Intelligent Systems Department

Wasserstein gradient flows: modeling and applications

Student: *Petr Mokrov*
Research Advisor: *Evgeny Burnaev*

May, 2022

Introduction / Background

Wasserstein gradient flow is an absolute continuous flow in the space $(\mathcal{P}_2(\mathbb{R}^N), \mathcal{W}_2)$. A Wasserstein gradient flow μ_t satisfies continuity equation:

$$\partial_t \mu_t + \mathrm{div}(b\mu_t) = 0\,,\ \mu_0 = \mu$$

for some vector field $b : [0, T] \times \mathbb{R}^N \to \mathbb{R}^N$

Example. Let $b : [0, T] \times \mathbb{R}^N \to \mathbb{R}^N$ be bounded and smooth. Let $X_t(y)$ be the unique solution of the Cauchy problem:

$$\dot{x} = b(x, t)\,,\ x(0) = y$$

Then the associated Wasserstein gradient flow could be defined as follows:

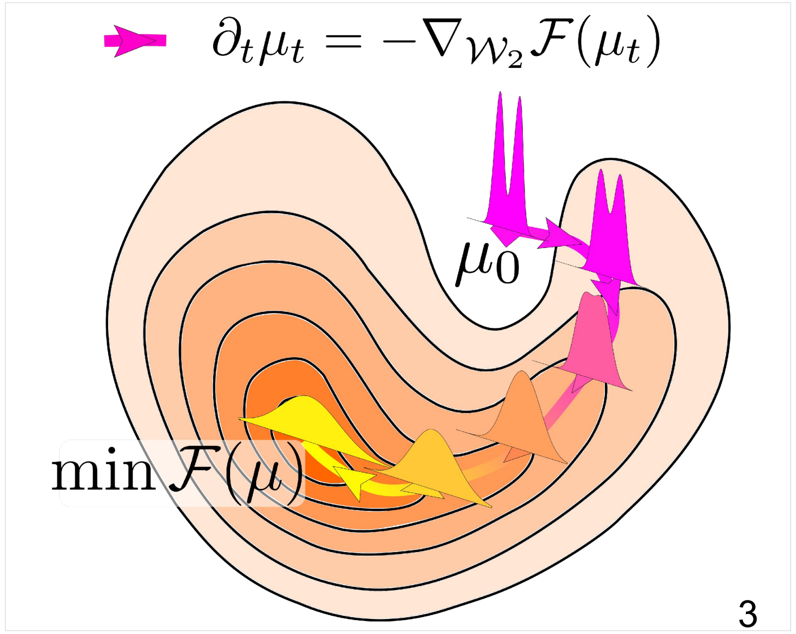
$$\mu_t = \mu \circ X_t^{-1}$$

Introduction / Background

Of the particular interest of my work are the Wasserstein gradient flows of the form:

$$\partial_t \mu_t + \operatorname{div}\left(\mu_t\left(-\boldsymbol{\nabla}_x \frac{\delta \mathcal{F}}{\delta \mu}(\mu_t)\right)\right) = 0\,,\ \mu_0 = \mu^0$$

Here $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^N) \to \mathbb{R}$ and $\frac{\delta \mathcal{F}}{\delta \mu}(\mu_t) : \mathbb{R}^N \to \mathbb{R}$ is the first variation of the functional at μ_t



Theoretical motivation.

- The term $\operatorname{div}\left(\mu_t\left(-\boldsymbol{\nabla}_x \frac{\delta \mathcal{F}}{\delta \mu}(\mu_t)\right)\right)$ can be understood as the gradient in Wasserstein space $\boldsymbol{\nabla}_{\mathcal{W}_2}\mathcal{F}(\mu_t)$ that's why μ_t forms *steepest* descent curve
- If the functional satisfies some convexity conditions, the μ_t converges to $\min\limits_{\mu \in \mathcal{P}_2(\mathbb{R}^N)} \mathcal{F}$

General problem

$$\partial_t \mu_t - \operatorname{div}\left(\mu_t \nabla_x \frac{\delta \mathcal{F}}{\delta \mu}(\mu_t)\right) = 0, \quad \mu_{t=0} = \mu^0, \; \mathcal{F} : \mathcal{P}(\mathbb{R}^N) \to \mathbb{R}$$

Wasserstein gradient flows appear in various applications:

- crowd modeling[1]
- generative modeling[2]
- reinforcement learning[3]
- population dynamics[4]
- nonlinear filtering[5]
- unnormalized posterior sampling[5]

There have been proposed several methods for modeling WGFs including time-space discretization[6], particle-based methods[7], forward Euler scheme[2]. My research is focused on modeling WGFs using JKO scheme[8] and it's application in synthetic and real-world tasks.

1: https://arxiv.org/abs/1002.0686
2: http://proceedings.mlr.press/v97/gao19b/gao19b.pdf
3: http://proceedings.mlr.press/v80/zhang18a.html
4: https://proceedings.mlr.press/v48/hashimoto16.html
5: https://arxiv.org/abs/2106.00736
6: https://doi.org/10.1016/0021-9991(70)90001-X
7: https://doi.org/10.1007/978-3-642-61544-3
8: https://doi.org/10.1137/S0036141096303359

Aim and objectives

The overall aim of the work is to

- develop scalable methods for modeling Wasserstein gradient flows based on JKO scheme and Optimal transport theory
- deploy the method in theoretical and applied tasks

The objectives of the present research run as follows:

1. To reformulate the JKO scheme in particular case of a WGF given by Fokker-Planck potential:

$$\mathcal{F}_{\text{FP}}(\rho) = \int_{\mathbb{R}^N} \Phi(x) d\rho(x) + \beta^{-1} \int_{\mathbb{R}^N} \log \rho(x) d\rho(x)$$

   based on ICNNs and Brenier's formulation of Wasserstein-2 distance
2. To formulate the JKO objective in a form which could be optimized via gradient descent and to implement the optimization algorithm
3. To analyze the performance of the proposed method on synthetic and real tasks with special attention to high-dimensional applications.

Methods.Theory

JKO scheme (Jordan, Knderlehrer, Otto, 1996)

The Wasserstein gradient flow[1] with the Fokker-Planck potential $\mathcal{F}_{\mathrm{FP}}$ could be approximated by the JKO scheme[1] which is the sequence $\{\mu_\tau^k\}_{k=0}^K$; $\mu_\tau^0 = \mu^0$ such that:

$$\mu_\tau^k \leftarrow \underset{\mu \in \mathcal{P}(\mathbb{R}^N)}{\arg\min} \frac{1}{2} \mathcal{W}_2^2(\mu_\tau^{k-1}, \mu) + \tau \mathcal{F}_{\mathrm{FP}}(\mu)$$

The Brenier's theorem[2] (Brenier, 1987) permits the following JKO reformulation:

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

$$\mu_\tau^{k+1} = \mu_\tau^k \circ \boldsymbol{\nabla}\psi_k^{-1}$$

1: https://doi.org/10.1007/s13373-017-0101-1
2: https://doi.org/10.1007/978-3-540-71050-9

Methods.Optimization

Input Convex Neural Networks[1] *(Amos, 2017):*

The convexity is ensured by special restrictions on the weights, activations and specific layers connection topology

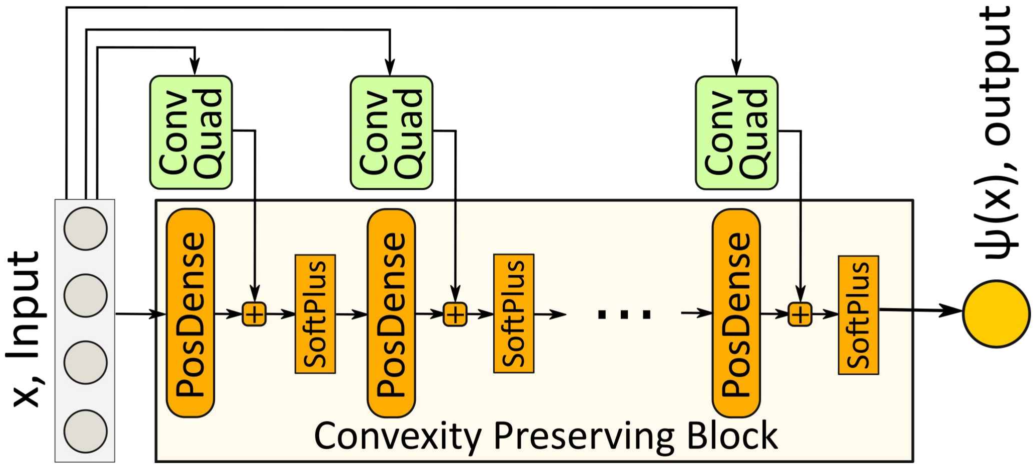


image source: Korotin et. al, 2021

Stochastic optimization for JKO via ICNNs *(Korotin, 2021):*

$$\widehat{F_{\mathrm{FP}}}(x_{1:n}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Phi(\nabla\psi_\theta(x_i)) - \beta^{-1}\log|\det \mathrm{Hess}(\psi_\theta)(x_i)|\right\} \; ; \; x_i \sim \mu_\tau^{k-1}$$

JKO sampling procedure:

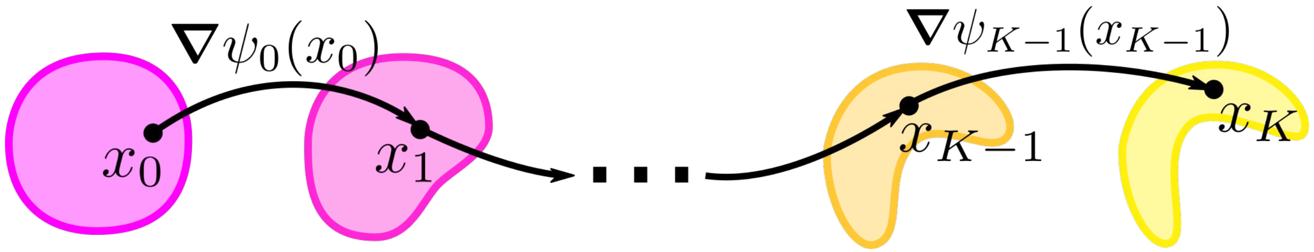
$$\text{Sample batch } Z \sim \mu^0; \quad \Rightarrow \quad X \leftarrow \nabla\psi_{K-1} \circ \cdots \circ \nabla\psi_0(Z) \sim \mu_\tau^K$$

1: https://proceedings.mlr.press/v70/amos17b/amos17b.pdf

Methods.Applications

Density estimation for JKO (Korotin, 2021)

Density estimation of the JKO scheme solution via change-of-variable formula $\mu_\tau^K(x_K) = \mu^0(x_0)\cdot\left[\prod_{i=0}^{K-1}\det\nabla^2\psi_i(x_i)\right]^{-1}$. It requires to solve the sequence of convex optimization problems $x_{i-1} = \arg\max_x \left(x^T x_i - \psi_i(x)\right)$



JKO based Metropolis-Hastings (Mokrov, 2021)

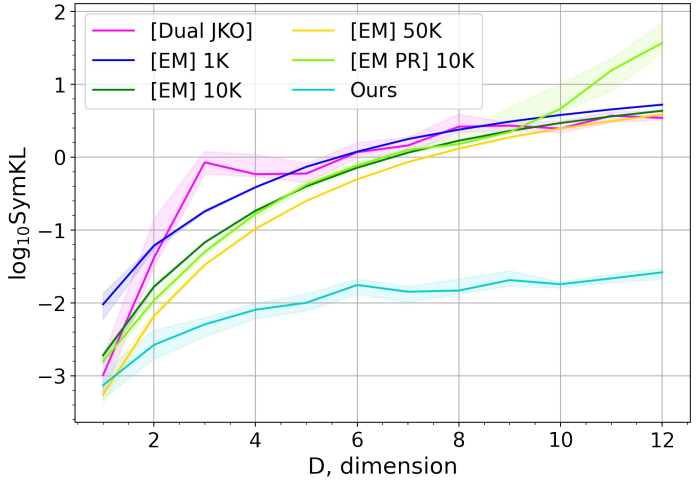
Metropolis-Hastings[1] algorithm for Nonlinear filtering application with special $\{\psi_i\}_{i=1}^K$ models-dependent proposals helping to avoid optimization problems solving.

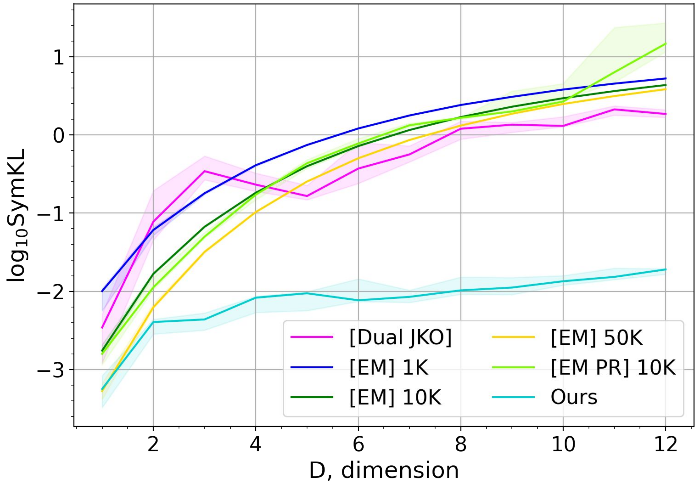
1: https://doi.org/10.1007/978-1-4757-3071-5_6

Results.Synthetic experiments

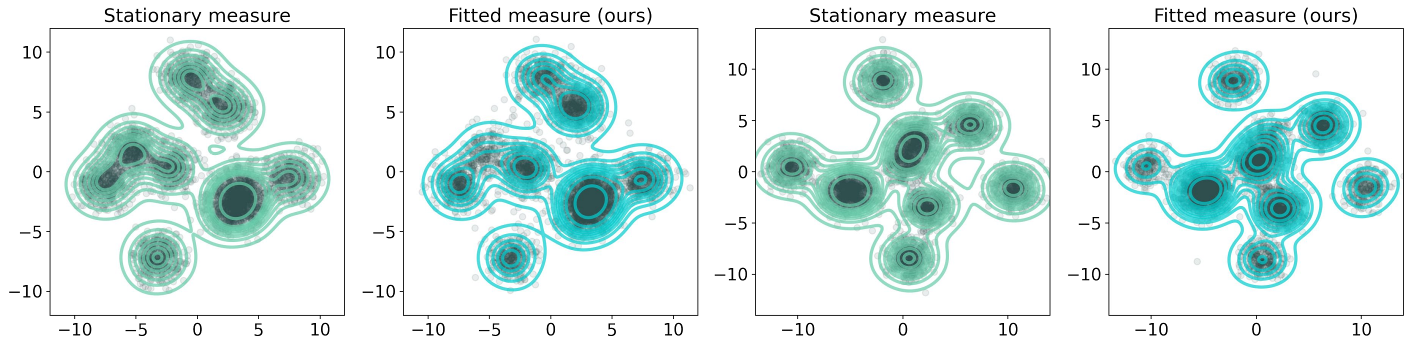
Ornstein-Uhlenbeck processes

The discrepancies between true and fitted measures when t = 0.5 (left) and t = 0.9 (right)

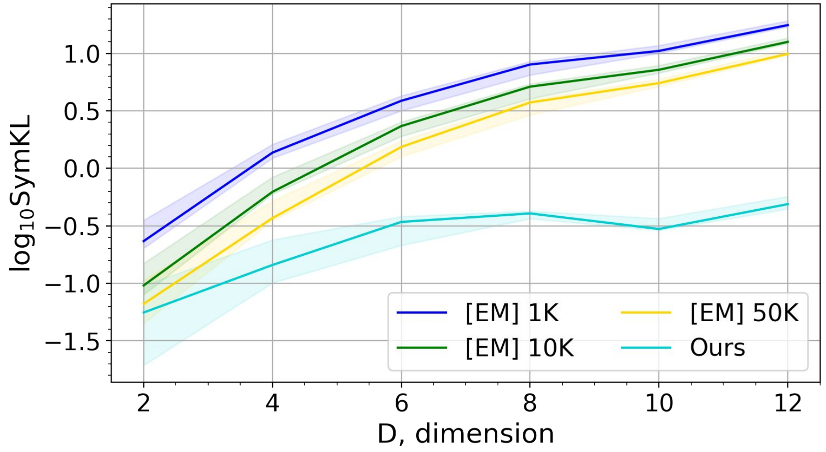




Convergence to stationary distribution



Visual discrepancy between fitted and true stationary distributions for the dimensionalities N = 32 (left) and N = 13 (right)



Convergence comparison in different dimensions

Results.Real-World applications

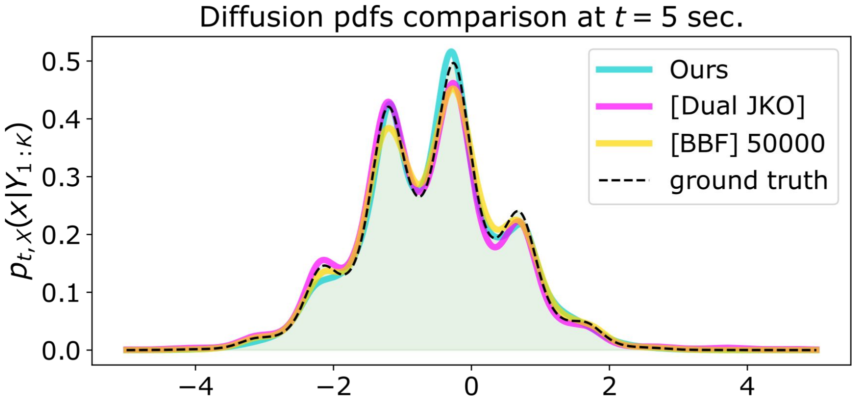
Unnormalized Posterior Sampling

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

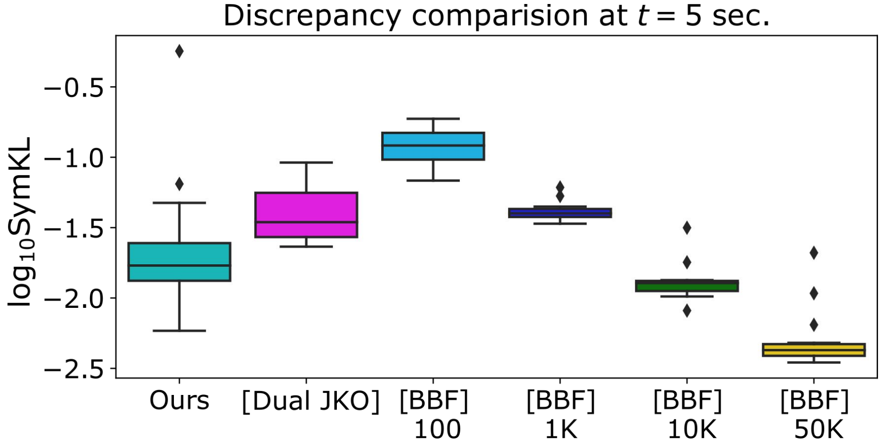
Comparison with SVGD[1] method on Bayesian logistic regression task for 9 benchmark datasets

1: https://arxiv.org/pdf/1608.04471.pdf

Nonlinear filtering



Visual (above) and boxplot (below) comparison of posterior distributions of nonlinear 1D Fokker-Planck diffusion with $\Phi(x) = \frac{1}{\pi}\sin(2\pi x) + \frac{1}{4}x^2$



Discussion of results

- Synthetic data experiments (convergence to stationary distribution, Ornstein-Uhlenbeck processes) shows that our model significantly outperforms particle-based methods (Euler-Maruyama iterations) and Dual JKO method in high-dimensional setting. Potentially, our method models the WGF with Fokker-Planck potential more accurately than competitive methods in more general settings.
- Our method shows the comparable performance to SVGD in the unnormalized posterior sampling experiment but more inference-friendly. Given trained JKO model we can generate as much samples from the posterior distribution as we want compared to SVGD approach (each new batch should be generated from scratch)
- Our method shows competitive performance in the Nonlinear filtering problem. In spite of particle-based method demonstrates superior quality in 1D problem, our approach will likely overcome the competitive methods in high dimensions as shown by the success in synthetic data experiments.

Scientific novelty

The works closest to our research are as follows:

- The work by Benamou et. al.[1] The authors derived the same formulation of the JKO with help of Brenier's convex pushforward transforms but utilized intricate discretization of the convex functions set. In opposite we exploit the ICNN parameterization.
- The work by Frogner et. al.[2] We partially replicate their synthetic experimental setup but use our original method.

For the last half a year there has been appeared several concurrent and incremental works: Alvarez-Melis et. al.[3], Bunne et. al.[4] etc which exploits the ideas similar to ours

1: https://arxiv.org/abs/1408.4536
2: http://proceedings.mlr.press/v108/frogner20a/frogner20a.pdf
3: https://arxiv.org/abs/2106.00774
4: https://arxiv.org/abs/2106.06345

Innovation

The WGF with Fokker-Planck functional could be potentially applied in:

- population dynamics[1] (in particular, scRNA-seq analysis).
- generative modelling[2] (diffusion-based)

However, these possibilities should be intensively studied in practice to validate the potential in industrial applications.

1: https://arxiv.org/pdf/2106.06345.pdf
2: https://arxiv.org/pdf/2112.02424.pdf

Conclusions

1. We reformulated the JKO scheme by substituting the optimization over set of probability measures with gradients of convex functions parameterized by ICNN.
2. We implemented the algorithm which allows to solve the ICNN-parameterized JKO via conventional gradient descent.
3. We analyzed the performance of our method when studying convergence to stationary distribution, Ornstein-Uhlenbeck processes and when applying our machinery for unnormalized posterior sampling as well as nonlinear filtering.

Outcomes

We have performed extensive theoretical and practical research of Wasserstein gradient flow with Fokker-Planck potential and published our analysis and results in the following paper:

P. Mokrov, A. Korotin, L. Li, A. Genevay, J. Solomon, E. Burnaev. Large-Scale Wasserstein Gradient Flows; in Advances in Neural Information Processing Systems, 2021[1]

Our code is available on github:
https://github.com/PetrMokrov/Large-Scale-Wasserstein-Gradient-Flows

1:https://openreview.net/forum?id=nlLjIuHsMHp

Outlook

1. The achieved results indicate that the JKO scheme is the powerful tool for modeling WGFs. Therefore, one possible future direction is to consider new applications of the method (such as generative modeling and population dynamics)

2. Besides, one can consider alternative approaches of modeling WGFs (based on CNF or Forward Euler Scheme) and apply them in appropriate applications.

Acknowledgements

I would like to express my gratitude to Alexander Korotin who helped me a lot with this project. In particular, he shared with me a lot of useful ideas and helped with writing and submitting of our article.