Mirror gradient flows: Euclidean and Wasserstein

Soumik Pal
University of Washington, Seattle

ICERM, May 10, 2023
Collaborators

Nabarun Deb, Young-Heon Kim, Geoff Schiebinger (UBC)
Motivation
Entropy regularized OT

- Marginals $e^{-f}$, $e^{-g}$ densities. Minimize over coupling $\Pi$.

$$\mathbb{W}_2^2(e^{-f}, e^{-g}) := \inf_{\gamma \in \Pi} \left[ \int \|y - x\|^2 \, d\gamma \right].$$

- Monge solutions are highly degenerate; supported on a graph.
- Entropy as a measure of degeneracy:

$$\text{Ent}(h) := \begin{cases} \int h(x) \log h(x) \, dx, & \text{for density } h, \\ \infty, & \text{otherwise.} \end{cases}$$

- Example: Entropy of $N(0, \sigma^2)$ is $-\log \sigma + \text{constant}$. 


Entropic regularization

Föllmer ’88, Cuturi ’13, Gigli ’19 ... suggested penalizing MK OT with entropy.

$$EOT_\epsilon(e^{-f}, e^{-g}) = \inf_{\gamma \in \Pi} \left[ \int \|y - x\|^2 \, d\gamma + \epsilon \text{Ent}(\gamma) \right].$$

Figure: Image by M. Cuturi
Structure of the solution

- The optimal coupling (Rüschendorf & Thomsen '93) \( \gamma_\epsilon \) must be of the form

\[
\gamma^\epsilon(x, y) = \exp \left( -\frac{1}{2\epsilon} \|y - x\|^2 - \frac{1}{\epsilon} u^\epsilon(x) - \frac{1}{\epsilon} v^\epsilon(y) - f(x) - g(y) \right).
\]

- \( u^\epsilon, v^\epsilon \) - Schrödinger potentials. Unique up to constant.
- Typically not explicit. Determined by marginal constraints

\[
\int \gamma^\epsilon(x, y) dy = e^{-f(x)}, \quad \int \gamma^\epsilon(x, y) dx = e^{-g(y)}.
\]
Sinkhorn/IPFP algorithm

- Initialize arbitrarily. Iteratively fit alternating marginals.
- At every **odd** step the \( X \) marginal is \( e^{-f} \).
- At every **even** step the \( Y \) marginal is \( e^{-g} \).
- Extract the sequence of \( X \)-marginals from **even** steps.

\[
(\rho_k^\epsilon, \; k = 1, 2, 3, \ldots).
\]

- How fast does \( \rho_k^\epsilon \) converge to \( e^{-f} \)? Cf. Marcel’s talk yesterday for \( \epsilon > 0 \) rates.
Our approach

- Embed the sequence in time steps $\epsilon$.
- Find the limiting absolutely continuous curve $(\rho_t, \ t \geq 0)$,

$$\rho_t = \lim_{\epsilon \to 0} \rho_t^\epsilon.$$

- Describe this curve as a “gradient flow”.
- Use gradient flow techniques to determine exponential rates of convergence under assumptions.
Euclidean mirror gradient flows
Diffeomorphisms by convex gradients

- \( u : \mathbb{R}^d \rightarrow \mathbb{R} \) differentiable strictly convex.
- \( x \leftrightarrow x^u = \nabla u(x) \) creates **mirror coordinates** by duality.
- Two notions of gradients. \( F : \mathbb{R}^d \rightarrow \mathbb{R} \).

\[
\nabla_x F(x), \quad \nabla_{x^u} F(x) := (\nabla^2 u(x))^{-1} \nabla_x F(x).
\]

- Usual case \( u(x) = \frac{1}{2} \|x\|^2 \).
Gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize $x_0$.
- Flow of the mirror coordinate.
  \[ \dot{x}_t^u = -\nabla_x F(x_t). \]
- Flow of the primal coordinate.
  \[ \dot{x}_t = -\nabla_{x^u} F(x_t). \]
- Gradient flow in a **Hessian Riemannian manifold** with a metric tensor given by the Hessian
  \[ (\nabla^2 u(x))^{-1} = \nabla^2 u^*(x^u). \]
- Widely used in optimization and ML.
Examples

- $d = 1$, $F(x) = x^2/2$, $x_0 = 1$.
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.
  \[ \dot{x}_t = -x_t, \quad x_t = e^{-t}. \]

- $u(x) = x^4$. Mirror flow converges in finite time.
  \[ \dot{x}_t = -\frac{1}{12x_t}, \quad x_t = \sqrt{(1 - t/6)^+}. \]

- $u(x) = 1/x$. Mirror flow converges polynomially.
  \[ \dot{x}_t = -\frac{1}{2}x_t^4, \quad x_t = (1 + 3t/2)^{-1/3}. \]
Polyak-Löjasiewicz condition

When can we guarantee exponential convergence?

Mirror Polyak-Löjasiewicz condition:

$$2\alpha (f(x) - f(x_{\text{min}})) \leq \|\nabla f(x)\|_{x^u}^2,$$

where

$$\|v\|_{x^u}^2 = v^T (\nabla^2 u(x))^{-1} v.$$

Then exponential convergence at rate $\alpha$. 
Wasserstein mirror gradient flows
Wasserstein gradient flow recap

- (Otto ’98) Wasserstein space $\mathcal{W}_2(\mathbb{R}^d)$ is a formal Riemannian manifold.
- Tangent space at $\rho$

$$\left\{ \nabla \phi, \phi \in C^\infty_c \right\}^{L^2(\rho)}.$$

- $F : \mathcal{W}_2 \to \mathbb{R}$. Wasserstein gradient is a Riemannian gradient.

$$\nabla_{\mathcal{W}} F(\rho) = \nabla \left( \frac{\delta F}{\delta \rho} \right).$$

- Wasserstein gradient flow solves continuity equation.

$$\dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla_{\mathcal{W}} F(\rho_t).$$
Mirror, mirror

- Special choice of mirror function on $\mathbb{W}_2$. Fix density $e^{-g}$.

$$U(\rho) := \frac{1}{2} \mathbb{W}_2^2 (\rho, e^{-g}) .$$

- (Generalized) Geodesically convex. Generates mirror coordinate:

$$\rho \iff x - \nabla u(\rho) = \nabla_{\mathbb{W}} U(\rho),$$

where $\nabla u(\rho)$ is the Brenier map transporting $\rho$ to $e^{-g}$.
Mirror flow PDE and continuity equations

- Mirror gradient flow PDE for the potential. Initialize at $u_0$.

\[ \nabla \dot{u}_t = -\nabla_{\mathcal{W}} F(\rho_t), \quad (\nabla u_t)_{\#\rho_t} = e^{-g}. \]

- Mirror gradient flow continuity equation. Initialize at $\rho_0$.

\[ \dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla_{xu_t} \frac{\delta F}{\delta \rho}(\rho_t) = - (\nabla^2 u_t)^{-1} \nabla_{\mathcal{W}} F(\rho_t). \]

where $\nabla u_t$ is the Brenier map from $\rho_t$ to $e^{-g}$.

- Unclear if solutions exist.
Example 1

- Entropy. $F(\rho) = \int \rho(x) \log \rho(x) dx$. Take $d = 1$.
- Take $\rho_0 = e^{-g} = N(0, 1)$.
- PDE for the Brenier potential

$$\nabla \dot{u}_t(x) = \log \rho_t(x) + 1.$$ 

- Solution $\rho_t = N(0, (1 + t)^2)$.
- Compare with the heat flow = Wasserstein grad flow. $\mu_t = N(0, 1 + t)$.
- Faster convergence for mirror flow.
Example 2

- Kullback-Leibler. Fix density $e^{-f}$.

\[ F(\rho) = \text{KL}(\rho \mid e^{-f}). \]

Geodesically convex if $f$ is convex.

- Mirror

\[ U(\rho) = \frac{1}{2} \mathcal{W}_2^2(\rho, e^{-g}), \quad \rho \leftrightarrow \nabla u. \]

- What is the mirror gradient flow of $F$?

- A PDE for the Brenier potentials and a cont.-eq. for the measures.
Parabolic Monge-Ampère

- Initialize convex $u_0$.

\[
\nabla \dot{u}_t = \nabla f(x) + \nabla \log \rho_t(x), \quad \text{where } (\nabla u_t)_{\rho_t} = e^{-g}.
\]

- Simplify

\[
\dot{u}_t(x) = f(x) - g(\nabla u_t(x)) + \log \det \nabla^2 u_t(x).
\]

- Parabolic dynamics added to the Monge-Ampère PDE. (See Kim-Streets-Warren ’12, Kitagawa ’12). Nice solutions exist under assumptions.
The Sinkhorn PDE

The continuity equation is another PDE.

\[ \dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla_{x_{ut}} (f + \log \rho_t). \]

- Gives an AC curve on the Wasserstein space. Converges to \( e^{-f} \) as \( t \to \infty \).
- Curious relation with linearized OT.
The limiting analysis of Sinkhorn iterations
Recap of Sinkhorn

- Initialize arbitrarily. Iteratively fit alternating marginals.
- At every **odd** step the $X$ marginal is $e^{-f}$.
- At every **even** step the $Y$ marginal is $e^{-g}$.
- Extract the sequence of $X$-marginals from **even** steps.

$$(\rho^\epsilon_k, \ k = 1, 2, 3, \ldots).$$

- Problem: Find the limiting absolutely continuous curve $(\rho_t, \ t \geq 0)$,

$$\rho_t = \lim_{\epsilon \to 0} \rho^\epsilon_{t/\epsilon}.$$
The limit is a mirror gradient flow

- **Theorem (DKPS ’23)** Under regularity assumptions on the parabolic MA,

\[ \dot{u}_t(x) = f(x) - g(\nabla u_t(x)) + \log \det \nabla^2 u_t(x). \]

the limiting curve of the $X$ marginals is a solution of the Sinkhorn PDE.

\[ \dot{\rho}_t + \nabla \cdot (\nu_t \rho_t) = 0, \quad \nu_t = -\nabla_{x u_t} (f + \log \rho_t). \]

- In particular, it is a mirror gradient flow of $F(\rho) = \text{KL}(\rho \mid e^{-f})$ with the mirror given by $U(\rho) = \frac{1}{2} \mathbb{W}_2^2(\rho, e^{-g})$.

- A symmetric statement holds for the sequence of $Y$ marginals.
Exponential rate of convergence

Theorem (DKPS ’23) Suppose $e^{-f}$ satisfies logarithmic Sobolev inequality. Also suppose that the solution of the parabolic MA satisfies

$$\inf_t \inf_x (\nabla^2 u_t(x))^{-1} \geq \lambda I,$$

then exponential convergence for the Sinkhorn PDE.

- There are conditions known where our assumptions are satisfied. See, e.g., Berman ’20.
- The proof is a standard gradient flow argument.
Our work is heavily influenced by two prior works.

- **Berman ’20.** Shows that the sequence of potentials from Sinkhorn iterations converge to the solution of the PMA.
- Our proofs require control of higher order errors than Berman’s.
- **Léger ’20.** Shows that discrete Sinkhorn potentials with positive $\epsilon > 0$ is a mirror descent of KL.
- But one cannot invert the relationship to get any gradient flow description of the evolution of the measures.
A McKean-Vlasov interpretation

Sinkhorn PDE is the marginal law of the following diffusion.

\[
dX_t = \left( -\frac{\partial f}{\partial X_t} (X_t) - \frac{\partial g}{\partial X_t} (X_t^u) + \frac{\partial h}{\partial X_t} (X_t) \right) dt + \sqrt{2 \frac{\partial X_t}{\partial X_t^u}} dB_t, \tag{1}
\]

where

- \( X_t \) has density \( \rho_t = e^{-h_t} \).
- \( (\nabla u_t) \# \rho_t = e^{-g} \).
- Diffusion matrix at time \( t \) is

\[
2 \frac{\partial x}{\partial X_t^u} = 2 \left( \nabla^2 u_t(x) \right)^{-1}.
\]

For \( f = g \), becomes \textbf{mirror Langevin diffusion} (Ahn-Chewi ’21). Generalizes Langevin.
Several open questions

- Replace KL by another divergence. Does this have any algorithmic potential?
- Other mirror functions than the squared Wasserstein distance.
- One can formally write the resulting Hessian geometry. But there are singularities.

\[ \langle v_1, v_2 \rangle_\rho = \int v_1^T(x) \left( \nabla^2 u_\rho(x) \right)^{-1} v_2(x) \rho(dx). \]

- Build a JKO like scheme for this Hessian geometry. See Rankin-Wong ’23 for some related constructions of the Bregman-Wasserstein divergences.
- Do particle systems that follow Euclidean mirror gradient flows converge to Wasserstein mirror gradient flows?

Thank you