Large-Scale Wasserstein Gradient Flows

NeurIPS 2021

Petr Mokrov[1, 2] Alexander Korotin[1] Lingxiao Li[3]
Aude Genevay[3] Justin Solomon[3] Evgeny Burnaev[1,4]

[1]Skolkovo Institute of Science and Technology
[2]Moscow Institute of Physics and Technology
[3]Massachusetts Institute of Technology
[4]Artificial Intelligence Research Institute

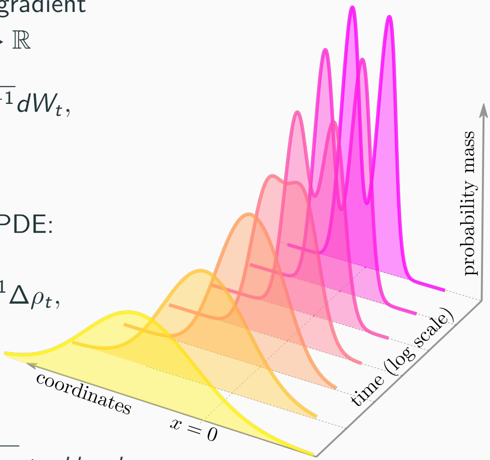
Problem Formulation: Fokker-Planck SDE

A **Langevin process**[1]
with the drift term given by the gradient
of a potential function $\Phi : \mathbb{R}^D \to \mathbb{R}$

$$dX_t = -\nabla\Phi(X_t)dt + \sqrt{2\beta^{-1}}dW_t,$$
$$\text{s.t. } X_0 \sim \rho^0$$

Corresponding **Fokker-Planck**[1] PDE:

$$\frac{\partial \rho_t}{\partial t} = \text{div}(\nabla\Phi(x)\rho_t) + \beta^{-1}\Delta\rho_t,$$
$$\text{s.t. } \rho_0 = \rho^0.$$



[1] Cédric Villani (2008). *Optimal transport: old and new.*

Wasserstein Gradient flows

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}$. The **Wasserstein gradient flow** $\{\rho_t\}_{t\in\mathbb{R}_+}$ is a continuous sequence of probability measures $\rho_t \in \mathcal{P}_2(\mathbb{R}^D)$ which satisfies the continuity equation

$$\begin{cases} \partial_t \rho_t - \boldsymbol{\nabla} \cdot \left(\rho_t \boldsymbol{\nabla}_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\right) = 0 \\ \rho_{t=0} = \rho^0 \end{cases}$$

- $\frac{\delta \mathcal{F}}{\delta \rho}$ is called the first variation.[a]

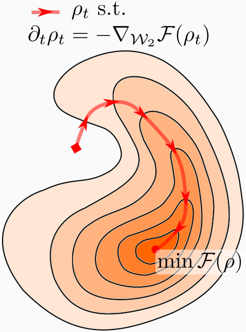
[a] Filippo Santambrogio (2016). *Euclidean, Metric, and Wasserstein Gradient Flows: an overview.* arXiv: 1609.03890 [math.AP].

Wasserstein Gradient flows

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}$. The **Wasserstein gradient flow** $\{\rho_t\}_{t\in\mathbb{R}_+}$ is a continuous sequence of probability measures $\rho_t \in \mathcal{P}_2(\mathbb{R}^D)$ which satisfies the continuity equation

$$\begin{cases} \partial_t \rho_t - \nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\right) = 0 \\ \rho_{t=0} = \rho^0 \end{cases}$$

- $\frac{\delta \mathcal{F}}{\delta \rho}$ is called the first variation.[a]
- $-\nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\right) = \nabla_{\mathcal{W}_2} \mathcal{F}(\rho)$

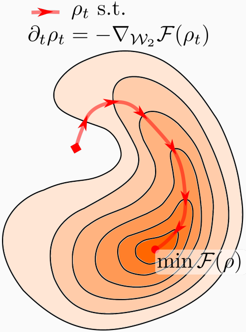


[a] Filippo Santambrogio (2016). *Euclidean, Metric, and Wasserstein Gradient Flows: an overview*. arXiv: 1609.03890 [math.AP].

Wasserstein Gradient flows

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^D) \to \mathbb{R}$. The **Wasserstein gradient flow** $\{\rho_t\}_{t\in\mathbb{R}_+}$ is a continuous sequence of probability measures $\rho_t \in \mathcal{P}_2(\mathbb{R}^D)$ which satisfies the continuity equation

$$\begin{cases} \partial_t \rho_t - \nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\right) = 0 \\ \rho_{t=0} = \rho^0 \end{cases}$$



- $\frac{\delta \mathcal{F}}{\delta \rho}$ is called the first variation.[a]
- $-\nabla \cdot \left(\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\right) = \nabla_{\mathcal{W}_2} \mathcal{F}(\rho)$
- The Fokker-Planck equation is the **WGF** with the functional

$$\mathcal{F}_{\mathrm{FP}}(\rho) = \underbrace{\int_{\mathbb{R}^D} \Phi(x) d\rho(x)}_{\text{potential energy}} + \beta^{-1} \underbrace{\int_{\mathbb{R}^D} \rho(x) \log \rho(x) dx}_{\text{neg. entropy}}$$

[a] Filippo Santambrogio (2016). *Euclidean, Metric, and Wasserstein Gradient Flows: an overview.* arXiv: 1609.03890 [math.AP].

JKO scheme

JKO scheme is the sequence $\{\rho_\tau^k\}_{k=1}^\infty \subset \mathcal{P}_2(\mathbb{R}^D)$ such that:

$$\rho_\tau^k \leftarrow \underset{\rho \in \mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \frac{1}{2}\mathcal{W}_2^2(\rho_\tau^{k-1}, \rho) + \tau \mathcal{F}(\rho)\,,$$

$$\rho_\tau^0 = \rho^0 \in \mathcal{P}_2(\mathbb{R}^D)$$

The parameter $\tau \in \mathbb{R}_+$ is the discretization step.

JKO scheme

JKO scheme is the sequence $\{\rho_\tau^k\}_{k=1}^\infty \subset \mathcal{P}_2(\mathbb{R}^D)$ such that:

$$\rho_\tau^k \leftarrow \arg\min_{\rho \in \mathcal{P}_2(\mathbb{R}^D)} \frac{1}{2}\mathcal{W}_2^2(\rho_\tau^{k-1}, \rho) + \tau\mathcal{F}(\rho)\,, \quad \rho_\tau^0 = \rho^0 \in \mathcal{P}_2(\mathbb{R}^D)$$

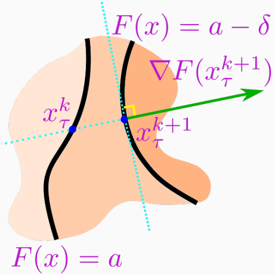
Similarity to Euclidean case

$$\underbrace{\begin{cases} \partial_t \rho_t + \nabla_{\mathcal{W}_2}\mathcal{F}(\rho) = 0 \\ \rho_{t=0} = \rho^0 \end{cases}}_{\text{Gradient flow in } \left(\mathcal{P}_2(\mathbb{R}^D), \mathcal{W}_2\right)} \sim \underbrace{\begin{cases} x'(t) = -\nabla F(x(t)) \\ x(0) = x_0 \in \mathbb{R}^n \end{cases}}_{\text{Gradient flow in Euclidean space } (\mathbb{R}^n, \|\cdot\|_2)}$$

The **Backward Euler Scheme** $\{x_\tau^k\}_{k=1}^\infty$ which models the Gradient flow in Euclidean space:

$$x_\tau^{k+1} = x_\tau^k - \tau\nabla F(x_\tau^{k+1}) \Leftrightarrow$$

$$\Leftrightarrow \underbrace{x_\tau^{k+1} = \arg\min_x \frac{1}{2}\|x - x_\tau^k\|^2 + \tau F(x)}_{\text{compare with JKO!}}$$



JKO scheme

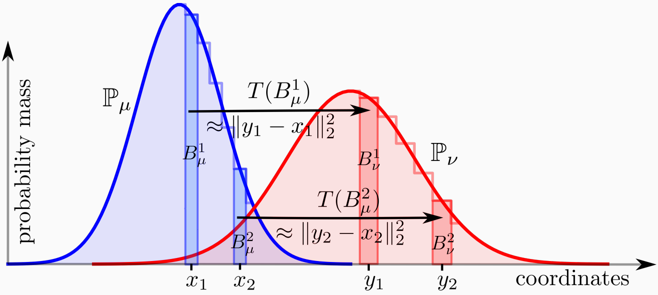
JKO scheme is the sequence $\{\rho_\tau^k\}_{k=1}^\infty \subset \mathcal{P}_2(\mathbb{R}^D)$ such that:

$$\rho_\tau^k \leftarrow \underset{\rho \in \mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \frac{1}{2}\mathcal{W}_2^2(\rho_\tau^{k-1}, \rho) + \tau\mathcal{F}(\rho),$$

$$\rho_\tau^0 = \rho^0 \in \mathcal{P}_2(\mathbb{R}^D)$$

(Squared) **Wasserstein-2 distance** between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^D)$

$$\mathcal{W}_2^2(\mu, \nu) = \inf_{\nu = T\sharp\mu} \int_{\mathbb{R}^D} \|x - T(x)\|_2^2 d\mu(x)$$



JKO scheme

JKO scheme is the sequence $\{\rho_\tau^k\}_{k=1}^\infty \subset \mathcal{P}_2(\mathbb{R}^D)$ such that:

$$\rho_\tau^k \leftarrow \underset{\rho\in\mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \frac{1}{2}\mathcal{W}_2^2(\rho_\tau^{k-1}, \rho) + \tau\mathcal{F}(\rho)\,,$$
$$\rho_\tau^0 = \rho^0 \in \mathcal{P}_2(\mathbb{R}^D)$$

Theorem[2] Given $\mathcal{F} = \mathcal{F}_{\text{FP}}(\rho) = \int_{\mathbb{R}^D} \Phi(x)d\rho(x) + \beta^{-1}\int_{\mathbb{R}^D} \rho(x)\log\rho(x)dx$ and $\mathcal{F}(\rho^0) < +\infty$ there exists unique solution of **JKO** $\{\rho_\tau^k\}_{k=1}^\infty$. Define $\rho_\tau : (0, +\infty) \times \mathbb{R}^n \to [0, \infty)$ as follows:

$$\rho_\tau(t) = \rho_\tau^k, \text{ for } t \in [k\tau, (k+1)\tau), k \in \mathbb{N}$$

Then, as $\tau \downarrow 0$: $\rho_\tau(t)$ weakly converges to the solution of the Wasserstein gradient flow associated with $\mathcal{F}$

[2] Richard Jordan, David Kinderlehrer, and Felix Otto (1998). "The Variational Formulation of the Fokker-Planck Equation". In: *SIAM J. Math. Anal.*

Brenier's theorem[4]

Theorem[4] Let μ be absolutely continuous. Then there exists unique μ - a.s. convex lower semicontinuous f, that the optimal T^* has the form: $T^*(x) = \nabla f(x)$. Therefore, in this case:

$$W_2^2(\mu, \nu) = \int_{\mathbb{R}^D} \|x - \nabla f(x)\|_2^2 d\mu(x)$$

Alternative formulation of JKO[3]

$$\psi_k = \underset{\psi \in \text{Conv}(\mathbb{R}^D)}{\arg\min}\ \tau \mathcal{F}_{\text{FP}}(\boldsymbol{\nabla}\psi \sharp \rho_\tau^k) + \frac{1}{2} \int_{\mathbb{R}^D} \|x - \boldsymbol{\nabla}\psi(x)\|_2^2 \mathrm{d}\rho_\tau^k(x)$$

$$\rho_\tau^{k+1} = \boldsymbol{\nabla}\psi_k \sharp \rho_\tau^k$$

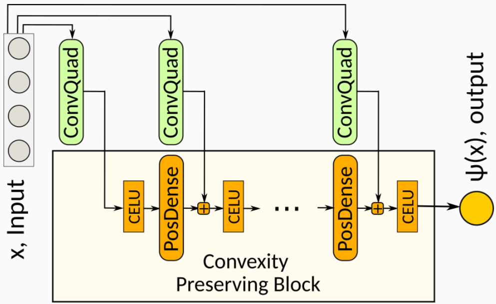
---

[3] Jean-David Benamou et al. (2014). *Discretization of functionals involving the Monge-Ampère operator*. arXiv: `1408.4536 [math.NA]`.

[4] Villani Cédric (2003). *Topics in optimal transportation / Cédric Villani*. eng. Graduate studies in mathematics. American mathematical society.

ICNN powered JKO

Consider the parametrization $\psi_\theta \in \mathrm{Conv}(\mathbb{R}^D), \theta \in \Theta$ given by ICNNs[5]



Typical ICNN architecture

Image source: Korotin et. al. (2019)

Each JKO optimization step reads as follows:

$$\theta^* \leftarrow \arg\min_\theta \left[\mathcal{F}_{\mathrm{FP}}(\nabla\psi_\theta \sharp \rho_\tau^k) + \frac{1}{2\tau} \int_{\mathbb{R}^D} \|x - \nabla\psi_\theta(x)\|_2^2 d\rho_\tau^k(x) \right]$$

[5] Brandon Amos, Lei Xu, and J Zico Kolter (2017). "Input convex neural networks". In: *Proceedings of the 34th International Conference on Machine Learning*.

Stochastic Optimization for JKO via ICNNs

ICNN powered JKO

$$\theta^* \leftarrow \arg\min_{\theta} \left[\mathcal{F}_{\text{FP}}(\nabla \psi_\theta \sharp \rho_\tau^k) + \frac{1}{2\tau} \int_{\mathbb{R}^D} \|x - \nabla \psi_\theta(x)\|_2^2 d\rho_\tau^k(x) \right]$$

$$\psi_k := \psi_{\theta^*} \, ; \, \rho_\tau^{k+1} = \boldsymbol{\nabla} \psi_k \sharp \rho_\tau^k$$

We need to optimize with respect to

$\mathcal{F}_{\text{FP}}(\nabla \psi_\theta \sharp \rho_\tau^k) = \int_{\mathbb{R}^D} \Phi(x) d\rho(x) + \beta^{-1} \int_{\mathbb{R}^D} \rho(x) \log \rho(x) dx$:

Theorem Let $\rho \in \mathcal{P}_2(\mathbb{R}^D)$ - absolute continuous, $T : \mathbb{R}^D \to \mathbb{R}^D$ is a diffeomorphism. Let $x_1, x_2, \ldots x_N \sim \rho$. Then

$$\widehat{\mathcal{F}_{\text{FP}}}(x_{1:N}) = \frac{1}{N} \sum_{k=1}^{N} \Phi\big(T(x_k)\big) - \beta^{-1} \frac{1}{N} \sum_{n=1}^{N} \log |\det \nabla T(x_n)|$$

is an estimator of $\mathcal{F}_{\text{FP}}(T \sharp \rho)$ up to constant.

Stochastic Optimization for JKO via ICNNs

Algorithm 1: Fokker-Planck JKO via ICNNs

Input : Initial measure ρ^0, batch size N, discr. step $\tau > 0$;
# of steps $K > 0$, temperature β^{-1}, target potential $V(x)$;

Output: trained ICNN models $\{\psi_k\}_{k=1}^K$ representing JKO steps

for $k = 0, 1, \ldots, K-1$ **do**

- $\psi_\theta \leftarrow$ basic ICNN model;
- **for** $i = 1, 2, \ldots$ **do**
  - Sample batch $Z \sim \rho^0$ of size N; $X \leftarrow \boldsymbol{\nabla}\psi_{k-1} \circ \cdots \circ \boldsymbol{\nabla}\psi_0(Z)$;
  - $\widehat{\mathcal{W}_2^2} \leftarrow \frac{1}{N} \sum_{x \in X} \|\boldsymbol{\nabla}\psi_\theta(x) - x\|_2^2$;
  - $\widehat{\mathcal{F}_{\text{FP}}} \leftarrow \frac{1}{N} \sum_{x \in X} V\big(\boldsymbol{\nabla}\psi_\theta(x)\big) - \beta^{-1} \frac{1}{N} \sum_{x \in X} \log\det \nabla^2 \psi_\theta(x)$
  - $\widehat{\mathcal{L}} \leftarrow \frac{1}{2\tau}\widehat{\mathcal{W}_2^2} + \widehat{\mathcal{F}_{\text{FP}}}$;
  - Perform a gradient step over θ by using $\frac{\partial \widehat{\mathcal{L}}}{\partial \theta}$;
- **end**
- $\psi_k \leftarrow \psi_\theta$

end

Density estimation via ICNN powered JKO

Let $\psi_0, \psi_1, \dots \psi_K$ be the convex potentials which minimize the corresponding JKO steps, i.e.

$$\rho_\tau^1 = \boldsymbol{\nabla}\psi_0 \sharp \rho^0\,;$$

$$\dots$$

$$\rho_\tau^K = \boldsymbol{\nabla}\psi_{K-1} \sharp \left[\boldsymbol{\nabla}\psi_{K-2} \sharp \{ \dots \boldsymbol{\nabla}\psi_0 \sharp \rho^0 \}\right]\,;$$

By change of variable formula, given $x_K \in \mathbb{R}^D$ the following holds true:

$$\rho_\tau^k(x_K) = \rho^0(x_0) \cdot \Big[\prod_{i=0}^{K-1} \det \nabla^2 \psi_i(x_i) \Big]^{-1}$$

where $x_0, x_1, \dots x_{K-1}$ are s.t. $x_K = \boldsymbol{\nabla}\psi_{K-1}(x_{K-1})$, $\dots$ $x_1 = \boldsymbol{\nabla}\psi_0(x_0)$

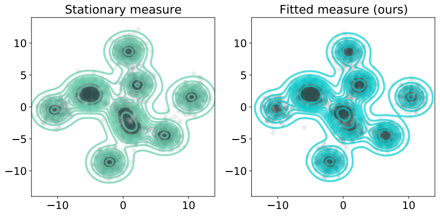
- If we sample x_K from ρ_τ^K we compute the density $\rho_\tau^K(x_K)$ on the fly!
- For arbitrary $x_K \in \mathbb{R}^D$ one need to solve the sequence of **convex** optimization problems:

$$x_i = \nabla\psi_{i-1}(x_{i-1}) \iff x_{i-1} = \arg\max_{x \in \mathbb{R}^D} \left[\langle x, x_i \rangle - \psi_{i-1}(x) \right]$$

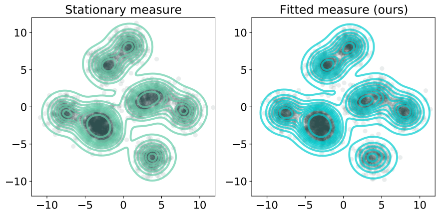
Study: Convergence to stationary distribution

The Fokker-Planck equation with potential $\mathcal{F}_{\text{FP}}(\rho) = \int_{\mathbb{R}^D} \Phi(x) d\rho(x) + \beta^{-1} \int_{\mathbb{R}^D} \rho(x) \log \rho(x) dx$ converges to stationary distribution

$$\rho^*(x) = Z^{-1} \exp(-\beta \Phi(x))$$



Projection to first two PC, $D = 13$

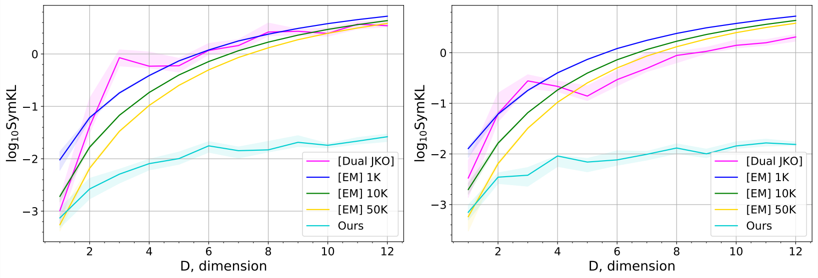


Projection to first two PC, $D = 32$

Examples of convergence to stationary mixture of gaussians distributions

Study: Ornstein-Uhlenbeck processes

- The potential $\Phi(x) = \frac{1}{2}(x - b)^T A(x - b)$, A is SPD matrix
- Given $\rho^0(X) \sim \mathcal{N}(\mu, \Sigma)$, distribution $\rho_t(x)$ has close-form solution (it is also normal distribution)



SymKL true vs fitted, $t = 0.5$ SymKL true vs fitted, $t = 0.9$

Discrepancy between true and predicted marginal distributions at different timesteps

Applications: Unnormalized Posterior Sampling

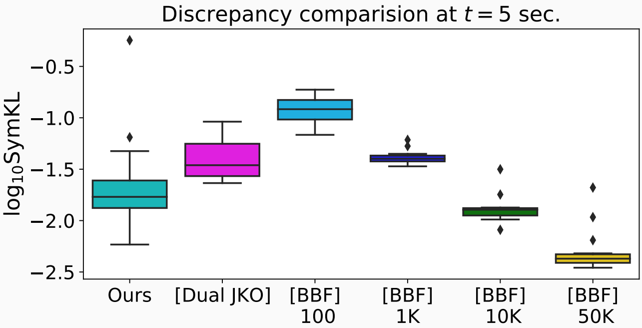
Comparison with SVGD[6] method on Bayesian Logistic Regression task for 9 benchmark datasets[6]

Dataset	Accuracy		Log-Likelihood	
	Ours	⌈SVGD⌋	Ours	⌈SVGD⌋
covtype	0.75	0.75	-0.515	-0.515
german	0.67	0.65	-0.6	-0.6
diabetis	0.775	0.78	-0.45	-0.46
twonorm	0.98	0.98	-0.059	-0.062
ringnorm	0.74	0.74	-0.5	-0.5
banana	0.55	0.54	-0.69	-0.69
splice	0.845	0.85	-0.36	-0.355
waveform	0.78	0.765	-0.485	-0.465
image	0.82	0.815	-0.43	-0.44

[6] Qiang Liu and Dilin Wang (2019). *Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm*. arXiv: `1608.04471 [stat.ML]`.

Applications: Nonlinear filtering

- In the problem of nonlinear filtering one need to compute the posterior distribution of nonlinear Fokker-Planck diffusion based on noisy observations from the process
- $\Phi(x) = \frac{1}{\pi}\sin(2\pi x) + \frac{1}{4}x^2$ (it is highly nonlinear process)
- Filtering takes $t_{\text{el}} = 9$ sec. (noisy observations each 0.5 sec.)

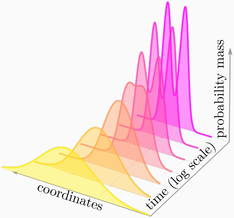


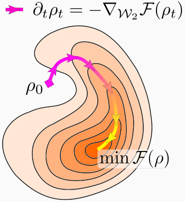
Thank you![7]

Large-Scale Wasserstein Gradient Flows

Modelling the Fokker-Planck equation via ICNN-powered JKO scheme.

https://arxiv.org/abs/2106.00736





https://github.com/PetrMokrov/Large-Scale-Wasserstein-Gradient-Flows

[7] The problem statement was developed in the framework of Skoltech-MIT NGP program. The work was supported by Ministry of Science and Higher Education grant No. 075-10-2021-068.