Generative modeling and Parabolic PDEs

Nabarun Deb
University of Chicago Booth School of Business

Indian Institute of Management, Bangalore

https://arxiv.org/pdf/2504.09279 (with Tengyuan Liang)

https://arxiv.org/pdf/2307.16421 (with Young-Heon Kim, Soumik Pal, Geoffrey Schiebinger)

Problem motivation

What is generative modeling?

- Suppose you have some complex data, perhaps images, speech, text, market trends — Generative modeling tries to learn the data generating process (DGP), typically a good approximation to it.
- After learning, the model replicates the DGP to generate new, yet realistic and diverse, data that resembles the original.

What is generative modeling?

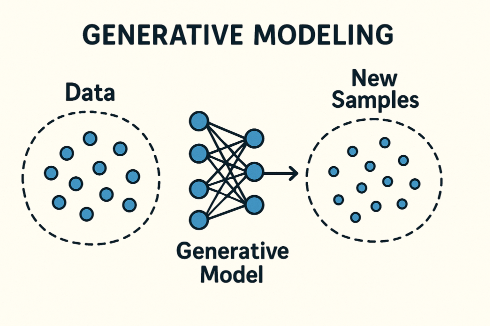
- Suppose you have some complex data, perhaps images, speech, text, market trends — Generative modeling tries to learn the data generating process (DGP), typically a good approximation to it.
- After learning, the model replicates the DGP to generate new, yet realistic and diverse, data that resembles the original.

Generative modeling is not copying, it is creating.

What is generative modeling?

- Suppose you have some complex data, perhaps images, speech, text, market trends — Generative modeling tries to learn the data generating process (DGP), typically a good approximation to it.
- After learning, the model replicates the DGP to generate new, yet realistic and diverse, data that resembles the original.

Generative modeling is not copying, it is creating.



What is "NOT" generative modeling

- Distribution/density estimation
  - Kernels, wavelets, deep neural net based density estimation
  - It is not easy to generate new samples from an arbitrary density function

What is “NOT” generative modeling

- Distribution/density estimation
  - Kernels, wavelets, deep neural net based density estimation
  - It is not easy to generate new samples from an arbitrary density function
- Bootstrapping
  - Generates random samples with replacement from a dataset. Powerful tool for estimating standard errors among other things
  - No “new” samples, simply copies existing data with different multiplicities

What is "NOT" generative modeling

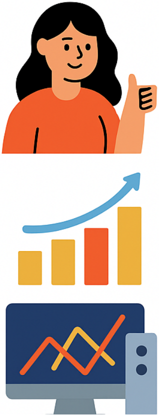
- Distribution/density estimation
  - Kernels, wavelets, deep neural net based density estimation
  - It is not easy to generate new samples from an arbitrary density function
- Bootstrapping
  - Generates random samples with replacement from a dataset. Powerful tool for estimating standard errors among other things
  - No "new" samples, simply copies existing data with different multiplicities
- Prediction models
  - Used when you have a specific question in mind — If my competitor increases price by 100 Rs, should I do the same?
  - Generative modeling would track entire price trajectories

Why do we care?

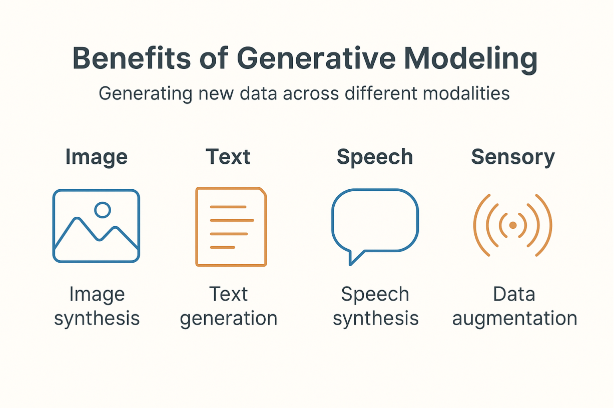
Learning to Generate

The Importance of Generative Modeling

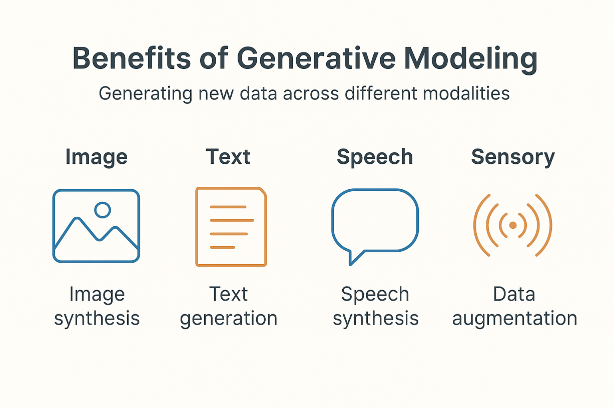
- Can we learn the structure of data to generate realistic samples?
- Applications in economics and business:
  - Simulating customer behavior and market dynamics
  - Stress-testing financial models under different scenarios
  - Creating synthetic data for training and risk managementt



Why do we care?



Why do we care?



Sensory data poses the most significant challenge for generative modeling — hard to get large scaled data sets — involves actual “contact” with smell+temperature

The Math Behind Generative Modeling: Learning Distributions

- Suppose $Z_1, Z_2, \cdots, Z_n \sim P$ (the data distribution)
- Generative modeling tries to learn P from the data in a way that makes it simple to simulate from P
- One strategy is to learn a function G (a denoiser map) such that

$$G(Z) \approx P$$

where Z is an "easy" distribution (like Normal).

The Math Behind Generative Modeling: Learning Distributions

- Suppose $Z_1, Z_2, \cdots, Z_n \sim P$ (the data distribution)
- Generative modeling tries to learn P from the data in a way that makes it simple to simulate from P
- One strategy is to learn a function G (a denoiser map) such that

$$G(Z) \approx P$$

  where Z is an "easy" distribution (like Normal). Usually such G is not unique and different criteria lead to different performance.

The Math Behind Generative Modeling: Learning Distributions

- Suppose $Z_1, Z_2, \cdots, Z_n \sim P$ (the data distribution)
- Generative modeling tries to learn P from the data in a way that makes it simple to simulate from P
- One strategy is to learn a function G (a denoiser map) such that

$$G(Z) \approx P$$

  where Z is an "easy" distribution (like Normal). Usually such G is not unique and different criteria lead to different performance.

How to generate new samples??

- First sample Z.
- Apply the learned denoiser $G(Z)$ to sample new data from P (approximately).

Does denoising work in practice?

GPT prompt: Generate an image of an IIM Bangalore male and female student inside a classroom

Does denoising work in practice?

GPT prompt: Generate an image of an IIM Bangalore male and female student inside a classroom

Does denoising work in practice?

GPT prompt: Can you now generate picture of a male and a female IIMB faculty member in an office party?

Does denoising work in practice?

GPT prompt: Can you now generate picture of a male and a female IIMB faculty member in an office party?

How did that happen?

GPT prompt: What kind of model was used in training?

How did that happen?

GPT prompt: What kind of model was used in training?

- The image was generated using OpenAI's DALL-E 3 model, which creates visuals based on text prompts.

How did that happen?

GPT prompt: What kind of model was used in training?

- The image was generated using OpenAI's DALL-E 3 model, which creates visuals based on text prompts.
- It was trained on billions of text-image pairs, using diffusion techniques that teach the model to generate images by gradually denoising from random patterns.

How did that happen?

GPT prompt: What kind of model was used in training?

- The image was generated using OpenAI's DALL-E 3 model, which creates visuals based on text prompts.
- It was trained on billions of text-image pairs, using diffusion techniques that teach the model to generate images by gradually denoising from random patterns.
- The training data includes publicly available and licensed sources, ensuring a broad and diverse visual vocabulary.

Two approaches towards Generative Modeling

Generative adversarial networks — one shot approach

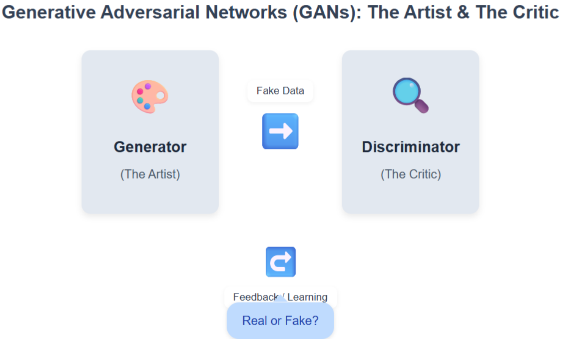
- Think of GANs as a dynamic duo in a constant competition:
  1. The Generator (Artist): Tries to create new, convincing "fakes" (e.g., realistic images, financial data).
  2. The Discriminator (Critic): Tries to distinguish between the "real" data and the "fake" data created by the Artist.

Generative adversarial networks — one shot approach

- Think of GANs as a dynamic duo in a constant competition:
  1. The Generator (Artist): Tries to create new, convincing "fakes" (e.g., realistic images, financial data).
  2. The Discriminator (Critic): Tries to distinguish between the "real" data and the "fake" data created by the Artist.
- They learn by competing: The Artist gets better at fooling the Critic, and the Critic gets better at spotting fakes.

Generative adversarial networks — one shot approach

- Think of GANs as a dynamic duo in a constant competition:
  1. The Generator (Artist): Tries to create new, convincing "fakes" (e.g., realistic images, financial data).
  2. The Discriminator (Critic): Tries to distinguish between the "real" data and the "fake" data created by the Artist.
- They learn by competing: The Artist gets better at fooling the Critic, and the Critic gets better at spotting fakes.



Mathematical formulation

- The generator has a candidate set of transformations or denoisers g_θ, indexed by some parameter θ (for e.g., a deep neural network).

Mathematical formulation

- The generator has a candidate set of transformations or denoisers g_θ, indexed by some parameter θ (for e.g., a deep neural network).

- The discriminator looks at the denoiser and computes a "distance" (typically integral probability metrics) between the denoised distribution and the data distribution. Remeber we want $g_\theta(Z)$ close to data distribution.

Mathematical formulation

- The generator has a candidate set of transformations or denoisers g_θ, indexed by some parameter θ (for e.g., a deep neural network).
- The discriminator looks at the denoiser and computes a "distance" (typically integral probability metrics) between the denoised distribution and the data distribution. Remeber we want $g_\theta(Z)$ close to data distribution.
- Large distance implies discriminator forces generator to choose a different parameter.

Mathematical formulation

- The generator has a candidate set of transformations or denoisers g_θ, indexed by some parameter θ (for e.g., a deep neural network).
- The discriminator looks at the denoiser and computes a "distance" (typically integral probability metrics) between the denoised distribution and the data distribution. Remeber we want $g_\theta(Z)$ close to data distribution.
- Large distance implies discriminator forces generator to choose a different parameter.

(A minimax game)

$$\inf_{g_\theta} \quad \sup_{f} \quad \left| \mathbb{E} f(g_\theta(Z)) - \mathbb{E}_{X \sim \mathrm{data}} f(X) \right|.$$

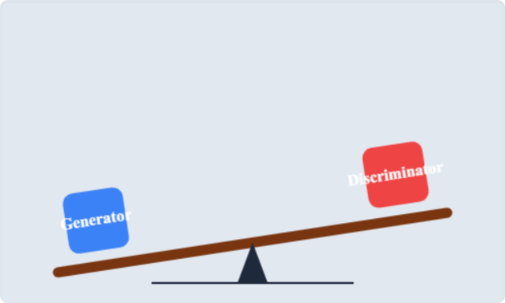
Here Z is the noise variable.

More on GANs

- Easy to sample: Once you have learned "the best" g_θ from the minimax game, sampling is just one-shot.

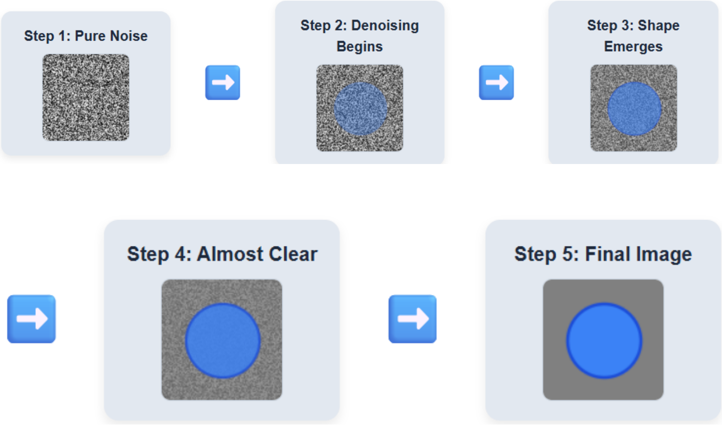
$$Z \sim \text{Noise}, \quad \text{Sample } g_\theta(Z).$$

- Hard to learn: The minimax game is hard to solve because of uncoupled data —



Leads to mode collapse where the generator produces very similar images.

Enter Diffusion models



More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.

More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.

- Unlike in GANs, these one-step denoisers involve "approximately coupled data" which makes learning easier; no mode collapse

More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.
- Unlike in GANs, these one-step denoisers involve “approximately coupled data” which makes learning easier; no mode collapse
- There’s no adversarial competition. It’s a single, guided process of refinement.

More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.
- Unlike in GANs, these one-step denoisers involve "approximately coupled data" which makes learning easier; no mode collapse
- There's no adversarial competition. It's a single, guided process of refinement.
- Diffusion models work by gradually "denoising" this random noise, step-by-step, until a clear, coherent image (or other data) emerges.

More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.
- Unlike in GANs, these one-step denoisers involve "approximately coupled data" which makes learning easier; no mode collapse
- There's no adversarial competition. It's a single, guided process of refinement.
- Diffusion models work by gradually "denoising" this random noise, step-by-step, until a clear, coherent image (or other data) emerges. Compared to GANs which are one-shot denoisers

More on Diffusion models

- Imagine starting with pure static or noise, like a blurry TV screen.
- Unlike in GANs, these one-step denoisers involve "approximately coupled data" which makes learning easier; no mode collapse
- There's no adversarial competition. It's a single, guided process of refinement.
- Diffusion models work by gradually "denoising" this random noise, step-by-step, until a clear, coherent image (or other data) emerges. Compared to GANs which are one-shot denoisers
- Harder to sample as they are not one-step; usually takes more time than GANs

Our goal

- GANs are easy to sample from (because one-shot) but are harder to learn (due to uncoupled nature of the learning problem)

Our goal

- GANs are easy to sample from (because one-shot) but are harder to learn (due to uncoupled nature of the learning problem)

- Diffusion models are harder to sample from (because sequential nature) but are easier to learn (because successive points in the sequence are "approximately coupled")

Our goal

- GANs are easy to sample from (because one-shot) but are harder to learn (due to uncoupled nature of the learning problem)

- Diffusion models are harder to sample from (because sequential nature) but are easier to learn (because successive points in the sequence are "approximately coupled")

New algorithm

- Combine ease of sampling with ease of learning

Our goal

- GANs are easy to sample from (because one-shot) but are harder to learn (due to uncoupled nature of the learning problem)

- Diffusion models are harder to sample from (because sequential nature) but are easier to learn (because successive points in the sequence are "approximately coupled")

New algorithm

- Combine ease of sampling with ease of learning
- A sequential algorithm where successive points are approximately coupled but you only need the last transformation to sample

Optimal Transport and connection to generative modeling

Wasserstein distance and optimal transport map

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

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

Wasserstein distance and optimal transport map

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

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

- The optimal γ_∞ is the law of (X, Y) where $Y = \nabla \phi_\infty(X)$ for some convex function $\phi_\infty : \mathbb{R}^d \to \mathbb{R}$.

Wasserstein distance and optimal transport map

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

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

- The optimal γ_∞ is the law of (X, Y) where $Y = \nabla \phi_\infty(X)$ for some convex function $\phi_\infty : \mathbb{R}^d \to \mathbb{R}$.

- We call $\nabla \phi_\infty$ the optimal transport (OT map) from e^{-f} to e^{-g}.

Wasserstein distance and optimal transport map

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

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

- The optimal γ_∞ is the law of (X, Y) where $Y = \nabla\phi_\infty(X)$ for some convex function $\phi_\infty : \mathbb{R}^d \to \mathbb{R}$.
- We call $\nabla\phi_\infty$ the optimal transport (OT map) from e^{-f} to e^{-g}.
- We will use the push-forward $\#$ notation, i.e., $\nabla\phi_\infty \# e^{-f} = e^{-g}$ will imply that if $Z \sim e^{-f}$ then $\nabla\phi_\infty(Z) \sim e^{-g}$.

OT map in sampling/generative modeling

- Target: Sample from e^{-g} (data distribution)
  Source: Some simple e^{-f} which is easy to sample from (this is the noise)

OT map in sampling/generative modeling

- Target: Sample from e^{-g} (data distribution)
  Source: Some simple e^{-f} which is easy to sample from (this is the noise)
- As $\nabla\phi_\infty(Z) \sim e^{-g}$, $\nabla\phi_\infty$ is a denoiser for generative modeling

OT map in sampling/generative modeling

- Target: Sample from e^{-g} (data distribution)
  Source: Some simple e^{-f} which is easy to sample from (this is the noise)

- As $\nabla\phi_\infty(Z) \sim e^{-g}$, $\nabla\phi_\infty$ is a denoiser for generative modeling

- **Ease of sampling**: What if we had $\nabla\phi_\infty(\cdot)$ or a good one-shot estimator (GANs)? Sample $Z_1, Z_2, \ldots \sim e^{-f}$ and return $\nabla\phi_\infty(Z_1), \nabla\phi_\infty(Z_2), \ldots$

- Estimating $\nabla\phi_\infty$ in one-shot can be hard (uncoupled data) — mode collapse in Generative adversarial nets Thanh-Tung and Tran (2020)

- **Ease of learning**: Many sequential approaches to generative modeling — flow-based, diffusion-based, (approximately coupled data) .. (see Kumar et al. (2019), Cheng et al. (2023), Huang et al. (2021), Karras et al. (2022), ...)

OT map in sampling/generative modeling

- Target: Sample from e^{-g} (data distribution)
  Source: Some simple e^{-f} which is easy to sample from (this is the noise)

- As $\nabla\phi_\infty(Z) \sim e^{-g}$, $\nabla\phi_\infty$ is a denoiser for generative modeling

- **Ease of sampling**: What if we had $\nabla\phi_\infty(\cdot)$ or a good one-shot estimator (GANs)? Sample $Z_1, Z_2, \ldots \sim e^{-f}$ and return $\nabla\phi_\infty(Z_1), \nabla\phi_\infty(Z_2), \ldots$

- Estimating $\nabla\phi_\infty$ in one-shot can be hard (uncoupled data) — mode collapse in Generative adversarial nets Thanh-Tung and Tran (2020)

- **Ease of learning**: Many sequential approaches to generative modeling — flow-based, diffusion-based, (approximately coupled data) .. (see Kumar et al. (2019), Cheng et al. (2023), Huang et al. (2021), Karras et al. (2022), ...)

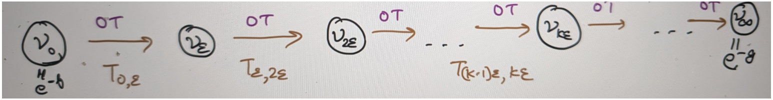
- One common theme — glue together OT maps over "small time jumps" over a path on probability measures.

An example flow: Fokker-Planck

- A popular path: $\{\nu_t\}_{t\geq 0}$ probability densities satisfying

$$\partial_t \nu_t = \nabla \cdot (\nu_t(\nabla g + \nabla \log \nu_t)) \implies \nu_\infty = e^{-g}.$$

- Illustration of flow —



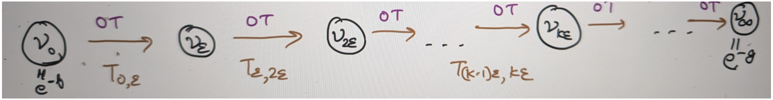
- Each $T_{(k-1)\varepsilon,k\varepsilon}$ is the OT map from $\nu_{(k-1)\varepsilon}$ to $\nu_{k\varepsilon}$.

An example flow: Fokker-Planck

- A popular path: $\{\nu_t\}_{t\geq 0}$ probability densities satisfying

$$\partial_t \nu_t = \nabla \cdot (\nu_t(\nabla g + \nabla \log \nu_t)) \implies \nu_\infty = e^{-g}.$$

- Illustration of flow —



- Each $T_{(k-1)\varepsilon,k\varepsilon}$ is the OT map from $\nu_{(k-1)\varepsilon}$ to $\nu_{k\varepsilon}$.

- How do we go from ν_0 to $\nu_{k\varepsilon}$?

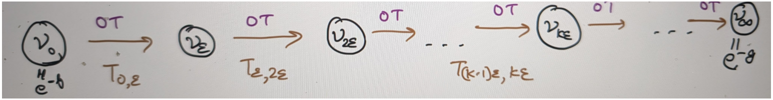
$$T = T_{(k-1)\varepsilon,k\varepsilon} \circ T_{(k-2)\varepsilon,(k-1)\varepsilon} \circ \ldots \circ T_{\varepsilon,2\varepsilon} \circ T_{0,\varepsilon}.$$

An example flow: Fokker-Planck

- A popular path: $\{\nu_t\}_{t\geq 0}$ probability densities satisfying

$$\partial_t \nu_t = \nabla \cdot (\nu_t(\nabla g + \nabla \log \nu_t)) \implies \nu_\infty = e^{-g}.$$

- Illustration of flow —



- Each $T_{(k-1)\varepsilon,k\varepsilon}$ is the OT map from $\nu_{(k-1)\varepsilon}$ to $\nu_{k\varepsilon}$.

- How do we go from ν_0 to $\nu_{k\varepsilon}$?

$$T = T_{(k-1)\varepsilon,k\varepsilon} \circ T_{(k-2)\varepsilon,(k-1)\varepsilon} \circ \ldots \circ T_{\varepsilon,2\varepsilon} \circ T_{0,\varepsilon}.$$

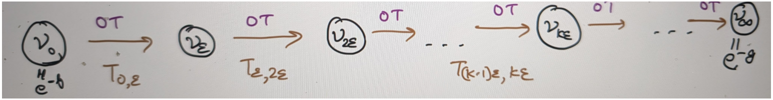
- But composition of OT map is not OT. So, for large $k\varepsilon$, T is not close to $\nabla\phi_\infty$ (the OT map from e^{-f} to e^{-g})

An example flow: Fokker-Planck

- A popular path: $\{\nu_t\}_{t\geq 0}$ probability densities satisfying

$$\partial_t \nu_t = \nabla \cdot (\nu_t(\nabla g + \nabla \log \nu_t)) \implies \nu_\infty = e^{-g}.$$

- Illustration of flow —



- Each $T_{(k-1)\varepsilon,k\varepsilon}$ is the OT map from $\nu_{(k-1)\varepsilon}$ to $\nu_{k\varepsilon}$.

- How do we go from ν_0 to $\nu_{k\varepsilon}$?

$$T = T_{(k-1)\varepsilon,k\varepsilon} \circ T_{(k-2)\varepsilon,(k-1)\varepsilon} \circ \ldots \circ T_{\varepsilon,2\varepsilon} \circ T_{0,\varepsilon}.$$

- But composition of OT map is not OT. So, for large $k\varepsilon$, T is not close to $\nabla\phi_\infty$ (the OT map from e^{-f} to e^{-g})

How about a flow on OT maps which recovers $\nabla\phi_\infty$ in the limit?

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

$$f(x) - g(\nabla\phi_\infty(x)) + \log \mathrm{Det}\left(\nabla^2\phi_\infty(x)\right) = 0.$$

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

$$f(x) - g(\nabla\phi_\infty(x)) + \log \operatorname{Det}\left(\nabla^2\phi_\infty(x)\right) = 0.$$

- Parabolic Monge-Ampère (PMA) is the dynamic version

$$\partial_t\phi_t(x) = f(x) - g(\nabla\phi_t(x)) + \log \operatorname{Det}\left(\nabla^2\phi_t(x)\right).$$

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

$$f(x) - g(\nabla\phi_\infty(x)) + \log \mathrm{Det}\left(\nabla^2\phi_\infty(x)\right) = 0.$$

- Parabolic Monge-Ampère (PMA) is the dynamic version

$$\partial_t\phi_t(x) = f(x) - g(\nabla\phi_t(x)) + \log \mathrm{Det}\left(\nabla^2\phi_t(x)\right).$$

- It is possible to identify the related continuity equation for a sequence of probability measures $\{\rho_t\}_{t\geq 0}$ such that $\nabla\phi_t \# \rho_t = e^{-g}$.

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

$$f(x) - g(\nabla\phi_\infty(x)) + \log \mathrm{Det}\left(\nabla^2\phi_\infty(x)\right) = 0.$$

- Parabolic Monge-Ampère (PMA) is the dynamic version

$$\partial_t\phi_t(x) = f(x) - g(\nabla\phi_t(x)) + \log \mathrm{Det}\left(\nabla^2\phi_t(x)\right).$$

- It is possible to identify the related continuity equation for a sequence of probability measures $\{\rho_t\}_{t\geq 0}$ such that $\nabla\phi_t \# \rho_t = e^{-g}$.
- Under regularity assumptions on f, g, strong convexity of ϕ_∞, and of the initializer (say $\tilde{\phi}_0$), the PMA admits a smooth solution $\{\tilde{\phi}_t\}_{t\geq 0}$ (see Kitagawa (2010), Kim et al. (2010), Berman (2020))

Parabolic Monge-Ampère

- A flow which directly operates on the space of OT maps
- Suppose $\nabla\phi_\infty \# e^{-f} = e^{-g}$, then usual (static) Monge-Ampère (MA) is just the change of variable formula —

$$f(x) - g(\nabla\phi_\infty(x)) + \log \operatorname{Det}\left(\nabla^2\phi_\infty(x)\right) = 0.$$

- Parabolic Monge-Ampère (PMA) is the dynamic version

$$\partial_t\phi_t(x) = f(x) - g(\nabla\phi_t(x)) + \log \operatorname{Det}\left(\nabla^2\phi_t(x)\right).$$

- It is possible to identify the related continuity equation for a sequence of probability measures $\{\rho_t\}_{t\geq 0}$ such that $\nabla\phi_t \# \rho_t = e^{-g}$.
- Under regularity assumptions on f, g, strong convexity of ϕ_∞, and of the initializer (say $\tilde{\phi}_0$), the PMA admits a smooth solution $\{\tilde{\phi}_t\}_{t\geq 0}$ (see Kitagawa (2010), Kim et al. (2010), Berman (2020))
- Importantly, $\nabla\tilde{\phi}_t \to \nabla\phi_\infty$ (PMA converges to actual OT) and the convergence is exponentially fast in t.

Illustration for generative modeling

- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).

Illustration for generative modeling

- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).
- Define $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$

Illustration for generative modeling

- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).
- Define $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$
- Set $\tilde{\phi}_0(x) = \|x\|^2/2$, implies $\tilde{\phi}_0^*(y) = \|y\|^2/2$.

Illustration for generative modeling

- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).
- Define $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$
- Set $\tilde{\phi}_0(x) = \|x\|^2/2$, implies $\tilde{\phi}_0^*(y) = \|y\|^2/2$.
- Illustration of flow —

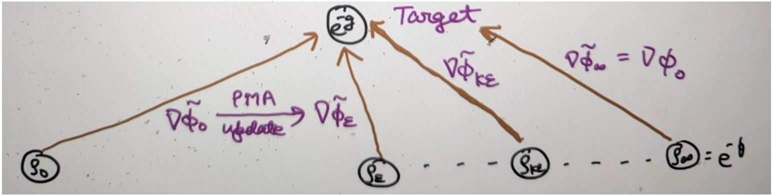
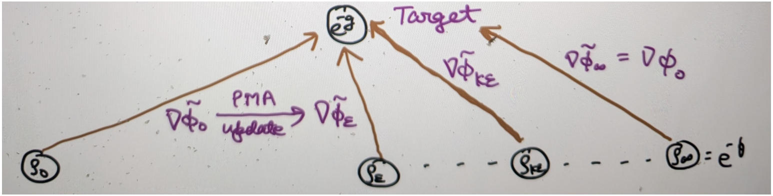


Illustration for generative modeling

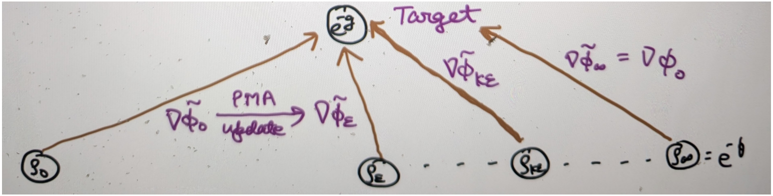
- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).
- Define $\rho_t = \nabla\tilde{\phi}_t^* \# e^{-g}$
- Set $\tilde{\phi}_0(x) = \|x\|^2/2$, implies $\tilde{\phi}_0^*(y) = \|y\|^2/2$.
- Illustration of flow —



- Each $\nabla\tilde{\phi}_{k\varepsilon}$ is the OT map from $\rho_{k\varepsilon}$ to e^{-g}.

Illustration for generative modeling

- Let $\tilde{\phi}_t^*$ denote the convex conjugate of $\tilde{\phi}_t$ (solution of PMA).
- Define $\rho_t = \nabla\tilde{\phi}_t^* \# e^{-g}$
- Set $\tilde{\phi}_0(x) = \|x\|^2/2$, implies $\tilde{\phi}_0^*(y) = \|y\|^2/2$.
- Illustration of flow —



- Each $\nabla\tilde{\phi}_{k\varepsilon}$ is the OT map from $\rho_{k\varepsilon}$ to e^{-g}.

Difference with existing approaches

- Discretizing PMA can be viewed as a new approach to generative modeling that combines ease of sampling with ease of learning.

Difference with existing approaches

- Discretizing PMA can be viewed as a new approach to generative modeling that combines ease of sampling with ease of learning.

- **Ease of sampling**: Generate samples $Z_1, ..., Z_n$ from e^{-f} (easy to generate). Construct $\nabla\tilde{\phi}_{k\varepsilon}$ for appropriate k, ϵ. Then $\nabla\tilde{\phi}_{k\varepsilon}(Z_i) \approx e^{-g}$ (No need for function composition).

Difference with existing approaches

- Discretizing PMA can be viewed as a new approach to generative modeling that combines ease of sampling with ease of learning.

- **Ease of sampling**: Generate samples $Z_1, \ldots, Z_n$ from e^{-f} (easy to generate). Construct $\nabla\tilde{\phi}_{k\varepsilon}$ for appropriate k, ϵ. Then $\nabla\tilde{\phi}_{k\varepsilon}(Z_i) \approx e^{-g}$ (No need for function composition).

- **Ease of learning**: Two successive iterations are optimally coupled with respect to a time varying cost (see D. and Liang (2025)). Each $\nabla\tilde{\phi}_{k\varepsilon}$ is close to $\nabla\tilde{\phi}_{(k-1)\varepsilon}$ and the updates can be tracked with score matching techniques as well.

Difference with existing approaches

- Discretizing PMA can be viewed as a new approach to generative modeling that combines ease of sampling with ease of learning.
- **Ease of sampling**: Generate samples $Z_1, \ldots, Z_n$ from e^{-f} (easy to generate). Construct $\nabla\tilde{\phi}_{k\varepsilon}$ for appropriate k, ϵ. Then $\nabla\tilde{\phi}_{k\varepsilon}(Z_i) \approx e^{-g}$ (No need for function composition).
- **Ease of learning**: Two successive iterations are optimally coupled with respect to a time varying cost (see D. and Liang (2025)). Each $\nabla\tilde{\phi}_{k\varepsilon}$ is close to $\nabla\tilde{\phi}_{(k-1)\varepsilon}$ and the updates can be tracked with score matching techniques as well.

Difference with existing approaches

- Discretizing PMA can be viewed as a new approach to generative modeling that combines ease of sampling with ease of learning.
- **Ease of sampling**: Generate samples $Z_1, \ldots, Z_n$ from e^{-f} (easy to generate). Construct $\nabla\tilde{\phi}_{k\varepsilon}$ for appropriate k, ϵ. Then $\nabla\tilde{\phi}_{k\varepsilon}(Z_i) \approx e^{-g}$ (No need for function composition).
- **Ease of learning**: Two successive iterations are optimally coupled with respect to a time varying cost (see D. and Liang (2025)). Each $\nabla\tilde{\phi}_{k\varepsilon}$ is close to $\nabla\tilde{\phi}_{(k-1)\varepsilon}$ and the updates can be tracked with score matching techniques as well.

A natural goal therefore is to discretize the PMA.

Time discretization for PMA using Sinkhorn algorithm scaling limits

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,

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

- As mentioned before, the optimal coupling above is degenerate 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,

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

- As mentioned before, the optimal coupling above is degenerate 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

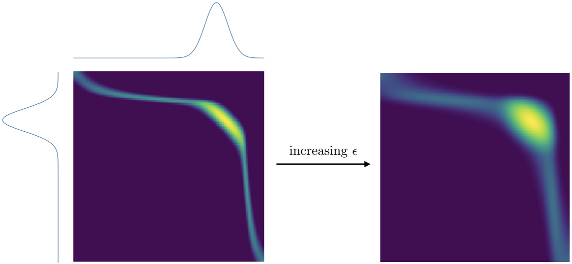


Figure: Image by M. Cuturi

- 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

$$\gamma^\epsilon(x,y) = \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi^\varepsilon(x) - \frac{1}{\varepsilon}\psi^\varepsilon(y) - f(x) - g(y)\right).$$

- $\phi^\epsilon, \psi^\epsilon$ - Schrödinger potentials. Unique up to constant.
- Typically not explicit. Determined by marginal constraints

$$\int \gamma^\epsilon(x,y)dy = e^{-f(x)}, \quad \int \gamma^\varepsilon(x,y)dx = e^{-g(y)}.$$

Structure of the solution

- The **optimal coupling** (Rüschendorf & Thomsen '93) γ^ϵ must be of the form

$$\gamma^\epsilon(x,y) = \exp\left(\frac{1}{\varepsilon}\langle x,y\rangle - \frac{1}{\varepsilon}\phi^\varepsilon(x) - \frac{1}{\varepsilon}\psi^\varepsilon(y) - f(x) - g(y)\right).$$

- $\phi^\epsilon, \psi^\epsilon$ - Schrödinger potentials. Unique up to constant.
- Typically not explicit. Determined by marginal constraints

$$\int \gamma^\epsilon(x,y)dy = e^{-f(x)}, \quad \int \gamma^\varepsilon(x,y)dx = e^{-g(y)}.$$

- This gives the fixed point system

$$\phi^\varepsilon(x) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x,y\rangle - \frac{1}{\varepsilon}\psi^\varepsilon(y) - g(y)\right) dy,$$

$$\psi^\varepsilon(y) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x,y\rangle - \frac{1}{\varepsilon}\phi^\varepsilon(x) - f(x)\right) dx.$$

Sinkhorn/IPFP algorithm

- An iterative approach to solving the fixed point system and produces a sequence of "couplings".

Sinkhorn/IPFP algorithm

- An iterative approach to solving the fixed point system and produces a sequence of "couplings".
- For $k \geq 1$,

$$\psi_k^\varepsilon(y) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi_{k-1}^\varepsilon(x) - f(x)\right) dx,$$

$$\phi_k^\varepsilon(x) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\psi_k^\varepsilon(y) - g(y)\right) dy.$$

Sinkhorn/IPFP algorithm

- An iterative approach to solving the fixed point system and produces a sequence of "couplings".
- For $k \geq 1$,

$$\psi_k^\varepsilon(y) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi_{k-1}^\varepsilon(x) - f(x)\right) dx,$$

$$\phi_k^\varepsilon(x) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\psi_k^\varepsilon(y) - g(y)\right) dy.$$

- The corresponding probability distribution

$$\gamma_k^\varepsilon(x, y) = \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi_{k-1}^\varepsilon(x) - \frac{1}{\varepsilon}\psi_k^\varepsilon(y) - f(x) - g(y)\right)$$

couples its X and Y marginals given by

$$p_X\gamma_k^\varepsilon(x) = \exp\left(\frac{1}{\varepsilon}(\phi_k^\varepsilon - \phi_{k-1}^\varepsilon)(x)\right), \quad p_Y\gamma_k^\varepsilon(y) = \exp(-g(y)).$$

Sinkhorn/IPFP algorithm

- An iterative approach to solving the fixed point system and produces a sequence of "couplings".
- For $k \geq 1$,

$$\psi_k^\varepsilon(y) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi_{k-1}^\varepsilon(x) - f(x)\right) dx,$$

$$\phi_k^\varepsilon(x) = \varepsilon \log \int \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\psi_k^\varepsilon(y) - g(y)\right) dy.$$

- The corresponding probability distribution

$$\gamma_k^\varepsilon(x, y) = \exp\left(\frac{1}{\varepsilon}\langle x, y\rangle - \frac{1}{\varepsilon}\phi_{k-1}^\varepsilon(x) - \frac{1}{\varepsilon}\psi_k^\varepsilon(y) - f(x) - g(y)\right)$$

couples its X and Y marginals given by

$$p_X\gamma_k^\varepsilon(x) = \exp\left(\frac{1}{\varepsilon}(\phi_k^\varepsilon - \phi_{k-1}^\varepsilon)(x)\right), \quad p_Y\gamma_k^\varepsilon(y) = \exp(-g(y)).$$

Do gradient of Sinkhorn potentials $\nabla\phi_k^\varepsilon$ approximate gradient of PMA $\nabla\tilde{\phi}_t$?

Some nice properties of Sinkhorn algorithm

- *Sample computation* — Suppose we only have data from either e^{-f} or e^{-g} or both, then $\phi_k^{\varepsilon}(x)$ and $\psi_k^{\varepsilon}(y)$ can be computed with empirical averages.

Some nice properties of Sinkhorn algorithm

- *Sample computation* — Suppose we only have data from either e^{-f} or e^{-g} or both, then $\phi_k^\varepsilon(x)$ and $\psi_k^\varepsilon(y)$ can be computed with empirical averages.

- *Fast computation* — see Cuturi (2013), Rubner et al. (1997), Pele and Werman (2009).

Some nice properties of Sinkhorn algorithm

- *Sample computation* — Suppose we only have data from either e^{-f} or e^{-g} or both, then $\phi_k^\varepsilon(x)$ and $\psi_k^\varepsilon(y)$ can be computed with empirical averages.
- *Fast computation* — see Cuturi (2013), Rubner et al. (1997), Pele and Werman (2009).
- *Gradient-free nature* — Note that updates of PMA

$$\partial_t \tilde{\phi}_t(x) = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x))$$

  require gradient computation of $\tilde{\phi}_t$.

Some nice properties of Sinkhorn algorithm

- *Sample computation* — Suppose we only have data from either e^{-f} or e^{-g} or both, then $\phi_k^\varepsilon(x)$ and $\psi_k^\varepsilon(y)$ can be computed with empirical averages.

- *Fast computation* — see Cuturi (2013), Rubner et al. (1997), Pele and Werman (2009).

- *Gradient-free nature* — Note that updates of PMA

$$\partial_t \tilde{\phi}_t(x) = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x))$$

  require gradient computation of $\tilde{\phi}_t$. However updating ϕ_k^ε, ψ_k^ε from the past iterates in Sinkhorn requires no gradient computation.

Some nice properties of Sinkhorn algorithm

- *Sample computation* — Suppose we only have data from either e^{-f} or e^{-g} or both, then $\phi_k^\varepsilon(x)$ and $\psi_k^\varepsilon(y)$ can be computed with empirical averages.

- *Fast computation* — see Cuturi (2013), Rubner et al. (1997), Pele and Werman (2009).

- *Gradient-free nature* — Note that updates of PMA

$$\partial_t \tilde{\phi}_t(x) = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x))$$

  require gradient computation of $\tilde{\phi}_t$. However updating ϕ_k^ε, ψ_k^ε from the past iterates in Sinkhorn requires no gradient computation.

- Not so nice - Instabilities for small ϵ.

Correct scaling for limits

- By Berman (2020), Léger (2020), Aubin-Frankowski et al. (2022), it follows:

$$(H_\epsilon^*)'(\gamma_{k+1}^\epsilon) - (H_\epsilon^*)'(\gamma_k^\epsilon) = -\mathrm{KL}'(p_X \gamma_k^\varepsilon | e^{-f}).$$

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

Correct scaling for limits

- By Berman (2020), Léger (2020), Aubin-Frankowski et al. (2022), it follows:

$$(H_\epsilon^*)'(\gamma_{k+1}^\epsilon) - (H_\epsilon^*)'(\gamma_k^\epsilon) = -\mathrm{KL}'(p_X \gamma_k^\varepsilon | e^{-f}).$$

  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:

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

Correct scaling for limits

- By Berman (2020), Léger (2020), Aubin-Frankowski et al. (2022), it follows:

$$(H_\epsilon^*)'(\gamma_{k+1}^\epsilon) - (H_\epsilon^*)'(\gamma_k^\epsilon) = -\mathrm{KL}'(p_X\gamma_k^\varepsilon | e^{-f}).$$

  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:

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

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

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

Correct scaling for limits

- By Berman (2020), Léger (2020), Aubin-Frankowski et al. (2022), it follows:

$$(H_\epsilon^*)'(\gamma_{k+1}^\epsilon) - (H_\epsilon^*)'(\gamma_k^\epsilon) = -\mathrm{KL}'(p_X\gamma_k^\varepsilon | e^{-f}).$$

  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:

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

  (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.

Correct scaling for limits

- By Berman (2020), Léger (2020), Aubin-Frankowski et al. (2022), it follows:

$$(H_\epsilon^*)'(\gamma_{k+1}^\epsilon) - (H_\epsilon^*)'(\gamma_k^\epsilon) = -\mathrm{KL}'(p_X\gamma_k^\varepsilon | e^{-f}).$$

  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:

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

  (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.

Study the approximation $\nabla \phi_k^\varepsilon \approx \nabla \tilde{\phi}_t$ when $k = t/\varepsilon$?

Main results

Recall that $\tilde{\phi}_t$ is used to denote solution of the PMA

$$\partial_t \tilde{\phi}_t = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x)).$$

Set $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$.

Main results

Recall that $\tilde{\phi}_t$ is used to denote solution of the PMA

$$\partial_t \tilde{\phi}_t = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x)).$$

Set $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$. Also $\phi_k^\varepsilon, \psi_k^\varepsilon$s are potentials from Sinkhorn and γ_k^ε is the corresponding coupling.

Main results

Recall that $\tilde{\phi}_t$ is used to denote solution of the PMA

$$\partial_t \tilde{\phi}_t = f(x) - g(\nabla \tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2 \tilde{\phi}_t(x)).$$

Set $\rho_t = \nabla \tilde{\phi}_t^* \# e^{-g}$. Also $\phi_k^\varepsilon, \psi_k^\varepsilon$s are potentials from Sinkhorn and γ_k^ε is the corresponding coupling.

Scaling limit for $\nabla \phi_{t/\varepsilon}^\varepsilon$ and $\gamma_{t/\varepsilon}^\varepsilon$

Under regularity assumptions on the PMA and appropriate initialization, we have

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \nabla(\phi_{t/\varepsilon}^\varepsilon - \tilde{\phi}_t)(x) = \frac{1}{2} \nabla f(x) + \nabla \log \rho_t(x).$$

Comparison with existing works

- In Berman (2020), it was shown that

$$\phi^{\varepsilon}_{t/\varepsilon} - \tilde{\phi}_t = O(\varepsilon)$$

which by reverse Poincaré type inequality implies

$$\nabla\phi^{\varepsilon}_{t/\varepsilon} - \nabla\tilde{\phi}_t = O(\sqrt{\varepsilon}).$$

This can be extended to $O(\varepsilon)$.

Comparison with existing works

- In Berman (2020), it was shown that

$$\phi^{\varepsilon}_{t/\varepsilon} - \tilde{\phi}_t = O(\varepsilon)$$

which by reverse Poincaré type inequality implies

$$\nabla\phi^{\varepsilon}_{t/\varepsilon} - \nabla\tilde{\phi}_t = O(\sqrt{\varepsilon}).$$

This can be extended to $O(\varepsilon)$.

- In Deb et al. (2023), we show that

$$\frac{1}{\varepsilon}(\phi^{\varepsilon}_{t/\varepsilon} - \phi^{\varepsilon}_{t/\varepsilon-1})(x) - f(x) \to -g(\nabla\tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2\tilde{\phi}_t(x))$$

in a weak sense.

Comparison with existing works

- In Berman (2020), it was shown that

$$\phi^{\varepsilon}_{t/\varepsilon} - \tilde{\phi}_t = O(\varepsilon)$$

which by reverse Poincaré type inequality implies

$$\nabla\phi^{\varepsilon}_{t/\varepsilon} - \nabla\tilde{\phi}_t = O(\sqrt{\varepsilon}).$$

This can be extended to $O(\varepsilon)$.

- In Deb et al. (2023), we show that

$$\frac{1}{\varepsilon}(\phi^{\varepsilon}_{t/\varepsilon} - \phi^{\varepsilon}_{t/\varepsilon-1})(x) - f(x) \to -g(\nabla\tilde{\phi}_t(x)) + \log \mathrm{Det}(\nabla^2\tilde{\phi}_t(x))$$

in a weak sense. Recall that

$$\mathrm{LHS} = \log \rho^{\varepsilon}_{t/\varepsilon}, \quad \text{and} \quad \mathrm{RHS} = \log \rho_t.$$

Then Deb et al. (2023) shows

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) \to 0.$$

Based on current bounds this can be improved to KL instead of Wasserstein.

Comparison with existing works

- Quantitatively, Deb et al. (2023) shows that

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\sqrt{\varepsilon}).$$

This can be extended to

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\varepsilon).$$

The metric can be improved to KL, the linearized optimal transport distance, etc.

Comparison with existing works

- Quantitatively, Deb et al. (2023) shows that

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\sqrt{\varepsilon}).$$

  This can be extended to

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\varepsilon).$$

  The metric can be improved to KL, the linearized optimal transport distance, etc.

- There is trade-off in that the improved bounds require two extra orders of regularity on the PMA.

Comparison with existing works

- Quantitatively, Deb et al. (2023) shows that

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\sqrt{\varepsilon}).$$

  This can be extended to

$$W_2(\rho^{\varepsilon}_{t/\varepsilon}, \rho_t) = O(\varepsilon).$$

  The metric can be improved to KL, the linearized optimal transport distance, etc.

- There is trade-off in that the improved bounds require two extra orders of regularity on the PMA.

- In Pooladian and Weed (2024), authors analyze Sinkhorn with space discretization and provide rates of convergence but with $k \sim (1/\varepsilon)^7$ as opposed to $k \sim (1/\varepsilon)$.

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

- Compare to Berman (2020), $\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + O(\varepsilon)$.

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

- Compare to Berman (2020), $\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + O(\varepsilon)$.
- We borrow and extend the coupling argument from Berman (2020).

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

- Compare to Berman (2020), $\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + O(\varepsilon)$.
- We borrow and extend the coupling argument from Berman (2020).
- A multivariate second order Laplace approximation.

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

- Compare to Berman (2020), $\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + O(\varepsilon)$.
- We borrow and extend the coupling argument from Berman (2020).
- A multivariate second order Laplace approximation.
- Typically to extract the coefficients

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + \varepsilon(\ldots) + \varepsilon^2(\ldots) + \ldots$$

  we need one extra order Laplace approximation which will introduce one extra PDE.

Proof technique

Main technical lemma

Under previous assumptions,

$$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t(x) + \varepsilon r_t(x) + O(\varepsilon^2),$$

where r_t depends on f, g, and $\tilde{\phi}_t$ (explicitly provided).

- Compare to Berman (2020), $\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + O(\varepsilon)$.
- We borrow and extend the coupling argument from Berman (2020).
- A multivariate second order Laplace approximation.
- Typically to extract the coefficients

  $$\phi^{\varepsilon}_{t/\varepsilon} = \tilde{\phi}_t + \varepsilon(\dots) + \varepsilon^2(\dots) + \dots$$

  we need one extra order Laplace approximation which will introduce one extra PDE.
- Solving the PDE for the coefficient of ε in terms of the solution of PMA $\tilde{\phi}_t$. Recall $\tilde{\phi}_t$ is the solution of the PMA.

Conclusion

- Discretizing parabolic Monge-Ampère could lead to a new perspective on generative modeling.
- There is a general family of parabolic PDEs. Can we design Sinkhorn-like algorithms for them?
- How to choose the source distribution in practice?
- What about random space discretization? How to choose $\varepsilon > 0$ based on data?
- Tracking these flows via particle systems ...

Conclusion

- Discretizing parabolic Monge-Ampère could lead to a new perspective on generative modeling.
- There is a general family of parabolic PDEs. Can we design Sinkhorn-like algorithms for them?
- How to choose the source distribution in practice?
- What about random space discretization? How to choose $\varepsilon > 0$ based on data?
- Tracking these flows via particle systems ...

Thank you. Questions?

Entropic regularization

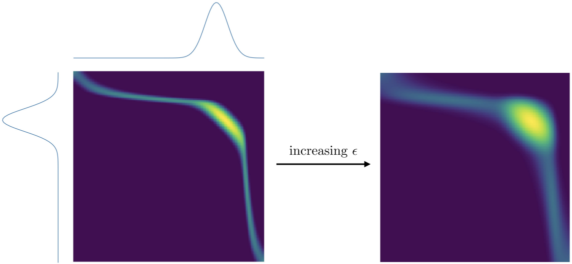


Figure: Image by M. Cuturi

- 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:

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

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:

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

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

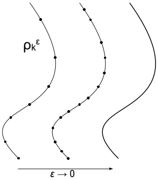
$$Z_{k+1}^\epsilon - Z_k^\epsilon = -\epsilon \nabla F(Z_k^\epsilon).$$

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

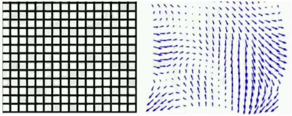


Figure: Image of a diffeomorphism by G. Peyré

- $u : \mathbb{R}^d \to \mathbb{R}$ 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 Z_0.
- Flow of the mirror coordinate.

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

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize Z_0.
- Flow of the mirror coordinate.

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

- Flow of the primal/canonical coordinate.

$$Z_{k+1} - Z_k = -\epsilon \nabla_{x^u} F(Z_k) \qquad \dot{x}_t = -\nabla_{x^u} F(Z_t) = -(\nabla^2 u(Z_t))^{-1} \nabla_x F(Z_t)$$

Mirror gradient flow ODEs

- Mirror gradient flows have two equivalent ODEs. Initialize Z_0.
- Flow of the mirror coordinate.

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

- Flow of the primal/canonical coordinate.

$$Z_{k+1} - Z_k = -\epsilon \nabla_{x^u} F(Z_k) \qquad \dot{x}_t = -\nabla_{x^u} F(Z_t) = -(\nabla^2 u(Z_t))^{-1} \nabla_x F(Z_t)$$

- 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 Z_0.
- Flow of the mirror coordinate.

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

- Flow of the primal/canonical coordinate.

$$Z_{k+1} - Z_k = -\epsilon \nabla_{x^u} F(Z_k) \qquad \dot{x}_t = -\nabla_{x^u} F(Z_t) = -(\nabla^2 u(Z_t))^{-1} \nabla_x F(Z_t)$$

- 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

- $d = 1$, $F(x) = x^2/2$, $Z_0 = 1$.
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.

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

Examples

- $d = 1$, $F(x) = x^2/2$, $Z_0 = 1$.
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.

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

- $u(x) = x^4$. Mirror flow converges in finite time.

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

Examples

- $d = 1,\ F(x) = x^2/2,\ Z_0 = 1.$
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.

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

- $u(x) = x^4$. Mirror flow converges in finite time.

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

- $u(x) = 1/x$. Mirror flow converges polynomially.

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

Examples

- $d = 1$, $F(x) = x^2/2$, $Z_0 = 1$.
- $u(x) = x^2/2$. Usual gradient flow converges exponentially.

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

- $u(x) = x^4$. Mirror flow converges in finite time.

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

- $u(x) = 1/x$. Mirror flow converges polynomially.

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

- 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.,

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

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon\to 0} \rho^\epsilon_{t/\epsilon} = \rho_t$ 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.,

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

Theorem (DKPS '23)

Under regularity assumptions, $\lim_{\epsilon \to 0} \rho^\epsilon_{t/\epsilon} = \rho_t$ 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: $\frac{d}{dt}Z_t = -\nabla F(Z_t)$. 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:

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

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 $\nabla u_\rho(X) \sim e^{-g}$.

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:

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

  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 $\nabla u(x)$.

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

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

- 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.

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

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

- 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: $\dot{x}_t = -(\nabla^2 u(Z_t))^{-1} \nabla_x F(Z_t)$

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.

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

where

- Z_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(Z_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,

$$Z_{t+h} - Z_t - hZ_t = o(|h|)$$

For interpretation

Euclidean gradient flows: Assuming smoothness,

$$Z_{t+h} - Z_t - hZ_t = o(|h|)$$

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.