Sliced-Wasserstein Gradient Flows

Clément Bonet\textsuperscript{1}, Nicolas Courty\textsuperscript{1}, François Septier\textsuperscript{1}, Lucas Drumetz\textsuperscript{2}

\textsuperscript{1}Université Bretagne Sud
\textsuperscript{2}IMT Atlantique

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

1 Gradient Flows on Euclidean Space

2 Wasserstein Gradient Flows

3 Sliced-Wasserstein Gradient Flows
   • SW-JKO Scheme
   • Empirical Results
Let $X = \mathbb{R}^p$, $d$ a distance (e.g. $d(x, y) = \|x - y\|_2$), $F : X \to \mathbb{R}$. 
Goal:

$$\min_x F(x)$$
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

$$\begin{cases}
\frac{dx}{dt}(t) = -\nabla F(x(t)) \\
x(0) = x_0
\end{cases}$$
Gradient Flows on $\mathbb{R}^p$

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

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

Solving the ODE in practice:
Gradient Flows on $\mathbb{R}^p$

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

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

Solving the ODE in practice:

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

$$x_{k+1} = x_k - \tau \nabla F(x_k)$$
If $F$ is differentiable, $x$ follows the Cauchy problem

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

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:
  $$x_{k+1} = x_k - \tau \nabla F(x_{k+1}) \iff 0 = \frac{x_{k+1} - x_k}{\tau} + \nabla F(x_{k+1})$$
  $$\iff x_{k+1} \in \arg\min_{x \in X} \frac{\|x - x_k\|^2}{2\tau} + F(x)$$
  $$\iff x_{k+1} = \text{prox}_{\tau F}(x_k)$$
Wasserstein Distance

Definition (Wasserstein Distance)

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

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

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]:

\[
\mu_{k+1}^\tau \in \argmin_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\tau} W_2^2(\mu, \mu_k^\tau) + F(\mu)
\]
Wasserstein Gradient Flows

Gradient Flow in \((\mathcal{P}_2(\mathbb{R}^d), W_2)\):
Iterated Minimization scheme (JKO Scheme) [Jordan et al., 1998]:
\[
\mu_{k+1}^\tau \in \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\tau} W_2^2(\mu, \mu_k^\tau) + F(\mu)
\]

Examples

- \(F(\mu) = \int \rho(x) \log \rho(x) \, dx + \int V(x) \rho(x) \, dx\) if \(d\mu = \rho \, d\text{Leb}\)
  Solution in the limit \(\tau \to 0\) to the PDE: (Fokker-Planck)
  \[
  \partial_t \rho_t = \text{div}(\rho_t \nabla V) + \Delta \rho_t
  \]
- \(F(\mu) = \frac{1}{2} SW_2^2(\mu, \nu) + \lambda H(\mu)\) [Bonnotte, 2013, Liutkus et al., 2019]
- \(F(\mu) = \frac{1}{2} \text{MMD}^2(\mu, \nu)\) [Arbel et al., 2019]
- \(F(\mu) = \frac{1}{2} \text{KSD}^2(\mu, \nu)\) [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 $F(\mu) = \int V(x)\rho(x)dx + \int \log(\rho(x))\rho(x)dx$, Gradient Flow solution of:

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

Associated SDE (Langevin Equation):

$$dX_t = -\nabla V(X_t)dt + \sqrt{2} \ dW_t$$
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]
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.
Theorem (Brenier’s Theorem)

Let $\mu, \nu \in P_2(\mathbb{R}^d)$, $\mu$ 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:

$$u_{k+1}^\tau \in \arg\min_{u \in \text{cvx}} \frac{1}{2\tau} \int \| \nabla u(x) - x \|^2_2 \rho_k^\tau(x) dx + F((\nabla u) \# \rho_k^\tau)$$

- 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:

$$\theta_{k+1}^\tau \in \arg\min_{\theta \in \{\theta, u_\theta \in \text{ICNN}\}} \frac{1}{2\tau} \int \| \nabla_x u_\theta(x) - x \|^2_2 \rho_k^\tau(x) dx + F((\nabla_x u_\theta) \# \rho_k^\tau)$$

- Backpropagate through gradient

$O(k^2)$ evaluations
Sliced-Wasserstein Distance

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

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

\[
SW_2^2(\mu, \nu) = \int_{S^{d-1}} W_2^2(P_{\#}\mu, P_{\#}\nu) \, \lambda(d\theta)
\]

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 $\left( P_2(\mathbb{R}^d), SW_2 \right)$:

$$\mu_{k+1}^\tau \in \arg\min_{\mu} \frac{1}{2\tau} SW_2^2(\mu, \mu_k^\tau) + F(\mu)$$
Sliced-Wasserstein Gradient Flows

Goal:

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

JKO scheme in $\mathcal{P}_2(\mathbb{R}^d)$, $SW_2$:

$$\mu_{k+1}^{\tau} \in \arg\min_{\mu} \frac{1}{2\tau} SW_2^2(\mu, \mu_k^{\tau}) + F(\mu)$$

- 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 \((P_2(\mathbb{R}^d), SW_2)\):

\[ \mu_{k+1}^\tau \in \arg\min_{\mu} \frac{1}{2\tau} SW_2^2(\mu, \mu_k^\tau) + F(\mu) \]

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

\[
\min_{(\rho_i)_{i \in \Sigma_N}} \frac{SW^2}{2\tau} \left( \sum_{i=1}^N \rho_i \delta_{x_i}, \mu_{k\tau} \right) + F \left( \sum_{i=1}^N \rho_i \delta_{x_i} \right)
\]
Use a discretized grid \((x_i)_{i=1}^N\), model \(\mu = \sum_{i=1}^N \rho_i \delta x_i\) and learn the weights:

\[
\min_{(\rho_i)_{i \in \Sigma_N}} \frac{SW^2_2}{2\tau} \left( \sum_{i=1}^N \rho_i \delta x_i, \mu_{\tau} \right) + F(\sum_{i=1}^N \rho_i \delta x_i)
\]

Learn particles, i.e. \(\mu = \frac{1}{N} \sum_{i=1}^N \delta x_i\) and solve

\[
\min_{(x_i)_{i}} \frac{SW^2_2}{2\tau} \left( \frac{1}{N} \sum_{i=1}^N \delta x_i, \mu_{\tau} \right) + F\left( \frac{1}{N} \sum_{i=1}^N \delta x_i \right)
\]
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:

\[
\min_{(\rho_i)_{i \in \Sigma_N}} \frac{SW_2^2(\sum_{i=1}^N \rho_i \delta_{x_i}, \mu_{\tau_k})}{2\tau} + F\left(\sum_{i=1}^N \rho_i \delta_{x_i}\right)
\]

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

\[
\min_{(x_i)_i} \frac{SW_2^2\left(\frac{1}{N} \sum_{i=1}^N \delta_{x_i}, \mu_{\tau_k}\right)}{2\tau} + F\left(\frac{1}{N} \sum_{i=1}^N \delta_{x_i}\right)
\]

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

\[
\min_{\theta} \frac{SW_2^2((g_\theta^{k+1})\#p_Z, \mu_{\tau_k})}{2\tau} + F((g_\theta^{k+1})\#p_Z)
\]
Fokker-Planck

\[ F(\mu) = \int V \, d\mu + \int \log(\rho(x)) \rho(x) \, dx \]

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

\[ F(\mu) = \int V \, d\mu + \int \log(\rho(x)) \rho(x) \, dx \]

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.
\[ W(\mu) = \frac{1}{2} \iint W(x - y) d\mu(x) d\mu(y) \]

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

**Figure**: Steady state of the aggregation equation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>Steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP ((\tau = 0.05))</td>
<td>20mn</td>
<td>200</td>
</tr>
<tr>
<td>Particles ((\tau = 0.05))</td>
<td>10mn</td>
<td>200</td>
</tr>
<tr>
<td>JKOICNN ((\tau = 0.1))</td>
<td>5h</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table**: Runtime on RTX2080TI.
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


