Large-Scale Wasserstein Gradient Flows

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

[1]Skolkovo Institute of Science and Technology (Moscow, Russia)
[2]Moscow Institute of Physics and Technology (Moscow, Russia)
[3]Massachusetts Institute of Technology (Cambridge, Massachusetts, USA)

Large-Scale Wasserstein Gradient Flows

Petr Mokrov и др. (2021). *Large-Scale Wasserstein Gradient Flows*. arXiv: 2106.00736 [cs.LG]

Accepted for NeurIPS 2021!

Concurrent works (last four months):

David Alvarez-Melis, Yair Schiff и Youssef Mroueh (2021). *Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks*. arXiv: 2106.00774 [stat.ML]

Charlotte Bunne и др. (2021). *JKOnet: Proximal Optimal Transport Modeling of Population Dynamics*. arXiv: 2106.06345 [cs.LG]

Hot topic!

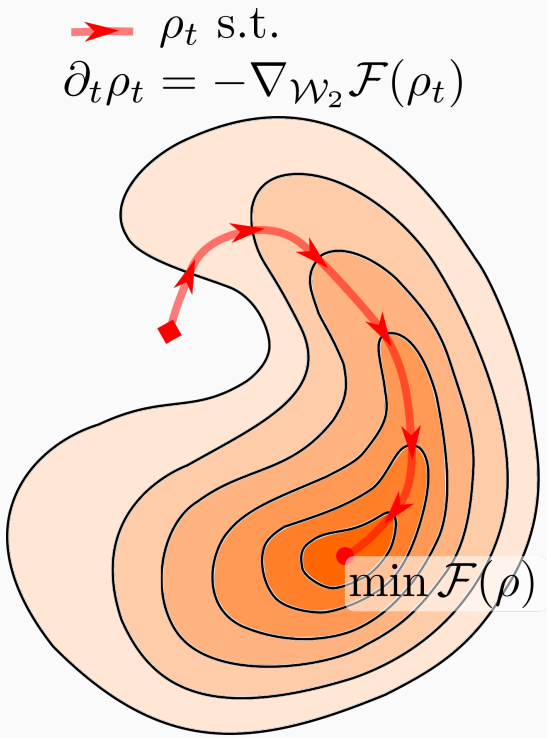


Table of Contents

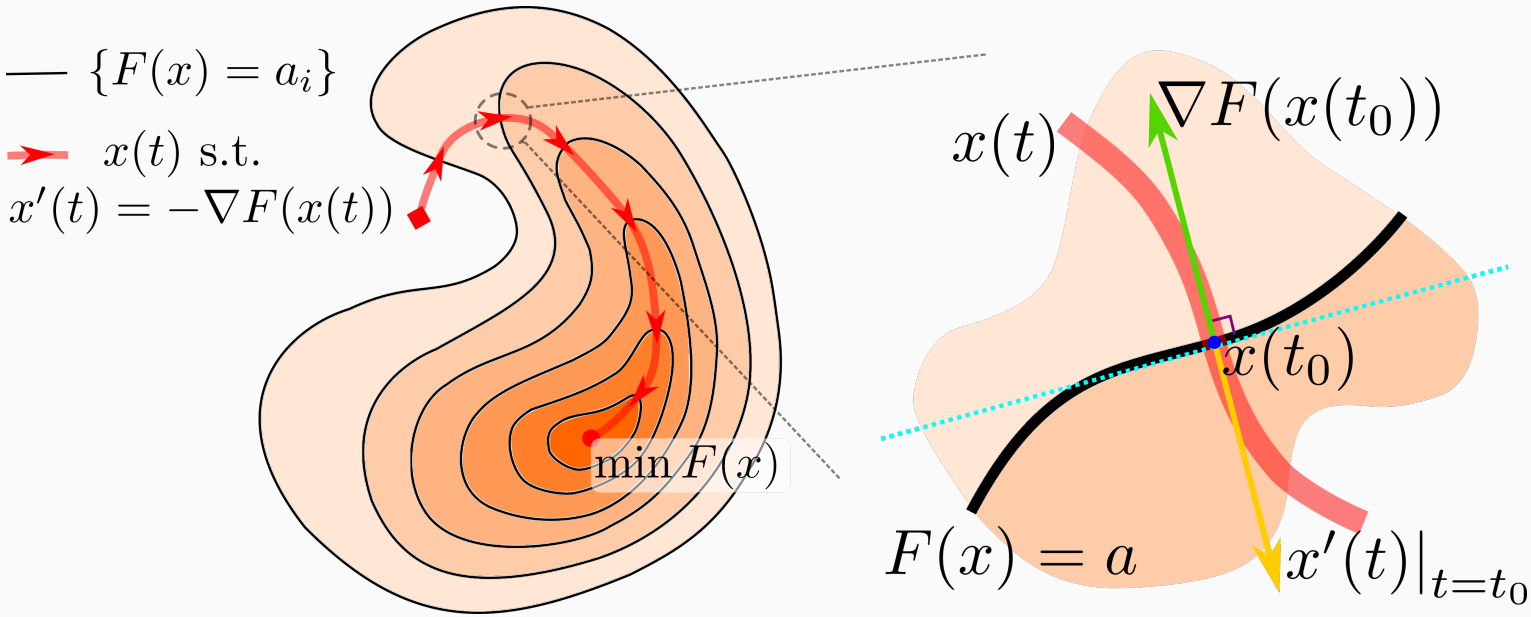
Gradient Flows

Gradient Flows in Euclidean space

Let $F : \mathbb{R}^n \to \mathbb{R}$ be smooth enough, $x_0 \in \mathbb{R}^n$.

The **gradient flow** is a curve $x(t)$ which satisfies the steepest descent scheme

$$\begin{cases} x'(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$



Gradient Flows in Euclidean space

Let $F : \mathbb{R}^n \to \mathbb{R}$ be smooth enough, $x_0 \in \mathbb{R}^n$.

The **gradient flow** is a curve $x(t)$ which satisfies the steepest descent scheme

$$\begin{cases} x'(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Several properties:

- The task is the Cauchy problem w.r.t. ∇F from the ODE theory
- There exists unique solution if ∇F is Lipschitz continuous[1]
- One can consider convex F and substitute $\nabla F(x(t))$ with $\partial F(x(t))$. The existence and uniqueness still holds :)[1]
- If F is λ-strong convex the convergence rate is exponential: $\|x(t) - x^*\| \leq \|x(0) - x^*\| e^{-\lambda t}$, x^* is unique minimizer of F.[1]

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

Gradient Flows in Euclidean space

Question: How to recover a gradient flow?

Let's build a discrete approximation sequence $x_1^\tau, x_2^\tau, \ldots$; such that $x_k^\tau \approx x(k\tau)$. Then, $\widehat{x'}(\tau k) = \frac{x_{k+1}^\tau - x_k^\tau}{\tau} \approx -\nabla F(x_k^\tau) \approx -\nabla F(x_{k+1}^\tau)$

- **explicit** (forward) Euler scheme:

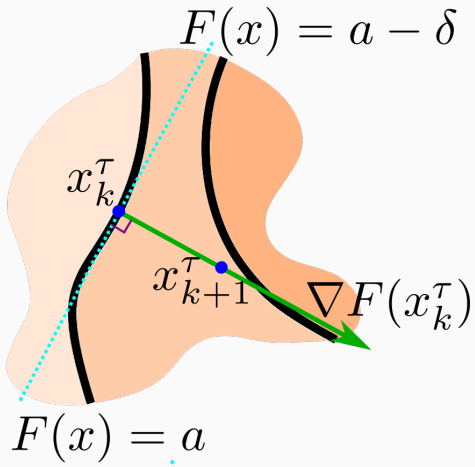
$$x_{k+1}^\tau = x_k^\tau - \tau \nabla F(x_k^\tau)$$

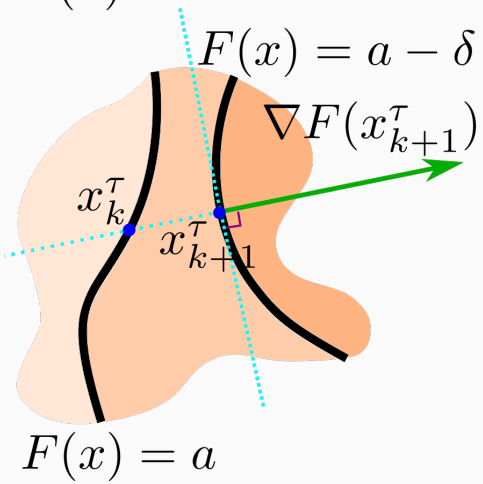
- **implicit** (backward) Euler scheme:

$$x_{k+1}^\tau = x_k^\tau - \tau \nabla F(x_{k+1}^\tau) \Rightarrow$$
$$\Rightarrow \frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \nabla F(x_{k+1}^\tau) = 0 \Rightarrow$$

- **Crucial! Minimizing Movement** scheme:

$$\Rightarrow x_{k+1}^\tau = \arg\min_x F(x) + \frac{\|x - x_k^\tau\|^2}{2\tau}$$





Gradient Flows in Wasserstein space

Question: Can we consider gradient flows in probability measures space?

Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{P}_2(\mathcal{X})$ is set of probability measures over $\mathcal{X}$ with finite second moment. Let $\mathcal{F} : \mathcal{P}_2(\mathcal{X}) \to \mathbb{R}$

First variation[2]
(nonstrict definition) The function $\frac{\delta \mathcal{F}}{\delta \rho}(\rho) : \mathcal{X} \to \mathbb{R}$ is called the *first variation* (if it exists and unique up to additive constant) if $\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{F}(\rho + \varepsilon\chi)|_{\varepsilon=0} = \int_{\mathcal{X}} \frac{\delta \mathcal{F}}{\delta \rho}(\rho) d\chi$ for every perturbation (measure on $\mathcal{X}$) χ such that $\rho + \varepsilon\chi \in \mathcal{P}_2(\mathcal{X})$.

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

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

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

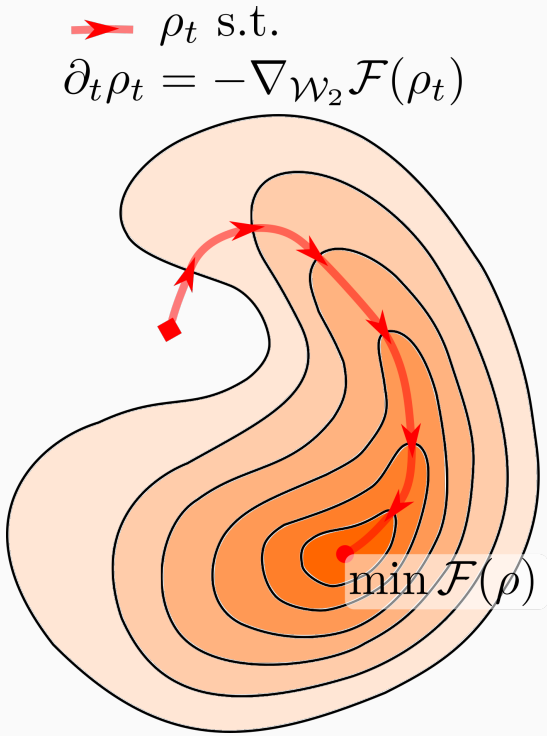
Gradient Flows in Wasserstein space

$$\partial_t \rho_t - \nabla \cdot (\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

Question Is there a "steepes descent" analogy in probability measures space $\mathcal{P}_2(\mathcal{X})$? - **Yes!**

There is so called Wasserstein-2 metric ($\mathcal{W}_2$) in $\mathcal{P}_2(\mathcal{X})$ such that in some sense:

$$-\nabla \cdot (\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = \nabla_{\mathcal{W}_2} \mathcal{F}(\rho)$$



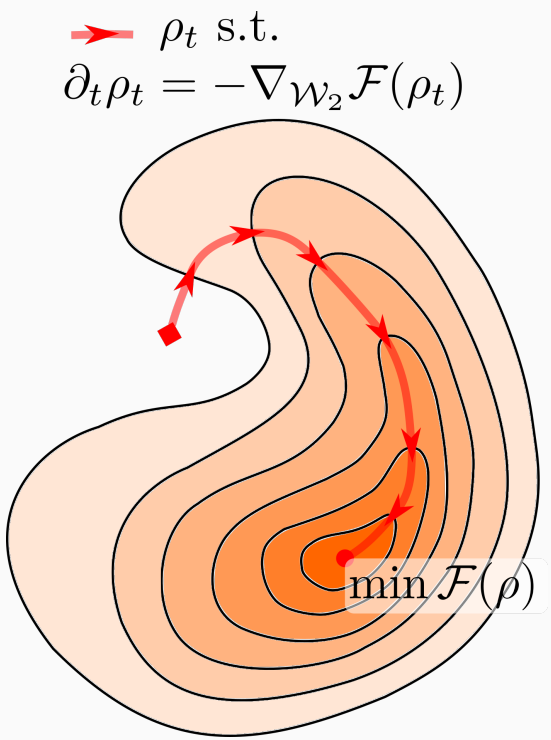
Gradient Flows in Wasserstein space

$$\partial_t \rho_t - \nabla \cdot (\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

Convergence[3]

Given $\mathcal{F}$ to be combination of $\mathcal{E}(\mu) = \int_{\mathcal{X}\times\mathcal{X}} W(x-y)d[\mu\otimes\mu](x,y) + \int_X V(x)d\mu(x)$ and $\mathcal{U}(\mu) = \int_X f(\mu(x))dx$ where f is convex superlinear, W, V are convex, suff. smooth, the gradient flow converges exponentially fast to a unique minimizer

$\mathcal{E}$ is called *potential* energy, $\mathcal{U}$ is called *internal* energy; if μ is not absolutely continuous w.r.t. Lebesgue measure $\mathcal{U}(\mu) := +\infty$



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

JKO

Minimizing Movement scheme in Wasserstein-2 space $(\mathcal{P}_2(\mathcal{X}), \mathcal{W}_2)$ is a sequence of $\{\rho_\tau^k\}_{k=1}^\infty \subset \mathcal{P}_2(\mathcal{X})$:

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

This equation is called **JKO** scheme[4]

Theorem Given $\mathcal{F} = \mathcal{E} + \mathcal{U}$ 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: $\rho_\tau(t) = \rho_\tau^k$, 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 gradient flow associated with $\mathcal{F}$

Question: But what is Wasserstein-2 distance? - **Next slides**

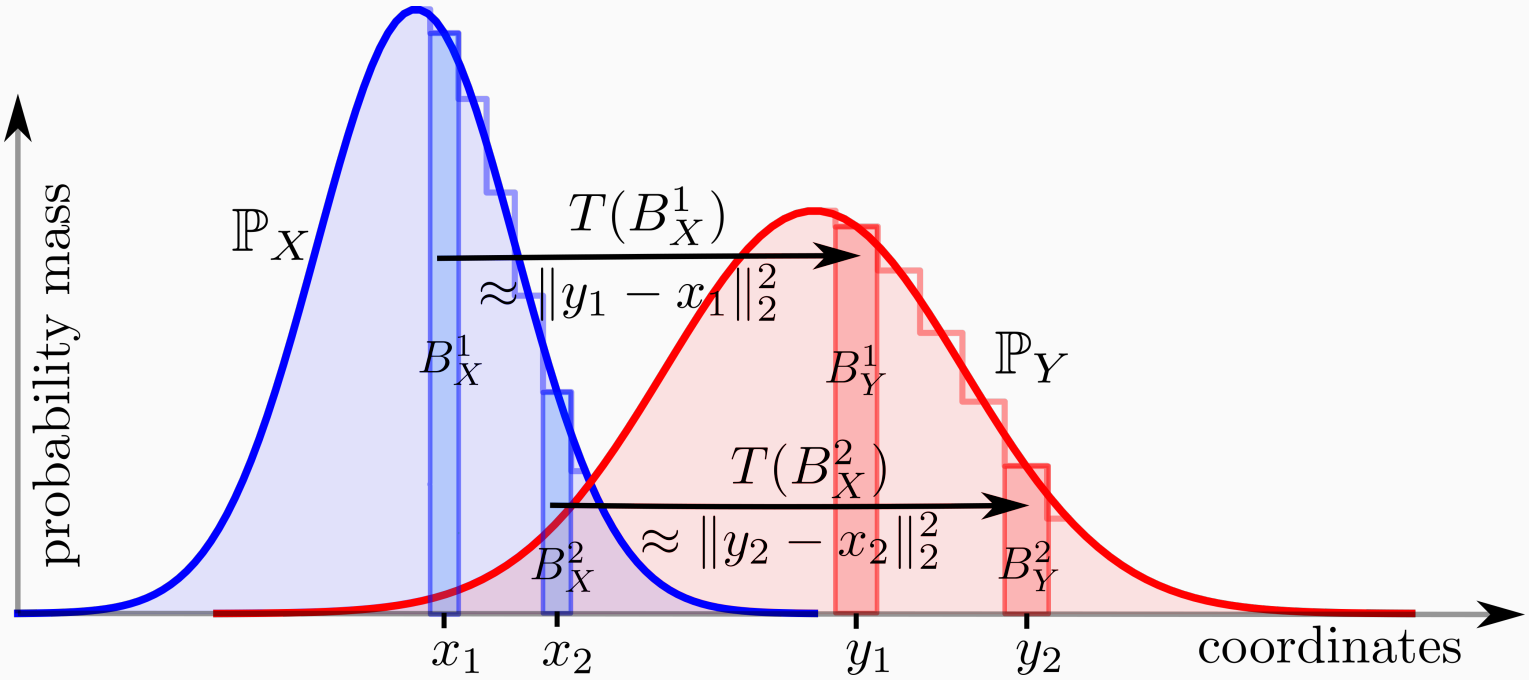
[4] Richard Jordan, David Kinderlehrer и Felix Otto (янв. 1998). "The Variational Formulation of the Fokker-Planck Equation". в: *SIAM J. Math. Anal.* 29.1, с. 1—17. ISSN: 0036-1410.

Wasserstein-2 distance

Pushforward transform Let $\mu, \nu \in \mathcal{P}_2(\mathcal{X})$, let $g : \mathcal{X} \to \mathcal{X}$ such, that $\forall B \in \mathcal{B}(\mathcal{X}) : \nu(g(B)) = \mu(B)$. Then ν is called *pushforward* of μ under g, it is denoted as $\nu = g\sharp\mu$.

(Squared) **Wasserstein-2 distance**

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



Wasserstein-2 distance

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

- The minimization problem above is called *Monge's optimal transportation problem*
- If μ and ν are absolutely continuous the minimizer T in the definition exists.
- There is alternative definition in terms of *Kantorovich's optimal transportation problem*:

$$W_2^2(\mu,\nu) = \inf_{\pi \in \prod(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2^2 d\pi(x,y)$$

$\prod(\mu,\nu)$ is set of prob. measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ an ν

Brenier's theorem[5]

Theorem 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:

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

Alternative formulation of JKO

$$\psi_k = \underset{\psi \in \mathsf{Conv}(\mathcal{X})}{\arg\min} \; \tau \mathcal{F}(\boldsymbol{\nabla} \psi \sharp \rho_\tau^k) + \frac{1}{2} \int_{\mathcal{X}} \|x - \boldsymbol{\nabla} \psi(x)\|_2^2 \mathrm{d}\rho_\tau^k(x)$$

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

[5] Villani Cédric (C 2003). *Topics in optimal transportation / Cédric Villani*. eng. Graduate studies in mathematics. Providence, Rhode Island: American mathematical society. ISBN: 0-8218-3312-X.

Gradient Flows: What for?

Stochastic Differential Equations

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)\,, \quad X(0) = X_0$$

Here $X(t)$ is a stochastic process, $f : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^n$ is the *drift* term, $g : \mathbb{R}_+ \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is the *diffusion* term, $W(t)$ is the standard Wiener process. The equation can be understood in integral sense:

$$X(t) = X_0 + \int_0^t f(s, X(s))ds + \overbrace{\int_0^t g(t, X(t))dW(t)}^{\text{integral Ito}}$$

Statement[6] Let f and g are smooth enough and for a solution $X(t)$ there exists the pdf $p(t, x)$. Then $p(t, X)$ satisfies the equation:

$$\frac{\partial p(t,x)}{\partial t} = \nabla \cdot (p(t,x)f(t,x)) + \frac{1}{2}\,\mathrm{tr}\left\{ \left\| \frac{\partial^2}{\partial x_i \partial x_j} \right\|_{i,j} [g(t,x)g(t,x)^T p(t,x)] \right\}$$

[6] А. А. Леваков (2009). *Стохастические дифференциальные уравнения*.

Stochastic Differential Equations

Recall the gradient flow formulation

$$\partial_t \rho_t - \boldsymbol{\nabla} \cdot (\rho_t \boldsymbol{\nabla}_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

We consider the functional $\mathcal{F}$ to be combination of the following functionals:

Functional	$\mathcal{V}(\rho)$	$\mathcal{W}(\rho)$	$\mathcal{U}(\rho)$
Formula	$\int_{\mathcal{X}} V(x) d\rho$	$\int_{\mathcal{X}\times\mathcal{X}} W(x-y) d\rho(x) d\rho(y)$	$\int_{\mathcal{X}} f(\rho(x)) dx$
Conditions	V - convex, suff. smooth	W - convex, suff. smooth	f - convex, superlinear
First variation	$\frac{\delta \mathcal{V}}{\delta \rho}(\rho) = V$	$\frac{\delta \mathcal{W}}{\delta \rho}(\rho) = 2\int_{\mathcal{X}} W(x,y) d\rho(y)$	$\frac{\delta \mathcal{U}}{\delta \rho} = f'(t)\vert_{t=\rho}$

Let's put the first variations in the gradient flow formulation!

Stochastic Differential Equations

$$\partial_t \rho_t - \boldsymbol{\nabla} \cdot (\rho_t \boldsymbol{\nabla}_\times \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

Let's put the first variations in the gradient flow formulation - we recover sever classical PDE associated with SDE![7]

Class	PDE $\partial_t \rho =$	Flow Functional $\mathcal{F}(\rho) =$
Heat Eqation	$\nabla^2 \rho$	$\int \rho \log \rho dx$
Advection	$\boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\int V(x) d\rho(x)$
Fokker-Planck	$\nabla^2 \rho + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\int \rho \log d\rho(x) + \int V(x) d\rho(x)$
Porous Media	$\nabla^2(\rho^m) + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\frac{1}{m-1} \int \rho^m(x) dx + \int V(x) d\rho$
Adv. + Diff.	$\boldsymbol{\nabla} \cdot [\rho(\nabla f'(\rho) + \nabla V$	$\int V(x) d\rho(x) + \int f(\rho(x)) dx +$
+ Interaction	$+(\nabla W) * \rho)]$	$+\frac{1}{2} \int \int W(x-y) d\rho(x) d\rho(y)$

[7] David Alvarez-Melis, Yair Schiff и Youssef Mroueh (2021). *Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks*. arXiv: 2106.00774 [stat.ML].

Stochastic Differential Equations

$$\partial_t \rho_t - \boldsymbol{\nabla} \cdot (\rho_t \boldsymbol{\nabla}_{\times} \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

Class	PDE $\partial_t \rho =$	Flow Functional $\mathcal{F}(\rho) =$
Heat Eqation	$\nabla^2 \rho$	$\int \rho \log \rho dx$
Advection	$\boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\int V(x) d\rho(x)$
Fokker-Planck	$\nabla^2 \rho + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\int \rho \log d\rho(x) + \int V(x) d\rho(x)$
Porous Media	$\nabla^2(\rho^m) + \boldsymbol{\nabla} \cdot (\rho \boldsymbol{\nabla} V)$	$\frac{1}{m-1} \int \rho^m(x) dx + \int V(x) d\rho$
Adv. + Diff.	$\boldsymbol{\nabla} \cdot [\rho\{\nabla f'(\rho) + \nabla V$	$\int V(x) d\rho(x) + \int f(\rho(x)) dx +$
+ Interaction	$+(\nabla W) * \rho\}]$	$+\frac{1}{2} \int \int W(x-y) d\rho(x) d\rho(y)$

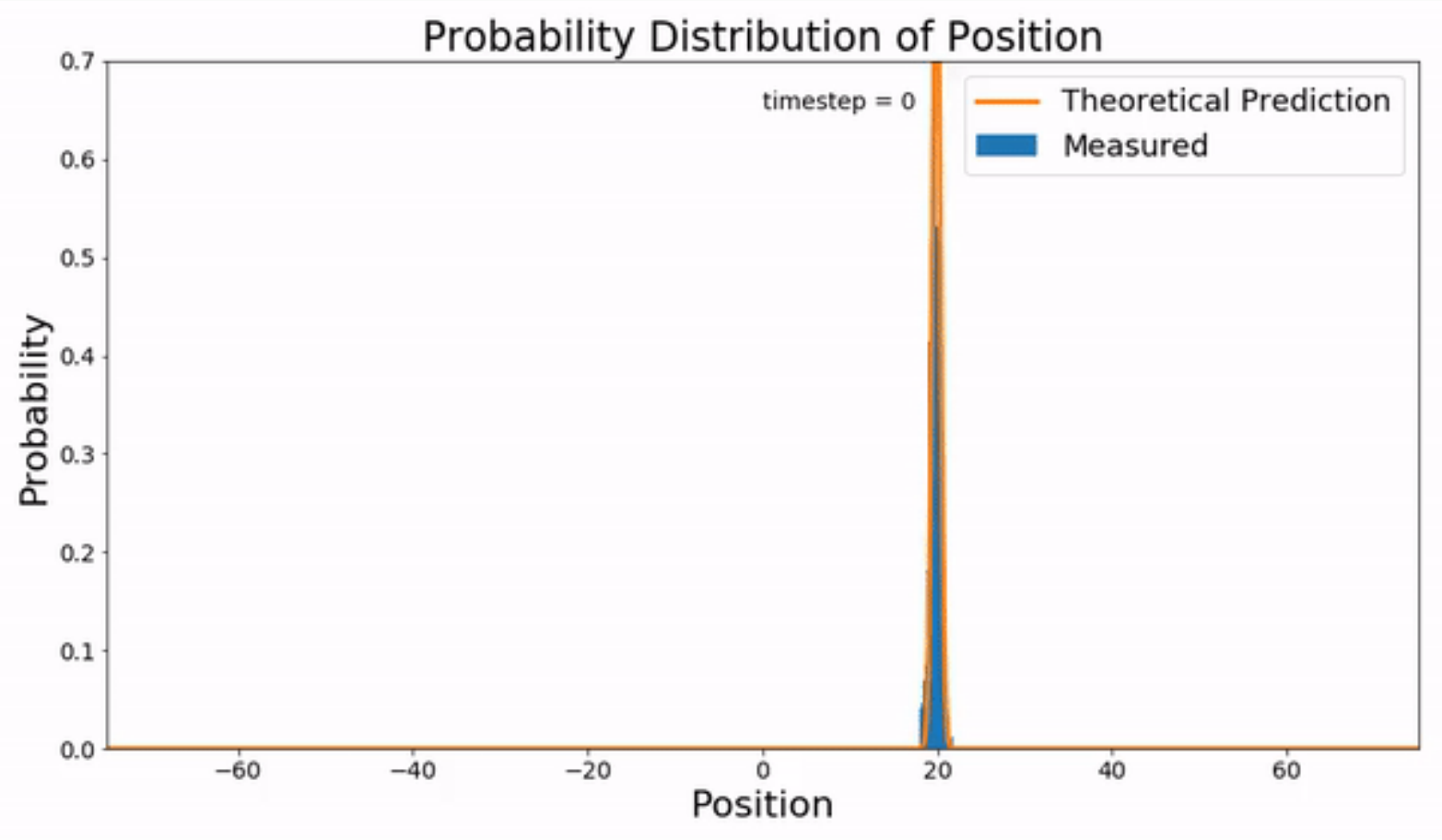
Can be solved via our method

Can be solved via our method in particular cases

Beyond our approach

Stochastic Differential Equations

Linear Fokker-Planck evolution example*



* Taken from Wikipedia: https://en.wikipedia.org/wiki/Fokker-Planck_equation

Stochastic Differential Equations

Particular use cases

- Modelling of SDEs arising in physics and biology (too general)
- **Population dynamics**[8]. The problem is to recover the potential $\mathcal{F}$ based on samples from the diffusion at different timesteps $t_1, t_2, \dots t_n$
- **Nonlinear filtering**. The problem is to model a stochastic process marginal distributions $\rho(t)$ given noisy observations from the process. (will be considered in further slides)
- GAN refining[9], Generative Modelling[10], ...

[8] Charlotte Bunne и др. (2021). *JKOnet: Proximal Optimal Transport Modeling of Population Dynamics*. arXiv: 2106.06345 [cs.LG].

[9] Abdul Fatir Ansari, Ming Liang Ang и Harold Soh (2021). "Refining Deep Generative Models via Discriminator Gradient Flow". в: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Zbc-ue9p_rE.

[10] Yang Song и др. (2021). "Score-Based Generative Modeling through Stochastic Differential Equations". в: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=PxTIG12RRHS.

Convergence to stationary distribution

$$\partial_t \rho_t - \nabla \cdot (\rho_t \nabla_x \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = 0\,, \quad \rho_{t=0} = \rho^0$$

Recall: Given $\mathcal{F}$ to be combination of potential energy $\mathcal{E}(\mu)$ and internal energy $\mathcal{U}(\mu)$ under several conditions the gradient flow converges exponentially fast to a unique minimizer

Observations:

- One can seek for optimizer of $\mathcal{F}$ following the appropriate Wasserstein gradient flow
- In particular case of Fokker-Planck equation $\mathcal{F}(\rho) = \beta^{-1} \int \rho \log \rho dx + \int V(x)\rho(x)dx$ the stationary distribution (minimizer) is $\frac{1}{Z}\exp(-\beta V(x))$, Z is partition function. One can converge to this distribution via WGF given potential V.

$\mathcal{E}(\mu) = \int_{\mathcal{X}\times\mathcal{X}} W(x-y)d[\mu\otimes\mu](x,y) + \int_{\mathcal{X}} V(x)d\mu(x)$; $\mathcal{U}(\mu) = \int_{\mathcal{X}} f(\mu(x))dx$

Convergence to stationary distribution

Particular use cases

- **Unnormalized posterior sampling** (will be considered in further slides)
- **Molecular discovery**[11] The idea is to increase *drug-likeness* (certain potential $\mathbb{E}_\rho V$) of molecules distribution while staying close to original distribution (certain discrepancy $D(\rho, \rho_0)$)
- Reinforcement Learning[12], ...

[11] David Alvarez-Melis, Yair Schiff и Youssef Mroueh (2021). *Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks*. arXiv: 2106.00774 [stat.ML].

[12] Ruiyi Zhang и др. (2018). "Policy Optimization as Wasserstein Gradient Flows". в: *Proceedings of the 35th International Conference on Machine Learning*. URL: https://proceedings.mlr.press/v80/zhang18a.html.

How to solve Wasserstein Gradient flows?

Basic Approaches

Both methods use the connection between SDE and WGF.

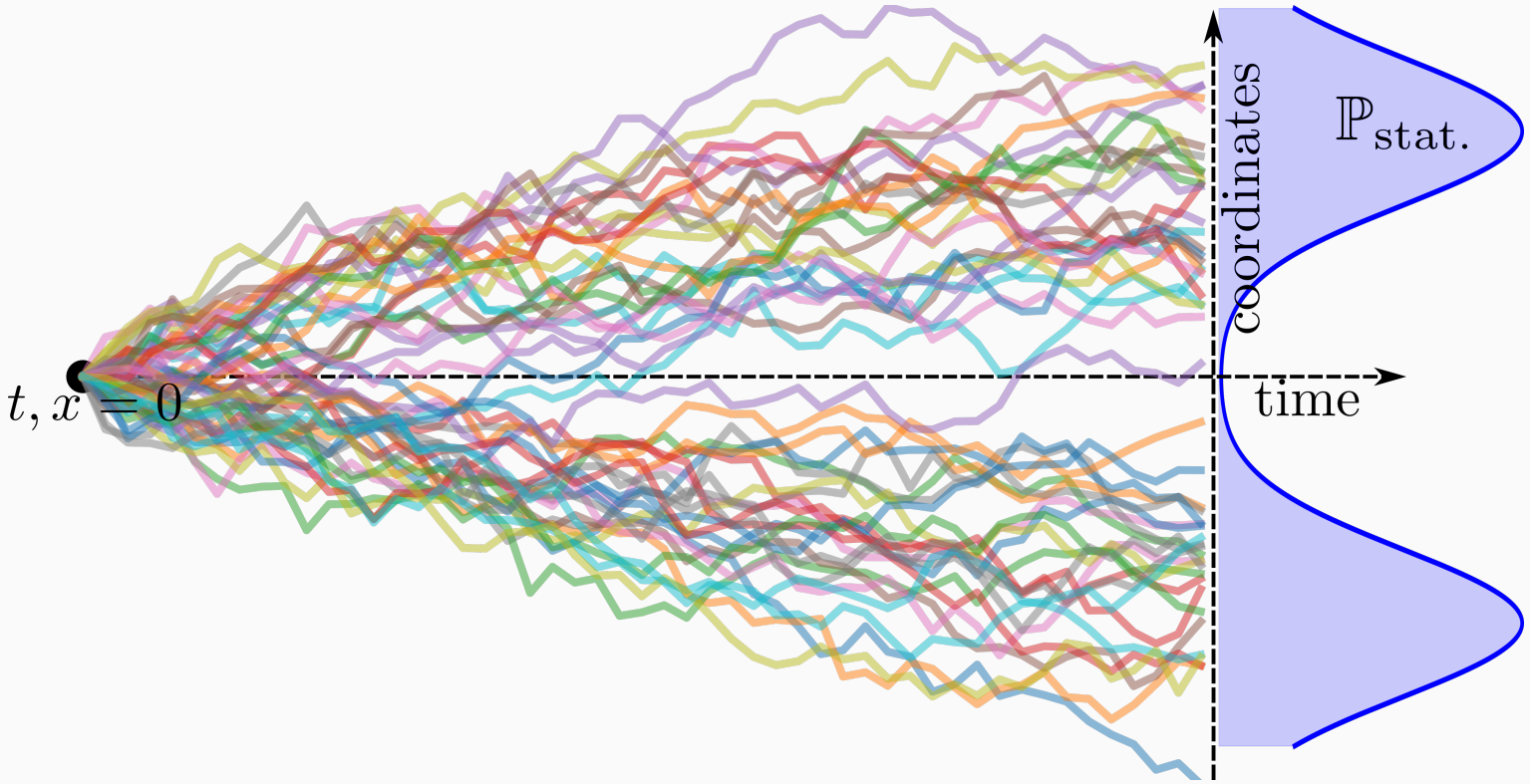
1. **Discretize space and time!** Cons.: Suffers from curse of dimensionality, not generalized to high dimensions.
2. **Euler-Maruyama approximation** Given the SDE $dX_t = f(t, X_t)dt + g(t, X_t)dW_t$, $X_{t=0} = X^0$ the idea is to approximate *trajectories* of the Stochastic Process via finite difference scheme:

$$X^\tau_{n+1} = X^\tau_n + f(n\tau, X^\tau_n)\tau + g(n\tau, X^\tau_n)\Delta W_n$$
$$\Delta W_n \sim \mathcal{N}(0, \tau I_{\text{dim}})$$

Pros. Simple and efficient for simulation and sampling
Cons. Doesn't provide density approximation (KDE perishes in multidimensional setting)

Basic Approaches

Demonstration of **Euler-Maruyama** simulation with multivariate stationary distribution



Regularized JKO[13]

Recall the **JKO** scheme:

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

Consider the *regularized* Wasserstein distance (note, we use Kantorovich form!):

$$\mathcal{W}_\gamma^2(\mu, \nu) = \inf_{\pi \in \prod(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2^2 d\pi(x, y) + \gamma \int_{\mathcal{X}\times\mathcal{X}} r(\pi(x, y)) dx dy$$

r assumed to be positive and strictly convex on $\mathbb{R}_+$. Examples are $r(x) = x^2$, $r(x) = x(\log(x) - 1)$. Note, that the operator $\nu \to \mathcal{W}_\gamma(\mu, \nu)$ is strictly convex!

[13] Charlie Frogner и Tomaso Poggio (2020). "Approximate Inference with Wasserstein Gradient Flows". в: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

Regularized JKO[14]

The *regularized* **JKO** objective then reads (for measures μ an ν):

$$\nu^* = \arg\min_{\nu \in \mathcal{P}_2(\mathcal{X})} \frac{1}{2}\mathcal{W}_\gamma^2(\mu, \nu) + \tau\mathcal{F}(\nu)$$

This optimization problem permits unconstrained dual formulation, i.e. there exists dual maximization objective $D_\gamma^\mu : L^2(\mathcal{X}) \times L^2(\mathcal{X}) \to \mathbb{R}$ s.t. strong duality holds[#]:

$$\max_{g,h \in L^2(\mathcal{X})} D_\gamma^\mu(g, h) = \min_{\nu \in \mathcal{P}_2(\mathcal{X})} \frac{1}{2}\mathcal{W}_\gamma^2(\mu, \nu) + \tau\mathcal{F}(\nu)$$

and optimal ν^* can be expressed in terms $\mathcal{F}$, g^*, h^*.

Cons.: biased optimization, doesn't work well in practice

[#] see original paper for reference

[14] Charlie Frogner и Tomaso Poggio (2020). "Approximate Inference with Wasserstein Gradient Flows". в: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

ICNN powered JKO

Recall the **alternative formulation of JKO**:

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

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

$$\Rightarrow \rho_\tau^K = \boldsymbol{\nabla}\phi_{K-1} \sharp \left[\boldsymbol{\nabla}\phi_{K-2} \sharp \{ \dots \boldsymbol{\nabla}\phi_0 \sharp \rho^0 \}\right]$$

Idea Consider the parameterization of the space of convex functions on $\mathcal{X}$: $\psi_\theta \in \mathsf{Conv}(\mathcal{X}), \theta \in \Theta$, then the optimization reads as follows:

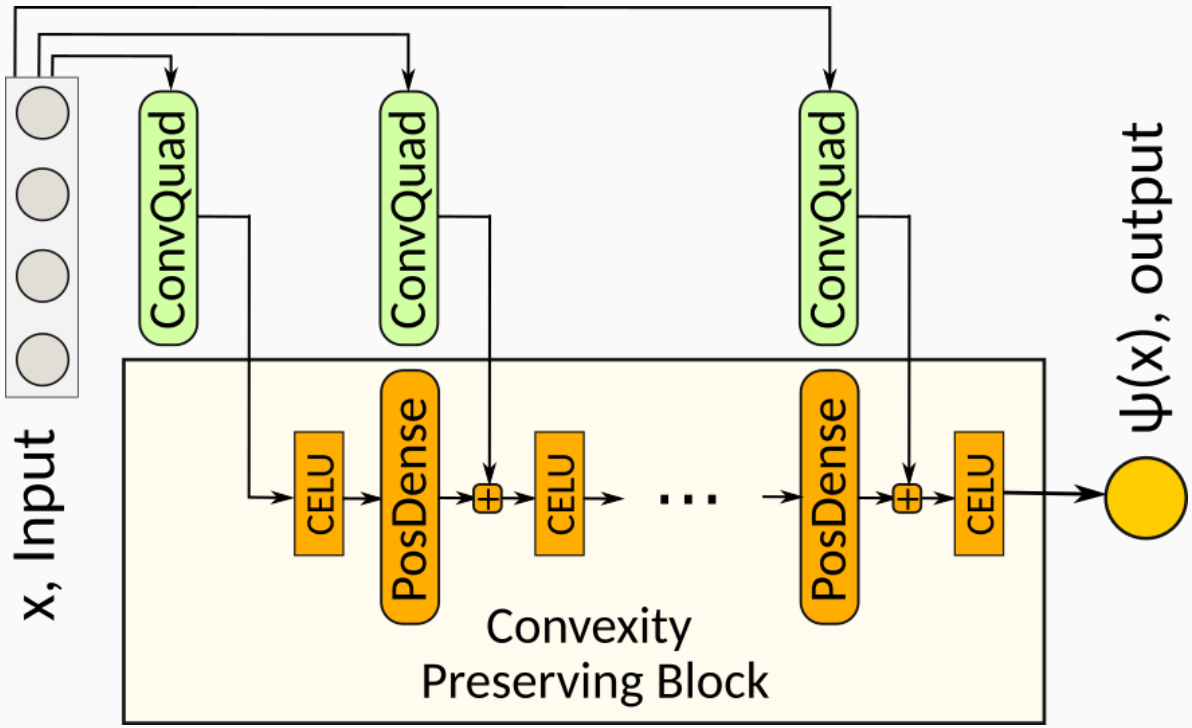
$$\theta^* \leftarrow \arg\min_\theta \left[\mathcal{F}(\nabla\psi_\theta \sharp \rho_\tau^k) + \frac{1}{2\tau}\int_{\mathcal{X}} \|x - \nabla\psi_\theta(x)\|_2^2 d\rho_\tau^k(x)\right]$$

Question 1: How to parameterize convex functions? - **Next slide**

Question 2: How to deal with $\mathcal{F}(\nabla\psi_\theta \sharp \rho_\tau^k)$ in practice? - **Coming soon**

Input Convex Neural Networks

Question: How to parameterize convex functions? - **ICNNs!**[15]



Convex Quadratic Layer (n-th neuron):

$$\mathrm{cq}_n(x) = \langle x, A_n x \rangle + \langle b_n, x \rangle + c_n, \qquad A_n = F_n^T F_n$$

[15] Alexander Korotin и др. (2021). "Wasserstein-2 Generative Networks". в: *International Conference on Learning Representations*.

Stochastic Optimization for JKO via ICNNs

Question: How to deal with $\mathcal{F}(\nabla\psi_\theta \sharp \rho_\tau^k)$ in practice?

Important In our study we consider $\mathcal{F}$ to be Fokker-Planck potential:

$$\mathcal{F}(\rho) \equiv \mathcal{F}_{\mathsf{FP}}(\rho) = \underbrace{\int_{\mathcal{X}} V(x) d\rho(x)}_{\text{potential energy}} + \beta^{-1} \underbrace{\int_{\mathcal{X}} \rho(x) \log \rho(x) dx}_{\text{neg. entropy}}$$

Theorem[16] Let $\rho \in \mathcal{P}_2(\mathcal{X})$ - absolute continuous, $T : \mathcal{X} \to \mathcal{X}$ is a diffeomorphism. Let $x_1, x_2, \dots x_N \sim \rho$. Then

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

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

[16] Petr Mokrov и др. (2021). *Large-Scale Wasserstein Gradient Flows*. arXiv: 2106.00736 [cs.LG].

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; $X \leftarrow \nabla\psi_{k-1} \circ \cdots \circ \nabla\psi_0(Z)$;

    $\widehat{\mathcal{W}_2^2} \leftarrow \frac{1}{N} \sum_{x \in X} \|\nabla\psi_\theta(x) - x\|_2^2$;

    $\widehat{\mathcal{F}_{\mathsf{FP}}} \leftarrow \frac{1}{N} \sum_{x \in X} V\big(\nabla\psi_\theta(x)\big) - \beta^{-1} \frac{1}{N} \sum_{x \in X} \log\det \nabla^2\psi_\theta(x)$

    $\widehat{\mathcal{L}} \leftarrow \frac{1}{2\tau}\widehat{\mathcal{W}_2^2} + \widehat{\mathcal{F}_{\mathsf{FP}}}$;

    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

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

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

Question: How to estimate the density ρ_τ^k?

By change of variable formula, given $x_k \in \mathcal{X}$ 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 $x_k \in \mathcal{X}$ one need to solve the optimization problems:

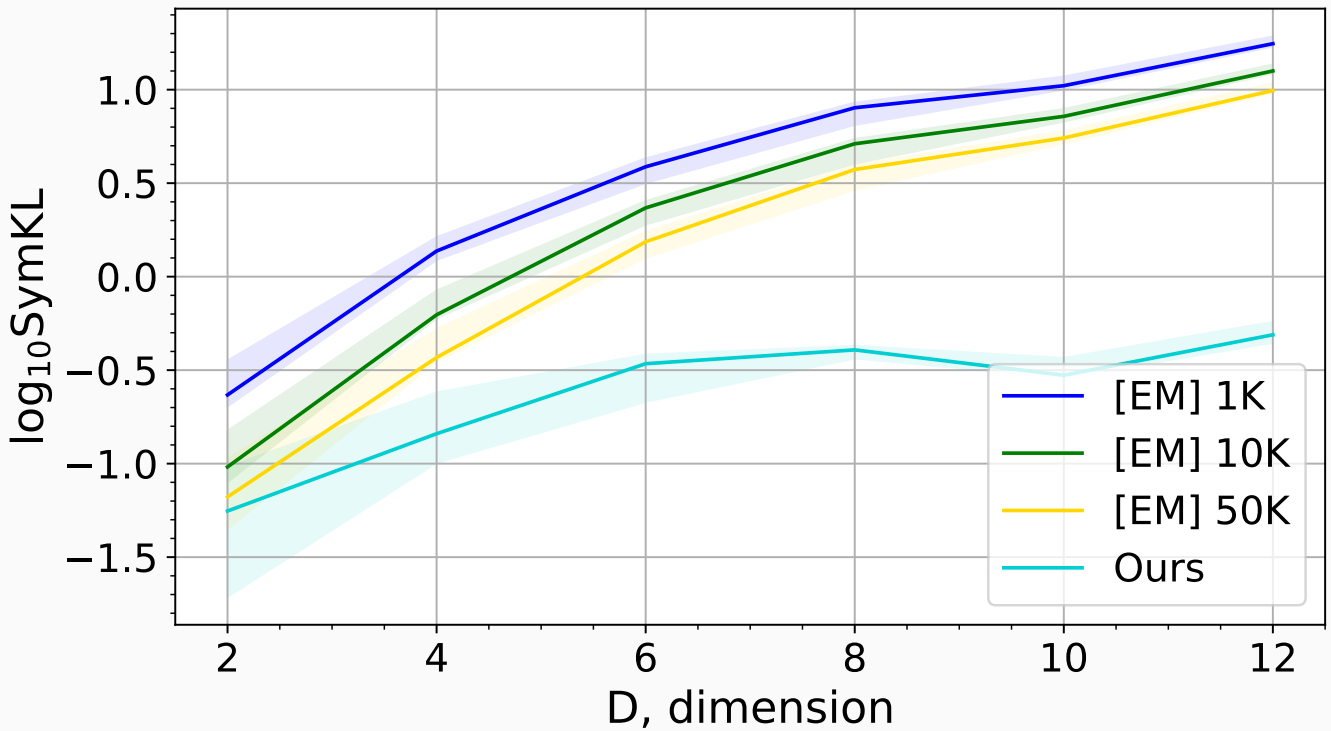
$$x_i = \nabla\psi_{i-1}(x_{i-1}) \iff x_{i-1} = \arg\max_{x \in \mathcal{X}} \left[\langle x, x_i \rangle - \psi_{i-1}(x) \right]$$

It is **convex optimization problem**!

Experiments

Convergence in high dimensions

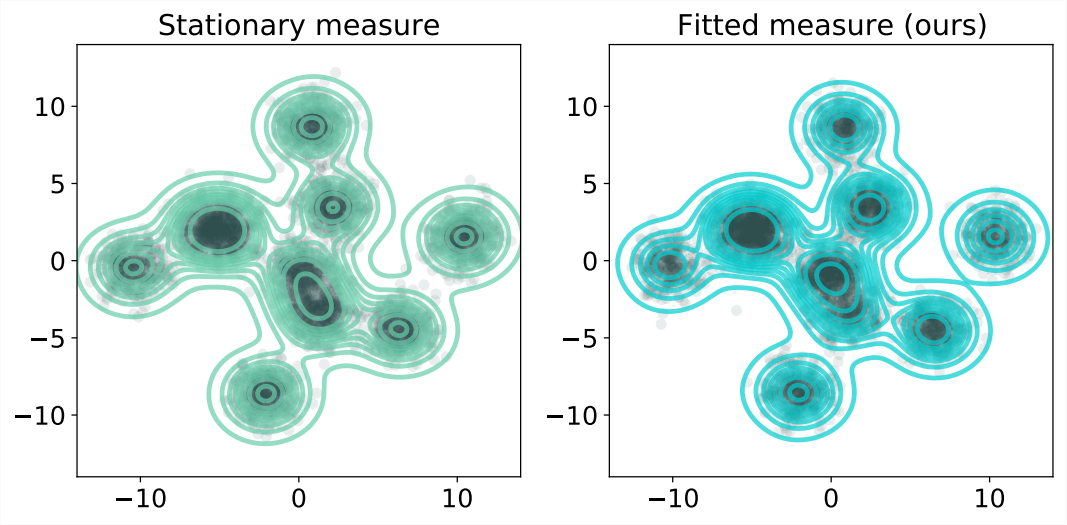
- Explore convergence of Fokker-Planck diffusion to the true stationary distribution $Z^{-1}\exp(-\beta V(x))$ in high dimensions
- $V(x)$ is a mixture of gaussians with $\sigma = 1$ and from 5 to 10 randomly generated centers
- We perform the JKO iterations for 4 sec. and use $\mathrm{SymKL}(\nu, \mu) = \mathrm{KL}(\nu\|\mu) + \mathrm{KL}(\mu\|\nu)$ to estimate the discrepancy.



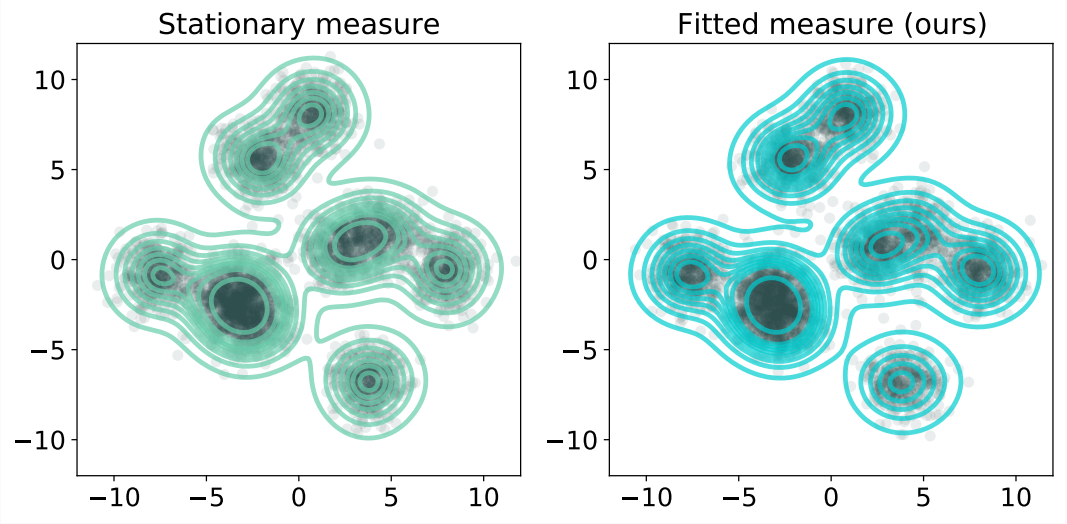
Convergence comparison in different dimensions

Convergence in high dimensions

- Visual discrepancy between fitted and true stationary distributions
- JKO iterations are performed until target metric (SymKL) stops changing.
- Project distributions to the first two PCA components of the true distribution



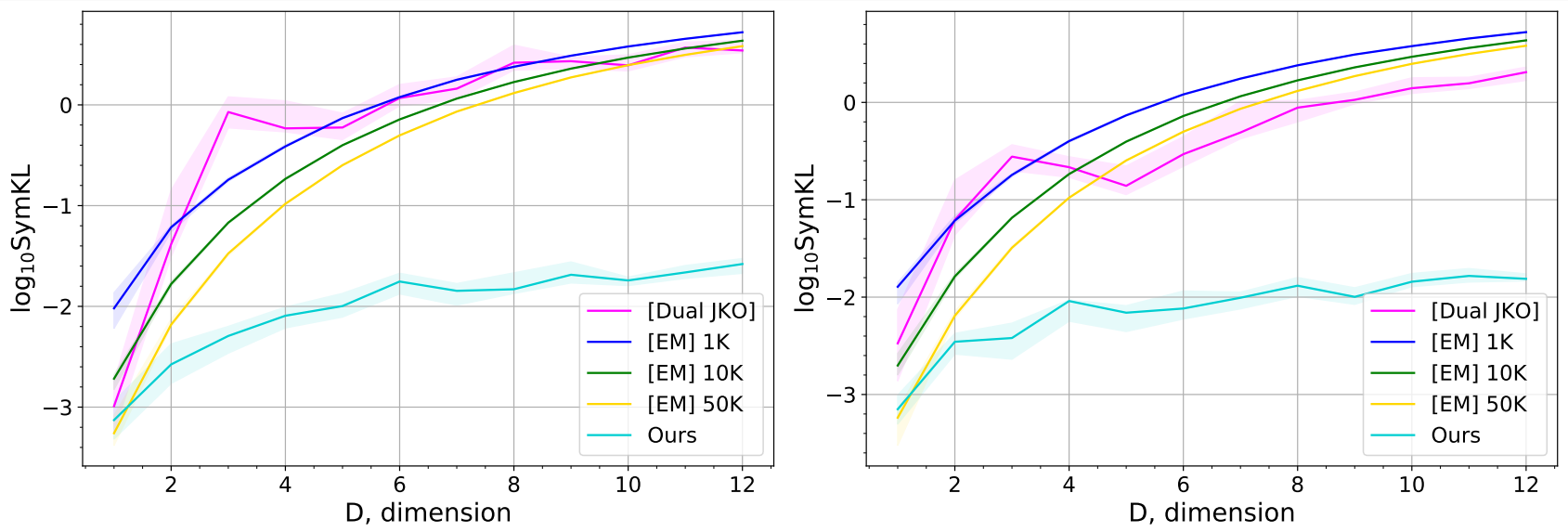
Projection to first two PC, $D = 13$



Projection to first two PC, $D = 32$

Ornstein–Uhlenbeck processes

- Potential $V(x) = \frac{1}{2}(x-b)^T A(x-b)$, 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)[17].
- We repeat experiment 15 times with randomly generated A, b



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

[17] Pat Vatiwutipong и Nattakorn Phewchean (2019). "Alternative way to derive the distribution ofthe multivariate Ornstein–Uhlenbeck process". в: *Advances in Difference Equations* 276.

Unnormalized Posterior Sampling

Given the model parameters $x \in \mathbb{R}^D$ with the prior distribution $p_0(x)$ and the conditional density $p(\mathcal{S}|x) = \prod_{m=1}^{M} p(s_m|x)$ of the data $\mathcal{S} = \{s_1, s_2, \dots s_M\}$ we are interested in posterior distribution:

$$p(x|\mathcal{S}) = \frac{p(\mathcal{S}|x)p_0(x)}{p(\mathcal{S})} \propto p(\mathcal{S}|x)p_0(x) = p_0(x) \cdot \prod_{m=1}^{M} p(s_m|x)$$

- Recall, that stationary distribution is $Z^{-1} \exp(-\beta V(x))$
- One can consider diffusion process with $V(x) = -\frac{1}{\beta} \log \left[p_0(x) \cdot p(\mathcal{S}|x)\right]$. The stationary distribution is $p(x|\mathcal{S})$
- Once the diffusion converge to the stationary distribution, we can both sample from $p(x|\mathcal{S})$ and estimate pdf of $p(x|\mathcal{S})$.

Unnormalized Posterior Sampling

- We apply our method to the Bayesian logistic regression
- There are used 8 benchmark datasets[18]
- We divide each dataset onto $\mathcal{S}_{\text{train}}$ and $\mathcal{S}_{\text{test}}$, derive posterior $p(x|\mathcal{S}_{\text{train}})$ and compute $\mathcal{S}_{\text{test}}$ predictive distribution log likelihood and accuracy

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

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

Nonlinear filtering[19]

Let X_t be diffusion process governed by Fokker-Planck equation with potential $V(x)$. At the time moments $t_1, t_2, \ldots, t_K$ we obtain the noisy observations of the process:

$$Y_k = X_{t_k} + v_k\,, \quad v_k \sim \mathcal{N}(0,1)$$

Our aim is to determine the process X distribution at the time $t \geq t_K$ given the observations $Y_{1:K}$: $p_{t,X}(x|Y_1, Y_2, \ldots Y_K)$

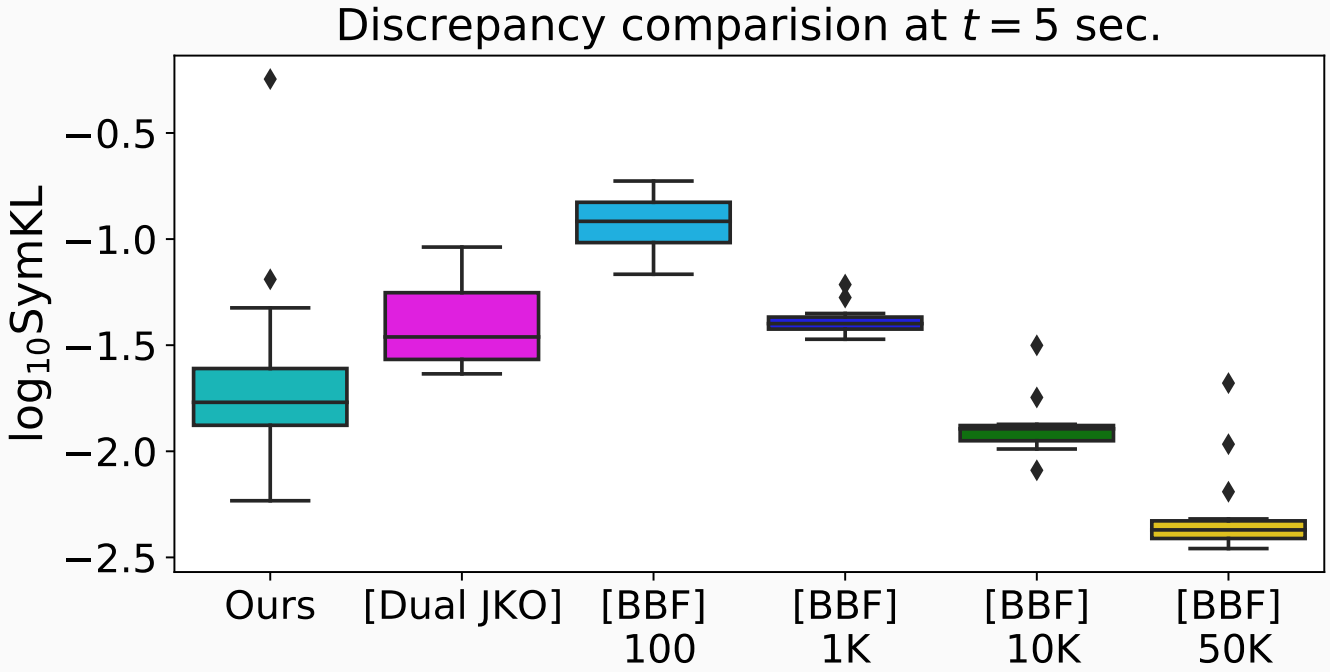
1. If $t > t_K$: $p_{t,X}(x|Y_1, \ldots Y_K)$ follows the diffusion process on time interval $[t_K, t]$ with initial distribution $p_{t_K,X}(x|Y_1, \ldots Y_K)$
2. If if $t_K = t$ then ($p_{t,Y}(y)$ is pdf of the Y at time moment t):

$$p_{t_K,X}(x|Y_{1:K}) = \frac{p_{t_K,Y}(Y_K|X_{t_K} = x)p_{t_K,X}(x|Y_{1:K-1})}{p_{t_K,Y}(Y_K|Y_{1:K-1})} \qquad \begin{pmatrix}\textbf{marginal}\\ \textbf{posterior}\end{pmatrix}$$

[19] Charlie Frogner и Tomaso Poggio (2020). "Approximate Inference with Wasserstein Gradient Flows". в: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*.

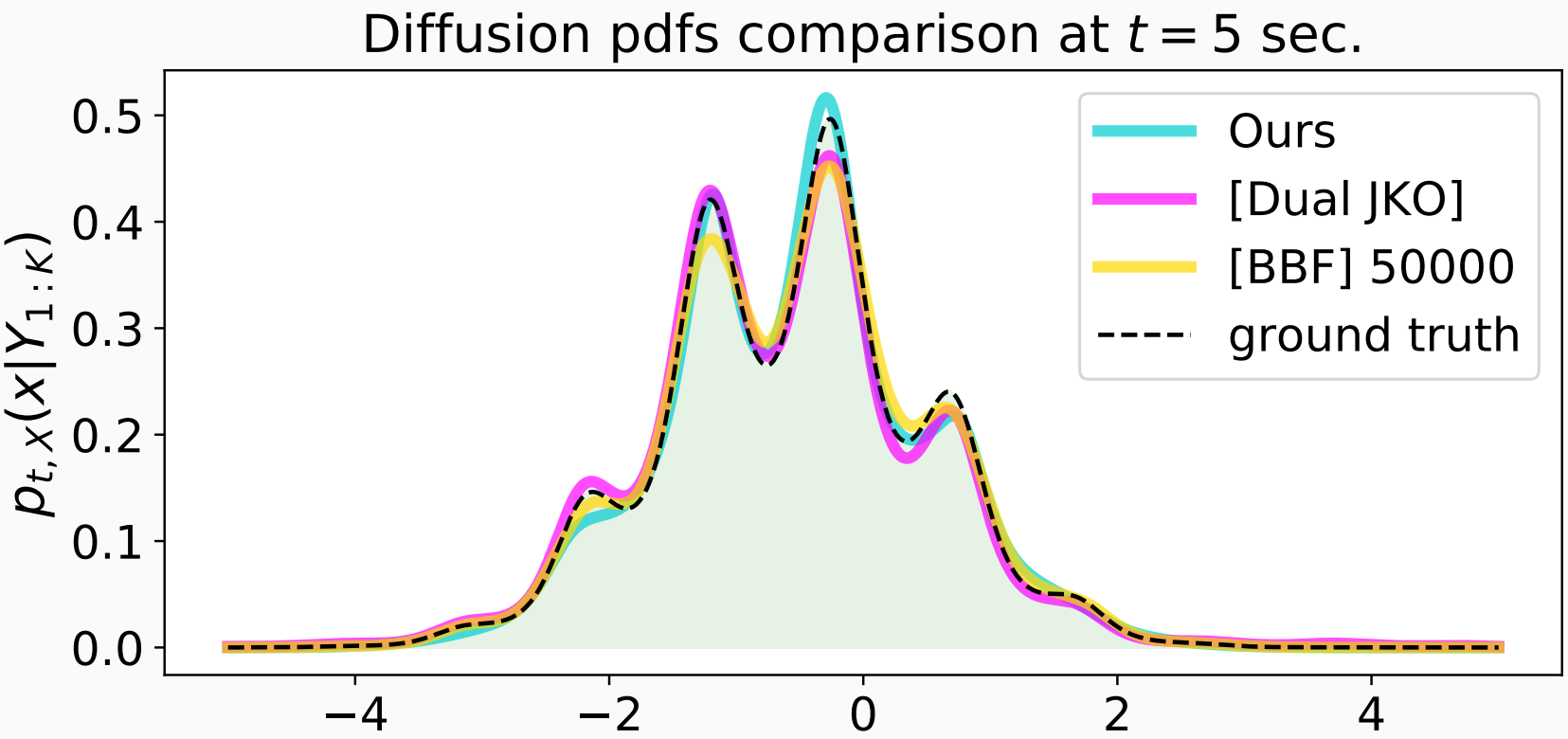
Nonlinear filtering

- $V(x) = \frac{1}{\pi}\sin(2\pi x) + \frac{1}{4}x^2$ (it is highly nonlinear process)
- Filtering takes $t_{\text{el}} = 9$ sec. (noise observations each 0.5 sec.)
- We use MCMC technique to sample from marginal posterior distribution.
- Reference method is numerical integration on fine grid
- We compare our method with Bayesian Bootstrap filter approach (which exploits particle simulation)



Nonlinear filtering

Probability density functions visual discrepancy demonstration



Large-Scale Wasserstein Gradient Flows - Conclusions

- Wasserstein gradient flow is a steepest descent curve in Wasserstein space of probability measures $(\mathcal{P}_2(\mathcal{X}), \mathcal{W}_2)$
- The WGFs have deep connection to SDEs which makes them especially interesting
- There are several methods aimed at modelling WGFs/SDEs which have problems with scalability/fullness
- There is a new scalable approach for modelling WGFs based on ICNN powered JKO which demonstrates it's applicability and efficiency in several experiments