# Sliced-Wasserstein Gradient Flows

Clément Bonet[1], Nicolas Courty[1], François Septier[1], Lucas Drumetz[2]

[1]Université Bretagne Sud
[2]IMT Atlantique

GDR ISIS - Optimal Transport and Statistical Learning
18/11/2021

# Contents

# Gradient Flows on $\mathbb{R}^p$

Let $X = \mathbb{R}^p$, $d$ a distance (*e.g.* $d(x, y) = \|x - y\|_2$), $F : X \to \mathbb{R}$.
Goal:

$$\min_x \; F(x)$$

# Gradient Flows on $\mathbb{R}^p$

Let $X = \mathbb{R}^p$, $d$ a distance (*e.g.* $d(x, y) = \|x - y\|_2$), $F : X \to \mathbb{R}$.
Goal:

$$\min_x \; F(x)$$

## Definition (Gradient Flow on $\mathbb{R}^p$)

A gradient flow is a curve $x : [0, T] \to X$ which decreases as much as possible along the functional $F$.
*i.e.* If $F$ is differentiable, $x$ follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

# Gradient Flows on $\mathbb{R}^p$

If $F$ is differentiable, $x$ follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Solving the ODE in practice:

# Gradient Flows on $\mathbb{R}^p$

If $F$ is differentiable, $x$ follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Solving the ODE in practice:

- Explicit Euler scheme ($x_k = x(k\tau)$):

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

# Gradient Flows on $\mathbb{R}^p$

If $F$ is differentiable, $x$ follows the Cauchy problem

$$\begin{cases} \frac{\mathrm{d}x}{\mathrm{d}t}(t) = -\nabla F(x(t)) \\ x(0) = x_0 \end{cases}$$

Solving the ODE in practice:

- Explicit Euler scheme ($x_k = x(k\tau)$):

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

- Implicit Euler scheme:

$$\begin{aligned} x_{k+1} = x_k - \tau \nabla F(x_{k+1}) &\iff 0 = \frac{x_{k+1} - x_k}{\tau} + \nabla F(x_{k+1}) \\ &\iff x_{k+1} \in \operatorname*{argmin}_{x \in X} \frac{\|x - x_k\|_2^2}{2\tau} + F(x) \\ &\iff x_{k+1} = \mathrm{prox}_{\tau F}(x_k) \end{aligned}$$

# Wasserstein Distance

## Definition (Wasserstein Distance)

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x - y\|_2^2 \, \mathrm{d}\gamma(x, y)$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d),\ \pi^1_{\#}\gamma = \mu,\ \pi^2_{\#}\gamma = \nu\}$.

# Wasserstein Gradient Flows

Gradient Flow in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$:
Iterated Minimization scheme (JKO Scheme) [Jordan et al., 1998]:

$$\mu_{k+1}^\tau \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \frac{1}{2\tau} W_2^2(\mu, \mu_k^\tau) + F(\mu)$$

# Wasserstein Gradient Flows

Gradient Flow in $(\mathcal{P}_2(\mathbb{R}^d), W_2)$:
Iterated Minimization scheme (JKO Scheme) [Jordan et al., 1998]:

$$\mu_{k+1}^\tau \in \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \ \frac{1}{2\tau} W_2^2(\mu, \mu_k^\tau) + F(\mu)$$

## Examples

- $F(\mu) = \int \rho(x) \log \rho(x) \ \mathrm{d}x + \int V(x)\rho(x)\mathrm{d}x$ if $\mathrm{d}\mu = \rho \mathrm{dLeb}$
  Solution in the limit $\tau \to 0$ to the PDE: (Fokker-Planck)

$$\partial_t \rho_t = \mathrm{div}(\rho_t \nabla V) + \Delta \rho_t$$

- $F(\mu) = \frac{1}{2} SW_2^2(\mu, \nu) + \lambda \mathcal{H}(\mu)$ [Bonnotte, 2013, Liutkus et al., 2019]
- $F(\mu) = \frac{1}{2} MMD^2(\mu, \nu)$ [Arbel et al., 2019]
- $F(\mu) = \frac{1}{2} \mathrm{KSD}^2(\mu, \nu)$ [Korba et al., 2021]

# Numerical Methods

- If an associated SDE is known, simulate from it [Liu et al., 2021, Liutkus et al., 2019, Arbel et al., 2019, Korba et al., 2021]

## Examples

Let $F(\mu) = \int V(x)\rho(x)\mathrm{d}x + \int \log(\rho(x))\rho(x)\mathrm{d}x$,
Gradient Flow solution of:

$$\partial_t \rho_t = \mathrm{div}(\rho_t \nabla V) + \Delta \rho_t$$

Associated SDE (Langevin Equation):

$$\mathrm{d}X_t = -\nabla V(X_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$$

# Numerical Methods

- If the SDE is known, simulate from it
- Solving the JKO scheme by discretizing the grid:
  - Entropic regularized scheme on a discretized grid [Peyré, 2015, Carlier et al., 2017]
  - Methods based on the dynamic formulation of the transport [Laborde, 2016, Carrillo et al., 2021]

# Numerical Methods

- If the SDE is known, simulate from it
- Solving the JKO scheme by discretizing the grid
- Using Neural Networks, *e.g.* JKOICNN [Alvarez-Melis et al., 2021, Mokrov et al., 2021, Bunne et al., 2021]

# JKOICNN

## Theorem (Brenier's Theorem)

*Let* $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$*,* $\mu$ *absolutely continuous with respect to the Lebesgue measure. Then, the optimal coupling* $\gamma^*$ *is unique and of the form* $\gamma^* = (Id, \nabla\varphi)_\# \mu$ *with* $\nabla\varphi$ *is a convex function.*

# JKOICNN

## Theorem (Brenier's Theorem)

*Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mu$ absolutely continuous with respect to the Lebesgue measure. Then, the optimal coupling $\gamma^*$ is unique and of the form $\gamma^* = (Id, \nabla\varphi)_\#\mu$ with $\nabla\varphi$ is a convex function.*

- Reformulate the problem as:

$$u_{k+1}^\tau \in \underset{u \in \mathrm{cvx}}{\mathrm{argmin}} \ \frac{1}{2\tau} \int \|\nabla u(x) - x\|_2^2 \ \rho_k^\tau(x)\mathrm{d}x + F\big((\nabla u)_\#\rho_k^\tau\big)$$

- Implicitly define $\rho_{k+1}^\tau = (\nabla u_{k+1}^\tau)_\#\rho_k^\tau$
- Use Input Convex Neural Networks (ICNN) [Amos et al., 2017] to model the convex functions:

$$\theta_{k+1}^\tau \in \underset{\theta \in \{\theta, u_\theta \in \mathrm{ICNN}\}}{\mathrm{argmin}} \ \frac{1}{2\tau} \int \|\nabla_x u_\theta(x) - x\|_2^2 \ \rho_k^\tau(x)\mathrm{d}x + F\big((\nabla_x u_\theta)_\#\rho_k^\tau\big)$$

- Backpropagate through gradient
- $O(k^2)$ evaluations

# Sliced-Wasserstein Distance

**Definition (Sliced-Wasserstein Distance [Rabin et al., 2011])**

Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,

$$SW_2^2(\mu, \nu) = \int_{S^{d-1}} W_2^2(P^\theta_\# \mu, P^\theta_\# \nu)\ \lambda(\mathrm{d}\theta)$$

where $P^\theta(x) = \langle x, \theta \rangle$, $\lambda$ uniform measure on $S^{d-1} = \{\theta \in \mathbb{R}^d,\ \|\theta\|_2 = 1\}$.

Properties:

- Distance
- Equivalent to $W_2$ for compact supported measures [Bonnotte, 2013]
- Metrizes the weak convergence as $W_2$ [Nadjahi et al., 2019]
- Easy to approximate

# Sliced-Wasserstein Gradient Flows

Goal:

$$\min_{\mu\in\mathcal{P}(\mathbb{R}^d)} \ F(\mu)$$

JKO scheme in $(P_2(\mathbb{R}^d), SW_2)$:

$$\mu_{k+1}^\tau \in \operatorname*{argmin}_\mu \ \frac{1}{2\tau} SW_2^2(\mu, \mu_k^\tau) + F(\mu)$$

Goal:

$$\min_{\mu\in\mathcal{P}(\mathbb{R}^d)} \ F(\mu)$$

JKO scheme in $(P_2(\mathbb{R}^d), SW_2)$:

$$\mu_{k+1}^\tau \in \operatorname*{argmin}_\mu \ \frac{1}{2\tau} SW_2^2(\mu,\mu_k^\tau) + F(\mu)$$

- Analysis of the SW-JKO scheme
  - Discrete solution at each step if *e.g.* $F$ convex and lsc.
  - Unique solution at each step if *e.g.* $\mu_k^\tau$ absolutely continuous or $F$ strictly convex.
  - $F$ non increasing along $(\mu_k^\tau)_k$.

# Sliced-Wasserstein Gradient Flows

Goal:

$$\min_{\mu\in\mathcal{P}(\mathbb{R}^d)} \ F(\mu)$$

JKO scheme in $(P_2(\mathbb{R}^d), SW_2)$:

$$\mu_{k+1}^\tau \in \operatorname*{argmin}_\mu \ \frac{1}{2\tau} SW_2^2(\mu, \mu_k^\tau) + F(\mu)$$

- Analysis of the SW-JKO scheme
  - Discrete solution at each step if *e.g.* $F$ convex and lsc.
  - Unique solution at each step if *e.g.* $\mu_k^\tau$ absolutely continuous or $F$ strictly convex.
  - $F$ non increasing along $(\mu_k^\tau)_k$.
- Pass to the limit
  - Does the gradient flow exist? In which sense?
  - Is the limit solution to a PDE?

# Solving the SW-JKO Scheme in Practice

- Use a discretized grid $(x_i)_{i=1}^N$, model $\mu = \sum_{i=1}^N \rho_i \delta_{x_i}$ and learn the weights:

$$\min_{(\rho_i)_i \in \Sigma_N} \ \frac{SW_2^2(\sum_{i=1}^N \rho_i \delta_{x_i}, \mu_k^\tau)}{2\tau} + F(\sum_{i=1}^N \rho_i \delta_{x_i})$$

# Solving the SW-JKO Scheme in Practice

- Use a discretized grid $(x_i)_{i=1}^N$, model $\mu = \sum_{i=1}^N \rho_i \delta_{x_i}$ and learn the weights:

$$\min_{(\rho_i)_i \in \Sigma_N} \ \frac{SW_2^2(\sum_{i=1}^N \rho_i \delta_{x_i}, \mu_k^\tau)}{2\tau} + F(\sum_{i=1}^N \rho_i \delta_{x_i})$$

- Learn particles, *i.e.* $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and solve

$$\min_{(x_i)_i} \ \frac{SW_2^2(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \mu_k^\tau)}{2\tau} + F(\frac{1}{N}\sum_{i=1}^N \delta_{x_i})$$

# Solving the SW-JKO Scheme in Practice

- Use a discretized grid $(x_i)_{i=1}^N$, model $\mu = \sum_{i=1}^N \rho_i \delta_{x_i}$ and learn the weights:

$$\min_{(\rho_i)_i \in \Sigma_N} \ \frac{SW_2^2(\sum_{i=1}^N \rho_i \delta_{x_i}, \mu_k^\tau)}{2\tau} + F(\sum_{i=1}^N \rho_i \delta_{x_i})$$

- Learn particles, *i.e.* $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ and solve

$$\min_{(x_i)_i} \ \frac{SW_2^2(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \mu_k^\tau)}{2\tau} + F(\frac{1}{N}\sum_{i=1}^N \delta_{x_i})$$

- Use a generative model (*e.g.* NF), *i.e.* $\mu = (g_\theta)_\# p_Z$ with $p_Z$ a standard distribution:

$$\min_\theta \ \frac{SW_2^2\big((g_\theta^{k+1})_\# p_Z, \mu_k^\tau\big)}{2\tau} + F\big((g_\theta^{k+1})_\# p_Z\big)$$

# Fokker-Planck

$$F(\mu) = \int V \mathrm{d}\mu + \int \log(\rho(x))\rho(x)\mathrm{d}x$$

with $V(x) = \frac{1}{2}(x-m)^T A(x-m)$, $\mu^* \propto e^{-V}$, *i.e.* $\mu^* = \mathcal{N}(m, A^{-1})$.

# Fokker-Planck

$$F(\mu) = \int V \mathrm{d}\mu + \int \log(\rho(x))\rho(x)\mathrm{d}x$$

with $V(x) = \frac{1}{2}(x-m)^T A(x-m)$, $\mu^* \propto e^{-V}$, *i.e.* $\mu^* = \mathcal{N}(m, A^{-1})$.

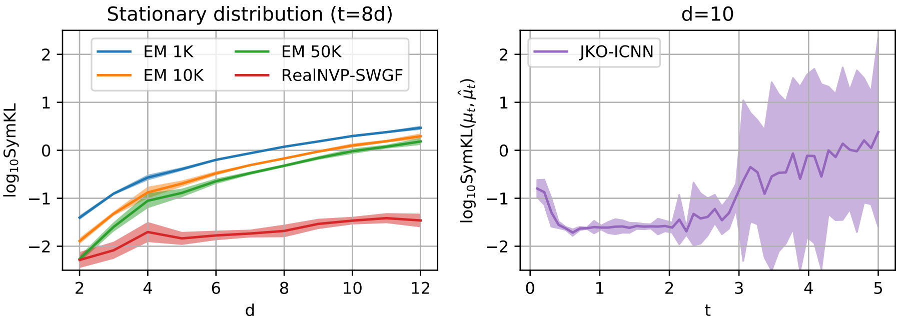
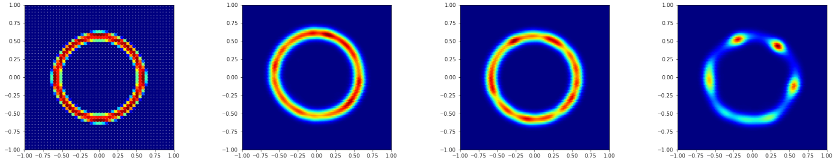



Figure: On the left, SymKL divergence between solutions at time $t = 8d$ (using $\tau = 0.1$ and 80 steps) and stationary measure. On the right, SymKL between the true WGF $\mu_t$ and the approximation with JKO-ICNN $\hat{\mu}_t$, run through 3 Gaussians with $\tau = 0.1$. We observe unstabilities at some point.

# Aggregation Equation

$$\mathcal{W}(\mu) = \frac{1}{2} \iint W(x-y)\mathrm{d}\mu(x)\mathrm{d}\mu(y)$$

with $W(x) = \frac{\|x\|^4}{4} - \frac{\|x\|^2}{2}$ [Carrillo et al., 2021].



(a) Discretized grid (b) MLP (c) Particles (d) JKO-ICNN

Figure: Steady state of the aggregation equation.

| MLP ($\tau = 0.05$) | Particles ($\tau = 0.05$) | JKOICNN ($\tau = 0.1$) |
|---|---|---|
| 20mn (200 steps) | 10mn (200 steps) | 5h (100 steps) |

Table: Runtime on RTX2080TI.

# Sliced-Wasserstein

Sliced-Wasserstein Flows [Liutkus et al., 2019]

$$F(\mu) = SW_2^2(\mu, \nu)$$

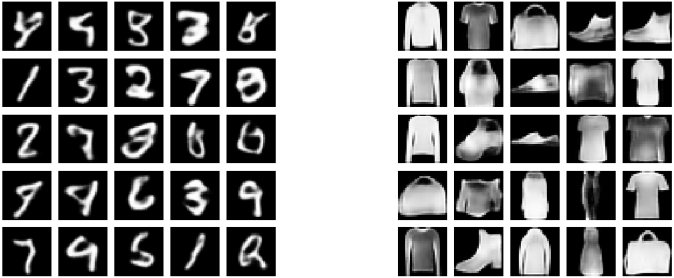


Figure: Generated sample obtained through a pretrained decoder $(d = 48)$.

# Conclusion

- Empirical study of Sliced-Wasserstein gradient flows
- Flexible implementations

Future work

- Theoretical study of SWGFs
- Use variant or approximation of SW in high dimension
- Other distance such as max-SW

# Conclusion

- Empirical study of Sliced-Wasserstein gradient flows
- Flexible implementations

Future work

- Theoretical study of SWGFs
- Use variant or approximation of SW in high dimension
- Other distance such as max-SW

Thank you!

# References I

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks, 2021.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *arXiv preprint arXiv:1906.04370*, 2019.

Nicolas Bonnotte. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Paris 11, 2013.

Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Jkonet: Proximal optimal transport modeling of population dynamics, 2021.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.

Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. *Foundations of Computational Mathematics*, pages 1–55, 2021.

# References II


Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1): 1–17, 1998.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel stein discrepancy descent. *arXiv preprint arXiv:2105.09994*, 2021.

Maxime Laborde. *Interacting particles systems, Wasserstein gradient flow approach*. PhD thesis, PSL Research University, 2016.

Shu Liu, Haodong Sun, and Hongyuan Zha. Approximating the optimal transport plan via particle-evolving method, 2021.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. Sliced-wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR, 2019.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows, 2021.

# References III


Kimia Nadjahi, Alain Durmus, Umut Şimşekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the sliced-wasserstein distance. *arXiv preprint arXiv:1906.04516*, 2019.

Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.