M}_{m,n} = y^* + c_1 \leq 2^{\alpha-1}(\alpha-1)^{-1} + 2c_1 + c_2 + \left|\frac{2^\alpha}{\alpha} + \frac{1}{\alpha(\alpha-1)} - \frac{2^{\alpha-1}}{\alpha-1}\right|.$$

The claim follows since by Jensen's inequality, $\mathbb{E}_X\left[\left(\sum_{i=1}^m \frac{\|x_i\|}{m}\right)^r\right] \leq \mathbb{E}_X\left[\sum_{i=1}^m \frac{\|x_i\|^r}{m}\right] = M_r(p)$. (Similarly for $\mathbb{E}_Y\left[\left(\sum_{j=1}^n \frac{\|y_j\|}{n}\right)^r\right]$.) □

Proof of Theorem 7. Without loss of generality, we assume that both

$$\int_{\|x\|\leq 1} p(x)\,\mathrm{d}x > 0, \quad \int_{\|x\|\leq 1} q(x)\,\mathrm{d}x > 0.$$

Let $\Omega_0 = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ and $\Omega_k = \{x \in \mathbb{R}^d : 2^{k-1} < \|x\| \leq 2^k\}$ for $k \geq 1$. For each $k \in \mathbb{N}$, the Lebesgue measure of $\{x : \mathrm{d}(x, \Omega_k) \leq 1\}$ is bounded by $C_d 2^{kd}$ for some $C_d > 0$. Let $\Lambda_2 = \frac{1}{n}\sum_{j=1}^n \left\|y_j\right\|^{\hat{\beta}_2}$, where $2 + \frac{2\alpha}{\alpha-1} < \hat{\beta}_2 < \frac{\beta_2}{d} - 1$. By Markov's inequality, the mass or proportion of Q_n that lies in Ω_k is bounded by

$$\begin{aligned}\Pr(x \sim Q_n : \|x\| > 2^{k-1}) &= \Pr(x \sim Q_n : \|x\|^{\hat{\beta}_2} > 2^{(k-1)\hat{\beta}_2}) \\ &\leq \frac{\mathbb{E}_{Q_n}\|x\|^{\hat{\beta}_2}}{2^{(k-1)\hat{\beta}_2}} = \Lambda_2 2^{-(k-1)\hat{\beta}_2}.\end{aligned}$$

Let $M = \max(\overline{M}, \overline{M}_{m,n})$, where $\overline{M}$ is the quantity in Lemma E3 with $R = 1$, and $\overline{M}_{m,n}$ is the random counterpart for (P_m, Q_n) as defined in Lemma E4. M is a random variable since $\overline{M}_{m,n}$ is random. Let $\mathscr{F}_M$ be the following class of functions

$$\mathscr{F}_{\alpha,M} = \left\{f_\alpha^*(\gamma) : \gamma \in \mathrm{Lip}_L(\mathbb{R}^d),\ \sup_{\|x\|=1} |\gamma(x)| \leq M\right\}. \tag{E4}$$

By formulas (3.4) and (4.3), functions in $\mathscr{F}_{\alpha,M}$ have Hölder norm on Ω_k bounded by $C_\alpha(M^{\frac{\alpha}{\alpha-1}} + L^{\frac{\alpha}{\alpha-1}} 2^{\frac{\alpha k}{\alpha-1}})$ for some $C_\alpha > 0$ that only depends on α. By Corollary 2.7.4 in [50] with $V = d$ and $r = 2$, we have

$$\begin{aligned}&\ln(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n)) \\ &\leq K\delta^{-d}\left(\sum_{k=0}^\infty (C_d 2^{kd})^{\frac{2}{d+2}} \left(C_\alpha(M^{\frac{\alpha}{\alpha-1}} + L^{\frac{\alpha}{\alpha-1}} 2^{\frac{\alpha k}{\alpha-1}})\right)^{\frac{2d}{d+2}} (\Lambda_2 2^{-(k-1)\hat{\beta}_2})^{\frac{d}{d+2}}\right)^{\frac{d+2}{2}} \\ &\leq K\delta^{-d}(M+L)^{\frac{d\alpha}{\alpha-1}} \Lambda_2^{d/2}\left(\sum_{k=0}^\infty 2^{\frac{2kd}{d+2} + \frac{2\alpha kd}{(\alpha-1)(d+2)} - \frac{\hat{\beta}_2 d(k-1)}{d+2}}\right)^{\frac{d+2}{2}} \\ &\leq K\delta^{-d}(M+L)^{\frac{d\alpha}{\alpha-1}} \Lambda_2^{d/2}.\end{aligned}$$

where the constant K can vary from line to line and does not depend on n, and the last step follows as the choice of $\hat{\beta}_2$ such that the series is summable over k independent of Q_n. Then we have

$$
\begin{aligned}
&\mathbb{E}_{X,Y}\left|D_\alpha^L(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \\
&= \mathbb{E}_{X,Y}\left|\sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le \overline{M}_{m,n}}}\left\{\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)]\right\} - \sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le \overline{M}}}\left\{\mathbb{E}_{P}[\gamma] - \mathbb{E}_{Q}[f_\alpha^*(\gamma)]\right\}\right| \\
&\le \mathbb{E}_{X,Y}\sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le M}}\left|\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)] - \left(\mathbb{E}_{P}[\gamma] - \mathbb{E}_{Q}[f_\alpha^*(\gamma)]\right)\right| \\
&\le \mathbb{E}_{X}\sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_{P}[\gamma] - \mathbb{E}_{P_m}[\gamma]\right| + \mathbb{E}_{X,Y}\sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le M}}\left|\mathbb{E}_{Q}[f_\alpha^*(\gamma)] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)]\right| \\
&\le \mathbb{E}_{X}\sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_{P}[\gamma] - \mathbb{E}_{P_m}[\gamma]\right| + \mathbb{E}_X\mathbb{E}_Y\mathbb{E}_{Y'}\mathbb{E}_\xi\sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le M}}\left|\frac{1}{n}\sum_{j=1}^n \xi_i\left(f_\alpha^*[\gamma(y_j)] - f_\alpha^*[\gamma(y_j')]\right)\right| \\
&\le \mathbb{E}_{X}\sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_{P}[\gamma] - \mathbb{E}_{P_m}[\gamma]\right| + 2\mathbb{E}_X\mathbb{E}_Y\mathbb{E}_\xi\sup_{\substack{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)\\ M(\gamma)\le M}}\left|\frac{1}{n}\sum_{j=1}^n \xi_i f_\alpha^*[\gamma(y_j)]\right| \\
&\le \mathbb{E}_{X}\sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_{P}[\gamma] - \mathbb{E}_{P_m}[\gamma]\right| + 2\mathbb{E}_{X,Y}\inf_{\theta>0}\left(4\theta + \frac{12}{\sqrt{n}}\int_\theta^\infty \sqrt{\ln \mathscr{N}(\mathscr{F}_{\alpha,M},\delta,L_2(Q_n))}\,\mathrm{d}\delta\right),
\end{aligned}
$$

where ξ_i's are the Rademacher variables.

First note that the first term $\mathbb{E}_X \sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_P[\gamma] - \mathbb{E}_{P_m}[\gamma]\right|$ is the convergence rate of the Wasserstein-1 distance and the bound follows the result of Theorem 1 in [18]:

$$
\mathbb{E}_X\sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)}\left|\mathbb{E}_{P}[\gamma] - \mathbb{E}_{P_m}[\gamma]\right| \le \frac{CM_r^{1/r}(p)}{m^{1/d}},
$$

with $r = \frac{d}{d-1}$. For the second term, we have

$$
\begin{aligned}
&\mathbb{E}_{X,Y}\inf_{\theta>0}\left(4\theta + \frac{12}{\sqrt{n}}\int_\theta^\infty \sqrt{\ln \mathscr{N}(\mathscr{F}_{\alpha,M},\delta,L_2(Q_n))}\,\mathrm{d}\delta\right) \\
&\le \mathbb{E}_{X,Y}\inf_{\theta>0}\left(4\theta + \frac{12}{\sqrt{n}}K(M+L)^{\frac{d\alpha}{2(\alpha-1)}}\Lambda_2^{d/4}\int_\theta^\infty \delta^{-\frac{d}{2}}\,\mathrm{d}\delta\right) \\
&\le \mathbb{E}_{X,Y}\inf_{\theta>0}\left(4\theta + \frac{12}{\sqrt{n}}K(M+L)^{\frac{d\alpha}{2(\alpha-1)}}\Lambda_2^{d/4}\cdot\frac{2}{2-d}\theta^{1-d/2}\right) \\
&\le \mathbb{E}_{X,Y}\left(4n^{-\frac{1}{d}} + 12K(M+L)^{\frac{d\alpha}{2(\alpha-1)}}\Lambda_2^{d/4}\cdot\frac{2}{2-d}n^{-\frac{1}{d}}\right) \\
&= 4n^{-\frac{1}{d}} + \frac{24K}{2-d}n^{-\frac{1}{d}}\cdot\mathbb{E}_{X,Y}\left[(M+L)^{\frac{d\alpha}{2(\alpha-1)}}\Lambda_2^{d/4}\right]
\end{aligned}
$$

where we pick $\theta = n^{-\frac{1}{d}}$. By the Cauchy-Schwartz inequality, we have

$$\mathbb{E}_{X,Y}\left[(M+L)^{\frac{d\alpha}{2(\alpha-1)}}\Lambda_2^{d/4}\right] \le \sqrt{\mathbb{E}_{X,Y}(M+L)^{\frac{d\alpha}{(\alpha-1)}}}\sqrt{\mathbb{E}_Y\Lambda_2^{d/2}}.$$

Notice that $\mathbb{E}_{X,Y}(M+L)^{\frac{d\alpha}{(\alpha-1)}}$ is bounded by Lemma E4 and the bound depends on $M_{\frac{d\alpha}{\alpha-1}}(p)$ and $M_{\frac{d\alpha}{\alpha-1}}(q)$. By Jensen's inequality, we have $\mathbb{E}_Y\Lambda_2^{d/2} \le (\mathbb{E}_Y\Lambda_2^d)^{1/2}$. And we have

$$\mathbb{E}_Y\Lambda_2^d = \mathbb{E}_Y\left(\frac{1}{n}\sum_{j=1}^n\left\|y_j\right\|^{\hat{\beta}_2}\right)^d \le \mathbb{E}_Y\left(\frac{1}{n}\sum_{j=1}^n\left\|y_j\right\|^{\hat{\beta}_2 d}\right) = M_{\hat{\beta}_2 d}(q),$$

where the inequality follows Jensen's inequality. Combining all these bounds, we obtain the result as in the statement of the theorem. □

PROPOSITION E1. For $d = 2$. Assume (P, Q) are distributions on $\mathbb{R}^d$ of heavy-tail (β_1, β_2), where $\beta_1 > 10$ and $\beta_2 > 18$. Suppose α satisfies $\frac{4\alpha}{\alpha-1} + 4 < \beta_1 - 2$ and $\frac{8\alpha}{\alpha-1} < \beta_2 - 10$, then if m and n are sufficiently large, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^L(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \le \frac{C_1\ln m}{m^{1/2}} + \frac{C_2\ln n}{n^{1/2}}, \tag{E5}$$

where C_1 depends on $M_{r_1}(p)$ for any $r_1 > 2$ and C_2 depends on $M_{\frac{4\alpha}{\alpha-1}+4}(p)$, $M_{\frac{4\alpha}{\alpha-1}+4}(q)$ and $M_{dr_2}(q)$ for any $2 + \frac{2\alpha}{\alpha-1} < r_2 < \frac{\beta_2-2}{4}$; both C_1 and C_2 are independent of m, n.

PROPOSITION E2. For $d = 1$. Assume (P, Q) are distributions on $\mathbb{R}^d$ of heavy-tail (β_1, β_2), where $\beta_1 > 7$ and $\beta_2 > 13$. Suppose α satisfies $\frac{2\alpha}{\alpha-1} + 4 < \beta_1 - 1$ and $\frac{6\alpha}{\alpha-1} < \beta_2 - 7$, then if m and n are sufficiently large, we have

$$\mathbb{E}_{X,Y}\left|D_\alpha^L(P_m\|Q_n) - D_\alpha^L(P\|Q)\right| \le \frac{C_1}{m^{1/2}} + \frac{C_2}{n^{1/2}}, \tag{E6}$$

where C_1 depends on $M_2(p)$ and C_2 depends on $M_{\frac{2\alpha}{\alpha-1}+4}(p)$, $M_{\frac{2\alpha}{\alpha-1}+4}(q)$ and $M_{dr_2}(q)$ for any $2 + \frac{2\alpha}{\alpha-1} < r_2 < \frac{\beta_2-1}{3}$; both C_1 and C_2 are independent of m, n.

Proof. The only difference from the proof of Theorem 7 is that we need to bound the random metric entropy differently since $\sqrt{\ln \mathcal{N}(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))}$ is no longer integrable at infinity, and the upper limit of the integral in Lemma E1 cannot be relaxed to ∞. Instead, we have

$$\begin{aligned}
&\mathbb{E}_{X,Y}\inf_{0<\theta<M_Y}\left(4\theta + \frac{12}{\sqrt{n}}\int_\theta^{M_Y}\sqrt{\ln \mathcal{N}(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))}\,\mathrm{d}\delta\right)\\
&\le \mathbb{E}_{X,Y}\inf_{0<\theta<M_Y}\left(4\theta + \frac{12}{\sqrt{n}}K(M+L)^{\frac{d\alpha}{(\alpha-1)}}\Lambda_2^{d/2}\int_\theta^{M_Y}\delta^{-\frac{d}{2}}\,\mathrm{d}\delta\right),
\end{aligned}$$

where $M_Y = \sup_{\gamma\in\mathscr{F}_{\alpha,M}}\sqrt{\frac{1}{n}\sum_{j=1}^n\left|\gamma(y_j)\right|^2} \le \sqrt{\frac{1}{n}\sum_{j=1}^n(M+L+L\left\|y_j\right\|)^2}$.

For $d = 2$, we have $\int_\theta^{M_Y} \delta^{-\frac{d}{2}}\,\mathrm{d}\delta = \ln M_y - \ln\theta$, and we can pick $\theta = \frac{\ln n}{\sqrt{n}}$, and use the inequality $\ln M_y \le M_y - 1$ and combine it with Lemmas E2 and E4 as in the proof of Theorem 7.

For $d = 1$, we have $\int_\theta^{M_Y} \delta^{-\frac{d}{2}}\,\mathrm{d}\delta = \frac{\sqrt{M_y}-\sqrt{\theta}}{2}$, and we can pick $\theta = \frac{1}{\sqrt{n}}$ to balance the two terms. □

F. Proofs of results in Section 6

Proof of Theorem 8. The proof is very similar to that of Theorem 7, therefore we only outline the improvement we can obtain. First, same as the beginning of the proof of Theorem 4.8 in [10], we can restrict the domain from $\mathscr{X}$ to $\mathscr{X}/G$ by invariance, so that we focus on Lipschitz functions on $\mathscr{X}/G$. Indeed, we have

$$
\begin{aligned}
&\mathbb{E}_{X,Y}\left|D_\alpha^{L,G}(P_m\|Q_n) - D_\alpha^{L}(P\|Q)\right| \\
&= \mathbb{E}_{X,Y}\left|\sup_{\substack{\gamma\in \mathrm{Lip}_L^G(\mathscr{X})\\ M(\gamma)\le \bar{M}_{m,n}}} \left\{\mathbb{E}_{P_m}[\gamma] - \mathbb{E}_{Q_n}[f_\alpha^*(\gamma)]\right\} - \sup_{\substack{\gamma\in \mathrm{Lip}_L^G(\mathscr{X})\\ M(\gamma)\le \bar{M}}} \left\{\mathbb{E}_{P}[\gamma] - \mathbb{E}_{Q}[f_\alpha^*(\gamma)]\right\}\right| \\
&\le \mathbb{E}_{X,Y} \sup_{\substack{\gamma\in \mathrm{Lip}_L^G(\mathscr{X})\\ M(\gamma)\le M}} \left|\frac{1}{m}\sum_{i=1}^m \gamma(x_i) - \frac{1}{n}\sum_{j=1}^n f_\alpha^*[\gamma(y_j)] - \left(\mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right)\right| \\
&= \mathbb{E}_{X,Y} \sup_{\substack{\gamma\in \mathrm{Lip}_L^G(\mathscr{X})\\ M(\gamma)\le M}} \left|\frac{1}{m}\sum_{i=1}^m \gamma(T_G(x_i)) - \frac{1}{n}\sum_{j=1}^n f_\alpha^*[\gamma(T_G(y_j))] - \left(\mathbb{E}_P[\gamma] - \mathbb{E}_Q[f_\alpha^*(\gamma)]\right)\right| \\
&\le \mathbb{E}_{X,Y} \sup_{\substack{\gamma\in \mathrm{Lip}_L^G(\mathscr{X}/G)\\ M(\gamma)\le M}} \left|\frac{1}{m}\sum_{i=1}^m \gamma(T_G(x_i)) - \frac{1}{n}\sum_{j=1}^n f_\alpha^*[\gamma(T_G(y_j))] - \left(\mathbb{E}_{P_{\mathscr{X}/G}}[\gamma] - \mathbb{E}_{Q_{\mathscr{X}/G}}[f_\alpha^*(\gamma)]\right)\right|.
\end{aligned}
$$

where $T_G : \mathscr{X} \to \mathscr{X}/G$ is the quotient map, and $P_{\mathscr{X}/G}, Q_{\mathscr{X}/G}$ are restrictions of P, Q on $\mathscr{X}/G$ since both P, Q are G-invariant, and $T_G(x_i)$ and $T_G(y_j)$ can be viewed as i.i.d. samples drawn from $P_{\mathscr{X}/G}, Q_{\mathscr{X}/G}$. Compared to the proof of Theorem 7, we have some minor differences. First, in the sub-Weibull setting, the bound provided by Markov's inequality has Weibull-type decay in k, and we can simply choose $\hat{\beta}_2 = 1$. Therefore, the summation in bounding $\ln(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))$ is summable in k. Moreover, to bound $\ln(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))$, we have δ^{-d} improved to δ^{-d^*} due to the intrinsic dimension assumption. Due to Assumption 3 on the group and the partition Ω_k's are circular about the origin, the Lebesgue measure induces a reduction by a factor of $1/|G|$ by working on $\mathscr{X}/G$ compared to $\mathscr{X}$, which then makes a reduction by $1/|G|$ in the bound of $\ln(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))$, and it eventually contributes to the factor $|G|$ in the bound in Theorem 8. On the other hand, we bound $\mathbb{E}_X \sup_{\gamma\in \mathrm{Lip}_L(\mathbb{R}^d)} \left|\mathbb{E}_P[\gamma] - \mathbb{E}_{P_m}[\gamma]\right|$ using the same procedure using metric entropy instead. Since the magnitude of Lipschitz functions grows slower than $\mathscr{F}_{\alpha,M}$, the procedure is straightforward. This finally creates a factor of $|G|$ in front of m in the final bound. For cases when the intrinsic dimension is 1 or 2, we can apply proofs of and after the above treatment. □

Proof of Theorem 9. Compared to the proof of that of Theorem 8, we do not need to make a factor $|G|$. Instead, in bounding $\ln(\mathscr{F}_{\alpha,M}, \delta, L_2(Q_n))$, we have δ^{-d^*} improved to $\delta^{-d^{**}}$ due to the intrinsic dimension assumption. □

Proof of Theorem 10 and Theorem 11. Since the variational form of W_1 is shift-invariant to $\gamma \in \mathrm{Lip}_L(\mathbb{R}^d)$, we can always assume $\gamma(0) = 0$. Thus, Lemmas E3 and E4 are not useful. Compared to the proof of Theorem 7, we can pick $\hat{\beta}_2 = 1$ and M can be set to 0. Finally, it is the limiting case of $\alpha \to \infty$. □

G. Values of different training objectives in Section 7.4

Training objective function values for Lipschitz-regularized and standard α-divergence

Table G1 summarizes the training objective (divergence) values for the standard and Lipschitz-regularized α-divergence in different GPA and GAN models in the experiment Section 7.4.

TABLE G1 *Final values of the training objective for GANs and GPAs under Lipschitz-regularized ($L = 1$) and standard α-divergences with $\alpha = 2$*

Model	Objective function (divergence) value
Lip-α GPA	0.0129187
α GPA	$3.0554032e + 26$
Lip-α GAN	0.358602
α GAN	$3.25298e + 7$

See the example in Section 7.4.