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

(f, Lip_L)-divergence

- The f-divergence between two probability measures P and Q induced by a convex function f satisfying $f(1) = 0$ is defined by

$$D_f(P\|Q) := E_Q[f(dP/dQ)]$$

  - KLD: $f = x \log x$, χ^2-divergence: $f = (x-1)^2$, Hellinger distance: $f = (\sqrt{x}-1)^2$, total variation: $f = \frac{1}{2}|x-1|$, ...

- A variational representation of the f-divergence via the Legendre transform

$$D_f(P\|Q) = \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_P[\phi] - E_Q[f^*(\phi)]\}$$

- Integral probability metrics (IPMs) maximize the differences of the respective expected values over a function space, Γ

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

  - 1-Wasserstein metric: $\Gamma = Lip_1(S)$, MMD distance: $\Gamma = RKHS$ unit ball, ...

$$D_f^{Lip_L}(P\|Q) := \sup_{\phi\in Lip_L, \nu\in\mathbb{R}} \{E_P[\phi-\nu] - E_Q[f^*(\phi-\nu)]\}$$

Lipschitz regularized gradient flows

- $\mathcal{F} : \mathcal{P} \to \mathbb{R}$ minimizing flow

$$\partial_t P_t - \nabla \cdot \left(P_t \nabla \frac{\delta \mathcal{F}(P_t)}{\delta P_t}\right) = 0, P_0 = P, P_\infty = Q$$

and corresponding particle dynamics $\frac{dY_t}{dt} = -\nabla \frac{\delta \mathcal{F}(P_t)}{\delta P_t}(Y_t)$.

  - $\mathcal{F} = D_{f_{KL}}(\cdot\|Q)$ with $f_{KL} = x\log x$: Fokker-Planck equation
  - $\mathcal{F} = D_{f_\alpha}(\cdot\|Q)$ with $f_\alpha = \frac{x^\alpha - 1}{\alpha(\alpha-1)}$: Weighted porous media equation

- Lipschitz regularized gradient flows $\mathcal{F} = D_f^{Lip_L}(\cdot\|Q)$ in variational representation: $\frac{\delta\mathcal{F}(P_t)}{\delta P_t} = \phi_t^*$

$$\partial_t P_t - \nabla\cdot(P_t\nabla\phi_t^*) = 0, P_0 = P, P_\infty = Q$$

with particle ODE system $\frac{dY_t}{dt} = -\nabla\phi_t^*(Y_t)$.

- Mobility $\mu_t(Y_t) = \frac{\partial\mathcal{D}}{\partial\mathcal{Z}}(Z_t)^T \frac{\partial\mathcal{D}}{\partial\mathcal{Z}}(Z_t)$ given by AE $\mathcal{Y} = \mathcal{D}\circ\mathcal{E}(\mathcal{Y}) = \mathcal{D}(\mathcal{Z})$ modifies the dynamics to $\frac{dY_t}{dt} = -\mu_t(Y_t)\nabla\phi_t^*(Y_t)$.

(f, Lip_L)-Generative particles algorithm

- Approximate distributions with N particles
  - $P \approx \hat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_{Y_0^{(i)}}$ and $Q \approx \hat{Q}_N = \frac{1}{N}\sum_{i=1}^N \delta_{X^{(i)}}$
- Parametrize the discriminator ϕ using a **Neural Network** and maximize the variational representation

$$N^{-1}\sum_{i=1}^N \phi(Y_n^{(i)}; W) - \left\{N^{-1}\sum_{i=1}^N f^*\left(\phi(X_n^{(i)}; W) - \nu\right) + \nu\right\}$$

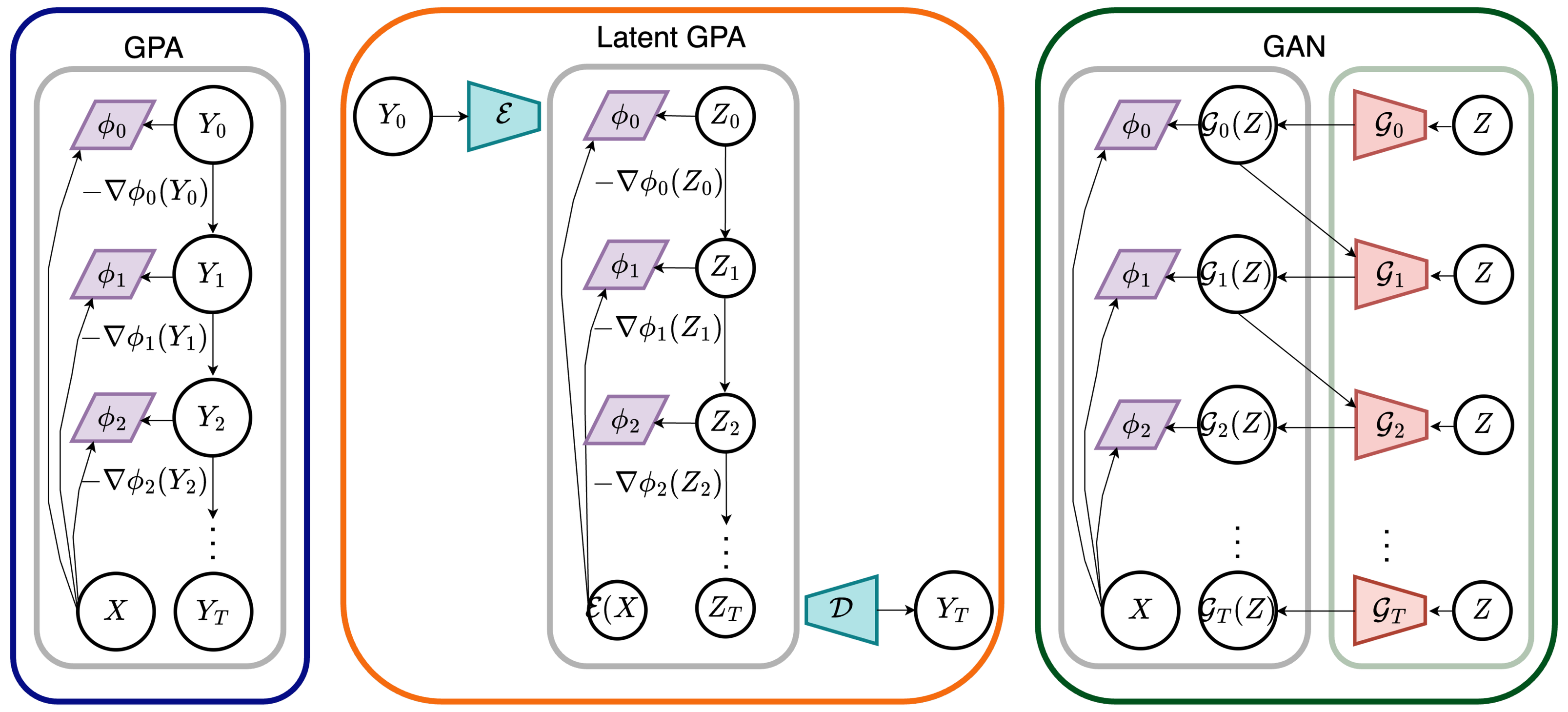
- Lipschitz continuity on ϕ: Spectral Normalization (Miyato et. al) (hard constraint)
- $\nabla\phi(Y_t)$: automatic differentiation of NN
- Solve the ODE with an explicit scheme

$$Y_{t+1} = Y_t - \Delta t \nabla\phi(Y_t)$$

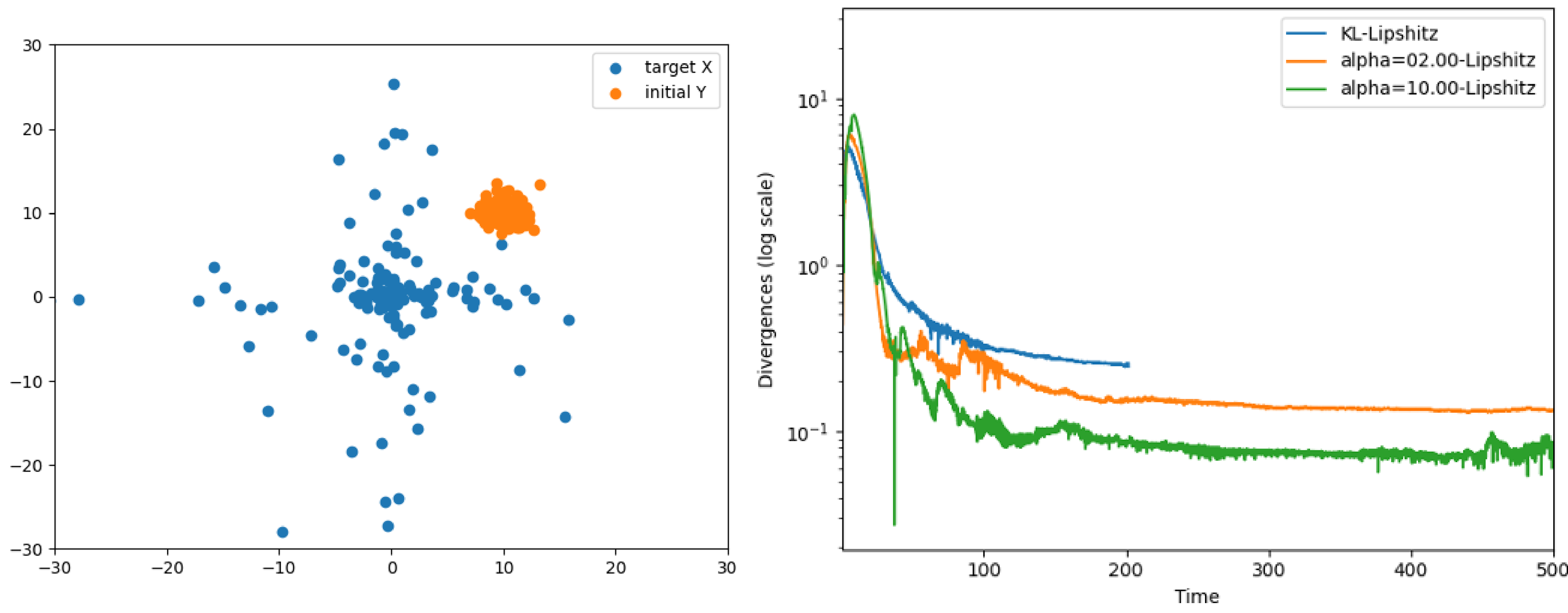
Latent generative particles

- **Pretrain an Auto-Encoder(AE)** $\mathcal{Y} = \mathcal{D}\circ\mathcal{E}(\mathcal{Y}) = \mathcal{D}(\mathcal{Z})$ on target samples $X^{\mathcal{Y}}$ where the latent space $\mathcal{Z}$ has a reduced dimension.
- Run GPA in the latent space
  - $Y_0^{\mathcal{Z}}$: Initial particles sampled in the latent space
  - $X^{\mathcal{Z}} = \mathcal{E}(X^{\mathcal{Y}})$: Target samples applied to the encoder $\mathcal{E}$
- After the GPA, reconstruct the particle $Y_T^{\mathcal{Z}}$ using the decoder $\mathcal{D}$.
- The convergence in the original space is guaranteed by the convergence in the latent space (Data-Processing Inequality).

$$D_f^{\Gamma_L}(\mathcal{D}_\# P^{\mathcal{Z}}\|Q^{\mathcal{Y}}) \le D_f^{\Gamma_{a_{\mathcal{D}}L}}(P^{\mathcal{Z}}\|\mathcal{E}_\# Q^{\mathcal{Y}})$$

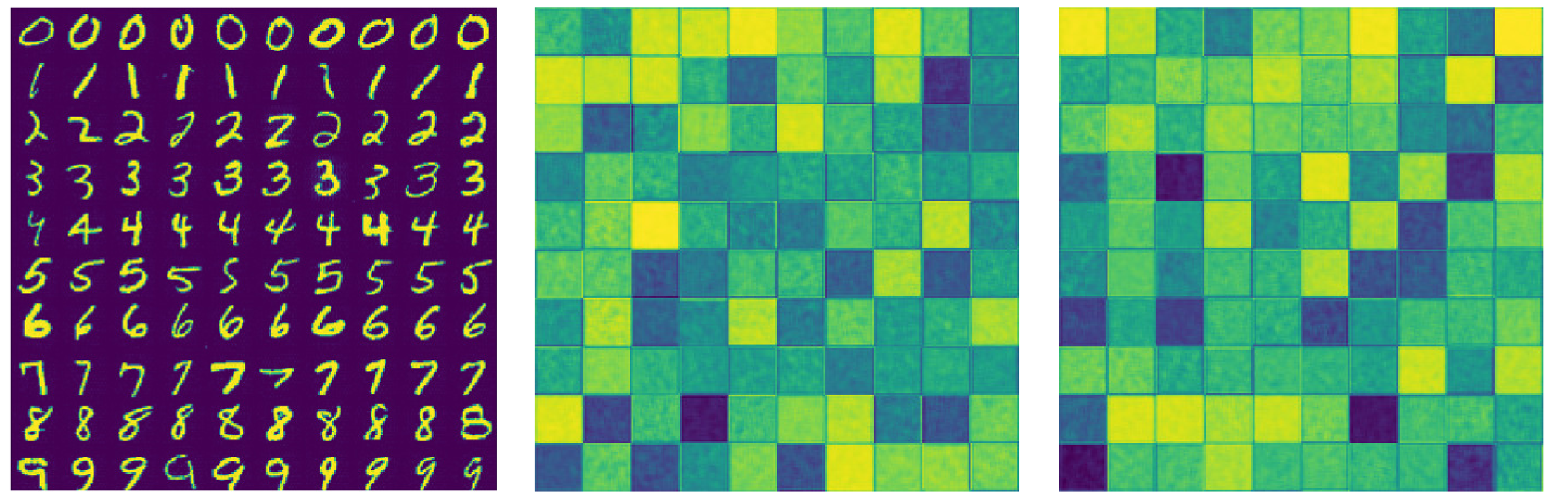


Learning heavy-tailed data with (f_α, Lip_1)-GPA



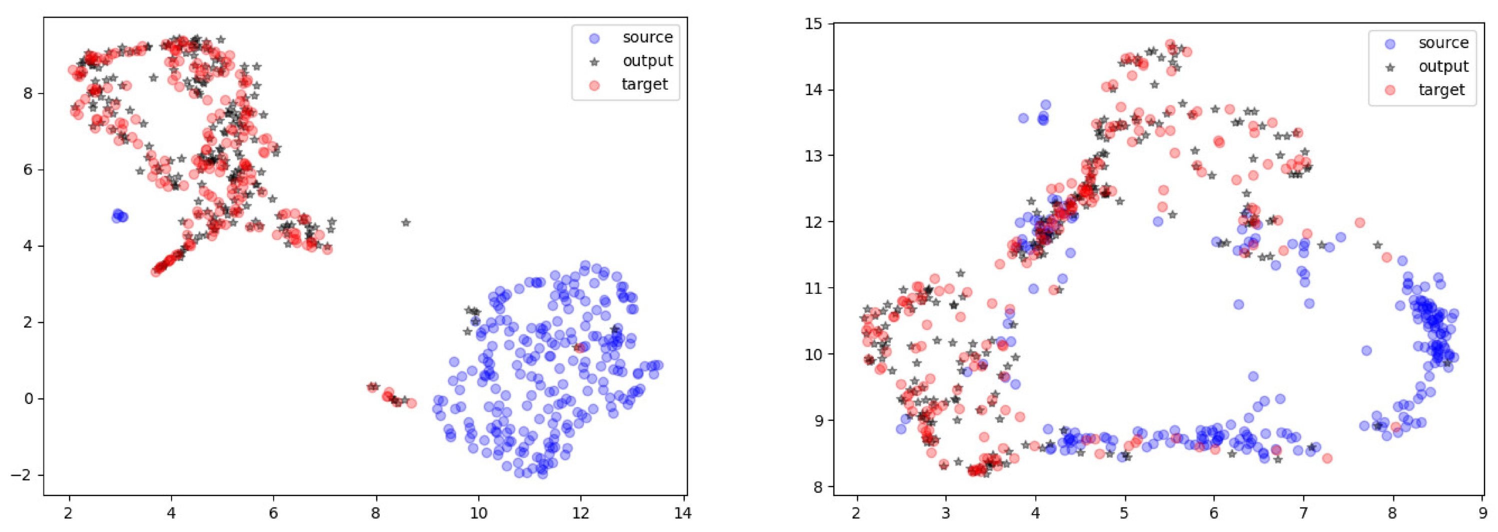
Learning $Student-t(0.5)$ from $N((10,10), I)$ by Lipschitz regularized GPA with different f. Blue: f_{KL}, Orange: f_α with $\alpha = 2$, Green: f_α with $\alpha = 10$.

Learning MNIST with a small number of data



Digit-conditioned generation of (f_{KL}, Lip_1)-**GPA (left)**, (f_{KL}, Lip_1)-**GAN (center)** and **WGAN-GP (right)** from 200 MNIST samples.

Merging Gene expression data in a latent space



Latent dimension $= 20$ **(Left)**, **Original dimension** $= 54,675$ **(Right)**

Acknowledgments

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