Lipschitz-Regularized Gradient Flows and Latent Generative Particles

Hyemin Gu[1], Panagiota Birmpa[2], Yannis Pantazis[3], Markos A. Katsoulakis[1] and Luc Rey-Bellet[1]

[1]University of Massachusetts Amherst, [2]Heriot-Watt University, [3]Foundation for Research & Technology - Hellas

SIAM 2024 Conference on Mathematics of Data Science

Lipschitz-regularized Gradient Flows

- **Lipschitz-regularized f-Divergence [1]**
  A flexible family of divergences in purpose of **comparing two mutually singular probability measures** $P, Q \in \mathcal{P}(\mathbb{R}^d)$ is defined as an infimal convolution of f-divergences (e.g KL, α, Shannon-Jensen) and 1-Wasserstein distance (Γ-Integral Probability Metric (IPM) where Γ is the 1-Lipschitz functions; denoted as Γ_1)

$$D_f^{\Gamma_L}(P\|Q) = \inf_{\gamma\in\mathcal{P}(\mathbb{R}^d)} D_f(\gamma\|Q) + L\cdot W^{\Gamma_1}(P,\gamma). \quad (1)$$

  The dual variational representation of (1) is

$$D_f^{\Gamma_L}(P\|Q) = \sup_{\phi\in\Gamma_L}\left\{\mathbb{E}_P[\phi] - \inf_{\nu\in\mathbb{R}}\{\nu + \mathbb{E}_Q[f^\star(\phi-\nu)]\}\right\} \quad (2)$$

  where $f^\star$ is the Legendre transform of f.

- **Lipschitz-regularized Gradient Flows [2]**
  Wasserstein gradient flows whose gradient dynamics are given by Lipschitz-regularized f-Divergences

$$\partial_t P_t = \mathrm{div}\left(P_t \nabla \frac{\delta D_f^{\Gamma_L}(P_t\|Q)}{\delta P_t}\right), P_0 = P. \quad (3)$$

  The first variation exists for any P, Q with $P \in \mathcal{P}_1(\mathbb{R}^d)$

$$\frac{\delta D_f^{\Gamma_L}(P\|Q)}{\delta P} = \phi^{L,*} = \underset{\phi\in\Gamma_L}{\mathrm{argmax}}\left\{E_P[\phi] - \inf_{\nu\in\mathbb{R}}(\nu + E_Q[f^\star(\phi-\nu)])\right\}. \quad (4)$$

  The Lagrangian formulation of the PDE (3) yields an ODE

$$\frac{d}{dt}Y_t = v_t^L(Y_t) = -\nabla\phi_t^{L,*}(Y_t), \quad Y_0 \sim P. \quad (5)$$

Generative Particles Algorithm (GPA)

- $(X^{(i)})_{i=1}^N$ from the "target" Q and $(Y_0^{(i)})_{i=1}^M$ from the "source" P are given.
- Learn **discriminator ϕ** which is parameterized by a **neural network** using the variational representation (2) and samples $(X^{(i)})_{i=1}^N$, $(Y_n^{(i)})_{i=1}^M$

$$\phi_n^{L,*} = \underset{\phi\in\Gamma_L^{NN}}{\mathrm{argmax}}\left\{\frac{\sum_{i=1}^M \phi(Y_n^{(i)})}{M} - \inf_{\nu\in\mathbb{R}}\left\{\nu + \frac{\sum_{i=1}^N f^\star(\phi(X^{(i)})-\nu)}{N}\right\}\right\} \quad (6)$$

- Explicitly impose the Lipschitz continuity of ϕ by **spectral normalization [3]**
- Obtain $\nabla\phi(Y_t)$ by automatic differentiation and solve the ODE (5) with an **explicit scheme**

$$Y_{n+1}^{(i)} = Y_n^{(i)} - \Delta t \nabla\phi_n^{L,*}(Y_n^{(i)}), \quad Y_0^{(i)} \sim P \quad i = 1, ..., M \quad (7)$$

- Iterate for n_T steps ($T = n_T\Delta t$); kinetic energy $\frac{1}{M}\sum_{i=1}^M |\nabla\phi_n^{L,*}(Y_n^{(i)})|^2 \to 0$.

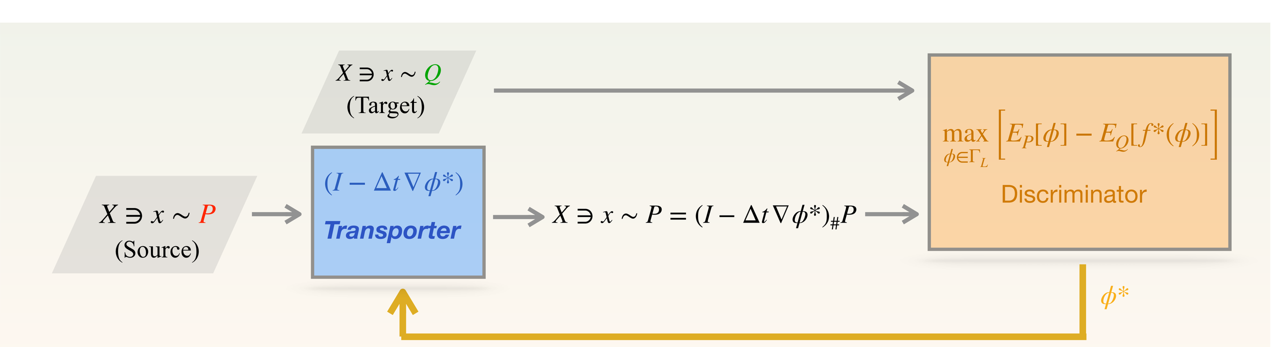


Figure: An iteration of GPA to transport the probability measure P

Generating Samples from Scarce Data Using GPA for Data Augmentation

Two approaches to ensure **generalization ability of GPA**

1. **Imbalanced sample sizes $M \gg N$**
2. **From training particles to generated particles**

(a) Fixed target samples with sample size $N = 200$ (b) $M = 600$ transported particles from $(f_{\mathrm{KL}}, \Gamma_5)$-GPA (c) 600 simultaneously transported particles from $(f_{\mathrm{KL}}, \Gamma_5)$-GPA

Figure: GPA for image generation given scarce target data (MNIST). **(b)** $M = 600$ initial particles from $Unif([0,1]^{784})$ were transported toward the target in the setting of $M \gg N$, which promotes sample diversity. **(c)** A new set of 600 initial particles from $Unif([0,1]^{784})$ were transported through the previously learned vector fields.

- **GPA for data augmentation**

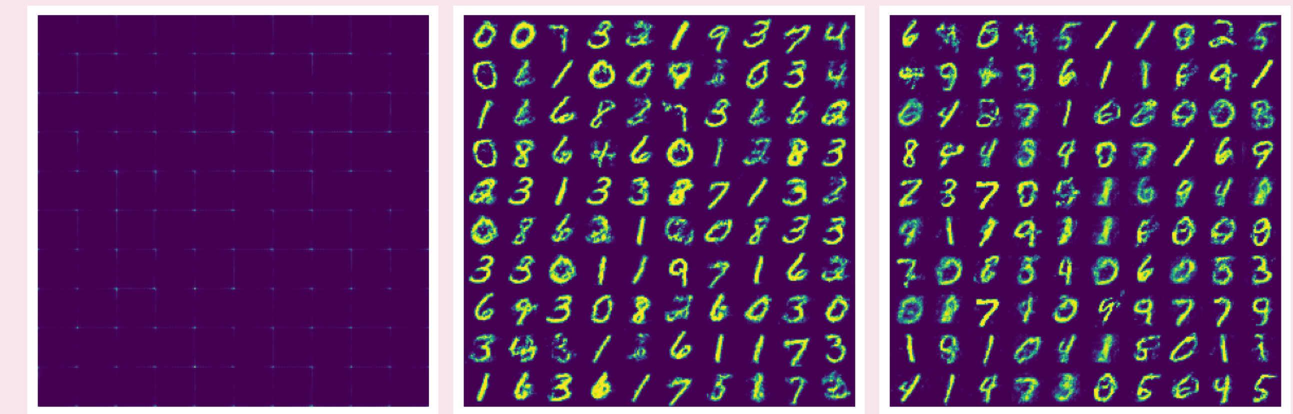


Figure: Evaluating GPA-Based Data Augmentation for Training WGAN on MNIST. WGAN trained with 200 original data (left), WGAN trained with 1400 original data (center), WGAN trained with 200 original data and 1200 GPA-augmented data (right). WGAN was not able to learn from 200 original samples from the MNIST data base WGAN trained with 1400 original data can now generate samples but in a moderate quality. We use the generated samples as in (b) and (c) in the previous figure for augmenting data to train a WGAN with a mixture of 1400 real, transported and generated samples in total. Such a GAN generated samples of similar quality compared to the GAN trained with 1400 original samples.

Numerical stability and L

The Lipschitz bound L on the discriminator space implies a pointwise bound $|\nabla\phi_n^{L,*}(Y_n^{(i)})| \leq L$. Hence the Lipschitz regularization imposes a speed limit L on the particles, ensuring the stability of the algorithm for suitable choices of L. Indeed, from a numerical analysis point of view, (7) is a particle-based explicit scheme for the PDE (3). In this context, the **Courant, Friedrichs, and Lewy (CFL) condition** for stability of discrete schemes for transport PDEs becomes

$$\sup_x |\nabla\phi_t^{L,*}(x)| \frac{\Delta t}{\Delta x} \leq 1. \quad (8)$$

We emphasize the importance of Lipschitz regularization in stabilizing dynamics when generating heavy-tailed distributions.

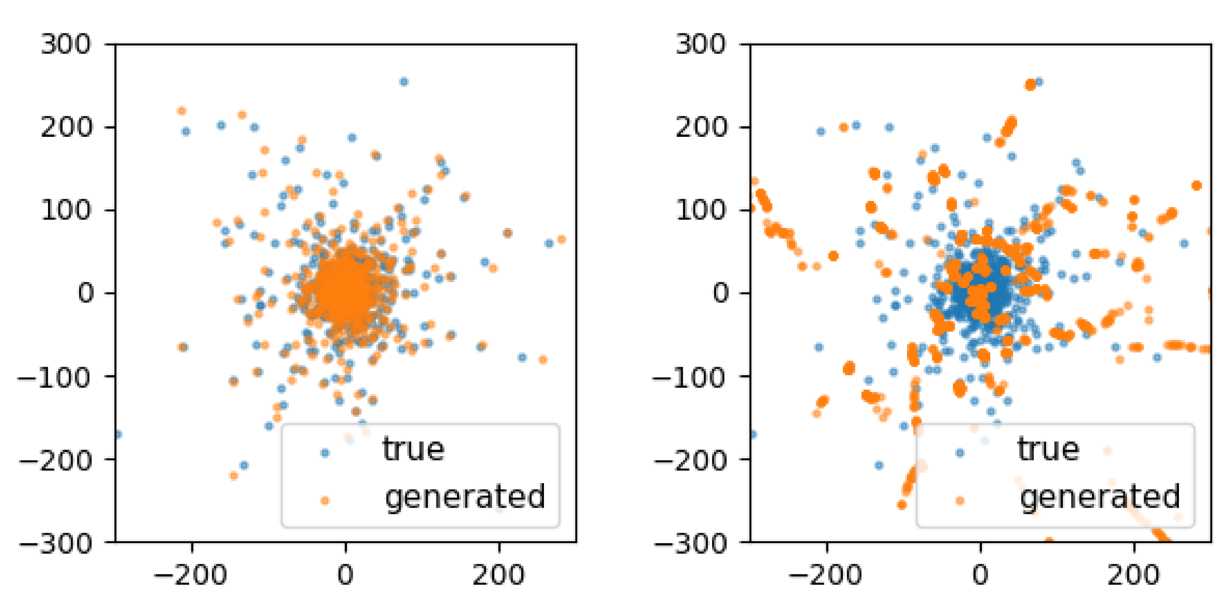
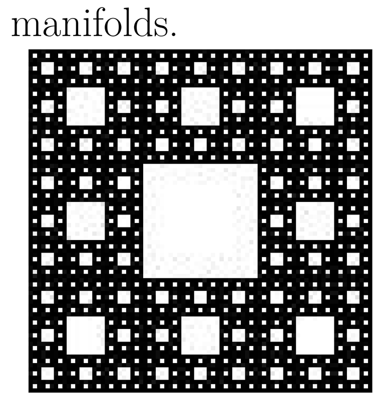


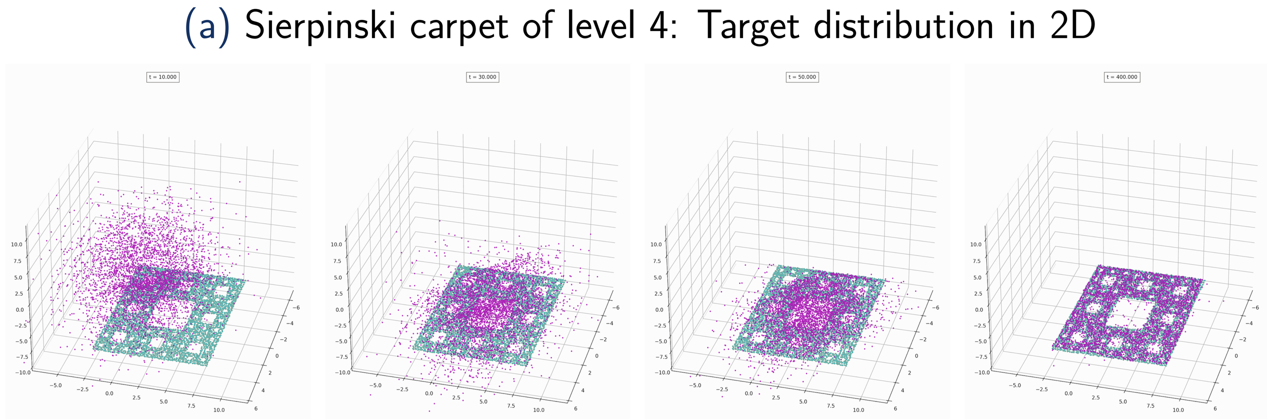
Figure: Learning a heavy-tailed distribution $f(x) \propto |x|^{-3}$ using GPA where Lipschitz-regularized with $L = 1$ (left), Unregularized (right).

Learning low-dimensional data manifold

While $D_f(P\|Q) < \infty$ and the existence of the first variation for f-divergences only if $P \ll Q$, **$D_f^{\Gamma_L}(P\|Q)$ does not require absolute continuity** and applies to any P with a finite first moment, regardless of the choice of the target Q [4]. Therefore, $D_f^{\Gamma_L}$ can be a suitable divergence for learning distributions with low-dimensional data manifolds.



(a) Sierpinski carpet of level 4: Target distribution in 2D



(b) GPA trajectories at 4 different time points

Figure: Learning 2D data manifold embedded in 3D using $(f_{\mathrm{KL}}, \Gamma_1)$-GPA. **(b)** 4,096 random samples are drawn from the 3D isotropic Gaussian source P and then transported (magenta). 4,096 target samples (cyan).

Latent Space Generative Particles

We leverage latent space formulations from recent generative flow papers [5] to achieve scalability in dimensions beyond the hundreds .

- **Idea:** A **pre-trained autoencoder** first projects the high-dimensional space to a lower dimensional latent space and then a generative model is trained in the latent space. Subsequently, the decoder maps the data generated in the latent space back to the original high-dimensional space.
- **Autoencoder performance guarantees**
  Given an autoencoder $\mathcal{E}: \mathbb{R}^d \to \mathbb{R}^{d'}$ with $a_{\mathcal{D}}$-Lipschitz continuous $\mathcal{D}: \mathbb{R}^{d'} \to \mathbb{R}^d$, which satisfies perfect reconstruction $\mathcal{D}_\#\mathcal{E}_\# Q^{\mathcal{Y}} = Q^{\mathcal{Y}}$,

$$D_f^{\Gamma_L}(\mathcal{D}_\# P^{\mathcal{Z}}\|\mathcal{D}_\#\mathcal{E}_\# Q^{\mathcal{Y}}) \leq D_f^{a_{\mathcal{D}}\Gamma_L}(P^{\mathcal{Z}}\|\mathcal{E}_\# Q^{\mathcal{Y}}). \quad (9)$$

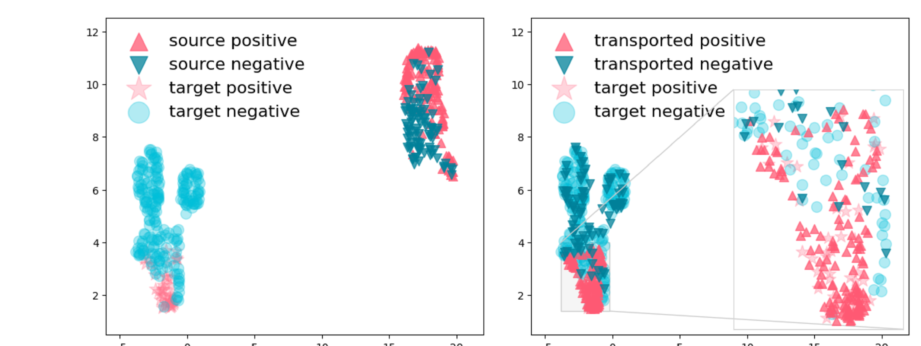


Figure: Gene expression dataset in $\mathbb{R}^{54,675}$ integration by GPA transportation. Two gene expression datasets without any transformation (left). Dataset integration using $(f_{\mathrm{KL}}, \Gamma_1)$-GPA in a latent space $\mathbb{R}^{50}$ obtained by PCA (right).

References

[1] Jeremiah Birrell, Paul Dupuis, Markos Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet.
(f,gamma)-divergences: Interpolating between f-divergences and integral probability metrics.
Journal of Machine Learning Research, 23:1–70, 01 2022.

[2] Hyemin Gu, Panagiota Birmpa, Yannis Pantazis, Luc Rey-Bellet, and Markos A. Katsoulakis.
Lipschitz-Regularized Gradient Flows and Generative Particle Algorithms for High-Dimensional Scarce Data.
SIAM J.Data Science, to appear, 2024.

[3] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida.
Spectral normalization for generative adversarial networks.
In *International Conference on Learning Representations*, 2018.

[4] Ziyu Chen, Hyemin Gu, Markos A. Katsoulakis, Luc Rey-Bellet, and Wei Zhu.
Learning heavy-tailed distributions with wasserstein-proximal-regularized α-divergences, 2024.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer.
High-resolution image synthesis with latent diffusion models.
CoRR, abs/2112.10752, 2021.

Acknowledgements

The research was partially supported by AFOSR FA9550-21-1-0354, NSF DMS-2008970 and TRIPODS CISE-1934846.