LIPSCHITZ REGULARIZED GRADIENT FLOWS AND LATENT GENERATIVE PARTICLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Lipschitz regularized f-divergences are constructed by imposing a bound on the Lipschitz constant of the discriminator in the variational representation. These divergences interpolate between the Wasserstein metric and f-divergences and provide a flexible family of loss functions for non-absolutely continuous (e.g. empirical) distributions, possibly with heavy tails. We first construct Lipschitz regularized gradient flows on the space of probability measures based on these divergences. Examples of such gradient flows are Lipschitz regularized Fokker-Planck and porous medium partial differential equations (PDEs) for the Kullback-Leibler and α-divergences, respectively. The regularization corresponds to imposing a Courant–Friedrichs–Lewy numerical stability condition on the PDEs. For empirical measures, the Lipschitz regularization on gradient flows induces a numerically stable transporter/discriminator particle algorithm, where the generative particles are transported along the gradient of the discriminator. The gradient structure leads to a regularized Fisher information which is the total kinetic energy of the particles and can be used to track the convergence of the algorithm. The Lipschitz regularized discriminator can be implemented via neural network spectral normalization and the particle algorithm generates approximate samples from possibly high-dimensional distributions known only from data. Notably, our particle algorithm can generate synthetic data even in small sample size regimes. A new data processing inequality for the regularized divergence allows us to combine our particle algorithm with representation learning, e.g. autoencoder architectures. The resulting particle algorithm in latent space yields markedly improved generative properties in terms of efficiency and quality of the synthetic samples. From a statistical mechanics perspective the encoding can be interpreted dynamically as learning a better mobility for the generative particles.

1 INTRODUCTION

In this paper we construct new algorithms that are capable of efficiently transporting arbitrary empirical distributions to a target data set. The transportation of the empirical distribution is constructed as a (discretized) gradient flow in probability space for Lipschitz-regularized f- divergences. Samples are viewed as particles and are transported along the gradient of the discriminator of the divergence towards the target data set. We take advantage of representation learning concepts, e.g. autoencoders, and make these algorithms efficient even in high-dimensional sample spaces by defining particle algorithms in latent space. Their accuracy is guaranteed by a new data processing inequality.

One of our main tools is Lipschitz regularized f-divergences which interpolate between the Wasserstein metric and f-divergences. Such divergences Dupuis & Mao (2022); Birrell et al. (2022a;c), discussed in Section 2 provide a flexible family of loss functions for non-absolutely continuous distributions. In Machine Learning one needs to build algorithms to handle target distributions Q which are singular, either by their intrinsic nature such as probability densities concentrated on low dimensional structures and/or because Q is usually only approximately known through N samples (the corresponding empirical distribution $\widehat{Q}_N$ is always singular). Another key ingredient in our construction is that we build gradient flows where mass is transported along the gradient of the optimal discriminator in the variational formulation of the divergences. The time discretization of such gradient flows for empirical distributions gives rise to a so-called transporter/discriminator particle

algorithm which transports an initial empirical distribution $\widehat{P}_N$ toward the target $\widehat{Q}_N$. The Lipschitz regularization provides numerically stable, mesh free, particle algorithms that can act as generative models for high-dimensional target distributions. Moreover the gradient structure yields a dissipation functional which corresponds to the kinetic energy of the particles (a Lipschitz regularized version of Fisher information) and which can be used to control the convergence of the algorithm. The third new element in our methods is the use of representation learning to reduce the sample space dimension. We construct latent particle algorithms by building a Lipschitz regularized gradient flow in latent space. The fidelity of the latent space particle algorithm is guaranteed by a new data processing inequality for Lipschitz regularized divergence which ensures that convergence in latent space implies convergence in real sample space.

We test and demonstrate our methods in four classes of problems. (a) Learning from synthetic data sets with heavy tails. (b) Image generation using particle algorithms even in small target sample size regimes. (c) Image-to-image transformations using latent particles. (d) Merging datasets from different breast cancer studies after transforming them using our latent particles methods aiming towards meta-analysis of gene expression data Taminau et al. (2014); Hughey & Butte (2015).

Related work. Our approach is inspired by the MMD and KALE gradient flows from Arbel et al. (2019); Glaser et al. (2021) based on an entropic regularization of the MMD metrics, and related work using the Kernelized Sobolev Discrepancy Mroueh et al. (2019). Furthermore, the recent work of Dupuis & Mao (2022); Birrell et al. (2022a) built the mathematical foundations for a large class of new divergences which contains the Lipschitz regularized f-divergences and used them to construct GANs, and in particular symmetry preserving GANs Birrell et al. (2022c)). Also related is the Sinkhorn divergence Genevay et al. (2016) which is a different entropic regularization of the 2-Wasserstein metrics. Lipschitz regularizations (or related spectral normalization) have been shown to improve the stability of GANs Miyato et al. (2018); Arjovsky et al. (2017); Gulrajani et al. (2017). Our particle algorithms share similarities with GANs Goodfellow et al. (2014); Arjovsky et al. (2017) (same discriminator but different generator step) and are also broadly related to continuous time normalizing flows Chen et al. (2018a); Köhler et al. (2020); Chen et al. (2018b). Gradient flows in probability spaces related the Kullback-Leibler (KL) divergence, such as the Fokker-Planck equations and Langevin dynamics Roberts & Tweedie (1996); Durmus & Éric Moulines (2017) or Stein variational gradient descent Liu & Wang (2016); Liu (2017); Lu et al. (2019), form the basis of a variety of sampling algorithms when the target distribution Q has a known density (up to normalization). The weighted porous media equations form another family of gradient flows based on α-divergences Markowich & Villani (2000); Otto (2001); Ambrosio et al. (2005); Dolbeault et al. (2008); Vázquez (2014) which are very useful in the presence of heavy tails. Our gradient flows are Lipschitz-regularizations of such classical PDE's (Fokker-Planck and porous media equations), see Appendix A for a PDE and numerical analysis perspective on such flows. Our latent generative particles approach is inspired by latent diffusion models using auto-encoders Rombach et al. (2021). and by autoencoders in coarse-graining for molecular dynamics, Vlachas et al. (2022); Wang & Gómez-Bombarelli (2019); Stieffenhofer et al. (2021).

2 LIPSCHITZ-REGULARIZED f-DIVERGENCES

In the paper Dupuis & Mao (2022), continuing with Birrell et al. (2022a) a new general class of divergences has been constructed which interpolate between f-divergences and integral probability metrics and inherit desirable properties from both. In this paper we focus on one specific family which we view as a Lipschitz regularization of the KL-divergence (or f-divergences) or as an entropic regularization of the 1-Wasserstein metric. We denote by $\mathcal{P}(\mathbb{R}^d)$ the space of all Borel probability measures on $\mathbb{R}^d$ by $\mathcal{P}_1(\mathbb{R}^d) = \left\{P \in \mathcal{P}(\mathbb{R}^d) : \int |x| dP(x) < \infty\right\}$. We denote by $C_b(\mathbb{R}^d)$ the bounded continuous function and by $\Gamma_L = \{f : \mathbb{R}^d \to \mathbb{R} : |f(x) - f(y)| \leq L|x-y| \text{ for all } x, y\}$ the Lipschitz continous functions with Lipschitz constant bounded by L (note that $a\Gamma_L = \Gamma_{aL}$).

f-divergences. If $f : [0, \infty) \to \mathbb{R}$ is strictly convex and lower-semicontinuous with $f(1) = 0$ the f-divergence of P with respect to Q is defined by $D_f(P\|Q) = E_Q[f(\frac{dP}{dQ})]$ if $P \ll Q$ and set to be

$+\infty$ otherwise. We have the variational representation (see e.g. Birrell et al. (2022a) for a proof)

$$D_f(P\|Q) = \sup_{\phi \in C_b(\mathbb{R}^d)} \left\{ E_P[\phi] - \inf_{\nu \in \mathbb{R}} \{\nu + E_Q[f^*(\phi - \nu)]\} \right\} \tag{1}$$

where $f^*(s) = \sup_{t\in\mathbb{R}} \{st - f(t)\}$ is the Legendre-Fenchel transform of f. We will use the KL-divergence with $f_{\mathrm{KL}}(x) = x \log x$ and the α-divergence: $f_\alpha = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$ with Legendre transforms $f^*_{\mathrm{KL}}(y) = e^{y-1}$ and $f^*_\alpha \propto y^{\frac{\alpha}{(\alpha-1)}}$ (see the Appendix). For KL the infimum over ν can be solved analytically and yields the Donsker-Varadhan with a $\log E_Q[e^\phi]$ term (see Birrell et al. (2022b) for more on variational representations).

Wasserstein metrics. The 1-Wasserstein metrics $W^{\Gamma_1}(P, Q)$ with transport cost $|x - y|$ is an integral probability metrics, see Arjovsky et al. (2017). By keeping the Lipschitz constant as a regularization parameter we set

$$W^{\Gamma_L}(P, Q) = \sup_{\phi \in \Gamma_L} \{E_P[\phi] - E_Q[\phi]\} \tag{2}$$

and note that we have $W^{\Gamma_L}(P, Q) = LW^{\Gamma_1}(P, Q)$.

Lipschitz-regularized f-divergences. The Lipschitz regularized f-divergences are defined directly in terms their variational representations, by replacing the optimization over bounded continuous functions in equation 1 by Lipschitz continuous functions in Γ_L.

$$D_f^{\Gamma_L}(P\|Q) := \sup_{\phi \in \Gamma_L} \left\{ E_P[\phi] - \inf_{\nu \in \mathbb{R}} \{\nu + E_Q[f^*(\phi - \nu)]\} \right\}. \tag{3}$$

Some of the important properties of Lipschitz regularized f-divergences, which summarizes results from Dupuis & Mao (2022); Birrell et al. (2022a) are given in Theorem 2.1. It is assumed there that f is super-linear (called admissible in Birrell et al. (2022a)), that is $\lim_{s\to\infty} f(s)/s = +\infty$. This excludes the case of α-divergences for $\alpha < 1$, for which the existence of optimizers is a more delicate problem, but parts of the theorems remain true.

Theorem 2.1. *Assume that f is superlinear and strictly convex. Then for $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$ we have*

1. **Divergence:** *$D_f^{\Gamma_L}(P\|Q)$ is a divergence, i.e. $D_f^{\Gamma_L}(P\|Q) \geq 0$ and $D_f^{\Gamma_L}(P\|Q) = 0$ if and only if $P = Q$. Moreover the map $(P, Q) \to D_f^{\Gamma_L}(P\|Q)$ is convex and lower-semicontinuous.*

2. **Infimal Convolution Formula:** $D_f^{\Gamma_L}(P\|Q) = \inf_{\gamma \in \mathcal{P}(\Omega)} \left\{ W^{\Gamma_L}(P, \gamma) + D_f(\gamma\|Q) \right\}$ *. In particular we have $0 \leq D_f^{\Gamma_L}(P\|Q) \leq \min\left\{D_f(P\|Q), W^{\Gamma_L}(P, Q)\right\}$.*

3. **Interpolation and limiting behavior of $D_f^{\Gamma_L}(P\|Q)$:**

$$\lim_{L\to\infty} D_f^{\Gamma_L}(P\|Q) = D_f(P\|Q) \quad \text{and} \quad \lim_{L\to 0} \frac{1}{L} D_f^{\Gamma_L}(P\|Q) = W^{\Gamma_1}(P, Q)\,. \tag{4}$$

4. **Optimizers:** *There exists a optimizer $\phi^{L,*} \in \Gamma_L$, unique up to a constant in supp$(P) \cup$ supp(Q).*

Remark 2.2. The optimizer $\gamma^{L,*}$ in the infimal convolution formula exists and is unique and we have $d\gamma^{L,*} \propto (f^*)'(\phi^{L,*})dQ$ (see Birrell et al. (2022a) for details). For example for KL we get $d\gamma^{L,*} \propto e^{\phi^{L,*}} dQ$.

3 Lipschitz-regularized gradient flows and their Transporter/Discriminator representation

Given a target probability distribution Q, we build an evolution equation for probability measures based on the Lipschitz regularized f-divergences $D_f^{\Gamma_L}(P\|Q)$ by considering the PDE

$$\partial_t P_t + \operatorname{div}\left(P_t \nabla \frac{\delta D_f^{\Gamma_L}(P\|Q)}{\delta P}(P_t) \right) = 0\,, \quad P_0 = P \in \mathcal{P}_1(\mathbb{R}^d) \tag{5}$$

where $\frac{\delta D_f^{\Gamma_L}(P\|Q)}{\delta P}$ is the first variation of $D_f^{\Gamma_L}(P\|Q)$ (to be discussed below in Theorem 3.1). An advantage of the Lipschitz regularized f-divergences is its ability to compare singular measures and so the equation 5 is to be understood in the sense of distributions (integrating against test functions).

In the limit $L \to \infty$ and if $P \ll Q$, equation 5 yields the Fokker-Planck equation (for KL divergence) and the weighted porous medium equation (for α-divergences) Otto (2001); Dolbeault et al. (2008), see Appendix A.

The following theorem was first proved in Dupuis & Mao (2022) for KL and can be generalized to the f-divergences considered in Theorem 2.1 (see the proof in Appendix D.

Theorem 3.1. *Assume f is superlinear and strictly convex and $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$. Then we have*

$$\frac{\delta D_f^{\Gamma_L}(P\|Q)}{\delta P}(P) = \phi^{L,*}\,. \tag{6}$$

In more details, let ρ be a signed measure of total mass 0 and let $\rho = \rho_+ - \rho_-$ where $\rho_\pm \in \mathcal{P}_1(\mathbb{R}^d)$ are mutually singular. If $P + \epsilon\rho \in \mathcal{P}_1(\mathbb{R}^d)$ for sufficiently small $|\epsilon|$ then $D_f^{\gamma_L}(P + \epsilon\rho\|Q)$ is differentiable at $\epsilon = 0$ and

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(D_f^{\Gamma_L}(P+\epsilon\rho\|Q) - D_f^{\Gamma_L}(P\|Q)\right) = \int \phi^{L,*}d\rho\,. \tag{7}$$

Combining Theorem 3.1 with equation 5 leads to a new class of PDEs:

Transporter/Discriminator PDE:

$$\partial_t P_t + \operatorname{div}(P_t\nabla\phi_t^{L,*}) = 0, \quad \text{where} \quad \phi_t^{L,*} = \operatorname*{arg\,max}_{\phi\in\Gamma_L}\{E_{P_t}[\phi] - E_Q[f^*(\phi)]\} \tag{8}$$

Remark 3.2. **(a)** The transporter/discriminator PDE equation 8 makes sense when P and Q are replaced by their empirical measures $\hat{P}_N$, $\hat{Q}_N$ based on N IID samples. This will be the basis of our numerical algorithm in Section 4 (see Algorithm 1). **(b)** Also equation 8 makes sense if P and Q are mutually singular (e.g. when Q is supported on a low-dimensional structure). We can view equation 8 as a Lipschitz regularization of classical PDEs which allows particle-based approximations based on data. In particular, the Lipschitz condition on $\phi \in \Gamma_L$ enforces a finite speed of propagation of at most L in the transport equation in equation 8. This is in sharp contrast with the Fokker Planck equation given in Appendix A which is a diffusion equation, see Appendix A.2 for more details and practical implications.

4 Lipschitz-regularized generative particles

In this section we build a numerical algorithm to solve the transporter/discriminator gradient flow when N IID samples from the target distribution are given. For a map $T : \mathbb{R}^d \to \mathbb{R}^d$ and $P \in \mathcal{P}(\mathbb{R}^d)$, the pushforward measure is denoted by $T_\# P$ (i.e. $T_\# P(A) = P(T^{-1}(A))$. The forward-Euler discretization of the system equation 8 yields:

Euler method for the Transporter/Discriminator PDE:

$$P_{n+1} = \left(I - \Delta t\nabla\phi_n^{L,*}\right)_\# P_n, \quad \text{where} \quad \phi_n^{L,*} = \operatorname*{arg\,max}_{\phi\in\Gamma_L}\{E_{P_n}[g] - E_Q[f^*(\phi)]\} \tag{9}$$

When only N IID samples $\{X^{(i)}\}_{i=1}^N$ of the target distribution Q are available we build a particle system by considering N IID samples $\{Y^{(i)}\}_{i=1}^N$ from some initial measure P ($M \neq N$ samples are also possible) and equation 9 becomes

Lipschitz regularized generative particles:

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \Delta t\nabla\phi_n^{L,*}(Y_n^{(i)})\,,\ \phi_n^{L,*} = \operatorname*{arg\,max}_{\phi\in\Gamma_L}\left\{\frac{\sum_{i=1}^N\phi(Y_n^{(i)})}{N} - \frac{\sum_{i=1}^N f^*(\phi(X^{(i)}))}{N}\right\} \tag{10}$$

The empirical measure $\hat{P}_n^N = N^{-1}\sum_{i=1}^N\delta_{Y_n^{(i)}}$ built from equation 10 gives a solution of the system equation 9 if we use as target the empirical measure $\hat{Q}^N = N^{-1}\sum_{i=1}^N\delta_{X^{(i)}}$ and as initial condition the empirical measure $\hat{P}^N = N^{-1}\sum_{i=1}^N\delta_{Y_0^{(i)}}$.

Algorithm 1: Lipschitz regularized generative particles algorithm

Require: f defined in equation 2 and its Legendre conjugate f^*, L: Lipschitz constant, T: number of updates for the particles, γ: time step size, N: number of particles

Require: $W = \{W^l\}_{l=1}^D$: parameters for the neural network $\phi : \mathbb{R}^d \to \mathbb{R}$, D: depth of the neural network, δ: learning rate of the neural network, T_{NN}: number of updates for the neural network.

Result: $\{Y_T^{(i)}\}_{i=1}^N$

Sample $\{X^{(i)}\}_{i=1}^N \sim Q$, a batch from the real data

Sample $\{Y_0^{(i)}\}_{i=1}^N \sim P_0 = P$, a batch of prior samples

Initialize W randomly and $W^l \leftarrow L^{1/D} * W^l / \|W^l\|_2$

for $n = 0$ **to** $(T-1)$ **do**

  for $m = 0$ **to** $T_{\text{NN}} - 1$ **do**

    $grad_W \leftarrow \nabla_W \left[N^{-1} \sum_{i=1}^N \phi(Y_n^{(i)}; W) - N^{-1} f^*(\phi(X_n^{(i)}; W)) \right]$

    $W \leftarrow W + \delta * grad_W$

    $W^l \leftarrow L^{1/D} * W^l / \|W^l\|_2$

  end

  $Y_{n+1}^{(i)} \leftarrow Y_n^{(i)} - \gamma \nabla \phi_n^L(Y_n^{(i)}; W), \quad i = 1, \cdots, N$

end

Remark 4.1. **(a)** The transport mechanism given by equation 10 is linear. However, nonlinear interactions between particles as introduced via the discriminator $\hat{\phi}_n^{L,*}$ are created due to the velocity field $\nabla \phi_n^{L,*}$ which depends on all particles that comprise $\hat{P}_n^N$ and $\hat{Q}^N$ at each step n. **(b)** Computationally, the discriminator optimization (over Lipschitz continuous functions) is implemented, for example, via spectral normalization for neural networks architectures. Moreover the gradient of the discriminator is computed only at the positions of the particles. **(c)** The Lipschitz bound L on the discriminator space implies a pointwise bound $|\nabla \phi_n^{L,*}(Y_n^{(i)})| \leq L$ and thus the particle speed is bounded by L. Hence the Lipschitz regularization imposes a speed limit L on the particles, ensuring the stability of the algorithm for suitable choices of L. This implicit grid is reminiscent of the Courant, Friedrichs, and Lewy (CFL) condition for the stability of discrete scheme. These are fundamental features for the performance and the stability of Algorithm 1 derived from equation 10 (see Sections 6) and Appendix A.

Kinetic energy of particles: The gradient structures implies, see Theorem A.1, that, for equation 8, the derivative of the regularized divergence satisfies $\frac{d}{dt} D_f^{\Gamma_L}(P_t \| Q) = -I_f^{\Gamma_L}(P_t \| Q)$ where $I_f^{\Gamma_L}(P_t \| Q) = E_{P_t}[|\nabla \phi^{L,*}|^2]$ which is interpreted as a Lipschitz-regularized Fisher Information. As $L \to \infty$ one recovers for example the Fisher Information used for the Fokker-Planck equation. For the Algorithm 1 the Lipshitz-regularized Fisher information

$$I_f^{\Gamma_L}(\hat{P}_n^N \| \hat{Q}^N) = \int |\nabla \phi_n^{L,*}|^2 \hat{P}_n^N(dx) = \frac{1}{N} \sum_{i=1}^N |\nabla \phi_n^{L,*}(Y_n^{(i)})|^2 \,, \tag{11}$$

is equal to the total kinetic energy of the particle since $\nabla \phi_n^{L,*}(Y_n^{(i)})$ is the velocity of the ith particle at time n. Clearly when the total kinetic energy $I_f^{\Gamma_L}(\hat{P}_n^N \| \hat{Q}^N)$ is zero, the algorithm will stop.

5 Latent generative particles: Gradient flows in latent space

A standard paradigm of machine learning is that target measures are often supported on low dimensional structures. We leverage this insight, in the form of an auto-encoder, to construct particle algorithms in a latent, lower dimensional, space. The resulting latent particle algorithms are both more accurate and efficient, even in high-dimensional sample spaces, and their performance is guaranteed by a new Data Processing Inequality in Theorem 5.1. Assume $Q = Q^{\mathcal{Y}}$ is supported on some low dimensional set $S \subset \mathcal{Y} = \mathbb{R}^d$, an encoder map $\mathcal{E} : \mathcal{Y} \to \mathcal{Z}$ where $\mathcal{Z} \subset \mathbb{R}^{d'}$, $d' < d$ and a

decoder map $\mathcal{D} : \mathcal{Z} \to \mathcal{Y}$ are invertible in S, i.e. $\mathcal{D} \circ \mathcal{E}(S) = \mathcal{D}(\mathcal{Z}) = S$. We denote by $\mathcal{E}_\# Q^{\mathcal{Y}}$ the image of the measure $Q^{\mathcal{Y}}$ by the map $\mathcal{E}$, i.e. for $A \subset \mathcal{Z}$, we define $\mathcal{E}_\# Q^{\mathcal{Y}}(A) = Q^{\mathcal{Y}}(\mathcal{E}^{-1}(A))$ and likewise for $\mathcal{D}_\# P^{\mathcal{Z}}$. The following theorem expresses how information remains controlled under encoding/decoding and guarantees the performance of the approximation $\mathcal{D}_\# P^{\mathcal{Z}}$ in the real space. The latter is achieved by an *a posteriori* estimate (12), in the sense of numerical analysis, where the approximation in the tractable latent space $\mathcal{Z}$ will bound the error in the real space $\mathcal{Y}$.

Theorem 5.1. *Suppose that*

1. **Perfect encoding.** *For $Q^{\mathcal{Y}}$ the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are such that $\mathcal{D}_\# \mathcal{E}_\# Q^{\mathcal{Y}} = Q^{\mathcal{Y}}$.*
2. **Lipschitz decoder.** *The decoder is Lipschitz continuous with Lipschitz constant $a_{\mathcal{D}}$.*

Then, for any $P^{\mathcal{Z}} \in \mathcal{P}(\mathcal{Z})$ we have

$$D_f^{\Gamma_L}\left(\mathcal{D}_\# P^{\mathcal{Z}} \| Q^{\mathcal{Y}}\right) \leq D_f^{\Gamma_{a_{\mathcal{D}} L}}\left(P^{\mathcal{Z}} \| \mathcal{E}_\# Q^{\mathcal{Y}}\right). \tag{12}$$

This theorem and more general versions thereof for other representation learning tools beyond autoencoders, is proved in Appendix E.2. The proof is a consequence of a new, tighter data processing inequality derived in Birrell et al. (2022a) that involves both transformations of probabilities and discriminator spaces Γ.

Remark 5.2. In practice an autoencoder is trained on data using the empirical measure $\widehat{Q}_N$ and suitable loss function and neural network architectures. Assumption 2 in Theorem 5.1 can easily be enforced using e.g. spectral normalization. Assumption 1 is a reasonable, but somewhat idealized, version of the requirement that the autoencoder captures adequately the features of the dataset Q. In particular the dimension of the latent space $\mathcal{Z}$ needs to be selected carefully (see Section 6).

Gradient flow in latent spaces. If $\phi_t^{\mathcal{Z}}$ is the discriminator in latent space leading to the gradient flow (8), $\partial_t P_t^{\mathcal{Z}} = \mathrm{div}(P_t^{\mathcal{Z}} \nabla \phi_t^{\mathcal{Z}})$ then, in the particle algorithm, each particle is transported following (the time-discretization of) the ODE $\dot{z}_t = -\nabla \phi_t^{\mathcal{Z}}(z_t)$, as in Section 4. The Algorithm 2 can be found in the Appendix E.3. Upon decoding we find the transport ODE in real space is

$$\dot{y}_t = \left(\frac{\partial \mathcal{D}}{\partial z}(z_t)\right)^T \dot{z}_t = -\left(\frac{\partial \mathcal{D}}{\partial z}(z_t)\right)^T \frac{\partial \mathcal{D}}{\partial z}(z_t) \nabla_y \phi_t^{\mathcal{Y}}(\mathcal{D}(z_t)) \tag{13}$$

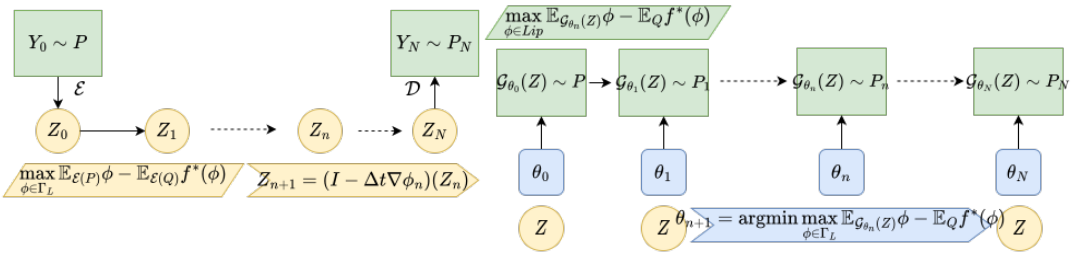
where $\frac{\partial \mathcal{D}}{\partial z}(z_t)$ is the Jacobian of $\mathcal{D}$ at the point z_t and the reconstructed discriminator $\phi^{\mathcal{Y}}$ is given by $\phi^{\mathcal{Z}} = \phi^{\mathcal{Y}} \circ \mathcal{D}$. Using (13) we can therefore interpret the encoding as learning a mobility $\mu_t = \frac{\partial \mathcal{D}}{\partial z}(z_t)^T \frac{\partial \mathcal{D}}{\partial z}(z_t)$, i.e., learning a better geometry in real space. This leads to a gradient flow in real space with non-trivial mobility, cf. (8),

$$\partial_t P_t^{\mathcal{Y}} = \mathrm{div}\left(\mu_t P_t^{\mathcal{Y}} \nabla \phi_t^{\mathcal{Y}}\right). \tag{14}$$

We note that the mobility concept is well-known in computational materials science where it is used to model kinetics of species and interfaces, see for instance Cahn (1965); Zhu et al. (1999); Wang et al. (2020). Finally, we note a similar computation to (13) in Mroueh et al. (2019) regarding the interpretation of GAN's as a gradient flow. The differences (and similarities) between (Lipschitz-regularized) Generative particle algorithm (GPA) and GAN are summarized in Figure 1 and Table 1, where in the latter we also include a comparison between mobilities.

	GPA	GPA in a latent space	GAN
Discriminator	$\phi^{\mathcal{Y}} \in \mathrm{Lip}(\mathcal{Y})$	$\phi^{\mathcal{Z}} \in \mathrm{Lip}(\mathcal{Z})$	$\phi^{\mathcal{Y}} \in \mathrm{Lip}(\mathcal{Y})$
Generator	$(I_{\mathcal{Y}} - \Delta t \nabla \phi^{\mathcal{Y}})_\# P_n^{\mathcal{Y}}$	$\left(\mathcal{D} \circ (I_{\mathcal{Z}} - \Delta t \nabla \phi^{\mathcal{Z}})\right)_\# P_n^{\mathcal{Z}}$	$\mathcal{G}_\theta(z)$, $z \sim \mathcal{N}(\mathbf{0}, I_{\mathcal{Z}})$
Updates	Particles $y \in \mathcal{Y}$	Particles $z \in \mathcal{Z}$	Generator parameters $\theta \in \mathbb{R}^{\lvert\theta\rvert}$
Mobility	$I_{\mathcal{Y}}$	$\frac{\partial \mathcal{D}}{\partial z}(z_t)^T \frac{\partial \mathcal{D}}{\partial z}(z_t)$	$\frac{\partial \mathcal{G}(\theta_t, z)}{\partial \theta}^T \frac{\partial \mathcal{G}(\theta_t, z)}{\partial \theta}$

Table 1: Comparison of features in GPA, GPA in the latent space and GAN. $\mathcal{Y} = \mathbb{R}^d$ and $\mathcal{Z} = \mathcal{E}(\mathcal{Y})$.



(a) GPA in latent space

(b) GAN

Figure 1: Workflow of different generative models. green: real space, yellow: latent space, blue: parameter space

6 EXPERIMENTS

We discuss examples that show distinguished properties of our Lipschitz regularized generative particles algorithm (GPA). First, we explore the choice of f, which enables us to learn heavy-tailed distributions such as $Student - t(0.5)$, see Figure 2. Distributions with exponential tails are explored in Appendix G.1. Second, particles from any empirical distribution can be transported to an empirical target distribution even if a relatively small number of samples is available from the target. Furthermore, combining our algorithm with representation learning, we introduce a bioinformatics application of our algorithm in the analysis of gene expression data. Gene expression datasets are not only high-dimensional but also small-sized thus it is crucial to increase the sample size by integrating together all available datasets from the same disease. However, this is not a straightforward process since it is well known that gene expression datasets may have different statistics even when they target the same disease; a phenomenon referred to as "batch effects" Tran et al. (2020). We propose to mitigate batch effects via the latent generative particle algorithm and match the statistics of the two datasets. Figure 3 presents the results on applying our algorithm between two breast cancer datasets from the Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo/`). More technical details and experiments can be found in the Appendix G.5. Third, we compare the image generation performance of our $(f_{\text{KL}}, \Gamma_1)$-GPA with $(f_{\text{KL}}, \Gamma_1)$-GAN and Wasserstein GAN on MNIST, and in particular in a small target sample size regime. See Figure 4 (a) - (f). CIFAR10 example in Appendix G.2 also testifies the ability of our algorithm to learn images without mode collapse in a small target sample size regime. In Appendix G.3, we compare the performance with other gradient flow based models and GANs on a low dimensional example. In addition, we show from the image to image transformation example on MNIST, that the latent GPA improves the quality of the generated images as well as the computational efficiency. See Figure 4 (g) - (i) and Table 2. We note that latent GPA achieved the best FID score among the tested models. In Appendix G.4, we have another example working in a latent space showing the importance of Q-perfect encoding property for the latent GPA.

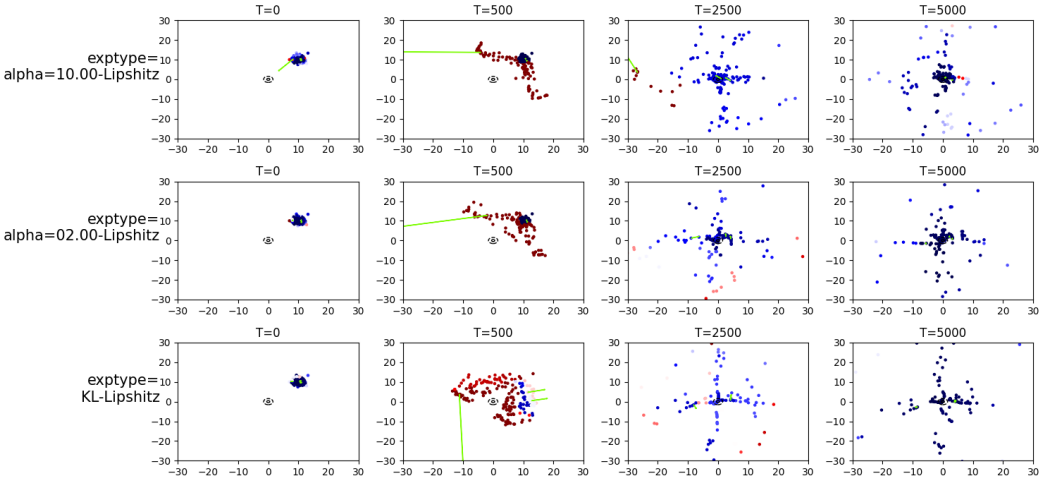
	Sample size	200	2000
Conditioning on digit labels	$(f_{\text{KL}}, \Gamma_1)$-GPA	4913.53	4926.91
	$(f_{\text{KL}}, \Gamma_1)$-GAN	5603.55	1270.13
	Wasserstein-GAN	5653.20	1879.18
Digit transform	$\mathbb{R}^{64}$	$\mathbb{R}^{128}$	$\mathbb{R}^{28\times 28}$
	946.98	846.99	3710.23

(a) Final FID for MNIST data.

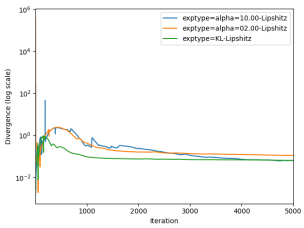
Space	Time
$\mathbb{R}^{64}$	949.408s
$\mathbb{R}^{128}$	999.862s
$\mathbb{R}^{28\times 28}$	2549.594s

(b) Computation time for digit transform

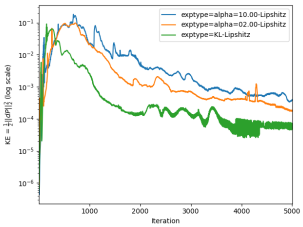
Table 2: MNIST performance summary. See Figure 4.



(a) $L = 1$) **(Top)** $f = f_\alpha$ with $\alpha = 10$, **(Middle)** $f = f_\alpha$ with $\alpha = 2$, **(Bottom)** $f = f_{\mathrm{KL}}$.

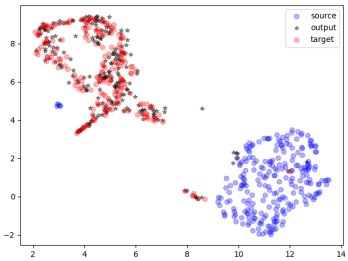


(b) Divergences

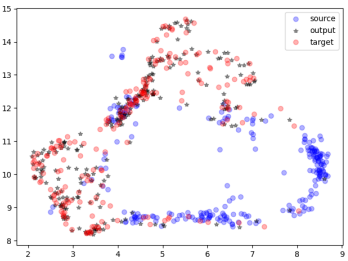


(c) Kinetic energy of particles

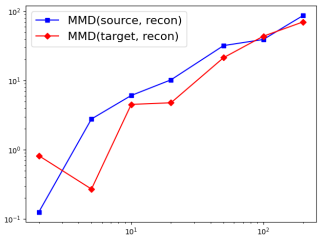
Figure 2: **(2D Student-t, $\nu = 0.5$) Convergence in selected choices of f:** 4 of 5 trials with f_{KL} ($\alpha = 1$) diverged, which implies the instability of the KL gradient flow. Instead, we use $f = f_\alpha$ with $\alpha = 10$ to handle the heavy tail. **green: $f = f_{\mathrm{KL}}$, orange: $f = f_\alpha$ with $\alpha = 2$, blue: $f = f_\alpha$ with $\alpha = 10$. (a)** Transportation of particles and then redistribution. The contour indicates the 50% region. **(b)** The Lipschitz-regularized f_α divergence with $\alpha = 10$ converged in a lower value compared to the $\alpha = 2$ case. **(c)** The kinetic energy decays as divergence converges. Learning rates are chosen as $\gamma = 0.1, \delta = 0.01$. f_{KL} results reflect the only converged experiment.



(a) Particles in the latent space



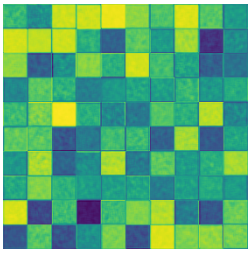
(b) Particles in the real space



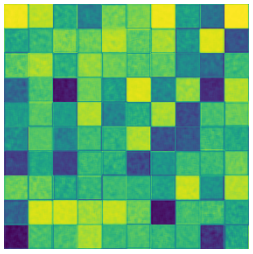
(c) MMD

Figure 3: **(Gene expression data; Merging Breast Cancer datasets).** We merged gene data using our latent GPA in significantly lower dimensions. Two distinct gene data sets but from the same disease decrease their dimension from 54,675D into $d' = 2, 5, 10, 20, 50, 100, 200$ using normalized PCA. Then, latent particles are transported using GPA. **blue: source, red: target, black: transported**. **(a)** Latent particles in the $\mathbb{R}^d$ with $d' = 20$ which are encoded by the PCA. **(b)** Transported samples are reconstructed to the real space. The 2D visualizations are obtained using the UMAP algorithm McInnes et al. (2018). **(c)** The MMD distance Gretton et al. (2012) between the reconstructed datasets. **blue: MMD**$(P_0^{\mathcal{Y}}, P_T^{\mathcal{Y}})$**, red: MMD**$(Q^{\mathcal{Y}}, P_T^{\mathcal{Y}})$, $T = 25K$. The transported distribution has smaller distances from the target distribution when $d' = 5, 10, 20, 50, 200$.

(a) $(f_{\mathrm{KL}}, \Gamma_1)$-GPA, 200 samples



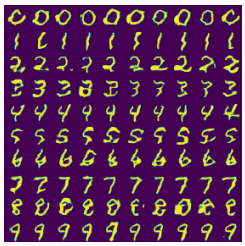
(b) $(f_{\mathrm{KL}}, \Gamma_1)$-GAN, 200 samples



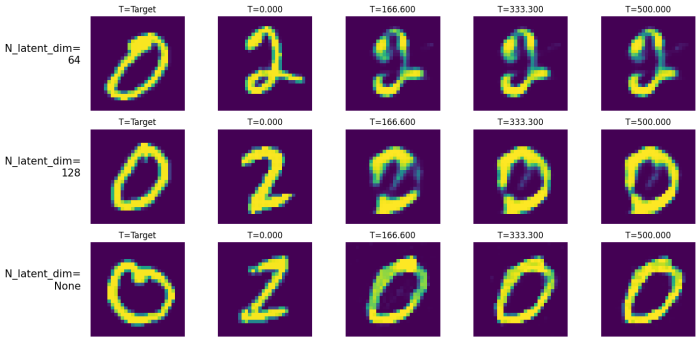
(c) Wasserstein-GAN, 200 samples

(d) $(f_{\mathrm{KL}}, \Gamma_1)$-GPA, 2000 samples

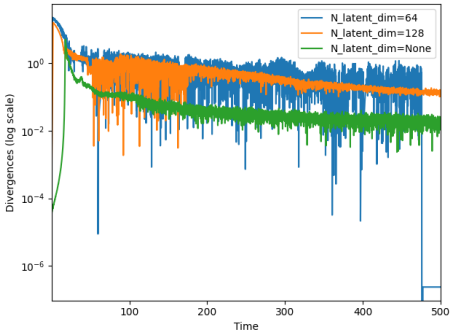
(e) $(f_{\mathrm{KL}}, \Gamma_1)$-GAN, 2000 samples



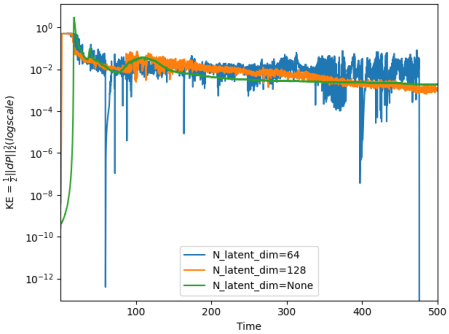
(f) Wasserstein-GAN, 2000 samples



(g) $(f_{\mathrm{KL}}, \Gamma_1)$-GPA. Transportations of particles in **(Top)** 64D, **(Middle)** 128D, **(Bottom)** 784D spaces



(h) $(f_{\mathrm{KL}}, \Gamma_1)$ divergences



(i) Kinetic energy of particles

Figure 4: **MNIST; (a)-(f): Learned digits by different generative models (column) with different number of training data (row).** The $(f_{\mathrm{KL}}, \Gamma_1)$-GPA was able to learn digits from a small data set, while the other methods failed. Using sufficiently large training data, GANs outperformed in capturing the scale, which can be observed by the more intense color contrast between a digit and its background. **(g)-(i): MNIST digit transform** $2 \to 0$**, latent and original generative particles.** Latent particles are obtained by auto-encoders trained with 60,000 training images. Particles are transported in $\mathbb{R}^{64}$, $\mathbb{R}^{128}$, and $\mathbb{R}^{28\times 28}$, respectively. **(g)** Time trajectories of particles in the real space. **(h)** $(f_{\mathrm{KL}}, \Gamma_1)$-divergence and **(i)** Kinetic energy of particles converged in 128D (orange) and 784D (green) spaces. 64D (blue) was unstable. See FID scores and computation time in Table 2.

7 REPRODUCIBILITY STATEMENT

Each figure contains main parameters such as divergence-specifying f, Lipschitz constant L, and dataset parameter. In Appendix Section F, experimental setting for the experiments in the main text and appendix are described in these aspects:

- Data sets
- Neural network architectures
- Mini-batch scheme
- Computational resources.

In Supplementary material, source code, all dependent libraries, and documentation (README.md) are attached. README.md specifies the required open-source libraries and the entire parameters set including random seeds for reproducing individual experiments.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

Michael Arbel, Anna Korba, Adil SALIM, and Arthur Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/944a5ae3483ed5c1e10bbccb7942a279-Paper.pdf.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, gamma)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022a.

Jeremiah Birrell, Markos A. Katsoulakis, and Yannis Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *IEEE Transactions on Information Theory*, pp. 1–1, 2022b. doi: 10.1109/TIT.2022.3160659.

Jeremiah Birrell, Markos A Katsoulakis, Luc Rey-Bellet, and Wei Zhu. Structure-preserving gans. *Accepted in ICML 2022, arXiv preprint arXiv:2202.01129*, 2022c.

John W. Cahn. Phase separation by spinodal decomposition in isotropic systems. *The Journal of Chemical Physics*, 42(1):93–99, 1965. doi: 10.1063/1.1695731. URL https://doi.org/10.1063/1.1695731.

José Carrillo, Jean Dolbeault, Ivan Gentil, and Ansgar Juengel. Entropy-energy inequalities and improved convergence rates for nonlinear parabolic equations. *Discrete and Continuous Dynamical Systems-Series B*, 6(5):1027–1050, 2006.

José Antonio Carrillo, Yanghong Huang, Francesco Saverio Patacchini, and Gershon Wolansky. Numerical study of a particle method for gradient flows. *Kinetic and Related Models*, 10(3): 613–641, 2017.

Changyou Chen, Chunyuan Li, Liqun Chen, Wenlin Wang, Yunchen Pu, and Lawrence Carin Duke. Continuous-time flows for efficient inference and density estimation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 824–833. PMLR, 10–15 Jul 2018a. URL https://proceedings.mlr.press/v80/chen18d.html.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Jean Dolbeault, Ivan Gentil, Arnaud Guillin, and Feng-Yu Wang. L^q-functional inequalities and weighted porous media equations. *Potential Anal.*, 28(1):35–59, 2008. ISSN 0926-2601. doi: 10.1007/s11118-007-9066-0. URL https://doi.org/10.1007/s11118-007-9066-0.

Paul Dupuis and Yixiang Mao. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *ESAIM Control Optim. Calc. Var.*, 28:Paper No. 10, 38, 2022. ISSN 1292-8119. doi: 10.1051/cocv/2022002. URL https://doi.org/10.1051/cocv/2022002.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017. doi: 10.1214/16-AAP1238. URL https://doi.org/10.1214/16-AAP1238.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf.

Pierre Glaser, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8018–8031. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/433a6ea5429d6d75f0be9bf9da26e24c-Paper.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2188385.2188410.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Jacob J Hughey and Atul J Butte. Robust meta-analysis of gene expression using the elastic net. *Nucleic acids research*, pp. gkv229, 2015.

Jonas Köhler, Leon Klein, and Frank Noe. Equivariant flows: Exact likelihood generative learning for symmetric densities. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kohler20a.html.

Randall LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems (Classics in Applied Mathematics Classics in Applied Mathemat)*. Society for Industrial and Applied Mathematics, USA, 2007. ISBN 0898716292.

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019. doi: 10.1137/18M1187611. URL https://doi.org/10.1137/18M1187611.

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the fokker-planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. 02 2018.

Youssef Mroueh, Tom Sercu, and Anant Raj. Sobolev descent. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2976–2985. PMLR, 2019.

F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000. ISSN 0022-1236. doi: https://doi.org/10.1006/jfan.1999.3557. URL https://www.sciencedirect.com/science/article/pii/S0022123699935577.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001. ISSN 0360-5302. doi: 10.1081/PDE-100002243. URL https://doi.org/10.1081/PDE-100002243.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341 – 363, 1996. doi: bj/1178291835. URL https://doi.org/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Marc Stieffenhofer, Tristan Bereau, and Michael Wand. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. *APL Materials*, 9:031107, 03 2021. doi: 10.1063/5.0039102.

Jonatan Taminau, Cosmin Lazar, Stijn Meganck, and Ann Nowé. Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis. *ISRN bioinformatics*, 2014, 2014.

Guiseppe Toscani and Cédric Villani. On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *Journal of Statistical Physics*, 98(5):1279–1309, 2000.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1):12–12, 2020. ISSN 1474-760X.

Juan Luis Vázquez. Barenblatt solutions and asymptotic behaviour for a nonlinear fractional heat equation of porous medium type. *Journal of the European Mathematical Society*, 16(4):769–803, 2014.

Pantelis Vlachas, Georgios Arampatzis, Caroline Uhler, and Petros Koumoutsakos. Multiscale simulations of complex systems by learning their effective dynamics. *Nature Machine Intelligence*, 03 2022. doi: 10.48550/arXiv.2006.13431.

Feng-Yu Wang. A generalization of Poincaré and log-Sobolev inequalities. *Potential Anal.*, 22(1): 1–15, 2005. ISSN 0926-2601. doi: 10.1007/s11118-004-4006-8. URL https://doi.org/10.1007/s11118-004-4006-8.

Feng-Yu Wang. Orlicz–poincaré inequalities. *Proceedings of the Edinburgh Mathematical Society*, 51(2):529–543, 2008.

Qiao Wang, Geng Zhang, Yajie Li, Zijian Hong, Da Wang, and Siqi Shi. Application of phase-field method in rechargeable batteries. *npj Computational Materials*, 6(1):1–8, 2020.

Wujie Wang and Rafael Gómez-Bombarelli. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials*, 5:125, 12 2019. doi: 10.1038/s41524-019-0261-5.

Jingzhi Zhu, Long-Qing Chen, Jie Shen, and Veena Tikare. Coarsening kinetics from a variable-mobility cahn-hilliard equation: Application of a semi-implicit fourier spectral method. *Physical Review E*, 60(4):3564, 1999.

A PDE, CONVERGENCE & LEARNING RATES

Here, we discuss how PDE perspectives and tools provide insights for the analysis, stability and convergence for the proposed generative particle algorithms. We focus on continuous time for convenience. By recalling Theorem 2.1,part 4. and $\gamma^{L,*} \to P$ as $L \to \infty$ (if P is absolutely continuous with respect to Q) in Remark 2.2, equation 8 becomes a Lipschitz-regularized f-divergences gradient flow (with its limit as $L \to \infty$), i.e.

$$\underbrace{\partial_t P_t = \text{div}\left(P_t \nabla f'\left(\frac{d\gamma^{L,*}}{dQ}\right)\right)}_{\text{Lip. regularized } f\text{-divergence flow}} \quad \underset{L\to\infty}{\longrightarrow} \quad \underbrace{\partial_t P_t = \text{div}\left(P_t \nabla f'\left(\frac{dP_t}{dQ}\right)\right)}_{f\text{-divergence flow}} \tag{15}$$

The right hand side of equation 15 is a nonlinear operator which encodes the Lipschitz regularization in the discriminator space. This defines a new class of PDE gradient flows where absolute continuity between P_t and Q for every $t \geq 0$, is not required, contrary to gradient flows of f-divergences (obtained as $L \to \infty$.) We discuss these connections next in the context of two well-known PDEs, the (linear) Fokker-Planck and the (non-linear) porous media equation. Rewriting the limiting equation in terms of the density $h_t = \frac{dP_t}{dQ}$ we have

1. Lipschitz-regularized Fokker-Planck. For the KL, $f(x) = x\log(x)$ we obtain

$$\partial_t P_t = \text{div}\left(P_t \nabla \log\left(\frac{d\gamma_t^{\Gamma_L,*}}{dQ}\right)\right) \quad \underset{L\to\infty}{\longrightarrow} \quad \partial_t h_t = (\Delta + \nabla \log(q)\cdot\nabla)h_t \tag{16}$$

2. Lipschitz-regularized Weighted Porous Medium equation (WPME). For the α-divergence with $f_\alpha(x) = \frac{1}{\alpha(\alpha-1)}x^\alpha$ we obtain a regularization of the porous media equation Otto (2001); Dolbeault et al. (2008)

$$\partial_t p_t + \frac{1}{\alpha-1}\text{div}\left(p_t \nabla\left(\frac{\eta_t^{\Gamma_L,*}}{q}\right)^{\alpha-1}\right) = 0 \quad \underset{L\to\infty}{\longrightarrow} \quad \frac{dh_t}{dt} = \frac{1}{\alpha}\left(\Delta + \nabla\log q\cdot\nabla\right)h_t^\alpha \tag{17}$$

A.1 CONVERGENCE TO EQUILIBRIUM AND FUNCTIONAL INEQUALITIES

Functional inequalities are fundamental methods for guaranteeing the convergence of gradient flow PDE to their equilibrium states and therefore are natural tools for studying convergence properties for the corresponding particle-based algorithms. A first step is to compute the rate of change of the divergence along solutions P_t of equation 8.

Theorem A.1. [Lipschitz regularized dissipation] *Along a trajectory of a smooth solution $\{P_t\}_{t\geq 0}$ of equation 8 with source probability distribution P we have the following rate of decay identity:*

$$\frac{d}{dt}D_f^{\Gamma_L}(P_t\|Q) = -I_f^{\Gamma_L}(P_t\|Q) \leq 0 \tag{18}$$

where we define the Lipschitz-regularized Fisher Information as

$$I_f^{\Gamma_L}(P\|Q) = \int |\nabla\phi^{L,*}|^2 P(dx) = E_P\left[\left|\nabla f'\left(\frac{d\gamma^{L,*}}{dQ}\right)\right|^2\right]. \tag{19}$$

Consequently, for any $T \geq 0$, we have $D_f^{\Gamma_L}(P_T\|Q) = D_f^{\Gamma_L}(P\|Q) - \int_0^T I_f^{\Gamma_L}(P_s\|Q)ds$.

Remark A.2. **(a)** For the generative particles the Lipschitz-regularized Fisher Information can be interpreted as their total kinetic energy, see Paragraph 11. **(b)** When $f = f_{\text{KL}}$, as $L \to \infty$, we recover the usual Fisher information $I_f^{\Gamma}(P\|Q) = E_P\left[|\nabla\log\left(\frac{p}{q}\right)|^2\right]$ which is used to prove convergence to the equilibrium state for the Fokker-Planck equation 16.

Functional inequalities, such as the classical Poincaré and the Logarithmic Sobolev-type inequalities, and many generalizations thereof see Markowich & Villani (2000); Otto & Villani (2000); Toscani & Villani (2000); Dolbeault et al. (2008); Wang (2005) are a powerful to prove convergence

to equilibrium (e.g exponential or polynomial convergence), building on dissipation estimates such as Theorem A.1. For example, if for some $\lambda > 0$, a Sobolev inequality $D_f^\Gamma(P\|Q) \leq \frac{1}{\lambda} I_f^\Gamma(P\|Q)$ holds (true when Q is sub-Gaussian), then we obtain exponential convergence to Q for any P_0: $D_f^\Gamma(P_t\|Q) \leq e^{-\lambda t} D_f^\Gamma(P_0\|Q)$. There exits various results, e.g. Carrillo et al. (2006); Dolbeault et al. (2008); Markowich & Villani (2000); Wang (2008), reviewed in the Section B, C, on functional inequalities when $L = \infty$ which have been used to prove convergence to equilibrium (at exponential or polynomial rate) for Fokker-Planck and/or the porous media equation when the target distribution are Gaussian distribution, stretched exponential and Student-t type distribution. The existence of functional inequalities for Lipschitz-regularized gradient flows is an open question.

A.2 Numerical analysis for Lipschitz regularized PDEs and generative particles

For a Lipschitz-regularized gradient flow equation 15, the transporter/discriminator representation equation 8 implies that the domain of dependence is determined by the velocity fields $\nabla\phi_t^{L,*}$ whose norm is bounded by the Lipschitz constant L. Therefore the domain of dependence of the solution is finite and is contained in a cone of slope L that emanates from any point (x, t) back to the time plane $t = 0$.

From a numerical analysis point of view, equation 9 is an explicit numerical scheme $\frac{p_{n+1}-p_n}{\Delta t} = \text{div}\left(p_n \nabla\phi_n^{L,*}(x)\right)$. For corresponding spatial discretization schemes there is an abundance of numerical methods which we can use to get some numerical analysis insight into our particle schemes equation 9. In particular, the Courant, Friedrichs, and Lewy (CFL) condition for stability of discrete schemes asserts that a numerical method can be convergent only if its numerical domain of dependence contains the true domain of dependence of the continuous PDE, LeVeque (2007). In our context, the CFL condition means $\sup_x |\nabla\phi_t^{L,*}(x)| \frac{\Delta t}{\Delta x} \leq C_{max}$ where $C_{max} = 1$ for such explicit schemes, LeVeque (2007). Clearly, the Lipschitz regularization enforces a CFL type condition with a learning rate Δt proportional to the inverse of L. It remains an open question how to rigorously extend this CFL analysis to particle-based algorithms where the spatial discretization grid Δx is known only implicitly as noted in Remark 4.1(b), see also related questions in Carrillo et al. (2017). Nevertheless, in our experiments we explore the inversely proportional relation between L and Δt suggested by the CFL analysis.

B Fokker-Planck Equation, and its convergence to equilibrium state

Generalized Fokker Planck as the gradient flow of f-divergences. Let p_t is the density of P_t. The associated gradient flow is given by the generalized Fokker-Planck equation

$$\frac{dp_t}{dt} = \nabla \cdot \left(p_t \nabla \frac{\delta D_f(P\|Q)}{\delta P}(p_t) \right) = \nabla \cdot \left(p_t \nabla f' \left(\frac{p_t}{q} \right) \right) \tag{20}$$

The Fokker Planck as gradient flow of KL. When $f = f_{\text{KL}}$, we obtain the known Fokker-Planck equation

$$\partial_t p_t = \Delta p_t - \nabla \cdot \left(p_t \frac{\nabla q}{q} \right).$$

B.1 Exponential decay when $q \propto e^{-V}$ and V is λ-convex

In this section for simplicity that the probability densities of both source and target distributions exist and are denoted by p, q. We consider the Cauchy problem of the Fokker-Planck equation given in Section A with

$$p(t = 0, \cdot) = p \geq 0 \text{ and } \int p = 1. \tag{21}$$

The next theorem in Markowich & Villani (2000) gives us the conditions that a probability measure satisfies in order to logarithmic Sobolev inequalities and consequently exponential decay.

Theorem B.1. *Let $q \in L^1(\mathbb{R}^d)$ and V be λ-convex (i.e. $D^2V(x) \geq \lambda I_d$ for all $x \in \mathbb{R}^d$), where I_d is the identity matrix of dimension d. Then, q satisfies a logarithmic Sobolev inequality with constant λ, i.e. $D_{\mathrm{KL}}(p\|q) \leq \frac{1}{2\lambda} I(p\|q)$, and the solution of the homogeneous Fokker-Planck equation goes to equilibrium in KL divergence, with a rate $e^{-2\lambda t}$ at least.*

Typical examples that satisfy the conditions of Theorem B.1 are

$$q(x) = \frac{e^{-|x|^\beta}}{\int e^{-|x|^\beta}}, \text{ for } x \in \mathbb{R}^d \quad \text{with} \qquad \beta \geq 2 \tag{22}$$

When $\beta = 2$, the target probability distribution with density q is the Gaussian with variance σ and zero mean, i.e.

$$q(x) = \frac{1}{(2\pi\sigma)^{d/2}} e^{-\frac{|x|^2}{2\sigma}},$$

(i.e. $V(x) = \frac{|x|^2}{2\sigma}$). By applying Theorem B.1, we get that for any initial probability distribution P which is absolutely continuous with respect to Q,

$$D_{\mathrm{KL}}(p_t\|q) \leq D_{\mathrm{KL}}(p_0\|q)e^{-2t/\sigma} \tag{23}$$

where we have also used that the Stam-Gross Logarithmic Sobolev inequality, i.e. $D_{\mathrm{KL}}(p\|q) \leq \frac{\sigma}{2} I(p\|q)$, see formula (14) in Markowich & Villani (2000).

B.2 Polynomial decay when $q \propto e^{-V}$ and V is degenerately convex at infinity

We consider a potential $V \in W^{2,\infty}_{\mathrm{loc}}$ such that $\int q = 1$ and degenerately convex at infinity, i.e.

$$U(u) - a \leq V(u) \leq U(u) + b \tag{24}$$

where a, b are nonnegative constants and U is convex degenerate, i.e.

$$D^2U(u) \geq c(1 + |u|)^{\beta-2}, \;\; c > 0 \;\text{ and }\; \beta \in (0, 2) \tag{25}$$

Without loss of generality we assume that U takes its unique minimum at 0. We further assume that for some $b, c, C_0 > 0$

$$\nabla V(u) \cdot u \geq c|u|^b - C_0 \tag{26}$$

A typical potential that satisfies equation 24, equation 25 and equation 26 is $V = |x|^\beta$ with $0 < \beta < 2$. Before we state the next theorem in Toscani & Villani (2000), we further define the following quantities

$$M_s(p) := \int p(x)(1 + |x|^2)^{s/2}, \;\; s > 2 \text{ and } \delta := \frac{2-\beta}{2(2-\beta) + (s-2)} \in (0, \frac{1}{2}) \tag{27}$$

Theorem B.2. *Let V be a potential satisfying assumptions equation 24, equation 25 and equation 26. Let p_0 be a probability density such that $D_{\mathrm{KL}}(p_0\|q) < \infty$, $M_s(p_0) < \infty$ given in equation 27 for $s > 2$. Let also $\{p_t\}_{t\geq 0}$ be a (smooth) solution of the Fokker-Planck equation with potential V and with initial datum p_0. Then, there is a constant C depending on $D_{\mathrm{KL}}(p_0\|q), M_s(p_0)$. and s such that for all $t > 0$,*

$$D_{\mathrm{KL}}(p_t\|e^{-V}) \leq \frac{C}{t^\kappa}, \;\; \text{with} \;\; \kappa = \frac{1-2\delta}{\delta} = \frac{s-2}{2-\beta}. \tag{28}$$

where δ is given in equation 27.

Note that as $\beta \to 2$, one recovers the usual logarithmic Sobolev inequality as discussed in Sect. B.1. We summarize the said examples in the following table.

C Weighted Porous Medium equations and their convergence to equilibrium state

C.1 Weighted Porous Medium equation

The gradient flow of f-divergences for likelihood ratio. One may rewrite equation 20 in terms of the likelihood ratio denoted by h_t and defined as

$$h_t = \frac{dp_t}{dq} \tag{29}$$

Table 3: Rate of convergence to equilibrium state $q \propto e^{-V}$ in KL divergence

Examples of $q \propto e^{-V}$	Rate of convergence in KL divergence
$q \propto e^{-\lvert x\rvert^{\beta}},\ \beta \geq 2$	at least $e^{-2\lambda t}$
Special case: $\mathcal{N}(0,\sigma)$	at least $e^{-2\lambda t}$, with $\lambda = \frac{1}{\sigma}$
$q \propto e^{-\lvert x\rvert^{\beta}},\ 0 < \beta < 2$	$\mathcal{O}(t^{-\kappa})$, κ as in equation 28

By using the operator identity (q being the multiplication operator by the function q), i.e.

$$\nabla q = q\left(\nabla + \nabla \log q\right)$$

we have that

$$\nabla \cdot p_t \nabla f'\left(\frac{p_t}{q}\right) = q(\nabla + \nabla \log q) h_t \nabla f'(h_t)$$

and thus we can rewrite equation 20 as

$$\frac{dh_t}{dt} = (\nabla + \nabla \log q) \cdot h_t \nabla f'\left(h_t\right) \tag{30}$$

Moreover if we denote ∇^* the adjoint of ∇ on $L^2(q)$ we have $\nabla^* = -(\nabla + \nabla \log q)$ and thus equation 30 has the form

$$\frac{dh_t}{dt} = -\nabla^* h_t \nabla f'(h_t) \tag{31}$$

Let now $f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$, we rewrite

$$h \nabla h^{\alpha-1} = \frac{1}{\beta}\nabla v^\beta = v^{\beta-1}\nabla r \Longrightarrow v = h^{\alpha-1} \text{ and } h = v^{\beta-1} \Longrightarrow \beta = \frac{\alpha}{\alpha-1}$$

and thus we obtain

$$\frac{dh_t}{dt} = \frac{1}{\alpha}\left(\Delta + \nabla \log q \cdot \nabla\right) h_t^\alpha \tag{32}$$

for $t \geq 0$ and $x \in \mathbb{R}^d$ corresponding to a non-negative initial condition $h(x,0) = h_0(x)$, $x \in \mathbb{R}^d$ is called weighted Porous Medium equation. For existence and uniqueness, see Dolbeault et al. (2008).

Remark C.1. The formula for f_α^* is given by

$$f_\alpha^*(y) = \begin{cases} \frac{(\alpha-1)^{\frac{\alpha}{(\alpha-1)}}}{\alpha} y^{\frac{\alpha}{(\alpha-1)}} \mathbf{1}_{y>0} + \frac{1}{\alpha(\alpha-1)} & , \alpha > 1 \\ \infty \mathbf{1}_{y \geq 0} + \left(\frac{1}{\alpha(1-\alpha)^{\frac{\alpha}{(1-\alpha)}}} |y|^{-\frac{\alpha}{(1-\alpha)}} \mathbf{1}_{y>0} - \frac{1}{\alpha(1-\alpha)}\right) \mathbf{1}_{y<0} & , \alpha \in (0,1) \end{cases} \tag{33}$$

Remark C.2. For completeness, we discuss a related gradient flow known as *granular media equation*. To be precise, the 2-Wasserstein gradient flow of $\mathcal{F}(p) = \frac{1}{2}\mathrm{MMD[p,q]}^2$ where $\mathrm{MMD[p,q]}$ is the Maximum mean discrepancy (MMD) Gretton et al. (2012). By recalling equation 2, MMD is defined as

$$\mathrm{MMD[p,q]} = \sup_{\mathrm{g} \in \mathrm{B_{RKHS}}(0,1)} \{\mathrm{E_Q[g]} - \mathrm{E_P[g]}\}$$

and its maximizer $\phi^*(z) = f_{q,p}(z) = \int k(x,z)q(x)dx - \int k(x,z)p(x)dx = k \star p(z) - k \star q(z)$ is called *witness function* between the probability densities q and p. In fact, g^* is the difference between the mean embeddings of q and p which finally makes MMD be re-written as the RKHS norm of the unnormalized g^*, i.e.

$$\mathrm{MMD[p,q]} = \|\phi^*\|_{\mathcal{H}} \tag{34}$$

Then the gradient flow equation associated to $\mathcal{F}$ leads to the granular media equation, i.e

$$\partial_t p_t(x) = div\left(p \nabla \cdot (k \star p - k \star q)\right) \equiv div\left(p_t \nabla \phi_t^*\right) \tag{35}$$

C.2 FUNCTIONAL INEQUALITIES FOR THE WEIGHTED POROUS MEDIUM EQUATION

In this section, we apply Theorem 4.5 in Dolbeault et al. (2008) to Weighted Porous Medium for the likelihood ratio $h_t = \frac{p_t}{q}$ and we prove polynomial decay in KL and χ^2-divergence. Before we state the result we first define the L^r-Poincaré inequality and L^r-logarithmic Sobolev inequality (see also Dolbeault et al. (2008)).

Definition C.3. Let q be a probability measure on a Riemannian manifold (M, g). Then the entropy is defined as follows: for any smooth function $f \in C^1(M)$

$$\mathbf{Ent}_q(f) := \int f \log\left(\frac{f}{\int f dq}\right) dq \tag{36}$$

while

$$\mathbf{Var}_q(f) := \int \left(f - \int f dq\right)^2 dq \tag{37}$$

Definition C.4. Let q be a probability measure on a Riemannian manifold (M, g). Let also ν be a positive measure on (M, g). We assume that $q \in (0, 1]$. We say that (q, ν) satisfies L^r-Poincaré inequality with constant C_{P} if and only if, for any nonnegative function $f \in C^1(M)$

$$\left[\mathbf{Var}_q(f^{2r})\right]^{1/r} \leq C_{\mathrm{P}} \int |\nabla f|^2 d\nu \tag{38}$$

We say that (q, ν) satisfies L^r-logarithmic Sobolev inequality with constant C_{LS} if and only if, for any nonnegative function $f \in C^1(M)$

$$\left[\mathbf{Ent}_q(f^{2r})\right]^{1/r} \leq C_{\mathrm{LS}} \int |\nabla f|^2 d\nu \tag{39}$$

Theorem C.5. *If (q, q) satisfies a $L^{2/3}$-Poincaré Sobolev inequality, for some constant $C_P > 0$, then for any non-negative initial condition $h_0 \equiv \frac{p_0}{q} \in L^2(q)$, we have for every $t \geq 0$*

$$\chi^2(p_t \| q) \leq \left(\left[\chi^2(p_0 \| q)\right]^{-1/2} + \frac{8}{9} C_P t\right)^{-2}. \tag{40}$$

Reciprocally, if the above inequality is satisfied for any g_0, then (q, q) satisfies a $L^{2/3}$-Poincaré Sobolev inequality with constant $C_P > 0$.

Theorem C.6. *Let $\alpha > 1$. If (q, q) satisfies a $L^{1/\alpha}$-logarithmic Sobolev inequality, for some constant $C_{LS} > 0$, then for any non-negative initial condition h_0 such that $D_{\mathrm{KL}}(p_0 \| q) < \infty$, we have for every $t \geq 0$*

$$D_{\mathrm{KL}}(p_t \| q) \leq \left([D_{\mathrm{KL}}(p_0 \| q)]^{1-\alpha} + \frac{4(\alpha - 1)}{\alpha} C_{LS} t\right)^{-1/(\alpha-1)}. \tag{41}$$

Reciprocally, if the above inequality is satisfied for any g_0, then (q, q) satisfies a $L^{1/\alpha}$-logarithmic Sobolev inequality with constant $C_{LS} > 0$.

Next we discuss two examples of probability distributions satisfy L^r-Poincaré inequality and L^r-logarithmic Sobolev inequality:

Let $r \in (0, 1]$ and $\beta \in [\frac{1}{2}, 1)$. The probability measure

$$dq = \frac{1}{2\Gamma\left(1 + \frac{1}{\beta}\right)} e^{-|x|^\beta} dx, \quad x \in \mathbb{R} \tag{42}$$

satisfies a L^r-Poincaré inequality and L^r-logarithmic Sobolev inequality.

Let $r \in [1/2, 1)$, then for $\beta > \frac{2r}{1-r}$ the probability measure

$$dq = \frac{\beta}{(1 + |x|)^{1+\beta}} dx, \quad x \in \mathbb{R} \tag{43}$$

satisfies a L^r-Poincaré inequality and a L^r-logarithmic Sobolev inequality.

Table 4: Rate of convergence to equilibrium state $q \propto e^{-V}$ in χ^2-divergence

Examples of $q \propto e^{-V}$	Rate of convergence in χ^2 divergence
$q = \frac{e^{-\lvert x\rvert^\beta}}{2\Gamma\left(1+\frac{1}{\beta}\right)}, \;\; 0 < r \leq 1\,,\; 1/2 \leq \beta < 1$	given in equation 40
$q = \frac{\beta}{(1+\lvert x\rvert)^{1+\beta}}, \;\; 1/2 \leq r < 1\,,\; \beta \geq \frac{2r}{1-r}$	given in equation 40

D FIRST VARIATION OF REGULARIZED DIVERGENCES

In this section we prove the following theorem

Theorem D.1. *Assume f is superlinear and strictly convex and $P, Q \in \mathcal{P}_1(\mathbb{R}^d)$.*

1. *For $x \notin \operatorname{supp}(P) \cap \operatorname{supp}(Q)$ define $\phi^{L,*}(y) = \sup_{x\in\operatorname{supp}(Q)} \left\{\phi^{L,*}(x) + L|x-y|\right\}$ then $\phi^{L,*}$ is Lipschitz continuous on $\mathbb{R}^d$.*
2. $\phi^{L,*} = \sup\{h(x) : h \in \Gamma_L,\, h(y) = \phi^{L,*}(y),$*for every* $y \in \operatorname{supp}(Q)\}$
3. *Let ρ be a signed measure of total mass 0 and let $\rho = \rho_+ - \rho_-$ where $\rho_\pm \in \mathcal{P}_1(K)$ are mutually singular. If $P + \epsilon\rho \in \mathcal{P}_1(K)$ for sufficiently small $|\epsilon|$ then $D_f^{\gamma_L}(P+\epsilon\rho\|Q)$ is differentiable at $\epsilon = 0$ and we have*

$$\lim_{\epsilon\to 0} \frac{1}{\epsilon}\left(D_f^{\Gamma_L}(P+\epsilon\rho\|Q) - D_f^{\Gamma_L}(P\|Q)\right) = \int \phi^{L,*} d\rho\,.$$

In other words we have

$$\frac{\delta D_f^{\Gamma_L}(P\|Q)}{\delta P}(P) = \phi^{L,*}$$

Proof. The proof of 1. is straightforward by using the triangular inequality of norms. For 2., since $h \in \Gamma_L$, we have that $h(x) \leq h(y) + \|x-y\|$. This implies that for $y \in \operatorname{supp}(Q)$ and $x \notin \operatorname{supp}(Q)$, $h(x) \leq \inf_{y\in\operatorname{supp}(Q)}\{h(y) + \|x-y\|\} = \inf_{y\in\operatorname{supp}(Q)}\{\phi^{L,*}(y) + \|x-y\|\} = \phi^{L,*}(x)$. Since $\phi^{L,*}(y) \in \Gamma_L$, this concludes the proof. For 3., we use the variational formula equation 3 for $D_f^{\Gamma_L}(P+\epsilon\rho\|Q)$ where we suppose that $P + \epsilon\rho \in \mathcal{P}_1(\mathbb{R}^d)$.

$$\begin{aligned} D_f^{\Gamma_L}(P+\epsilon\rho\|Q) &= \sup_{\phi\in\Gamma_L}\left\{E_{P+\epsilon\rho}[\phi] - \inf_{\nu\in\mathbb{R}}\left\{\nu + E_Q[f^*(\phi-\nu)]\right\}\right\} \\ &\geq \int \phi^{*,L}\, d(P+\epsilon\rho) - \inf_{\nu\in\mathbb{R}}\left\{\nu + \int f^*(\phi^{*,L}-\nu)dQ\right\} \\ &= \epsilon\int \phi^{*,L} d\rho + D_f^{\Gamma_L}(P\|Q) \end{aligned} \tag{44}$$

Thus

$$\liminf_{\epsilon\to 0^+} \frac{1}{\epsilon}\left(D_f^{\Gamma_L}(P+\epsilon\rho\|Q) - D_f^{\Gamma_L}(P\|Q)\right) \geq \int \phi^{*,L} d\rho$$

For the other direction: Let us define $F(\epsilon) = D_f^{\Gamma_L}(P+\epsilon\rho\|Q)$. By Theorem 18 and 71 in Birrell et al. (2022a), $F(\epsilon)$ is convex, lower semicontinuous and finite on $[0, \epsilon_0]$. Due to the convexity of F, F is differentiable on $(0, \epsilon_0)$ except for a countable number of points. Let $\epsilon \in (0, \epsilon_0)$ such that F is differentiable and $\delta > 0$ small. Also, let $\phi_\epsilon^{*,L}$ be the optimizer of $D_f^{\Gamma_L}(P+\epsilon\rho\|Q)$ satisfying $\phi_\epsilon^{*,L}(0) = 0$ so that

$$D_f^{\Gamma_L}(P+\epsilon\rho\|Q) = \int \phi_\epsilon^{*,L}\, d(P+\epsilon\rho) - \inf_{\nu\in\mathbb{R}}\left\{\nu + \int f^*(\phi_\epsilon^{*,L}-\nu)dQ\right\}$$

By using the same argument as before in the proof, we have that

$$D_f^{\Gamma_L}(P+(\epsilon+\delta)\rho\|Q) - D_f^{\Gamma_L}(P+\epsilon\rho\|Q) \geq \delta\int \phi_\epsilon^{*,L} d\rho \tag{45}$$

and

$$D_f^{\Gamma_L}(P+(\epsilon-\delta)\rho\|Q) - D_f^{\Gamma_L}(P+\epsilon\rho\|Q) \geq -\delta \int \phi_\epsilon^{*,L} d\rho \tag{46}$$

which gives us that

$$\begin{aligned}
\int \phi_\epsilon^{*,L} d\rho &\leq \lim_{\delta\to 0} \frac{1}{\delta}\left(D_f^{\Gamma_L}(P+(\epsilon+\delta)\rho\|Q) - D_f^{\Gamma_L}(P+\epsilon\rho\|Q)\right) \\
&= F'(\epsilon) \\
&= \lim_{\delta\to 0} \frac{1}{\delta}\left(D_f^{\Gamma_L}(P+\epsilon\rho\|Q) - D_f^{\Gamma_L}(P+(\epsilon-\delta)\rho\|Q)\right) \\
&\leq \int \phi_\epsilon^{*,L} d\rho
\end{aligned} \tag{47}$$

Consequently,

$$F'(\epsilon) = \int \phi_\epsilon^{*,L} d\rho \tag{48}$$

Let $F'_+(0) = \lim_{\epsilon\to 0^+} \frac{1}{\epsilon}\left(F(\epsilon) - F(0)\right)$. By convexity, for any sequence $\{\epsilon_n\}_{n\in\mathbb{N}}$ such that $\epsilon_0 > \epsilon_n \downarrow 0$, we have

$$F'_+(0) = \lim_{n\to\infty} F'(\epsilon_n) = \lim_{n\to\infty} \int \phi_{\epsilon_n}^{*,L} d\rho$$

By applying the Arzelá-Ascoli to $\phi_{\epsilon_n}^{*,L}$, and then doing a diagonalization argument, there exists a subsequence of $\{n_k\}_{k\geq 0} \subset \{n\}_{n\geq 0}$, such that $\phi_{\epsilon_{n_k}}^{*,L}$ converges pointwise to a function $\phi_0^{*,L} \in \mathrm{Lip}^L(\mathbb{R}^d)$. For simplicity, from now on we denote n the convergent subsequence.

At this point, we recall that for any $\epsilon \in (0,\epsilon_0)$, $\phi_\epsilon^{*,L}(0) = 0$. For any x, $|\phi_\epsilon^{*,L}(x) - \phi_\epsilon^{*,L}(0)| \leq L\|x\|_d$ which implies that

$$|\phi_\epsilon^{*,L}(x)| \leq L\|x\|_d$$

Thus by the dominated convergence theorem

$$F'_+(0) = \lim_{n\to\infty} \int \phi_{\epsilon_n}^{*,L} d\rho = \int \phi_0^* d\rho$$

By the lower semicontinuity of $D_f^{\Gamma_L}(\cdot\|Q)$, we have

$$\begin{aligned}
D_f^{\Gamma_L}(P\|Q) &\leq \liminf_{n\to\infty} D_f^{\Gamma_L}(P+\epsilon_n\rho\|Q) \\
&= \liminf_{n\to\infty} \left\{ E_{P+\epsilon_n\rho}[\phi_{\epsilon_n}^{*,L}] - \inf_{\nu\in\mathbb{R}} \left\{\nu + E_Q[f^*(\phi_{\epsilon_n}^{*,L} - \nu)]\right\} \right\} \\
&= \liminf_{n\to\infty} E_{P+\epsilon_n\rho}[\phi_{\epsilon_n}^{*,L}] - \limsup_{n\to\infty} \inf_{\nu\in\mathbb{R}} \left\{\nu + E_Q[f^*(\phi_{\epsilon_n}^{*,L} - \nu)]\right\} \\
&\leq E_P[\phi_0^{*,L}] - \inf_{\nu\in\mathbb{R}} \left\{\nu + E_Q[f^*(\phi_0^{*,L} - \nu)]\right\} \\
&\leq D_f^{\Gamma_L}(P\|Q)
\end{aligned} \tag{49}$$

where for the second inequality we use the dominated convergence theorem, equation 48 and that by Fatou's lemma

$$\limsup_{n\to\infty} \int \phi_{\epsilon_n}^{*,L} dQ \geq \liminf_{n\to\infty} \int \phi_{\epsilon_n}^{*,L} dQ \geq \int \phi_0^{*,L} dQ$$

Since both sides of the inequality coincide, $\phi_0^{*,L}$ must be the optimizer. By Theorem 3.1, part 1. and Theorem D.1, part 2., we have that $\phi_0^{*,L}(x) \leq \phi^{*,L}$ for all x. Thus

$$F'_+(0) = \int \phi_0^* d\rho \leq \int \phi^* d\rho.$$

which concludes the proof. □

E DETAILS ON LATENT GENERATIVE PARTICLES

E.1 GENERALIZATION OF ENCODER AND DECODER FUNCTIONS

In the Section 5, we built Theorem 5.1 based on an encoder map $\mathcal{E} : \mathcal{Y} \to \mathcal{Z}$ (for instance $\mathcal{Y} = \mathbb{R}^d$, $\mathcal{Z} \subset \mathbb{R}^{d'}$, $d' << d$) and a decoder map $\mathcal{D} : \mathcal{Z} \to \mathcal{Y}$. For the generalization purpose, we identify these mappings as dirac kernels $K_{\mathcal{E}}(y, dz) = \delta_{\mathcal{E}(y)}(dz)$ and $K_{\mathcal{D}} = \delta_{\mathcal{D}(z)}(dy)$.

1. A pullback function induced by the kernel $K_{\mathcal{E}}$ is given as
$$K_{\mathcal{E}}[f](y) := \int f(z')K_{\mathcal{E}}(y, dz') = f(\mathcal{E}(y)) \tag{50}$$
for $f \in \mathcal{M}_b(\mathcal{Z}), y \in \mathcal{Y}$
2. A push-forward measure induced by the kernel $K_{\mathcal{E}}$ maps $\mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Z})$ and defines a push-forward measure which is given as
$$K_{\mathcal{E}}[P](B) := \int K_{\mathcal{E}}(y, B)P(dy) = P \circ \mathcal{E}^{-1}(B) \tag{51}$$
for $P \in \mathcal{P}(\mathcal{Y})$ and a $\mathcal{Z}$-measurable set B.

Likewise, the kernel $K_{\mathcal{D}}$ induces a pullback function and a push-forward measure in the opposite direction. In the previous formulation, the $Q^{\mathcal{Y}}$-perfect encoding property $\mathcal{D}_{\#}\mathcal{E}_{\#}Q^{\mathcal{Y}} = Q^{\mathcal{Y}}$ can be rewritten as $Q^{\mathcal{Y}} = K_{\mathcal{D}}[K_{\mathcal{E}}[Q^{\mathcal{Y}}]]$. Given any $P^{\mathcal{Y}} \in \mathcal{P}(\mathcal{Y})$, we have the latent probability measure $P^{\mathcal{Z}} = K_{\mathcal{E}}[P^{\mathcal{Y}}] \in \mathcal{P}(\mathcal{Z})$ and the reconstructed probability measure $\tilde{P^{\mathcal{Y}}} = K_{\mathcal{D}}[K_{\mathcal{E}}[P^{\mathcal{Y}}]] \in \mathcal{P}(\mathcal{D}(\mathcal{Z}))$ where $\mathcal{D}(\mathcal{Z}) \subset \mathcal{Y} = \mathbb{R}^d$. In general, $\tilde{P^{\mathcal{Y}}} \neq P^{\mathcal{Y}}$.

Transition probability kernels are defined in the form of conditional distributions: $K_p(y, dz) = p(dz|y)$ from $\mathcal{Y}$ to $\mathcal{Z}$ and $K_q(z, dy) = q(dy|z)$ from $\mathcal{Z}$ to $\mathcal{Y}$. The kernel-induced pullback functions $K_p[f](y) = \int f(z')p(dz'|y) = \mathbb{E}_{Z|Y=y\sim p(dz|y)}[f(Z)|Y = y]$ or $K_q[g](z) = \int f(y')q(dy'|z) = \mathbb{E}_{Y|Z=z\sim q(dy|z)}[g(Y)|Z = z]$ are interpreted as conditional expectations. In addition, the kernels induce push forward measures $P^{\mathcal{Z}}(dz) = p(dz|y)P^{\mathcal{Y}}(dy)$ for $P^{\mathcal{Y}} \in \mathcal{P}(\mathbb{R}^d)$ or $R^{\mathcal{Y}}(dy) = q(dy|z)R^{\mathcal{Z}}(dy)$ for $R^{\mathcal{Z}} \in \mathcal{P}(\mathbb{R}^{d'})$. For the Q-perfect encoding property, we require these kernels to satisfy $dQ^{\mathcal{Y}}(dy) = q(dy|z)p(dz|y)dQ^{\mathcal{Y}}(dy)$.

E.2 DATA PROCESSING INEQUALITY AND PROOF OF THEOREM 5.1

The proof of Theorem 5.1 is a consequence of a, new, tighter data processing inequality derived in Birrell et al. (2022a) that involves both transformations of probabilities and discriminator spaces Γ.

Theorem E.1 (**Data processing inequality for** (f, Γ) **-divergences).** *Given a real valued convex function f, $P, Q \in \mathcal{P}(\Omega)$, and a probability kernel K from $(\Omega, \mathcal{M})$ to $(N, \mathcal{N})$, if $\Gamma \subset \mathcal{N}$ is nonempty, then*
$$D_f^{\Gamma}(K[P]\|K[Q]) \leq D_f^{K[\Gamma]}(P\|Q). \tag{52}$$

Proof. From the variational formulation of divergences, we have
$$D_f^{\Gamma}(K[P]\|K[Q]) = \sup_{\phi\in\Gamma,\nu\in\mathbb{R}} \int\int(\phi(y)-\nu)K(x, dy)P(dx) - \int\int f^*(\phi(y)-\nu)K(x, dy)Q(dx). \tag{53}$$
Since f^* is convex, Jensen's inequality gives
$$\int f^*(\phi(y) - \nu)K(x, dy) \geq f^*\left(\int(\phi(y) - \nu)K(x, dy)\right) \tag{54}$$
for all $x \in \Omega$. Hence,
$$D_f^{\Gamma}(K[P]\|K[Q]) \leq \sup_{\phi\in\Gamma,\nu\in\mathbb{R}} \mathbb{E}_P[K[\phi] - \nu] - \mathbb{E}_Q[f^*(K[\phi] - \nu)] = D_f^{K[\Gamma]}(P\|Q). \tag{55}$$

□

Now we state and prove the generalized version of the Theorem 5.1.

Theorem E.2. *Suppose that*

1. **Perfect encoding.** *For $Q^{\mathcal{Y}}$ the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are such that $K_{\mathcal{D}}[K_{\mathcal{E}}[Q^{\mathcal{Y}}]] = Q^{\mathcal{Y}}$.*
2. $K_{\mathcal{D}}[\Gamma_{\mathcal{Y}}] \subset \Gamma_{\mathcal{Z}}$**.** *The pullback functions induced by the decoder kernel is included in the real function space.*

Then, for any $P^{\mathcal{Z}} \in \mathcal{P}(\mathbb{R}^{d'})$ we have

$$D_f^{\Gamma_{\mathcal{Y}}}(K_{\mathcal{D}}[P^{\mathcal{Z}}]||Q^{\mathcal{Y}}) \leq D_f^{\Gamma_{\mathcal{Z}}}(P^{\mathcal{Z}}||K_{\mathcal{E}}[Q^{\mathcal{Y}}]). \tag{56}$$

Proof. Since the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$ perfectly reconstruct $Q^{\mathcal{Y}}$,

$$D_f^{\Gamma_{\mathcal{Y}}}(K_{\mathcal{D}}[P^{\mathcal{Z}}]\|Q^{\mathcal{Y}}) = D_f^{\Gamma_{\mathcal{Y}}}(K_{\mathcal{D}}[P^{\mathcal{Z}}]\|K_{\mathcal{D}}[K_{\mathcal{E}}[Q^{\mathcal{Y}}]]).$$

From data processing inequality,

$$D_f^{\Gamma_{\mathcal{Y}}}(K_{\mathcal{D}}[P^{\mathcal{Z}}]\|K_{\mathcal{D}}[K_{\mathcal{E}}[Q^{\mathcal{Y}}]]) \leq D_f^{K_{\mathcal{D}}[\Gamma_{\mathcal{Y}}]}(P^{\mathcal{Z}}\|K_{\mathcal{E}}[Q^{\mathcal{Y}}]).$$

By the assumption that $K_{\mathcal{D}}[\Gamma_{\mathcal{Y}}] \subset \Gamma_{\mathcal{Z}}$,

$$D_f^{K_{\mathcal{D}}[\Gamma_{\mathcal{Y}}]}(P^{\mathcal{Z}}\|K_{\mathcal{E}}[Q^{\mathcal{Y}}]) \leq D_f^{\Gamma_{\mathcal{Z}}}(P^{\mathcal{Z}}\|K_{\mathcal{E}}[Q^{\mathcal{Y}}]).$$

□

E.3 Latent generative particles algorithm

Algorithm 2: Latent Lipschitz regularized generative particles algorithm

Require: f defined in equation 2 and its Legendre conjugate f^*, L: Lipschitz constant, T: number of updates for the particles, γ: time step size, N: number of particles

Require: $W = \{W^l\}_{l=1}^D$: parameters for the neural network $\phi : \mathbb{R}^{d'} \to \mathbb{R}$, D: depth of the neural network, δ: learning rate of the neural network, T_{NN}: number of updates for the neural network.

Require: $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}^{d'}$: trained encoder, $\mathcal{D} : \mathbb{R}^{d'} \to \mathbb{R}^d$: trained decoder.

Result: $\{Y_T^{(i)}\}_{i=1}^N$

Sample $\{\bar{X}^{(i)} = \mathcal{E}(X^{(i)}) \in \mathbb{R}^{d'}\}_{i=1}^N$ where $\{X^{(i)}\} \sim Q$ is a batch from the real data

Sample $\{\bar{Y_0}^{(i)} = \mathcal{E}(Y^{(i)}) \in \mathbb{R}^{d'}\}_{i=1}^N$ where $\{Y^{(i)}\} \sim P_0 = P$ is a batch of prior samples

Apply Lipschitz regularized generative particles algorithm 1 on $\bar{X}^{(i)}$ and $\bar{Y_0}^{(i)}$

Reconstruct $Y_T^{(i)} = \mathcal{D}(\bar{Y_T}^{(i)})$

F Experimental setting

Neural network architectures. We use the discriminator ϕ (compared to GAN setting) which is implemented using a neural network. In Table 5 we provide the architecture of the neural networks used to produce the experimental results. The Lipschitz constraint on ϕ is implemented by spectral normalization (the weight matrix in each layer of the D layers in total has spectral norm $\|W^l\|_2 = L^{1/D}$).

Mini-batch scheme. To handle larger data sets with $N > 200$ we use mini-batch training where we use, at each iteration step, N_{mb} randomly chosen data points from the target distribution to move N_{mb} particles. A proper choice of mini batch size N_{mb} should be chosen to learn the distribution as well as reduce the computation time. Default setting is $N_{\text{mb}} = 200$.

CNN Discriminator
5×5 Conv SN, 2×2 stride ($1 \to ch_1$) leaky ReLU Dropout, rate 0.3
5×5 Conv SN, 2×2 stride ($ch_1 \to ch_2$) leaky ReLU Dropout, rate 0.3
5×5 Conv SN, 2×2 stride ($ch_2 \to ch_3$) leaky ReLU Dropout, rate 0.3
Flatten with dimension ℓ_3
$W^4 \in \mathbb{R}^{\ell_3 \times d}$ with SN, $b^4 \in \mathbb{R}^d$ ReLU
$W^5 \in \mathbb{R}^{d \times 1}$ with SN, $b^5 \in \mathbb{R}$ Linear

(a) Image data (MNIST, CIFAR10)

FNN Discriminator
$W^1 \in \mathbb{R}^{d \times \ell_1}$ with SN, $b^1 \in \mathbb{R}^{\ell_1}$ ReLU
$W^2 \in \mathbb{R}^{\ell_1 \times \ell_2}$ with SN, $b^2 \in \mathbb{R}^{\ell_2}$ ReLU
$W^3 \in \mathbb{R}^{\ell_2 \times \ell_3}$ with SN, $b^3 \in \mathbb{R}^{\ell_3}$ ReLU
$W^4 \in \mathbb{R}^{\ell_3 \times 1}$ with SN, $b^4 \in \mathbb{R}$ Linear

(b) Low dimensional data with dimension d

Table 5: Neural network architectures of the discriminator $\phi : \mathbb{R}^d \to \mathbb{R}$

Dataset	f	L	data parameter	NN structure	learning rate	Δt	N_Q
2D t	KL, $\alpha = 2, 10$	1	$\nu = 0.5$	FFN $(32, 32, 32)$	0.01	0.1	200
Gene data	KL	1	$d = 2, 5, 10$ $d = 20, 50, 100$ $d = 200$	FFN $(32, 32, 32)$ FFN $(64, 64, 64)$ FFN $(128, 128, 128)$	0.1	5.0	245
MNIST	KL	1	conditioned	CNN $(128, 128, 128)$	0.05	1.0	200, 2K
MNIST	KL	1	$2 \to 0$	CNN $(64, 64, 64)$	0.01	0.1	200
	KL	1	$d' = 64$ $d' = 128$	FNN $(128, 256, 256)$ FNN $(256, 512, 256)$	0.01	0.1	200
2D Gaussian	KL	1,10	$\sigma_Q = 2.0$	FFN $(32, 32, 32)$	0.05	1.0	200
2D $e^{-\|x\|^\beta}$	KL $\alpha = 2$	1	$\beta = 0.4$	FFN $(32, 32, 32)$	0.05	0.1	200
CIFAR10	KL	1	label=0	CNN $(128, 128, 128)$	0.01	0.1	200
2D Gaussian Mixture	KL	1, 10, 100, None	$\sigma_Q = 0.5$	FFN $(32, 32, 32)$	0.005	1.0	200
12D Gaussian Mixture	KL	1	$\sigma_Q = 0.5$ $d' = 2$	FFN $(32, 32, 32)$ FFN $(32, 32, 32)$	0.005	0.5 0.05	5K 200

Table 6: Data sets and important parameters

Data sets and important parameters. See Table 6. More details can be found in Supplementary material README.md.

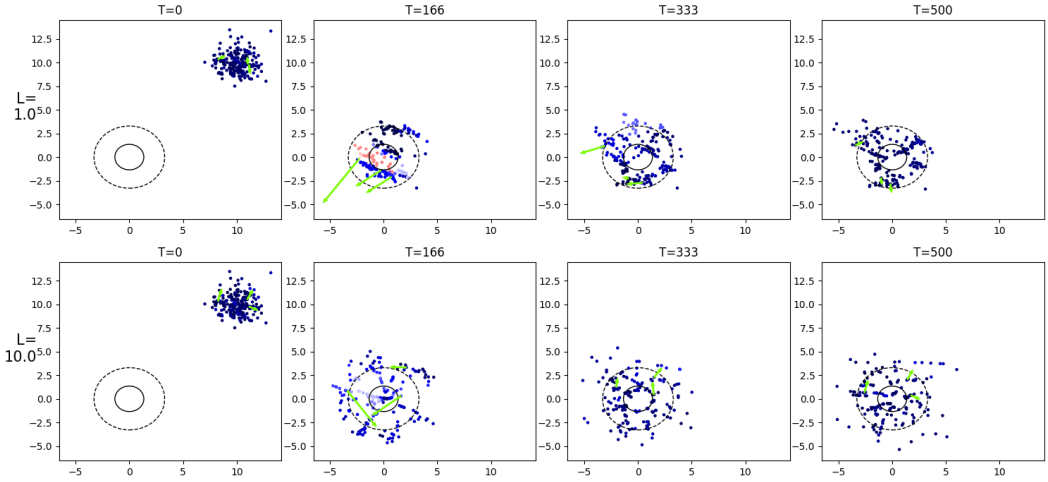
Computational resources. Low dimensional examples G.1 are computed in the tensorflow-CPU environment: *tensorflow-gpu=2.8.0 with CPU model Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz $\sim$ 2.11 GHz*. High dimensional examples G.2, G.3, G.4 are computed in the tensorflow-GPU environment: *tensorflow-gpu=2.7.0 with GPU model Tesla K80 in Google cloud platform*.

G ADDITIONAL EXPERIMENTS

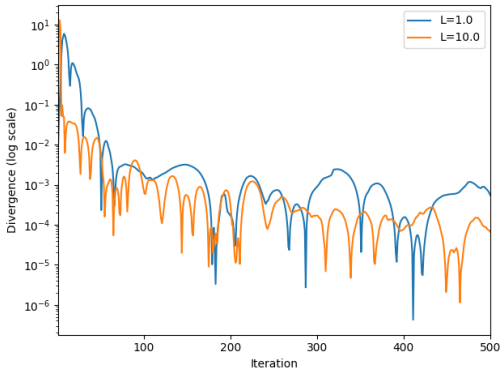
G.1 LOW DIMENSIONAL EXAMPLES WITH DIFFERENT TAILS

Our low dimensional examples illustrate the interplay between the Lipschitz regularization parameter L which stabilizes the propagation of particles and the choice of divergences f to capture the tail behavior of the target Q. We have explored exponential tails in Figure 5 for the 2D Gaussians with different L's, and Figure 6 for the 2D stretched exponential $\exp(-|x|^{0.4})$.

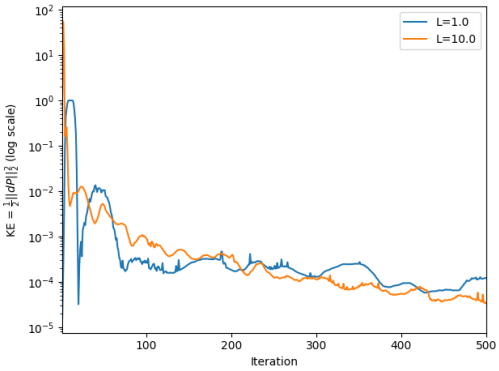
In the figures, the plots (a) exhibit the trajectory of particle at the labeled time step T with its velocity (green arrows). The color map indicates the speed of each particle: Blue (Low) $\rightarrow$ Red (High), which is fixed for all figures. The contours indicate 25%, 50% percentile regions for the stretched exponential and 50%, 90% percentile regions for the Gaussian target distribution. The empirical Lipschitz-regularized f divergence (loss function) and the empirical Lipschitz-regularized Fisher Information are given in plots (b) and (c), respectively.



(a) **(Top)** $f = f_{\text{KL}}$ with $L = 1.0$, **(Bottom)** $f = f_{\text{KL}}$ with $L = 10.0$.

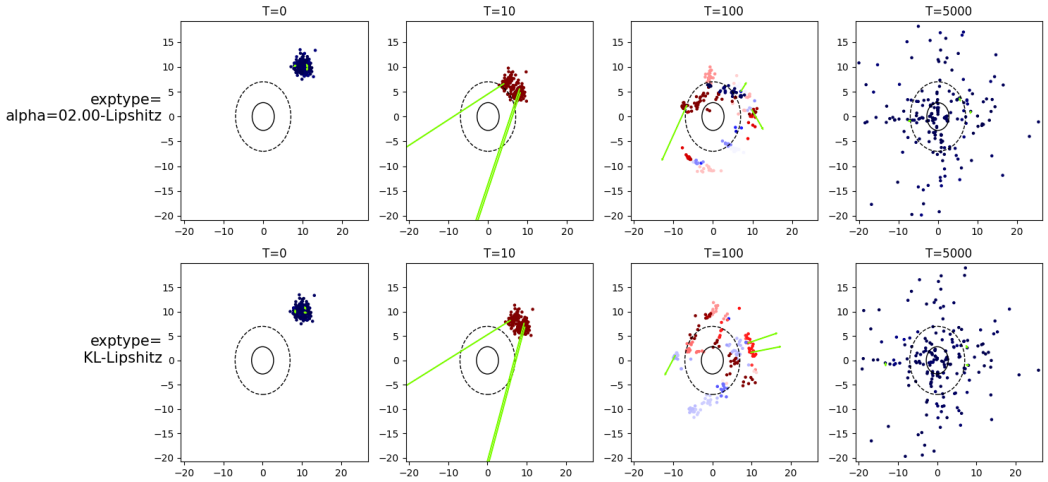


(b) KL-divergence

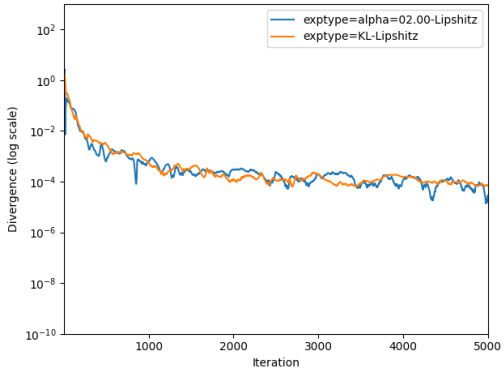


(c) Fisher information

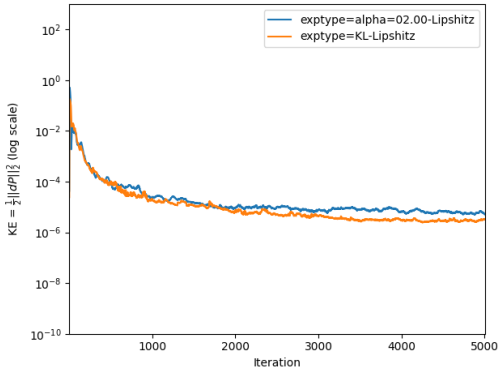
Figure 5: **(2D Gaussian Q, $\sigma_Q = 2$) The Lipschitz constraint L as a regularizer.** When learning Gaussians, particles can blow up without a regularization L. In fact, the size of L depends on data and affects the learning rate Δt. This observation is in alignment with the interpretation of L with the CFL condition in Section A.2. In this example, $\sigma_Q = 2.0 > \sigma_P = 1.0$. For $L \geq 1$, we have convergence. However, $L = 100, \infty$ blew up within fixed learning rates $\gamma = 1.0, \delta = 0.05$ in Algorithm 1. $L = 100$ requires a careful choice of learning rates $\gamma = 10.0, \delta = 0.005$ to have a similar result. For most cases, $L = 1$ produces stable results.



(a) **(Top)** $f = f_\alpha$ with $\alpha = 2$ and $L = 1$, **(Bottom)** $f = f_{\text{KL}}$ with $L = 1$.



(b) Divergences



(c) Kinetic energy of particles

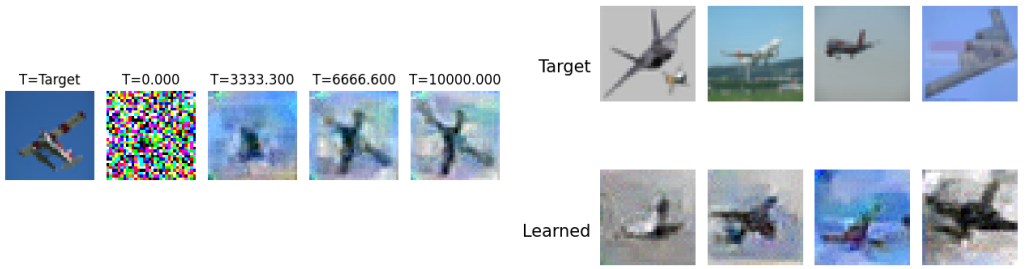
Figure 6: **(2D stretched exponential** $\propto \exp(-|x|^{0.4})$**,** $\beta = 0.4$**) The convergence in** $f = f_{\text{KL}}$ **and** $f = f_\alpha$ **with** $\alpha = 2$**.** This example has a heavier tail compared to the Gaussian ($\beta = 2 > 0.4$). **(a)** The trajectories of particles from KL flow 16 and $\alpha = 2$ flow 17 regularized by $L = 1$ look similar. The particles are transported to the 50 percentile region at an early stage and then redistributed. **(b)** For both f, the losses decrease without lots of fluctuations. The choice of $L = 1$ along with the learning rates $\gamma = 1.0, \delta = 0.05$ produces stable results. **(c)** The kinetic energies of the particle systems decay fast and converge. Indeed, the convergence of stretched exponentials is guaranteed by the classical functional inequalities of $f = \alpha$, KL gradient flows in Appendix B, C.

G.2 Additional high dimensional example

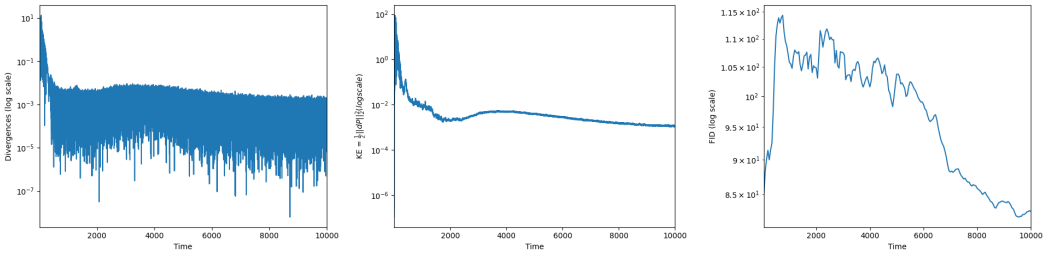
CIFAR10. We applied $(f_{\text{KL}}, \Gamma_1)$ generative particle algorithm and generated high dimensional image data. Using Lipschitz regularized KL-divergence, we transported samples from the *logistic distribution* P to samples from the target image *distribution for airplanes* Q. L is fixed to 1 throughout as we found, empirically, that the algorithm is stable for this value and unstable for larger L (for example when $L = 10$ the algorithm diverges). Training data which are samples from the target distribution Q are described as: $N = 200$ number of color images whose pixels are in $[0, 1]^{32\times32\times3}$. On purpose, we used a small number of samples $N = 200$ which corresponds to 3.3% of the entire 60,000 data, to generate new images, see Figure 7.

Overall, we observe from our experiments that Lipschitz regularized particle methods generate realistic looking samples from a relatively small number of "real" samples from Q. We suspect that learning the empirical distribution of Q using generative particles allows us to approximate individual samples from Q see equation 10. In fact, our setting rigorously allows comparing non absolutely

continuous empirical distributions while the particles are driven towards the empirical distribution of Q by the gradient flow, see equation 10.



(a) Time trajectories (b) Target samples and output samples



(c) $(f_{\mathrm{KL}}, \Gamma_1)$ divergence (d) Kinetic energy of particles (e) FID

Figure 7: **(CIFAR10, 200 Training samples) Generative particles algorithm on 2D images are less likely to meet with mode collapse even with small number of training data. (a)** Time trajectories of $(f_{\mathrm{KL}}, \Gamma_1)$-generative particle algorithm. Particles from the logistic distribution in the space $[0, 1]^{3072}$ are transported to the label 0 (airplane) data. **(b)** The generated samples vary in shape without mode collapse. **(c)** The particles are transported through the $(f_{\mathrm{KL}}, \Gamma_1)$-flow. **(d)** Lipschitz regularized fisher information gives a more stable performance measure compared to the $(f_{\mathrm{KL}}, \Gamma_1)$-divergence. **(e)** The final FID score is 82.61. Learning rates are chosen as $\gamma = 0.1, \delta = 0.01$.

G.3 Comparison with other generative models

2D Mixture of Gaussians. The target distribution Q consists of 4 wells of 2D Gaussian distributions with the same standard deviation 0.5 and centered at $(0, 0), (4, 0), (0, 4), (4, 4)$, respectively. We transport 200 samples from an initial Gaussian distribution P centered on the well at $(4, 4)$ and with standard deviation 0.5.

We compare the performance of $(f_{\mathrm{KL}}, \Gamma_L)$-generative particle algorithm with the corresponding GAN, as well as gradient flows such as KALE flow Glaser et al. (2021) and MMD flow Arbel et al. (2019), and the KL unadjusted Langevin algorithm Durmus & Éric Moulines (2017). We can compare the behaviors of our generative particle algorithm and GAN, and its dependence on the Lipschitz regularization L. The $(f_{\mathrm{KL}}, \Gamma_L)$-generative particle algorithm and KALE flow use different regularization techniques where the former provides an interpolation between KL divergence and Wasserstein metric (Γ_1) and the latter provides an interpolation between KL divergence and Maximum Mean Discrepancy (MMD). We observed that the former captured the four wells while the latter failed. The comparison with KL ULA and MMD flow suggests that the use of regularization enables the convergence without further techniques such as adding noise. See Figure 8.

2D Mixture of Gaussians embedded in 12D. The target distribution Q is a mixture of 2D Gaussians (as in the previous example) but supported in a plane embedded in 12 dimensional space. In Figure 9 we compare the performance of $(f_{\mathrm{KL}}, \Gamma_1)$-generative particle algorithm with the corre-

sponding GAN and Wasserstein GAN, in the sense of transportation and redistribution of the mass in the plane supporting the distribution and its orthogonal subspace. $L = 1$ is fixed. A Gaussian distribution centered at $(5, \cdots, 5) \in \mathbb{R}^{12}$ and with standard deviation 0.5 is picked as the initial distribution P for the $(f_{\text{KL}}, \Gamma_1)$-generative particle algorithm.

G.4 Additional latent generative particles emxample

2D Mixture of Gaussians embedded in 12D (Revisited). The Figure 9 (b) shows that Lipschitz regularized generative particles algorithm struggles in making particles converge to the 2D plane where the mixture of Gaussians lies. It is interpreted that our algorithm produces noisy outputs compared to GANs. The main reason is for each step GPA minimizes $D_f^{\Gamma_L}(P\|Q)$ using the first order explicit Euler method while GANs minimize $D_f^{\Gamma_L}(P\|Q)$ by training a nonlinear generator g_{θ_t}. To make up for this behavior, we capture the sub-manifold (2D plane) from the latent space obtained by auto-encoders. Then our algorithm transports particles to capture the four wells. It is already shown in the Figure 8 that our algorithm performs well in this task. The key for the success of latent GPA is the Q-perfect encoding property in the Theorem 5.1. We compare the two Figures 10 and 11 where the latent spaces are obtained by training auto-encoders using different losses: the former uses Mean Square Error (MSE) in the real space, and the latter uses MSE in the real space regularized by KL divergence ($2 *$ MSE $+$ KL). The former auto-encoder is over-fitted and fails to achieve Q-perfect encoding property. In this case, the divergence converged in the latent space but the reconstructed outputs did not capture the four wells. However, the outputs from the latter captures the four wells in the 2D plane as well as the 10D orthogonal subspace.

G.5 Microarray gene expression data

The flexibility on the choice of source distributions and the small sample size regime enable our generative particles algorithm to be used for medical data-processing purpose. In addition, using latent generative particles scheme, we can effectively handle high-dimensional data, such as gene expression data, typically in the dimension of $5 \sim 6 \times 10^5$ depending on the probes in a significantly reduced dimensions.We suggest batch normalization/data merging as an application of our algorithm.

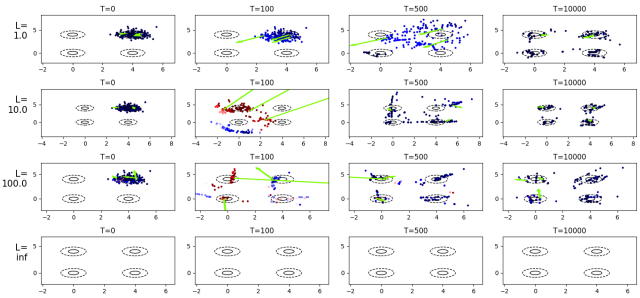
Dataset. We tested on publicly available gene expression data sets from Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo/`):

- Breast cancer: Accession number GSE47109 (206 samples), GSE10843 (245 samples)

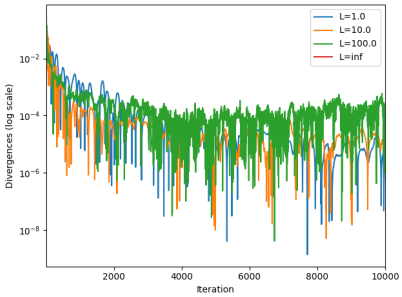
The former forms the source dataset, and the latter forms the target dataset. The source and the target data lie in the same dimensional space $\mathbb{R}^{54,675}$.

Auto-encoder. We applied PCA on the combined matrix for the source data and the target data which have been firstly normalized to mean zero and variance one. The normalized PCA can be interpreted as a linear auto-encoder. The PCA decoder is Lipschitz continuous with $L = \sqrt{d'}$. (Let $\mathbf{y} = \sum_{i=1}^{d'} z_i \mathbf{v}_i, \mathbf{y}' = \sum_{i=1}^{d'} z_i' \mathbf{v}_i \in \mathbb{R}^d$. The decoder $\mathcal{D}(\mathbf{z}) = \sum_{i=1}^{d'} z_i \mathbf{v}_i$ satisfies $\|\mathcal{D}(\mathbf{z}) - \mathcal{D}(\mathbf{z}')\| = \|\sum_{i=1}^{d'}(z_i - z_i')\mathbf{v}_i\| \leq \|\mathbf{z} - \mathbf{z}'\|\sqrt{\sum_{i=1}^{d'}\|\mathbf{v}_i\|^2} = \sqrt{d'}\|\mathbf{z} - \mathbf{z}'\|$ by Cauchy-Schwarz inequality.)

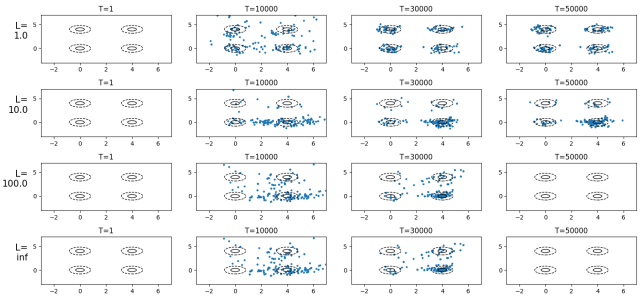
Outputs See Figures 12, 13 for the transported particles in the latent space for varying $d' = 2, 5, 10, 20, 50, 100, 200$ and the reconstructed space.



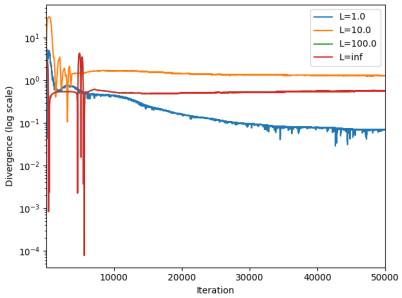
(a) Trajectory of $(f_{\mathrm{KL}}, \Gamma_L)$-generative particle algorithm



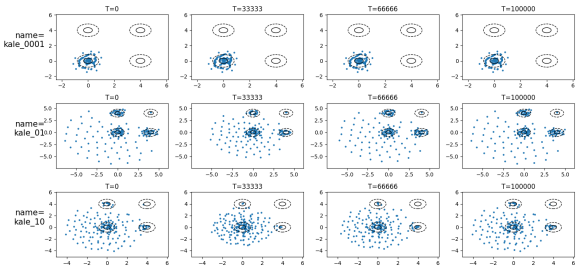
(b) Divergences of $(f_{\mathrm{KL}}, \Gamma_L)$- generative particle algorithm



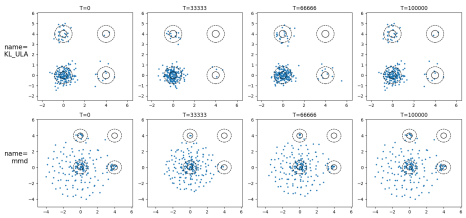
(c) Trajectory of $(f_{\mathrm{KL}}, \Gamma_L)$-GAN



(d) Divergences of $(f_{\mathrm{KL}}, \Gamma_L)$-GAN

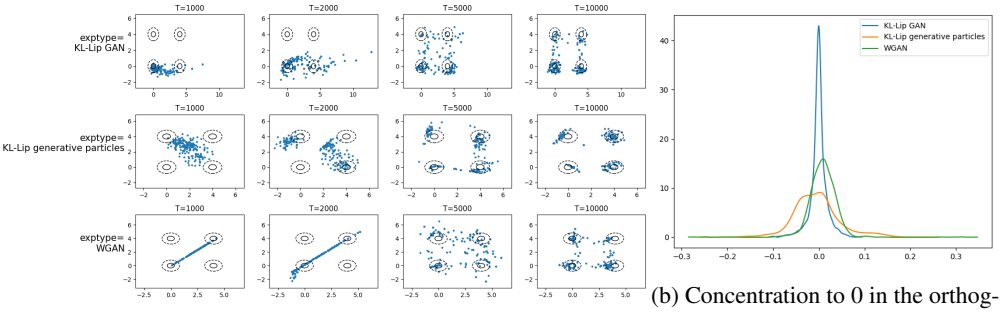


(e) Trajectory of KALE flow Glaser et al. (2021) (Different regularizations)



(f) Trajectory of KL ULA Durmus & Éric Moulines (2017), MMD flow Arbel et al. (2019)

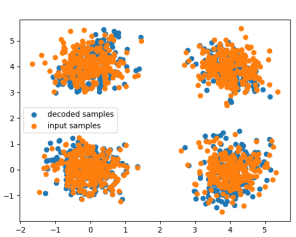
Figure 8: **(2D Mixture of Gaussians) Comparison between different methods. (a)** $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm with different values for L. The particles are transported to the 4 wells faster as L gets larger, however for large L the algorithm become unstable $(L \geq 100)$. Learning rates are chosen as $\gamma = 1.0, \delta = 0.005$. **(c)** Comparison with $(f_{\mathrm{KL}}, \Gamma_L)$-GAN which has the same discriminator as $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm but a different generator mechanism. $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm converges 3-5 times faster compared to $(f_{\mathrm{KL}}, \Gamma_L)$-GAN. The time step for second column of (c) corresponds to the last column of (a). **(b)+(d)** The values of the loss function for $(f_{\mathrm{KL}}, \Gamma_L)$-GAN are less oscillatory when compared to $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm, however their corresponding divergence is at least one order of magnitude larger than the $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm. **(e)** KALE flow can be compared with $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm in the sense of being a different regularization technique. The KALE gradient flow regularizes the RKHS norm of ϕ^* while $(f_{\mathrm{KL}}, \Gamma_L)$-generative particles algorithm regularizes the norm of $\nabla\phi^*$. The KALE flow Glaser et al. (2021) fails to capture the 4 wells in a reasonable amount of time. Here a Gaussian kernel with $\sigma = 0.5$ is chosen for the RKHS kernel. Learning rate is chosen as 0.001. **(f)** Bottom: MMD Arbel et al. (2019) gradient flow (without extra noise). A Gaussian kernel with $\sigma = 0.5$ is used for the RKHS. Top: For comparison KL gradient flow trained with the unadjusted Langevin algorithm (ULA) Durmus & Éric Moulines (2017). Learning rate is chosen as 0.001.



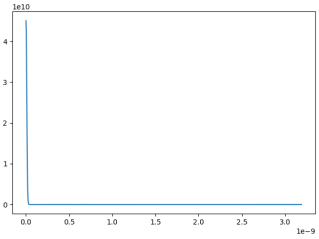
(a) Trajectories of different methods in the projected subspace

(b) Concentration to 0 in the orthogonal subspace

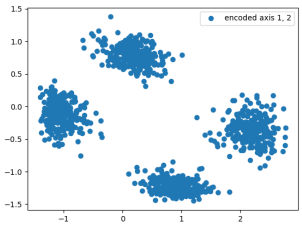
Figure 9: **(2D Mixture of Gaussians embedded in 12D, 5000 target samples) Comparison between different methods. (a)** The time trajectories in the 2D projected subspace of $(f_{\mathrm{KL}}, \Gamma_1)$-GAN (top), $(f_{\mathrm{KL}}, \Gamma_1)$-generative particles algorithm (center), Wasserstein-GAN (bottom). **(b)** Particle distributions, projected onto components orthogonal to the support plane. Values concentrated around 0 indicate convergence to the sub-manifold supporting the distribution. 5000 target samples are used with mini-batch scheme with $N_{\mathrm{mb}} = 200$. $L = 1$ is fixed. Learning rates are chosen as $\gamma = 0.25, \delta = 0.005$ for the $(f_{\mathrm{KL}}, \Gamma_1)$-generative particles algorithm.



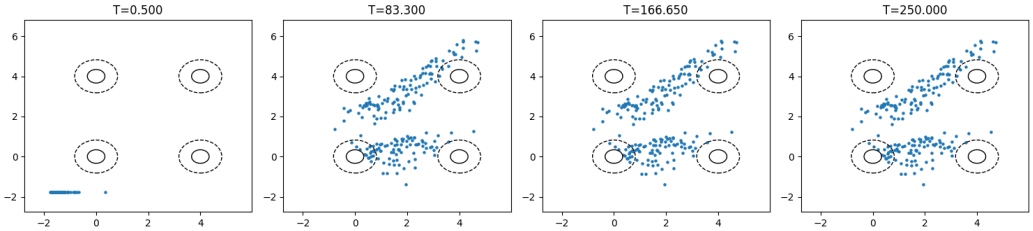
(a) Output of autoencoder in the projected subspace



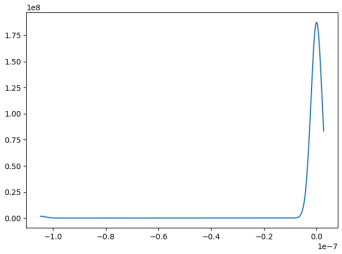
(b) Output of autoencoder in the orthogonal subspace



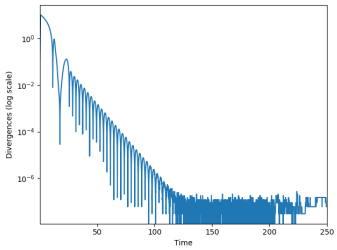
(c) Coarse grained variable



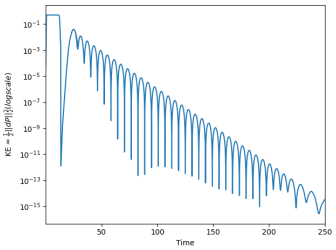
(d) GPA trajectories



(e) GPA trajectory in the orthogonal subspace

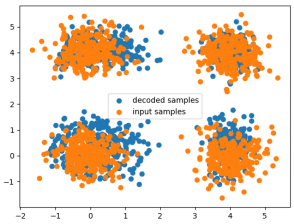


(f) Divergence

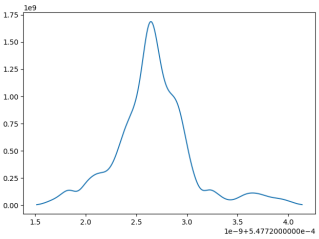


(g) Kinetic energy of latent particles

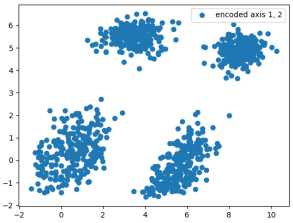
Figure 10: **(2D Mixture of Gaussians embedded in 12D, latent generative particles algorithm with latent dimension 2) Auto-encoder without Q-perfect encoding property.** The auto-encoder is trained using MSE loss and is over-fitted to the training samples. **(a)-(b)** The reconstructed outputs for the training samples of the auto-encoder. **(c)** The latent particles for the training samples of the auto-encoder. It captures the four wells. **(d)** Trajectories of reconstructed outputs from the latent generative particles algorithm in the 2D projected plane. The generated samples are unknown to the decoder. The reconstructed outputs fail to capture the four wells. **(e)** Final trajectory of the reconstructed outputs from the latent generative algorithm in the 10D orthogonal subspace. **(f)-(g)** Performance measures of particles transportation in the latent space. The source particles converges to the targets in the latent space. The learning rates are chosen as $\gamma = 1.0, \delta = 0.005$.



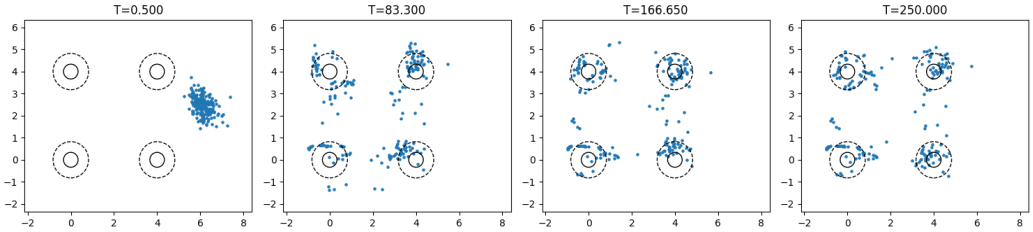
(a) Output of autoencoder in the projected subspace



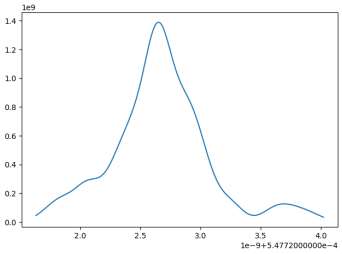
(b) Output of autoencoder in the orthogonal subspace



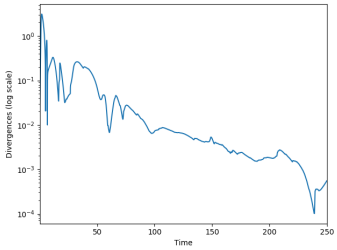
(c) Coarse grained variable



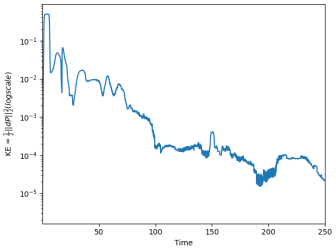
(d) GPA trajectories



(e) GPA trajectory in the orthogonal subspace

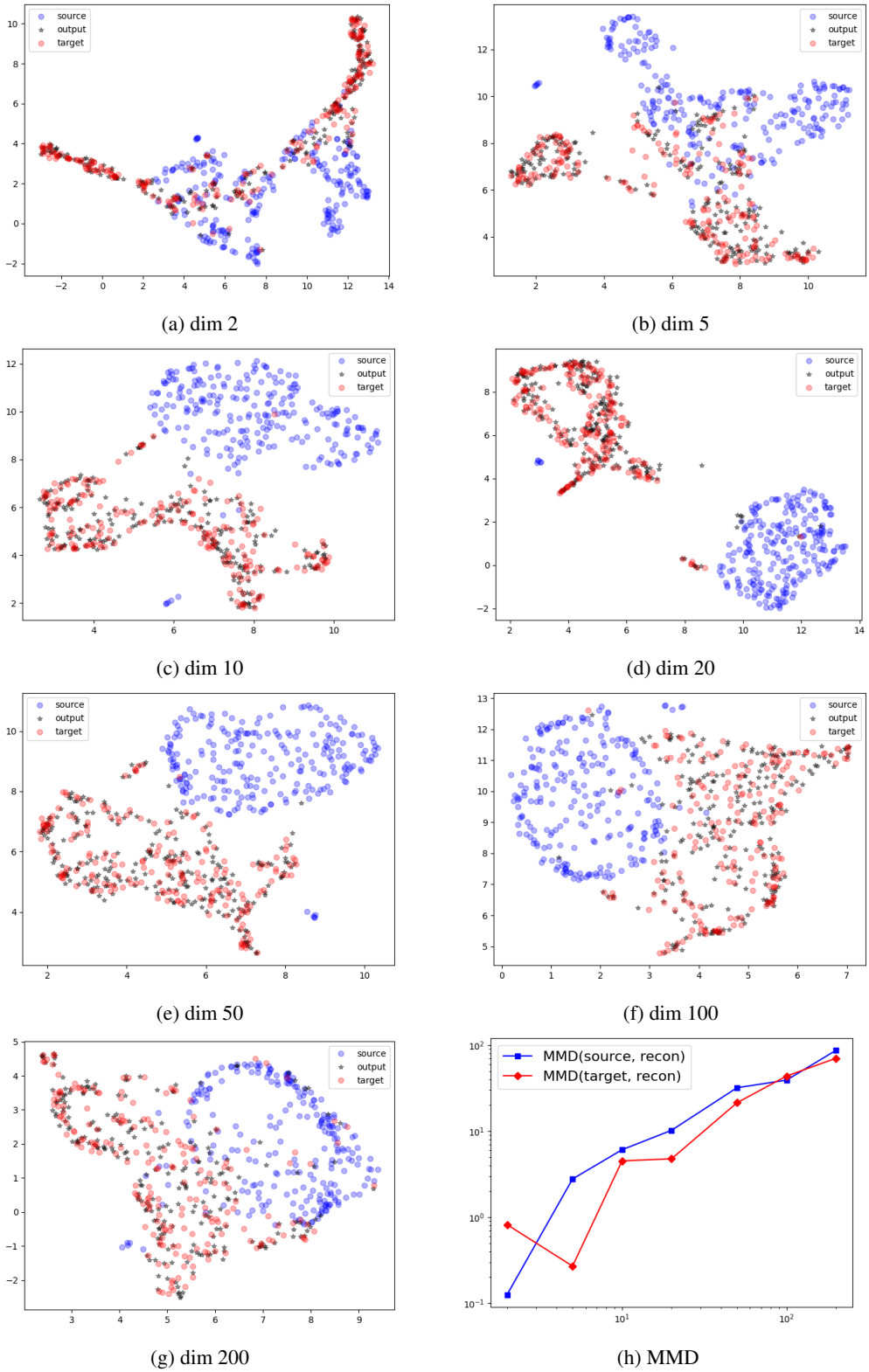


(f) Divergence



(g) Kinetic energy of latent particles

Figure 11: **(2D Mixture of Gaussians embedded in 12D, latent generative particles algorithm with latent dimension 2). Auto-encoder with Q-perfect encoding property.** The auto-encoder is trained using 2*MSE+KL loss and reconstructs the target Q in an almost perfect manner. **(a)-(b)** The reconstructed outputs for the training samples of the auto-encoder. **(c)** The latent particles for the training samples of the auto-encoder. It captures the four wells. **(d)** Trajectories of reconstructed outputs from the latent generative particles algorithm in the 2D projected plane. The generated samples are unknown to the decoder. The reconstructed outputs captures the four wells. **(e)** Final trajectory of the reconstructed outputs from the latent generative algorithm in the 10D orthogonal subspace which are highly concentrated to 0. The problematic behavior from the original generative particles algorithm (See Figure 9) is settled. **(f)-(g)** Performance measures of particles transportation in the latent space. The source particles converges to the targets in the latent space. The learning rates are chosen as $\gamma = 1.0, \delta = 0.005$.



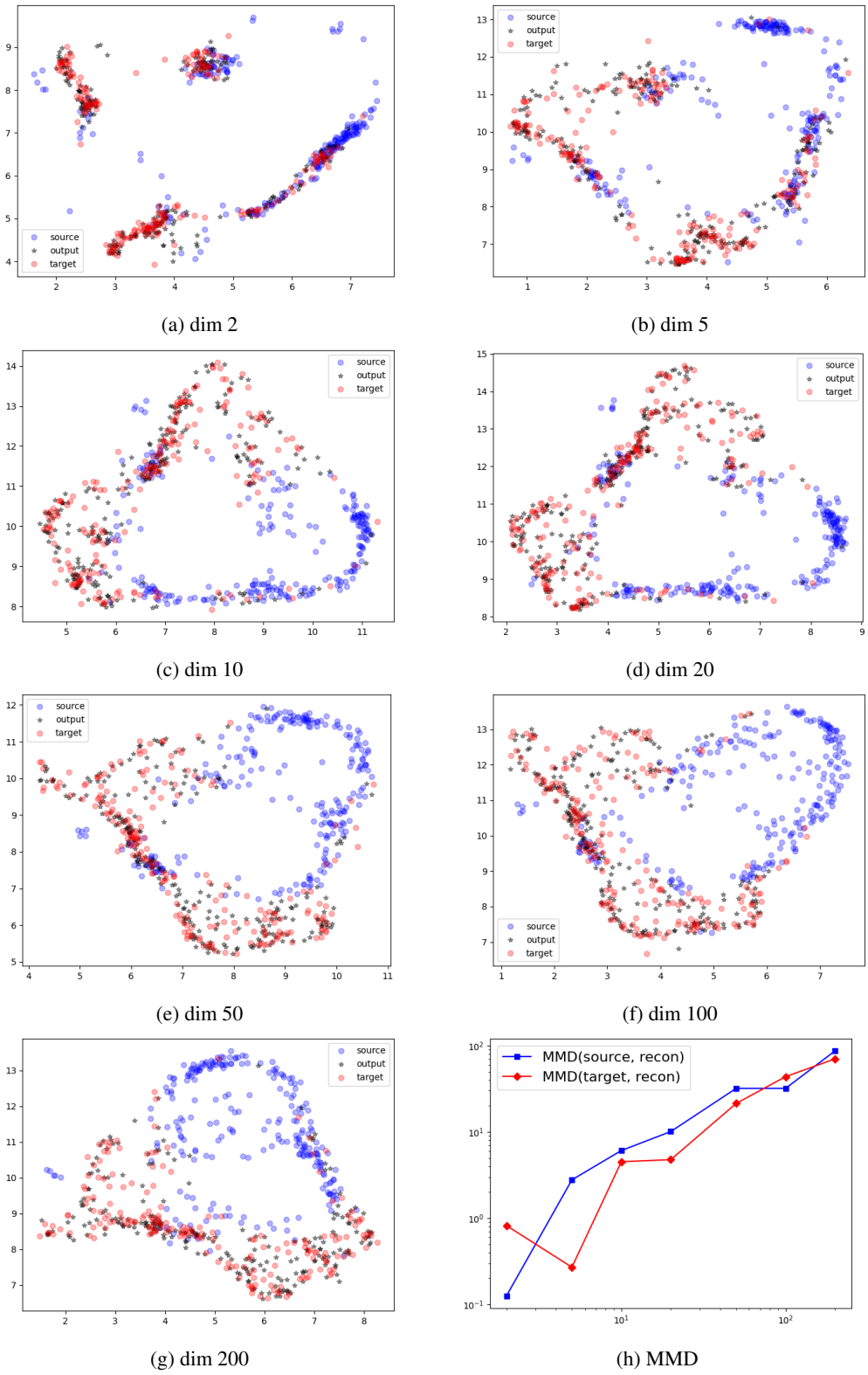
(a) dim 2 (b) dim 5

(c) dim 10 (d) dim 20

(e) dim 50 (f) dim 100

(g) dim 200 (h) MMD

Figure 12: **(Gene expression data, BreastCancer) Latent samples. blue: source, red: target, black: transported. (h)** The distance between the latent distributions. **blue: MMD**$(P_0^{\mathcal{Z}}, P_T^{\mathcal{Z}})$**, red: MMD**$(Q^{\mathcal{Z}}, P_T^{\mathcal{Z}})$**, black: MMD**$(P_0^{\mathcal{Z}}, Q^{\mathcal{Z}})$ with $T = 25,000$.



(a) dim 2 (b) dim 5

(c) dim 10 (d) dim 20

(e) dim 50 (f) dim 100

(g) dim 200 (h) MMD

Figure 13: **(Gene expression data, BreastCancer) Reconstructed samples. blue: source, red: target, black: transported. (h)** The distance between the reconstructed distributions. **blue: MMD**$(P_0^{\mathcal{Y}}, P_T^{\mathcal{Y}})$**, red: MMD**$(Q^{\mathcal{Y}}, P_T^{\mathcal{Y}})$**, black: MMD**$(P_0^{\mathcal{Y}}, Q^{\mathcal{Y}})$ with $T = 25,000$.