Wasserstein Mirror Gradient Flows as the Limit of the Sinkhorn Algorithm

Nabarun Deb
University of Chicago

Summer School on Optimal Transport,
Stochastic Analysis and Applications to Machine Learning

Korea Advanced Institute of Science and Technology (KAIST)

Joint work with Young-heon Kim, Soumik Pal, Geoffrey Schiebinger

Entropy regularized OT

- Marginals e^{-f}, e^{-g} densities. Minimize over coupling Π, i.e., all $\gamma \in \Pi$ the first and second marginals of γ are e^{-f} and e^{-g} respectively,

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

- Monge solutions are highly degenerate; supported on a graph, and hard to compute.

Entropy regularized OT

- Marginals e^{-f}, e^{-g} densities. Minimize over coupling Π, i.e., all $\gamma \in \Pi$ the first and second marginals of γ are e^{-f} and e^{-g} respectively,

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

- Monge solutions are highly degenerate; supported on a graph, and hard to compute.
- Entropy as a measure of degeneracy:

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

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

Entropic regularization

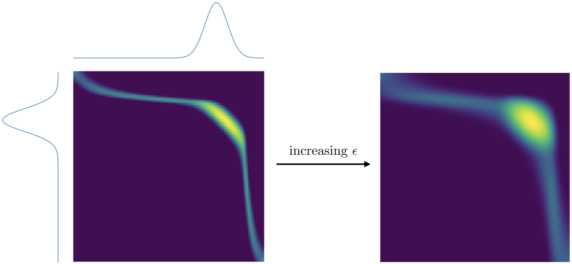


Figure: Image by M. Cuturi

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

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

Structure of the solution

- The **optimal coupling** (Rüschendorf & Thomsen '93) γ^ϵ 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^ϵ, v^ϵ - 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 a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.

Sinkhorn/IPFP algorithm

- Initialize a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.
- At every odd step, say γ_{2k+1}^ϵ, the X marginal is e^{-f}.

Sinkhorn/IPFP algorithm

- Initialize a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.
- At every odd step, say γ_{2k+1}^ϵ, the X marginal is e^{-f}.
- At every even step, say γ_{2k}^ϵ the Y marginal is e^{-g}. So, e.g.,

$$\gamma_1^\epsilon(x, y) = e^{-f(x)} \frac{\gamma_0^\epsilon(x, y)}{\int_y \gamma_0^\epsilon(x, y)\, dy},$$

Sinkhorn/IPFP algorithm

- Initialize a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.
- At every odd step, say γ_{2k+1}^ϵ, the X marginal is e^{-f}.
- At every even step, say γ_{2k}^ϵ the Y marginal is e^{-g}. So, e.g.,

$$\gamma_1^\epsilon(x,y) = e^{-f(x)} \frac{\gamma_0^\epsilon(x,y)}{\int_y \gamma_0^\epsilon(x,y)\,dy}, \quad \gamma_2^\epsilon(x,y) = e^{-g(y)} \frac{\gamma_1^\epsilon(x,y)}{\int_x \gamma_1^\epsilon(x,y)\,dx}$$

Sinkhorn/IPFP algorithm

- Initialize a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.
- At every odd step, say γ_{2k+1}^ϵ, the X marginal is e^{-f}.
- At every even step, say γ_{2k}^ϵ the Y marginal is e^{-g}. So, e.g.,

$$\gamma_1^\epsilon(x,y) = e^{-f(x)} \frac{\gamma_0^\epsilon(x,y)}{\int_y \gamma_0^\epsilon(x,y)\, dy}, \quad \gamma_2^\epsilon(x,y) = e^{-g(y)} \frac{\gamma_1^\epsilon(x,y)}{\int_x \gamma_1^\epsilon(x,y)\, dx}$$

- Extract the sequence of X-marginals from even steps.

$$\left(\rho_k^\epsilon,\ k = 1, 2, 3, \ldots\right).$$

In fact, ρ_k^ϵ characterizes the corresponding γ_k^ϵ via a variational problem.

Sinkhorn/IPFP algorithm

- Initialize a distribution γ_0^ϵ on $\mathbb{R}^d \times \mathbb{R}^d$ "appropriately". Iteratively fit alternating marginals.
- At every odd step, say γ_{2k+1}^ϵ, the X marginal is e^{-f}.
- At every even step, say γ_{2k}^ϵ the Y marginal is e^{-g}. So, e.g.,

$$\gamma_1^\epsilon(x,y) = e^{-f(x)} \frac{\gamma_0^\epsilon(x,y)}{\int_y \gamma_0^\epsilon(x,y)\, dy}, \quad \gamma_2^\epsilon(x,y) = e^{-g(y)} \frac{\gamma_1^\epsilon(x,y)}{\int_x \gamma_1^\epsilon(x,y)\, dx}$$

- Extract the sequence of X-marginals from even steps.

$$\left(\rho_k^\epsilon,\ k = 1, 2, 3, \ldots\right).$$

In fact, ρ_k^ϵ characterizes the corresponding γ_k^ϵ via a variational problem.

- How fast does ρ_k^ϵ converge to e^{-f} when $\varepsilon \to 0$ appropriately scaled with $k \to \infty$? For the case $\varepsilon > 0$, see Ghosal and Nutz, 2022, Conforti et al., 2023, ...

The "Scaling" limit

- By Berman '20 and Léger '20, it follows:

$$(H_\epsilon^*)'(\rho_{k+1}^\epsilon) - (H_\epsilon^*)'(\rho_k^\epsilon) = -\mathrm{KL}'(\rho_k | e^{-f}).$$

Here $H_\epsilon(\cdot)$ is itself characterized by a variational problem, H_ϵ^* is the dual, and $'$ is used for first variation.

The "Scaling" limit

- By Berman '20 and Léger '20, it follows:

$$(H_\epsilon^*)'(\rho_{k+1}^\epsilon) - (H_\epsilon^*)'(\rho_k^\epsilon) = -\mathrm{KL}'(\rho_k|e^{-f}).$$

Here $H_\epsilon(\cdot)$ is itself characterized by a variational problem, H_ϵ^* is the dual, and $'$ is used for first variation. No missing ϵ on RHS

- This reminds us of usual gradient descent:

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

The "Scaling" limit

- By Berman '20 and Léger '20, it follows:

$$(H_\epsilon^*)'(\rho_{k+1}^\epsilon) - (H_\epsilon^*)'(\rho_k^\epsilon) = -\mathrm{KL}'(\rho_k | e^{-f}).$$

Here $H_\epsilon(\cdot)$ is itself characterized by a variational problem, H_ϵ^* is the dual, and $'$ is used for first variation. No missing ϵ on RHS

- This reminds us of usual gradient descent:

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

(Cauchy problem) By Santambrogio '16, with $k = t/\epsilon$ and $\epsilon \to 0$, we have $x_{t/\epsilon}^\epsilon \to \tilde{x}_t$ where

$$\frac{d}{dt}\tilde{x}_t = -\nabla F(\tilde{x}_t).$$

The "Scaling" limit

- By Berman '20 and Léger '20, it follows:

$$(H_\epsilon^*)'(\rho_{k+1}^\epsilon) - (H_\epsilon^*)'(\rho_k^\epsilon) = -\mathrm{KL}'(\rho_k|e^{-f}).$$

Here $H_\epsilon(\cdot)$ is itself characterized by a variational problem, H_ϵ^* is the dual, and $'$ is used for first variation. No missing ϵ on RHS

- This reminds us of usual gradient descent:

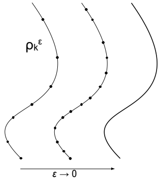
$$x_{k+1}^\epsilon - x_k^\epsilon = -\epsilon \nabla F(x_k^\epsilon).$$

(Cauchy problem) By Santambrogio '16, with $k = t/\epsilon$ and $\epsilon \to 0$, we have $x_{t/\epsilon}^\epsilon \to \tilde{x}_t$ where

$$\frac{d}{dt}\tilde{x}_t = -\nabla F(\tilde{x}_t).$$

$\tilde{x}_t \to \tilde{x}_\infty$ (optimizer of F) usually exponentially fast if F is λ-convex. Helps to speed up convergence, understand regularization, etc.

Our approach



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

$$\rho_t = \lim_{\epsilon \to 0} \rho^{\epsilon}_{t/\epsilon}.$$

- **Describe this curve as a "mirror gradient flow".**
- Use gradient flow techniques to determine exponential rates of convergence under assumptions.
- Come up with a Mckean-Vlasov diffusion whose marginals follow the same mirror gradient flow.

Euclidean mirror gradient flows

Diffeomorphisms by convex gradients

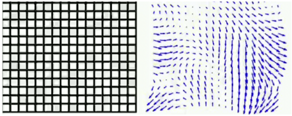


Figure: Image of a diffeomorphism by G. Peyré

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

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

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

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize x_0.
- Flow of the mirror coordinate.

$$\nabla u(x_{k+1}) - \nabla u(x_k) = -\epsilon \nabla F(x_k) \qquad \dot{x}_t^u = \frac{d}{dt}\nabla u(x_t) = -\nabla_x F(x_t)$$

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize x_0.
- Flow of the mirror coordinate.

$$\nabla u(x_{k+1}) - \nabla u(x_k) = -\epsilon \nabla F(x_k) \qquad \dot{x}_t^u = \frac{d}{dt}\nabla u(x_t) = -\nabla_x F(x_t)$$

- Flow of the primal/canonical coordinate.

$$x_{k+1} - x_k = -\epsilon \nabla_{x^u} F(x_k) \qquad \dot{x}_t = -\nabla_{x^u} F(x_t) = -(\nabla^2 u(x_t))^{-1} \nabla_x F(x_t)$$

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize x_0.
- Flow of the mirror coordinate.

$$\nabla u(x_{k+1}) - \nabla u(x_k) = -\epsilon \nabla F(x_k) \qquad \dot{x}_t^u = \frac{d}{dt}\nabla u(x_t) = -\nabla_x F(x_t)$$

- Flow of the primal/canonical coordinate.

$$x_{k+1} - x_k = -\epsilon \nabla_{x^u} F(x_k) \qquad \dot{x}_t = -\nabla_{x^u} F(x_t) = -(\nabla^2 u(x_t))^{-1} \nabla_x F(x_t)$$

- Gradient flow in a Hessian Riemannian manifold with a metric tensor given by the Hessian

$$\left(\nabla^2 u(x)\right)^{-1} = \nabla^2 u^*(x^u).$$

- What to expect? Interpret Sinkhorn as a mirror descent on the space of probability measures. What are F and u?

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize x_0.
- Flow of the mirror coordinate.

$$\nabla u(x_{k+1}) - \nabla u(x_k) = -\epsilon \nabla F(x_k) \qquad \dot{x}_t^u = \frac{d}{dt} \nabla u(x_t) = -\nabla_x F(x_t)$$

- Flow of the primal/canonical coordinate.

$$x_{k+1} - x_k = -\epsilon \nabla_{x^u} F(x_k) \qquad \dot{x}_t = -\nabla_{x^u} F(x_t) = -(\nabla^2 u(x_t))^{-1} \nabla_x F(x_t)$$

- Gradient flow in a Hessian Riemannian manifold with a metric tensor given by the Hessian

$$\left(\nabla^2 u(x)\right)^{-1} = \nabla^2 u^*(x^u).$$

- What to expect? Interpret Sinkhorn as a mirror descent on the space of probability measures. What are F and u?

Examples

- $d = 1$, $F(x) = x^2/2$, $x_0 = 1$.
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.

$$\dot{x}_t = -x_t, \quad x_t = e^{-t}.$$

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.

$$\dot{x}_t = -\frac{1}{12x_t}, \quad x_t = \sqrt{(1 - t/6)^+}.$$

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.

$$\dot{x}_t = -\frac{1}{2}x_t^4, \quad x_t = (1 + 3t/2)^{-1/3}.$$

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.

$$\dot{x}_t = -\frac{1}{2}x_t^4, \quad x_t = (1 + 3t/2)^{-1/3}.$$

- For analogy, we say a mirror gradient flow is characterized by an objective function F and a mirror map u.

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon\to 0} \rho_{t/\epsilon}^\epsilon = \rho_t$ where ρ_t is the Wasserstein mirror flow with

- Objective function: $F(\rho) = \mathrm{KL}(\rho | e^{-f})$
- Mirror map: $U(\rho) = \frac{1}{2} W_2^2(\rho, e^{-g})$

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = \rho_t$ where ρ_t is the Wasserstein mirror flow with

- Objective function: $F(\rho) = \mathrm{KL}(\rho | e^{-f})$
- Mirror map: $U(\rho) = \frac{1}{2} W_2^2(\rho, e^{-g})$

- How do we describe Wasserstein mirror flows?

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = \rho_t$ where ρ_t is the Wasserstein mirror flow with

- Objective function: $F(\rho) = \mathrm{KL}(\rho | e^{-f})$
- Mirror map: $U(\rho) = \frac{1}{2} W_2^2(\rho, e^{-g})$

- How do we describe Wasserstein mirror flows?
  Parabolic PDE+continuity equation

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = \rho_t$ where ρ_t is the Wasserstein mirror flow with

- Objective function: $F(\rho) = \mathrm{KL}(\rho | e^{-f})$
- Mirror map: $U(\rho) = \frac{1}{2} W_2^2(\rho, e^{-g})$

- How do we describe Wasserstein mirror flows?
  Parabolic PDE+continuity equation
- Do we still (potentially??) need to make sense of the Hessian of $U(\cdot)$?

The limit of Sinkhorn is a mirror gradient flow

- Recall that we wanted to study the limit of ρ_k^ϵ (X marginals from Sinkhorn) for $k = t/\epsilon$, i.e.,

$$\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = ??$$

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon \to 0} \rho_{t/\epsilon}^\epsilon = \rho_t$ where ρ_t is the Wasserstein mirror flow with

- Objective function: $F(\rho) = \mathrm{KL}(\rho | e^{-f})$
- Mirror map: $U(\rho) = \frac{1}{2} W_2^2(\rho, e^{-g})$

- How do we describe Wasserstein mirror flows?
  Parabolic PDE+continuity equation
- Do we still (potentially??) need to make sense of the Hessian of $U(\cdot)$?
  No!

Wasserstein mirror gradient flows

Wasserstein gradient flow recap

- (Otto '98) Wasserstein space $\mathbb{W}_2(\mathbb{R}^d)$ is a formal Riemannian manifold.
- Tangent space at ρ
$$\overline{\{\nabla\phi,\ \phi \in C_c^\infty\}}^{\mathbf{L}^2(\rho)}.$$
- $F : \mathbb{W}_2 \to \mathbb{R}$. Wasserstein gradient is a Riemannian gradient.
$$\nabla_{\mathbb{W}} F(\rho) = \nabla \left(\frac{\delta F}{\delta \rho} \right).$$

Here $\frac{\delta F}{\delta \rho}$ denotes the first variation, i.e., $\frac{d}{dt} F(\rho + t\nu)\Big|_{t=0} = \int \frac{\delta F}{\delta \rho}\, d\nu$.

Wasserstein gradient flow recap

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

$$\overline{\{\nabla\phi,\ \phi \in C_c^\infty\}}^{\mathbf{L}^2(\rho)}.$$

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

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

  Here $\frac{\delta F}{\delta \rho}$ denotes the first variation, i.e., $\frac{d}{dt} F(\rho + t\nu)\Big|_{t=0} = \int \frac{\delta F}{\delta \rho}\, d\nu$.

- Wasserstein gradient flow solves continuity equation.

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

  v_t often called velocity. Belongs in the tangent space.

Wasserstein gradient flow recap

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

$$\overline{\{\nabla\phi,\ \phi \in C_c^\infty\}}^{\mathbf{L}^2(\rho)}.$$

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

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

  Here $\frac{\delta F}{\delta \rho}$ denotes the first variation, i.e., $\frac{d}{dt}F(\rho + t\nu)\Big|_{t=0} = \int \frac{\delta F}{\delta\rho}\, d\nu$.

- Wasserstein gradient flow solves continuity equation.

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

  v_t often called velocity. Belongs in the tangent space.

- A gradient descent analogy: $\frac{d}{dt}x_t = -\nabla F(x_t)$. Effectively usual gradient replaced with $\nabla_{\mathbb{W}}$ to get v_t.

Mirror, mirror on the ...

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

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

- (Generalized) Geodesically convex. Generates mirror coordinate:

$$\rho \iff \underbrace{x - \nabla u_\rho(x)}_{\text{Kantorovich potential}} = \nabla_{\mathbb{W}} U(\rho),$$

where $\nabla u_\rho(\cdot)$ is the Brenier map transporting ρ to e^{-g}, i.e., u_ρ is convex and $(\nabla u_\rho)\#\rho = e^{-g}$ or, if $X \sim \rho$, then $\nabla u_\rho(X) \sim e^{-g}$.

Mirror, mirror on the ...

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

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

- (Generalized) Geodesically convex. Generates mirror coordinate:

$$\rho \Longleftrightarrow \underbrace{x - \nabla u_\rho(x)}_{\text{Kantorovich potential}} = \nabla_{\mathbb{W}} U(\rho),$$

  where $\nabla u_\rho(\cdot)$ is the Brenier map transporting ρ to e^{-g}, i.e., u_ρ is convex and $(\nabla u_\rho)\#\rho = e^{-g}$ or, if $X \sim \rho$, then $\nabla u_\rho(X) \sim e^{-g}$.

- Recall Euclidean mirror descent: Given a convex mirror map u, the mirror coordinates are given by $\nabla u(x)$.

Mirror, mirror on the ...

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

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

- (Generalized) Geodesically convex. Generates mirror coordinate:

$$\rho \iff \underbrace{x - \nabla u_\rho(x)}_{\text{Kantorovich potential}} = \nabla_{\mathbb{W}} U(\rho),$$

  where $\nabla u_\rho(\cdot)$ is the Brenier map transporting ρ to e^{-g}, i.e., u_ρ is convex and $(\nabla u_\rho)\#\rho = e^{-g}$ or, if $X \sim \rho$, then $\nabla u_\rho(X) \sim e^{-g}$.
- Recall Euclidean mirror descent: Given a convex mirror map u, the mirror coordinates are given by $\nabla u(x)$.
- Natural analog would be to describe two equivalent fows — one for probability measures $(\rho_t)_{t\geq 0}$ (primal coordinate) and another for Brenier porentials $(\nabla u_{\rho_t})_{t\geq 0} \equiv (\nabla u_t)_{t\geq 0}$ (mirror coordinate)

Mirror flow PDE and continuity equations

- Mirror gradient flow PDE for the potential (mirror coordinate). Initialize at u_0.

$$\frac{\partial}{\partial t}\nabla_{\mathbb{W}} U(\rho_t) = -\nabla_{\mathbb{W}} F(\rho_t)$$
$$\Longrightarrow \nabla \dot{u}_t = \nabla_{\mathbb{W}} F(\rho_t), \quad \nabla u_t \# \rho_t = e^{-g}.$$

Euclidean case: $\frac{\partial}{\partial t}\nabla u(x_t) = -\nabla F(x_t)$.

Mirror flow PDE and continuity equations

- Mirror gradient flow PDE for the potential (mirror coordinate). Initialize at u_0.

$$\frac{\partial}{\partial t}\nabla_{\mathbb{W}} U(\rho_t) = -\nabla_{\mathbb{W}} F(\rho_t)$$
$$\implies \nabla \dot{u}_t = \nabla_{\mathbb{W}} F(\rho_t), \quad \nabla u_t \# \rho_t = e^{-g}.$$

Euclidean case: $\frac{\partial}{\partial t}\nabla u(x_t) = -\nabla F(x_t)$.

- Mirror gradient flow continuity equation (primal coordinates). Initialize at ρ_0.

$$\dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \; v_t = -\left(\nabla^2 u_t\right)^{-1} \nabla_{\mathbb{W}} F(\rho_t) = -\nabla_{x^{u_t}} \frac{\delta F}{\delta \rho}(\rho_t).$$

where ∇u_t is the Brenier map from ρ_t to e^{-g}, $\nabla u_t \# \rho_t = e^{-g}$.

Mirror flow PDE and continuity equations

- Mirror gradient flow PDE for the potential (mirror coordinate). Initialize at u_0.

$$\frac{\partial}{\partial t}\nabla_{\mathbb{W}} U(\rho_t) = -\nabla_{\mathbb{W}} F(\rho_t)$$
$$\implies \nabla \dot{u}_t = \nabla_{\mathbb{W}} F(\rho_t), \quad \nabla u_t \# \rho_t = e^{-g}.$$

Euclidean case: $\frac{\partial}{\partial t}\nabla u(x_t) = -\nabla F(x_t)$.

- Mirror gradient flow continuity equation (primal coordinates). Initialize at ρ_0.

$$\dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \; v_t = -\left(\nabla^2 u_t\right)^{-1} \nabla_{\mathbb{W}} F(\rho_t) = -\nabla_{x^{u_t}} \frac{\delta F}{\delta \rho}(\rho_t).$$

where ∇u_t is the Brenier map from ρ_t to e^{-g}, $\nabla u_t \# \rho_t = e^{-g}$.
Euclidean case: $\dot{x}_t = -(\nabla^2 u(x_t))^{-1} \nabla_x F(x_t)$

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

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

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 (Sinkhorn flow)

- The mirror flow of $F(\rho) = \mathrm{KL}(\rho | e^{-f})$ can be faster than usual Fokker-Planck.
- Take $\rho_0 = e^{-g} = N(0, \eta^2)$, for $\eta > 0$.
- Take $e^{-f} = N(0, 1)$.
- Both Fokker-Planck and Wassertein mirror flow admit Gaussian solutions of the form

$$N(0, \sigma_{F,t}^2), \quad N(0, \sigma_{M,t}^2).$$

- If $\eta < 1$, then

$$\lim_{t \to \infty} \frac{|1 - \sigma_{F,t}^2|}{|1 - \sigma_{M,t}^2|} = \infty,$$

exponentially.

Example 3 (Sinkhorn flow)

- The mirror flow of $F(\rho) = \mathrm{KL}(\rho|e^{-f})$ can be faster than usual Fokker-Planck with multivariate Gaussians.
- Take $\rho_0 = N(0, I_d)$ and $e^{-g} = N(0, \Theta)$.
- Take $e^{-f} = N(0, \Sigma)$. Assume Σ and Θ commute, both are invertible.

Example 3 (Sinkhorn flow)

- The mirror flow of $F(\rho) = \mathrm{KL}(\rho | e^{-f})$ can be faster than usual Fokker-Planck with multivariate Gaussians.
- Take $\rho_0 = N(0, I_d)$ and $e^{-g} = N(0, \Theta)$.
- Take $e^{-f} = N(0, \Sigma)$. Assume Σ and Θ commute, both are invertible.
- Both Fokker-Planck and Wassertein mirror flow admit Gaussian solutions of the form

$$N(0, \Sigma_{F,t}), \quad N(0, \Sigma_{M,t}).$$

- If $\|\Sigma^{-1}\Theta\|_{\mathrm{op}} < 1$, then

$$\lim_{t\to\infty} \frac{\|\Sigma - \Sigma_{F,t}\|_{\mathrm{op}}}{\|\Sigma - \Sigma_{M,t}\|_{\mathrm{op}}} = \infty,$$

exponentially.

Interpreting mirror flow velocity

- Consider Wasserstein gradient flow of F, i.e.,

$$\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla \left(\frac{\delta F}{\delta \rho} \right)_{\rho = \rho_t}.$$

If T_{t+h} is the transport map from ρ_t to ρ_{t+h}, then

$$T_{t+h} = \mathrm{Id} + h\mathbf{v_t} + o(|h|).$$

Interpreting mirror flow velocity

- Consider Wasserstein gradient flow of F, i.e.,

$$\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla \left(\frac{\delta F}{\delta \rho} \right)_{\rho = \rho_t}.$$

If T_{t+h} is the transport map from ρ_t to ρ_{t+h}, then

$$T_{t+h} = \mathrm{Id} + h\mathbf{v_t} + o(|h|).$$

- Consider Wasserstein mirror flow of F, i.e.,

$$\partial_t \rho_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla_{x^{u_t}} \left(\frac{\delta F}{\delta \rho} \right).$$

If T_t is the transport map from e^{-g} to ρ_t, then

$$T_{t+h} = T_t + h\mathbf{v_t}(\mathbf{T_t}) + o(|h|).$$

Recall Linearized OT

Given probability measures μ_1, μ_2, ν, let $T_1 \# \nu = \mu_1$ and $T_2 \# \nu = \mu_2$ (T_1, T_2 are optimal transport maps).

LOT defn.

$$\mathrm{LOT}_\nu(\mu_1, \mu_2) = \| T_1 - T_2 \|_{L^2(\nu)}.$$

Recall Linearized OT

Given probability measures μ_1, μ_2, ν, let $T_1 \# \nu = \mu_1$ and $T_2 \# \nu = \mu_2$ (T_1, T_2 are optimal transport maps).

LOT defn.

$$\mathrm{LOT}_\nu(\mu_1, \mu_2) = \| T_1 - T_2 \|_{L^2(\nu)}.$$

For Wasserstein mirror flows ...

LOT metric derivative

$$\lim_{h \to 0^+} \frac{1}{h} \mathrm{LOT}_{e^{-g}}(\rho_{t+h}, \rho_t) = \| v_t \|_{L^2(\rho_t)} .$$

Recall Linearized OT

Given probability measures μ_1, μ_2, ν, let $T_1 \# \nu = \mu_1$ and $T_2 \# \nu = \mu_2$ (T_1, T_2 are optimal transport maps).

LOT defn.

$$\mathrm{LOT}_\nu(\mu_1, \mu_2) = \| T_1 - T_2 \|_{L^2(\nu)}.$$

For Wasserstein mirror flows ...

LOT metric derivative

$$\lim_{h \to 0^+} \frac{1}{h} \mathrm{LOT}_{e^{-g}}(\rho_{t+h}, \rho_t) = \| v_t \|_{L^2(\rho_t)} \, .$$

For usual gradient flow, the above holds with usual Wasserstein distance.

Recap of Sinkhorn

- Initialize "appropriately". 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.

$$\left(\rho_k^\epsilon,\ k = 1, 2, 3, \ldots\right).$$

Recap of Sinkhorn

- Initialize "appropriately". 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.

$$\left(\rho_k^{\epsilon},\ k = 1, 2, 3, \ldots\right).$$

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

$$\rho_t = \lim_{\epsilon \to 0} \rho^{\epsilon}_{t/\epsilon}.$$

- **Describe this curve as a "Wasserstein mirror gradient flow".**
- Use gradient flow techniques to determine exponential rates of convergence under assumptions.
- Come up with a Mckean-Vlasov diffusion whose marginals follow the same mirror gradient flow.

The limit of Sinkhorn is a mirror gradient flow

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

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

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

  $$\dot{\rho}_t + \nabla \cdot (v_t \rho_t) = 0, \quad v_t = -\nabla_{X^{u_t}} \left(f + \log \rho_t\right).$$

  Moreover,

  $$\mathbb{W}_2^2(\rho^{\epsilon}_{t/\epsilon}, \rho_t) = O(\varepsilon).$$

The limit of Sinkhorn is a mirror gradient flow

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

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

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

$$\dot{\rho}_t + \nabla\cdot\left(v_t\rho_t\right) = 0, \quad v_t = -\nabla_{x^{u_t}}\left(f + \log\rho_t\right).$$

  Moreover,

$$\mathbb{W}_2^2(\rho^{\epsilon}_{t/\epsilon}, \rho_t) = O(\varepsilon).$$

- In particular, it is a mirror gradient flow of $F(\rho) = \mathrm{KL}(\rho \mid e^{-f})$ 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.

The limit of Sinkhorn is a mirror gradient flow

- Theorem (DKPS '23) Under regularity assumptions on the parabolic MA,
$$\dot{u}_t(x) = f(x) - g\left(\nabla u_t(x)\right) + \log\det\nabla^2 u_t(x).$$
the limiting curve of the X marginals is a solution of the Sinkhorn PDE.
$$\dot{\rho}_t + \nabla\cdot(v_t\rho_t) = 0, \quad v_t = -\nabla_{x^{u_t}}\left(f + \log\rho_t\right).$$
Moreover,
$$\mathbb{W}_2^2(\rho^{\epsilon}_{t/\epsilon}, \rho_t) = O(\varepsilon).$$

- In particular, it is a mirror gradient flow of $F(\rho) = \mathrm{KL}(\rho \mid e^{-f})$ 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.
- The assumptions hold when e^{-f} and e^{-g} are supported on a Torus, f and g have two uniformly continuous derivatives.
- The parabolic PDE occurs in Berman '20 where the author studies limit of the Sinkhorn potentials.

Exponential rate of convergence

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

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

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.

A McKean-Vlasov interpretation

Consider the mirror flow for an *objective function* $F(\cdot)$ and with mirror map $\frac{1}{2}W_2^2(\cdot, e^{-g})$.

A McKean-Vlasov interpretation

Consider the mirror flow for an *objective function* $F(\cdot)$ and with mirror map $\frac{1}{2}W_2^2(\cdot, e^{-g})$.

"Sinkhorn like" PDE is the marginal law of the following diffusion.

$$dX_t = \left(-\frac{\partial}{\partial x^{u_t}}\frac{\delta F}{\delta \rho_t}(X_t) - \frac{\partial g}{\partial x^{u_t}}\left(X_t^{u_t}\right)\right) dt + \sqrt{2\frac{\partial X_t}{\partial X_t^{u_t}}} dB_t, \tag{0.1}$$

where

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

$$2\frac{\partial x}{\partial x^{u_t}} = 2\left(\nabla^2 u_t(x)\right)^{-1}.$$

Different from mirror Langevin diffusion (Ahn-Chewi '21), as u_t depends on law(X_t).

Several open questions

- Replace KL by another divergence. Does this have any algorithmic potential?
- How to choose e^{-g} in practice?
- Other mirror functions than the squared Wasserstein distance.
- One can 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?
- For more details https://arxiv.org/pdf/2307.16421.pdf

Several open questions

- Replace KL by another divergence. Does this have any algorithmic potential?
- How to choose e^{-g} in practice?
- Other mirror functions than the squared Wasserstein distance.
- One can 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?
- For more details https://arxiv.org/pdf/2307.16421.pdf

Thank you. Questions?

For interpretation

Euclidean gradient flows: Assuming smoothness,

$$x_{t+h} - x_t - hx_t = o(|h|)$$

For interpretation

Euclidean gradient flows: Assuming smoothness,

$$x_{t+h} - x_t - hx_t = o(|h|)$$

Wasserstein gradient flows: Recall

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

Assuming smoothness,

$$W_2(\rho_{t+h},\ (\mathrm{Id} + hv_t)_{\#}\rho_t) = o(|h|),$$

Requires v_t in the tangent space (satisfied for gradient flows)

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.