Why Should We Care About Wasserstein Gradient Flows?

CSML Reading Group
23 April 2026

Rui-Yang Zhang

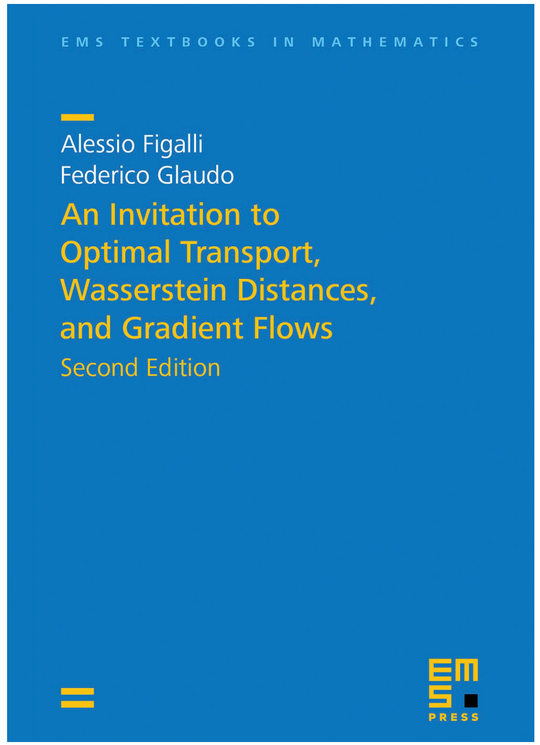
Motivation

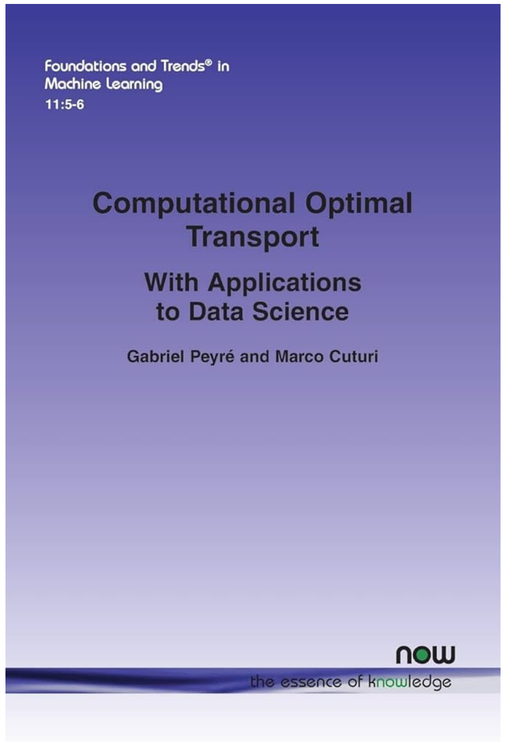
Motivation

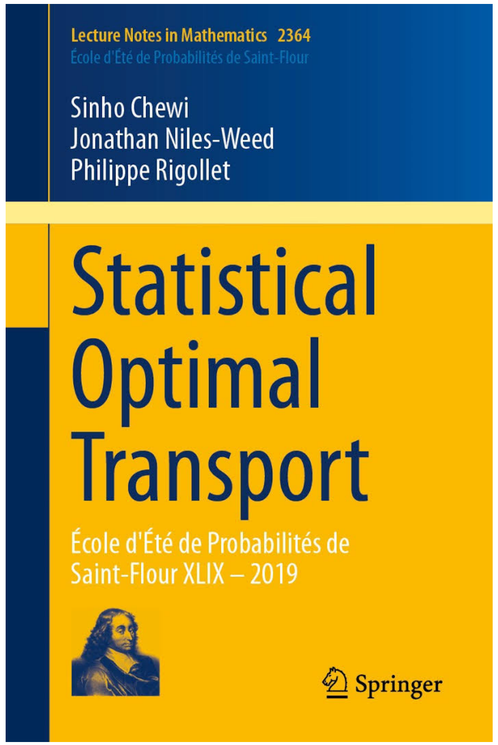
- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.

Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.





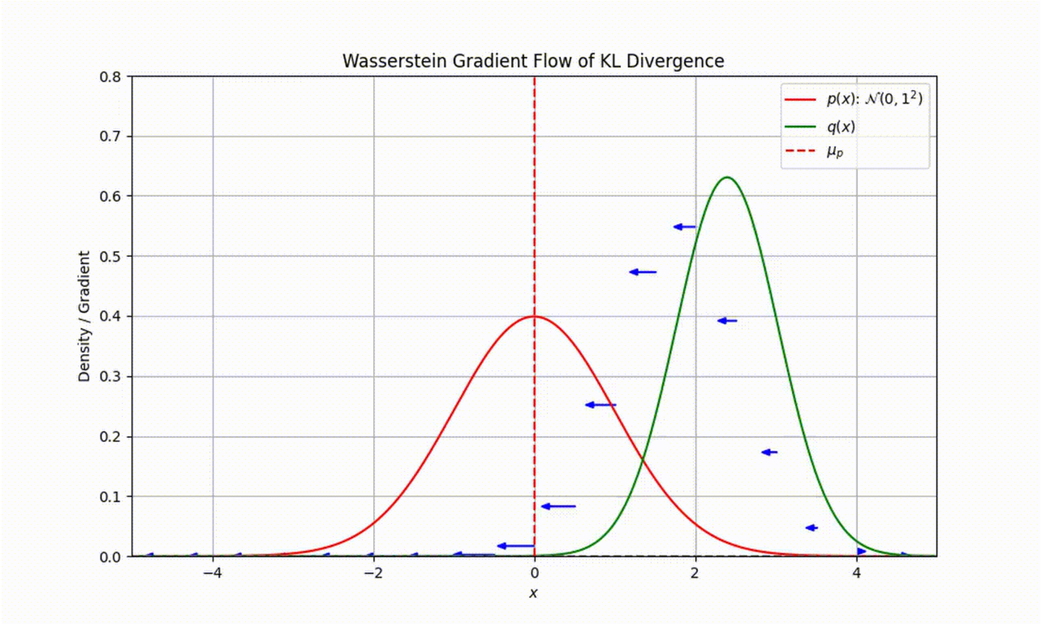


Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.

Motivation

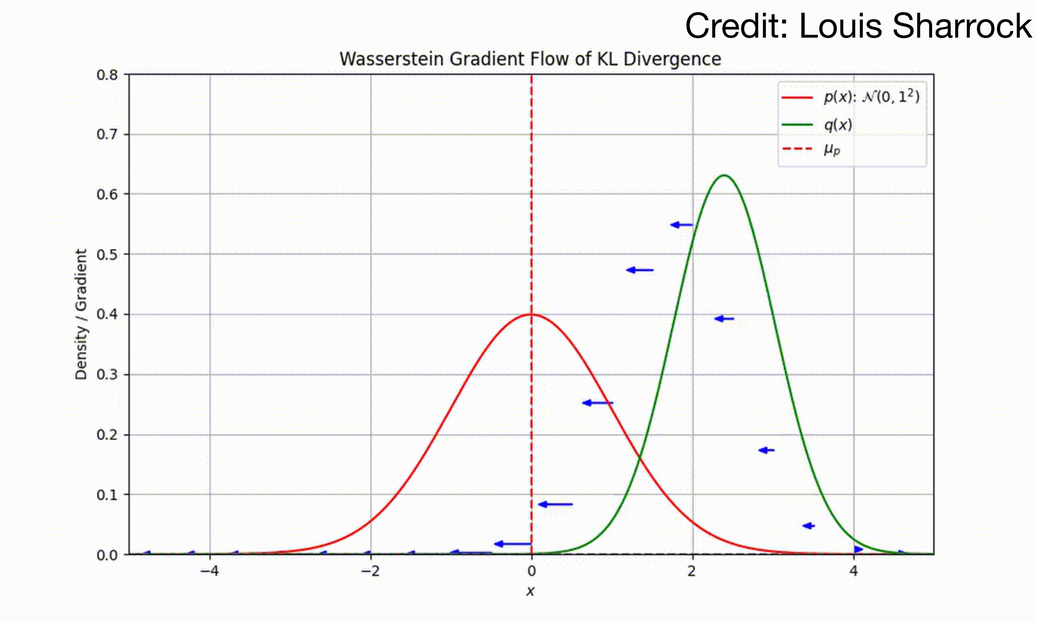
- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.



Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.

Credit: Louis Sharrock



Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.
- There have been some successes, theoretically and methodologically, of WGF for CSML.

Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.
- There have been some successes, theoretically and methodologically, of WGF for CSML.
- A growing community.

Motivation

- Wasserstein gradient flow (and
  is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.
- There have been some succes
  WGF for CSML.
- A growing community.

Gradient Flows For Sampling, Inference, and Learning (In Person)

Date: Friday 01 December 2023, 10.00AM
Location: London

Royal Statistical Society, 12 Errol Street, London EC1Y 8LX

Section Group Meeting

2nd RSS/Turing Workshop on Gradient Flows for Sampling, Inference, and Learning

Date: Monday 24 March 2025, 10.00AM
Location: The Alan Turing Institute

The Alan Turing Institute, British Library, 96 Euston Rd, London NW1 2DB

Section Group Meeting

Motivation

- Wasserstein gradient flow (and computational / statistical optimal transport) is an emerging field in CSML.
- A new perspective on sampling: “Sampling as Optimisation in the Space of Measures”.
- There have been some successes, theoretically and methodologically, of WGF for CSML.
- A growing community.
- This talk focuses on WGF’s relevance for CSML Methodologies.

Why Should We Care About Gradient Flows?

September 2024 · Rui-Yang Zhang, Christopher Nemeth

Optimisation is a fundamental task in modern-day statistics and machine learning. A large set of problems in machine learning and statistics can be easily phrased as an optimisation problem - given some objective function f defined on a domain $\mathcal{X}$, we wish to find a point $x \in \mathcal{X}$ that minimises f (or maximises $-f$). Sometimes, we do not even need to find the global minimum of f, and a sufficiently close local minimum would be good too.

https://chris-nemeth.github.io/blogs/caring_about_gf/

Also shout out to Liam for some feedback.

Bayes Posterior Sampling

Overdamped Langevin

Overdamped Langevin

$$dX_t = \nabla \log \pi(X_t) dt + \sqrt{2} dW_t$$

Overdamped Langevin

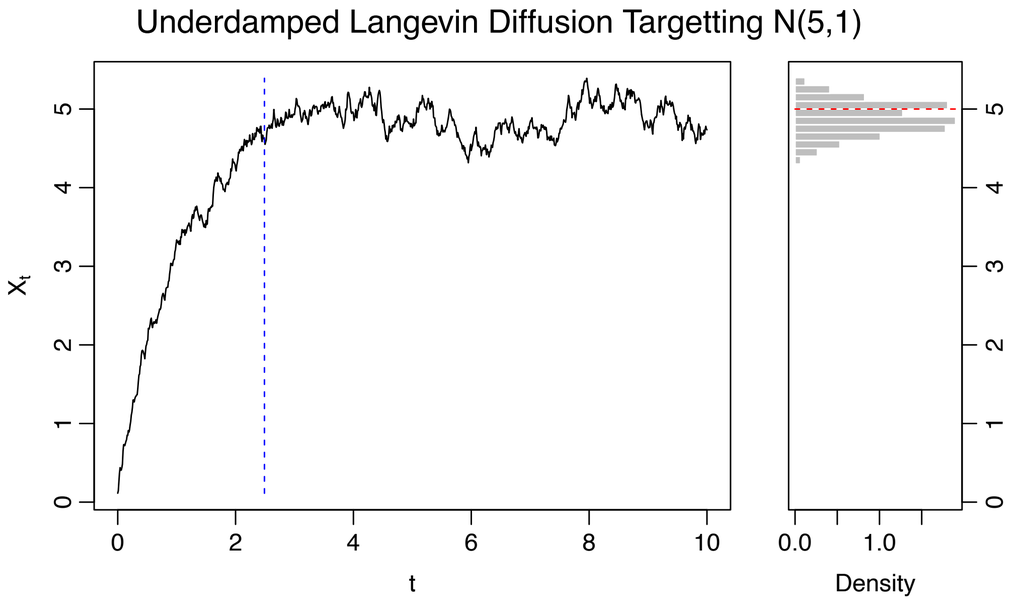
$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

The law of $X_t \,|\, X_0 = x$ converges to π for any initial condition x, for suitably well target distribution π.

Overdamped Langevin

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

The law of $X_t \,|\, X_0 = x$ converges to π for any initial condition x, for suitably well target distribution π.

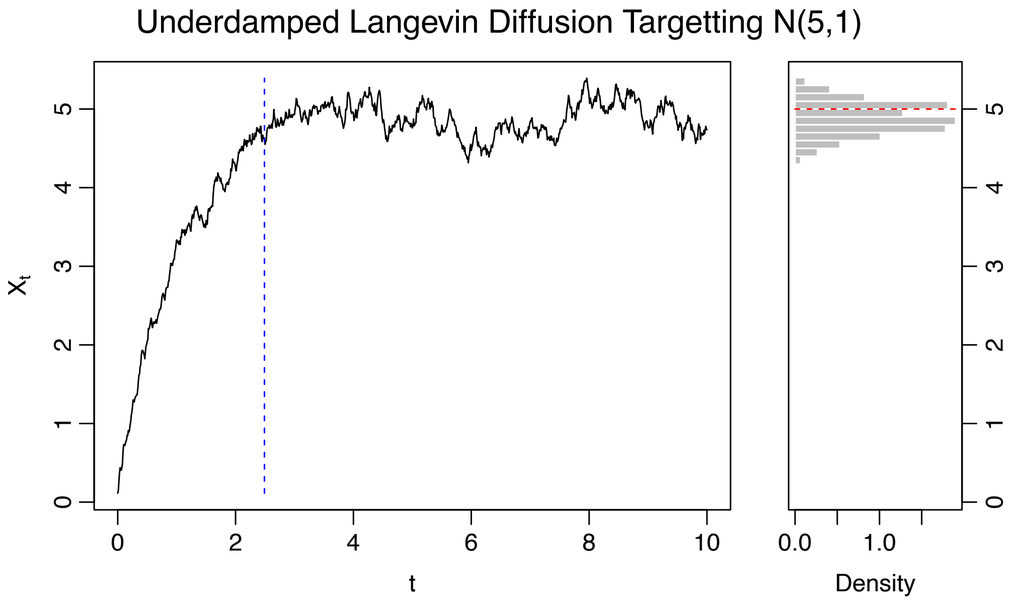


Overdamped Langevin

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

The law of $X_t \,|\, X_0 = x$ converges to π for any initial condition x, for suitably well target distribution π.

This follows from the Fokker-Planck equation.



Overdamped Langevin

Bernoulli **2**(4), 1996, 341–363

Exponential convergence of Langevin distributions and their discrete approximations

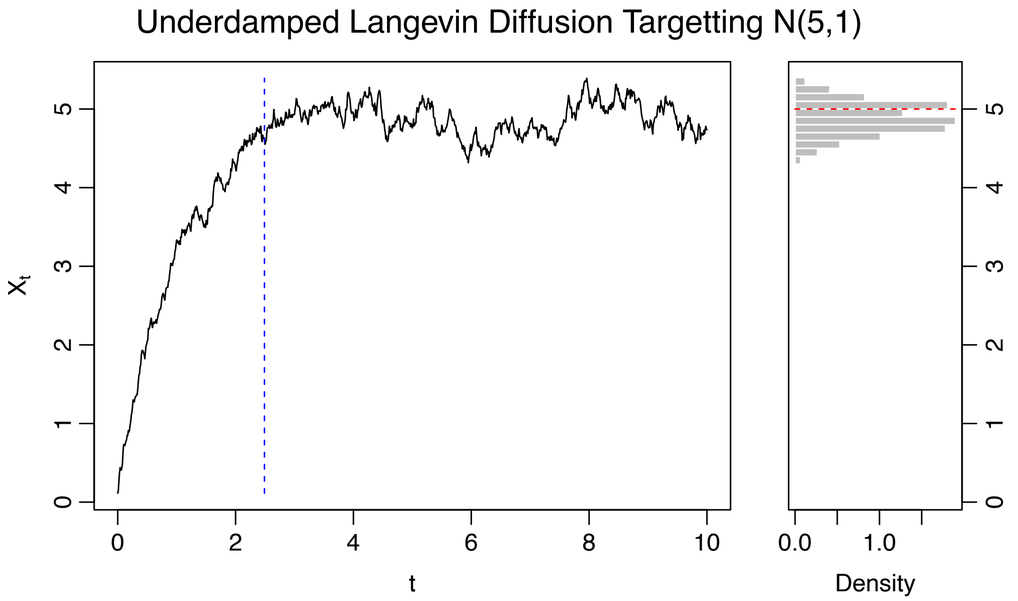
GARETH O. ROBERTS[1]* and RICHARD L. TWEEDIE[2]

[1]*Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, 16 Mill Lane, Cambridge CB2 1SB, UK*
[2]*Department of Statistics, Colorado State University, Fort Collins CO 80523, USA*

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

The law of $X_t \,|\, X_0 = x$ converges to π for any initial condition x, for suitably well target distribution π.

This follows from the Fokker-Planck equation.



$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$\partial_t p(x, t) = -\partial_x[a(x, t)p(x, t)] + \partial^2_{xx}[b^2(x, t)p(x, t)/2].$$

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$\partial_t p(x,t) = -\partial_x[a(x,t)p(x,t)] + \partial^2_{xx}[b^2(x,t)p(x,t)/2].$$

NB: While X_t is a random variable, its law evolve deterministically.

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$\partial_t p(x,t) = -\partial_x[a(x,t)p(x,t)] + \partial^2_{xx}[b^2(x,t)p(x,t)/2].$$

NB: While X_t is a random variable, its law evolve deterministically.

For our overdamped Langevin, we have

$$\partial_t p(x,t) = -\partial_x[\nabla \log \pi(x,t)p(x,t)] + \partial^2_{xx}[p(x,t)].$$

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$\partial_t p(x,t) = -\partial_x[a(x,t)p(x,t)] + \partial_{xx}^2[b^2(x,t)p(x,t)/2].$$

NB: While X_t is a random variable, its law evolve deterministically.

For our overdamped Langevin, we have

$$\partial_t p(x,t) = -\partial_x[\nabla \log \pi(x,t)p(x,t)] + \partial_{xx}^2[p(x,t)].$$

Set $p = \pi$, RHS gives $-\partial_x[\nabla \log \pi \cdot \pi] + \partial_{xx}^2[\pi] = 0$,

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

For a (one-dimensional) SDE $dX_t = a(X_t, t)dt + b(X_t, t)dW_t$,

the Fokker-Planck equation describes the evolution of X_t's law $p(x, t)$,

$$\partial_t p(x,t) = -\partial_x[a(x,t)p(x,t)] + \partial_{xx}^2[b^2(x,t)p(x,t)/2].$$

NB: While X_t is a random variable, its law evolve deterministically.

For our overdamped Langevin, we have

$$\partial_t p(x,t) = -\partial_x[\nabla \log \pi(x,t)p(x,t)] + \partial_{xx}^2[p(x,t)].$$

Set $p = \pi$, RHS gives $-\partial_x[\nabla \log \pi \cdot \pi] + \partial_{xx}^2[\pi] = 0$,

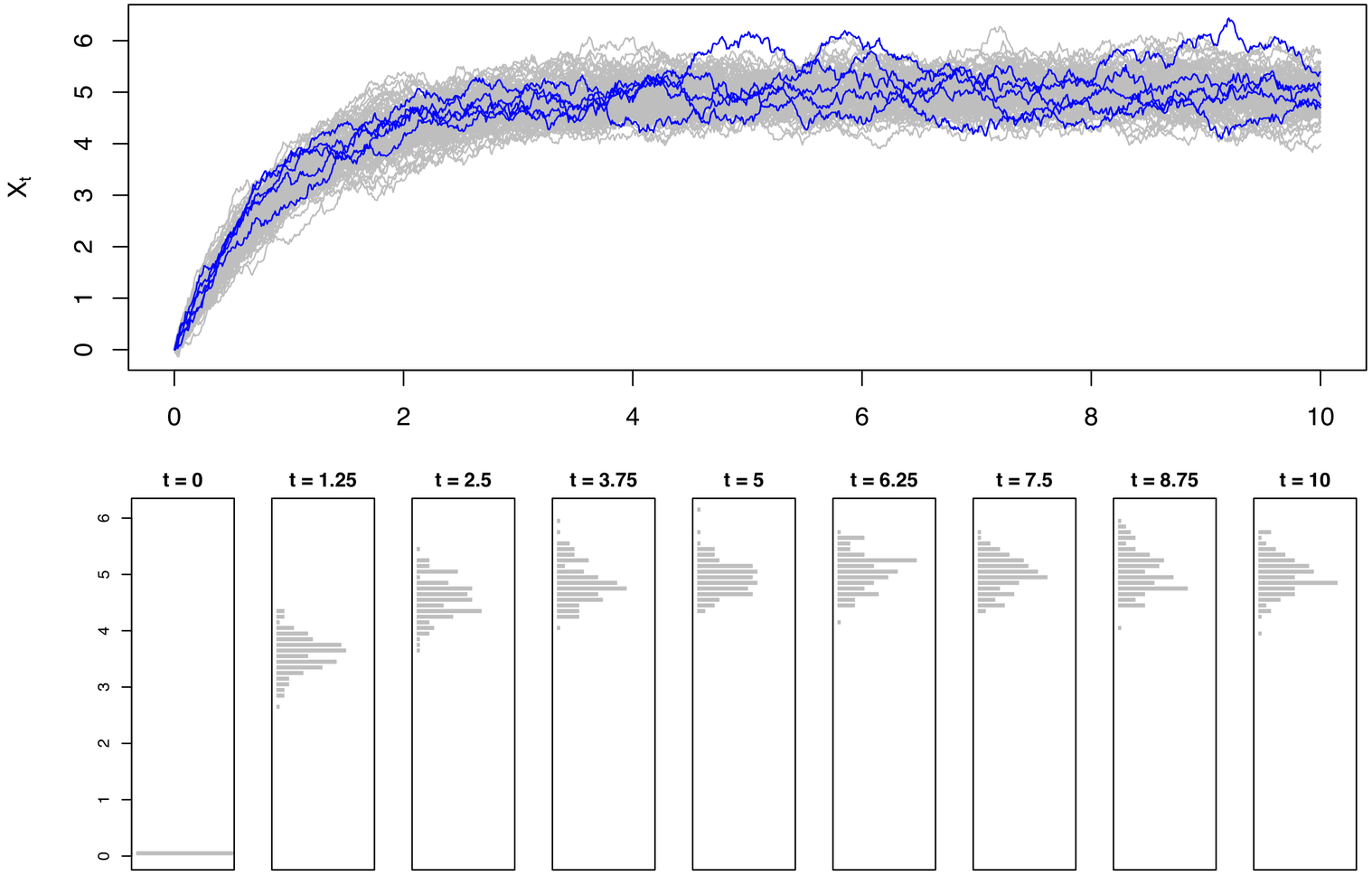
Making π a stationary point of $p_x(t) := p(x, t)$.

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin

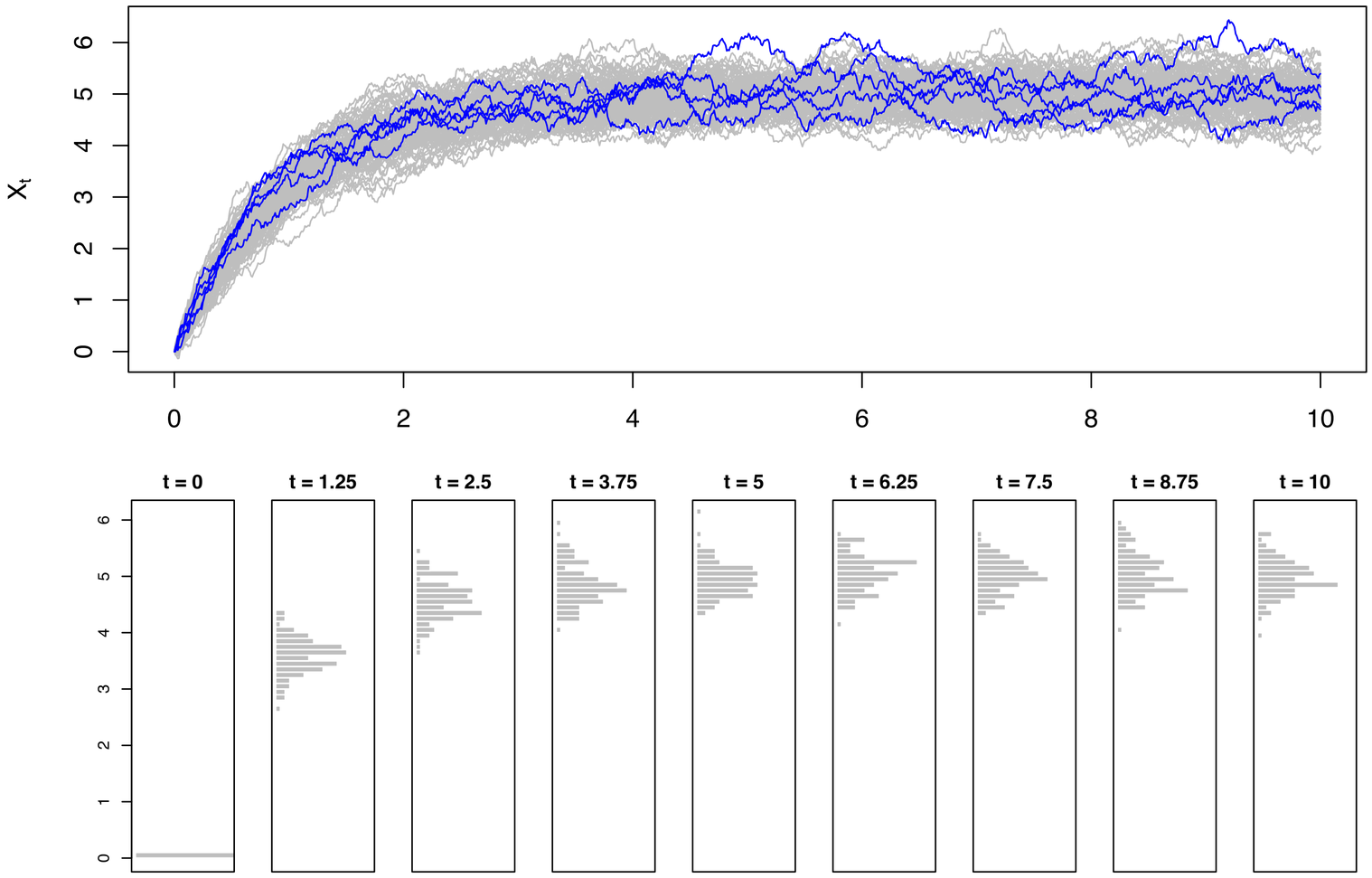
$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin



$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

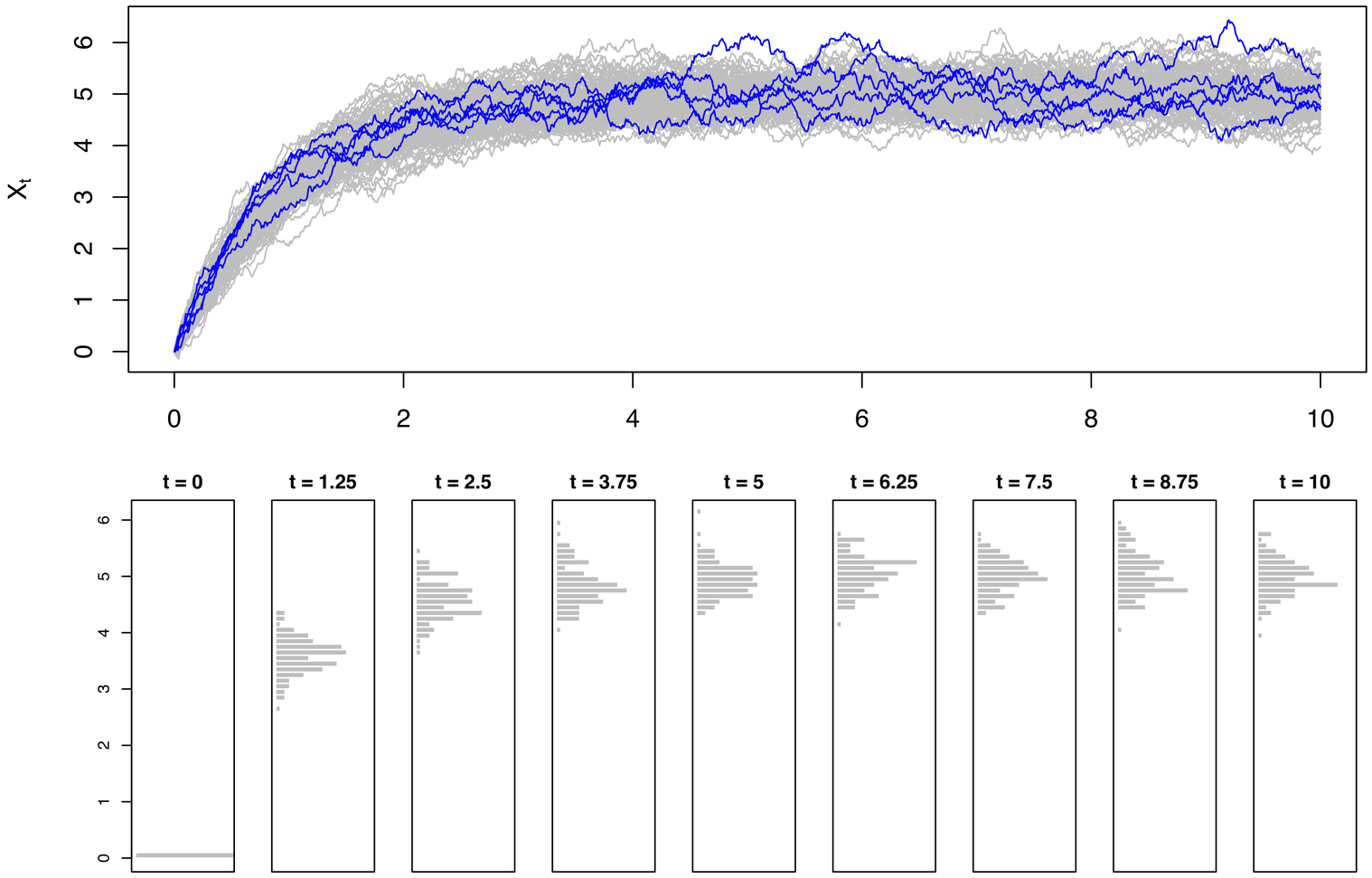
Overdamped Langevin



- The probability measure μ_t of X_t is moving from initial position $\mu_0(x) = \delta(x - 0)$ to a stationary distribution $\mu_\infty(x) = \pi(x)$.

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

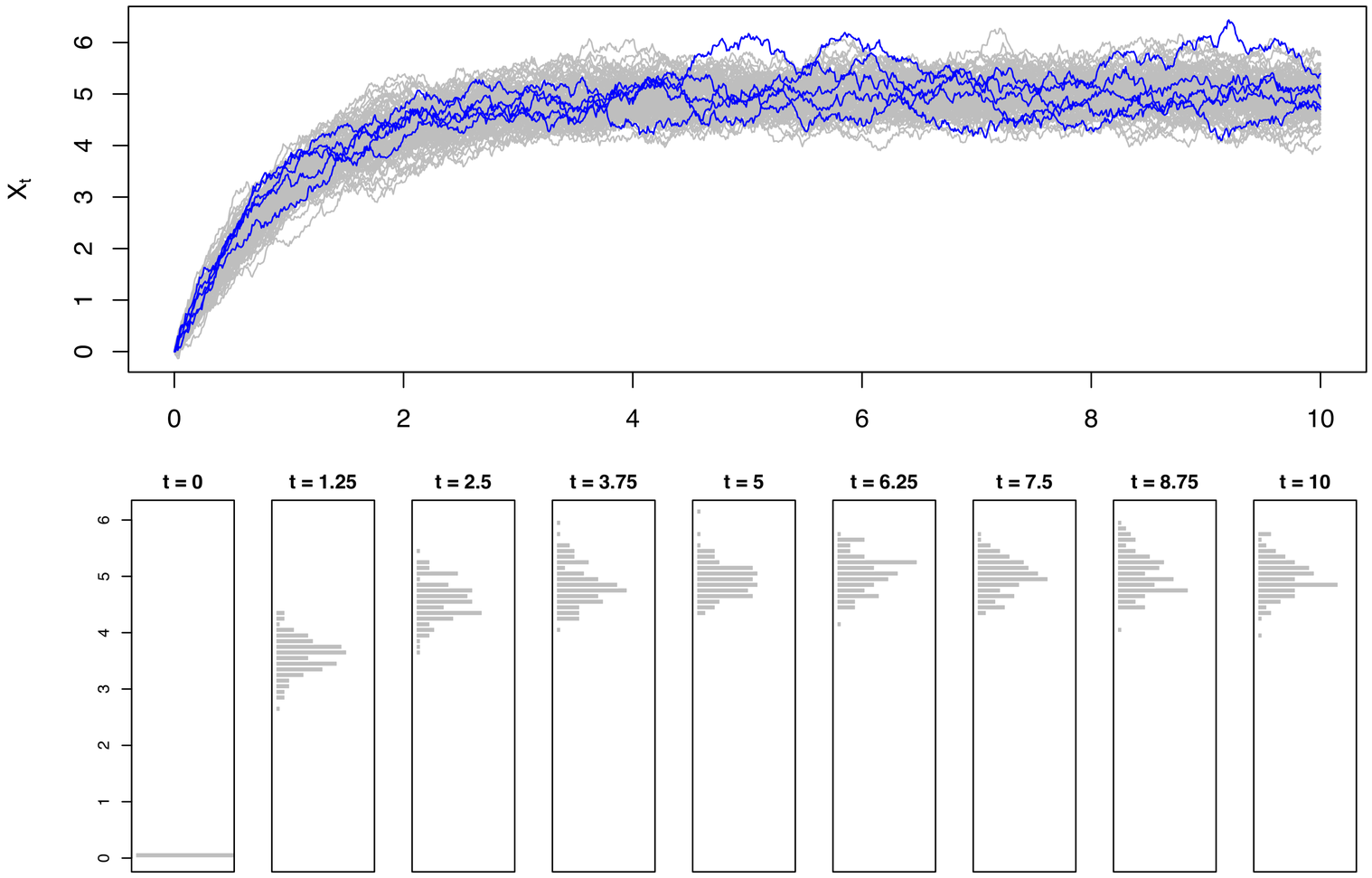
Overdamped Langevin



- The probability measure μ_t of X_t is moving from initial position $\mu_0(x) = \delta(x - 0)$ to a stationary distribution $\mu_\infty(x) = \pi(x)$.
- Almost as if it is '*flowing*' down a '*gradient*' in some space of probability measures from μ_0 to a local optimum π.

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

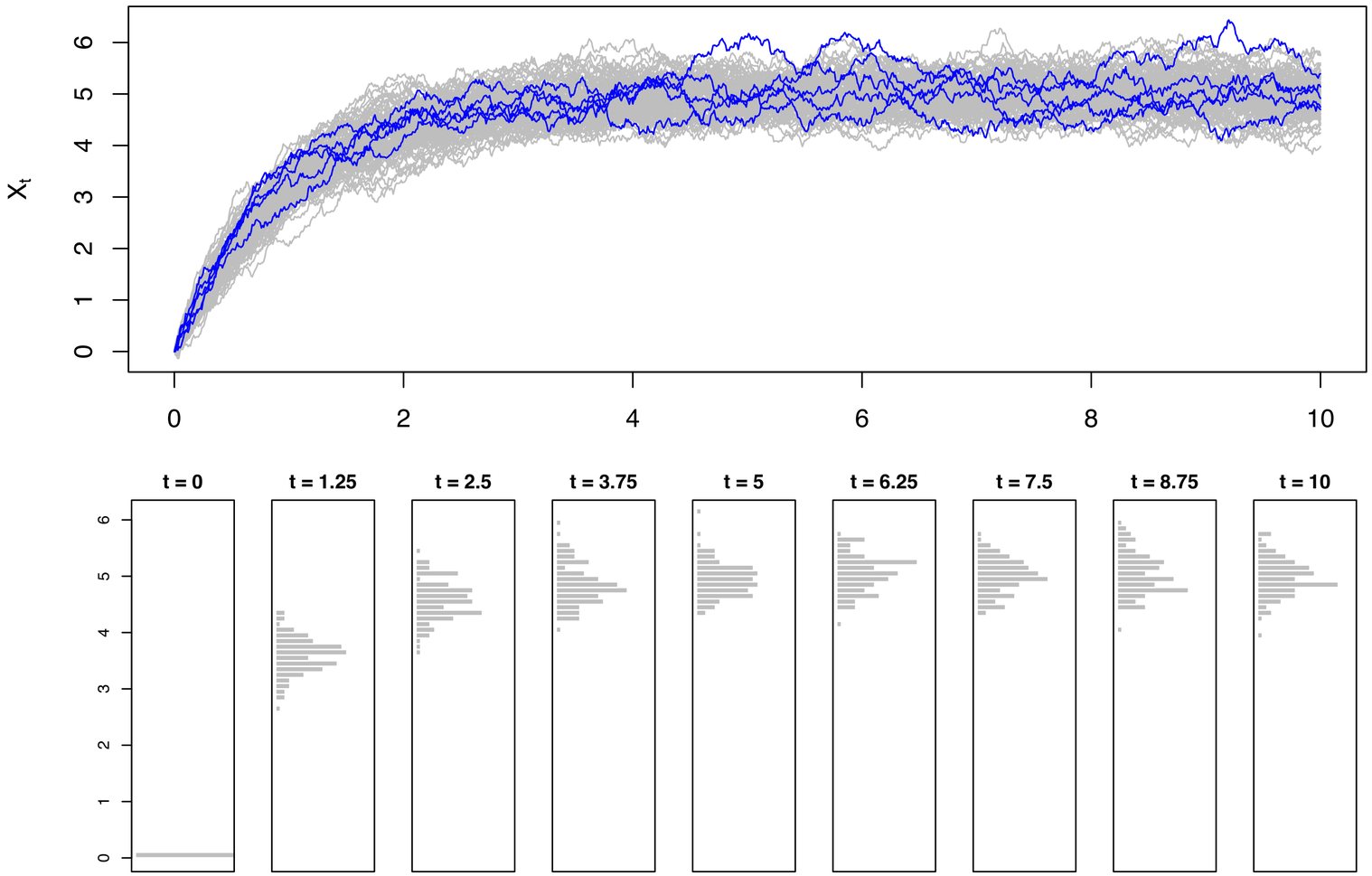
Overdamped Langevin



- The probability measure μ_t of X_t is moving from initial position $\mu_0(x) = \delta(x - 0)$ to a stationary distribution $\mu_\infty(x) = \pi(x)$.
- Almost as if it is '*flowing*' down a '*gradient*' in some space of probability measures from μ_0 to a local optimum π.
- Turns out, the underdamped Langevin is doing a gradient flow in the Wasserstein space of measures following the objective functional $\mathrm{KL}(\cdot \| \pi)$.

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$$

Overdamped Langevin



- The probability measure μ_t of X_t is moving from initial position $\mu_0(x) = \delta(x - 0)$ to a stationary distribution $\mu_\infty(x) = \pi(x)$.
- Almost as if it is '*flowing*' down a '*gradient*' in some space of probability measures from μ_0 to a local optimum π.
- Turns out, the underdamped Langevin is doing a gradient flow in the Wasserstein space of measures following the objective functional $\mathrm{KL}(\cdot \| \pi)$.

The variational formulation of the Fokker--Planck equation
R Jordan, D Kinderlehrer, F Otto
SIAM journal on mathematical analysis, 1998 • SIAM

> The Fokker--Planck equation, or forward Kolmogorov equation, describes the evolution of the probability density for a stochastic process associated with an Ito stochastic differential equation. It pertains to a wide variety of time-dependent systems in which randomness plays a role. In this paper, we are concerned with Fokker--Planck equations for which the drift term is given by the gradient of a potential. For a broad class of potentials, we construct a time discrete, iterative variational scheme whose solutions converge to the

SHOW MORE

☆ Save 99 Cite Cited by 2682 Related articles All 14 versions »

Wasserstein Gradient Flow

Wasserstein Gradient Flow

Wasserstein Gradient Flow

Required Ingredients

Wasserstein Gradient Flow

Required Ingredients

- A space of probability measures $\mathscr{P}(\mathbb{R}^d)$.

Wasserstein Gradient Flow

Required Ingredients

- A space of probability measures $\mathscr{P}(\mathbb{R}^d)$.
- A metric W in the space.

Wasserstein Gradient Flow

Required Ingredients

- A space of probability measures $\mathscr{P}(\mathbb{R}^d)$.
- A metric W in the space.
- A notion of gradient at a point ∇_W.

Wasserstein Gradient Flow

Required Ingredients

- A space of probability measures $\mathscr{P}(\mathbb{R}^d)$.
- A metric W in the space.
- A notion of gradient at a point ∇_W.

GOAL: μ_{t+1}"="$\mu_t - h\nabla_W F(\mu_t)$ and $\mu_\infty \approx \pi := \underset{\mu\in\mathscr{P}(\mathbb{R}^d)}{\operatorname{argmin}} F(\mu)$

Wasserstein Gradient Flow

Required Ingredients

- A space of probability measures $\mathscr{P}(\mathbb{R}^d)$.
- A metric W in the space.
- A notion of gradient at a point ∇_W.

GOAL: μ_{t+1}"="$\mu_t - h\nabla_W F(\mu_t)$ and $\mu_\infty \approx \pi := \underset{\mu\in\mathscr{P}(\mathbb{R}^d)}{\operatorname{argmin}} F(\mu)$

How to move in the space of probability measures?

Wasserstein Gradient Flow

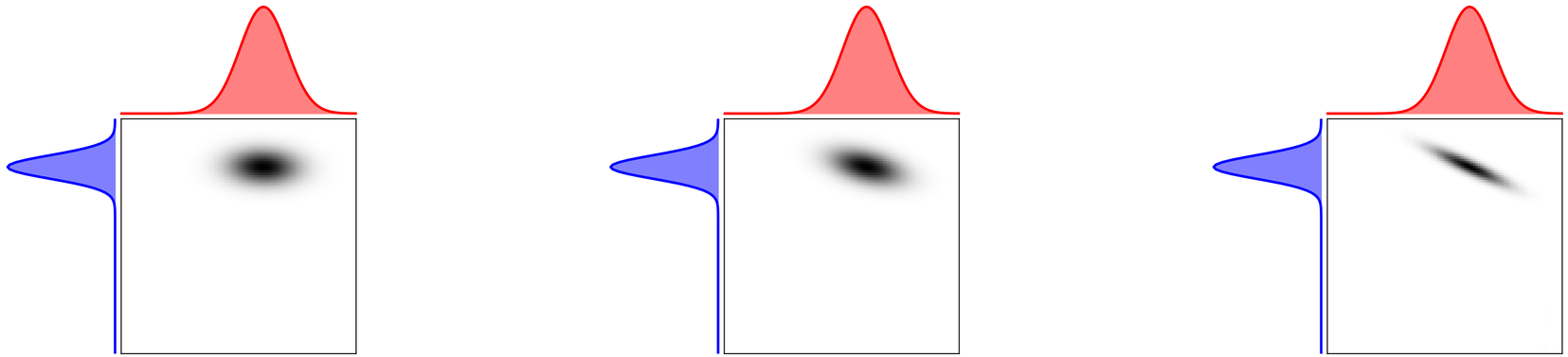
Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.

Wasserstein Gradient Flow

Coupling

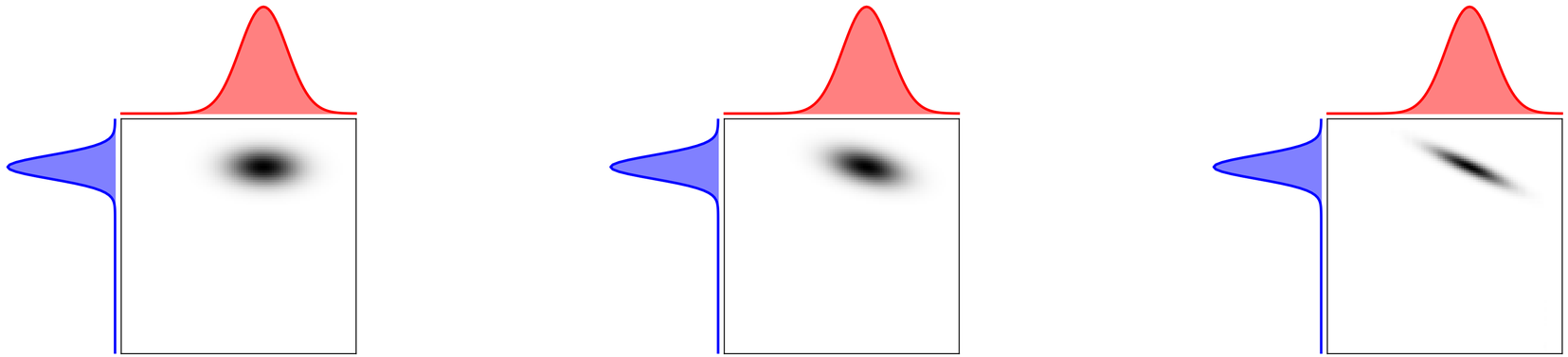
- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.



Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.



Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.

- Such a joint distribution is a **coupling** $\gamma \in \Pi(\mu, \nu)$ of all possible joints.

Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.
- Such a joint distribution is a **coupling** $\gamma \in \Pi(\mu, \nu)$ of all possible joints.
- We can find an ‘optimal’ coupling which minimises a loss, e.g.

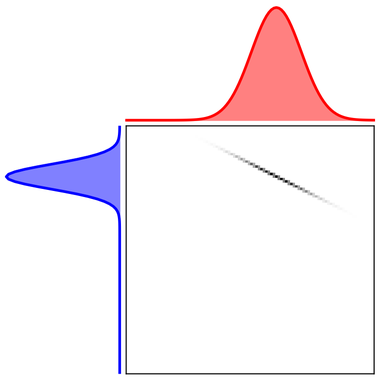
$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x_1 - x_2\|^2 \gamma(dx_1, dx_2).$$

Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.
- Such a joint distribution is a **coupling** $\gamma \in \Pi(\mu, \nu)$ of all possible joints.
- We can find an ‘optimal’ coupling which minimises a loss, e.g.

$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x_1 - x_2\|^2 \gamma(dx_1, dx_2).$$



Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.
- Such a joint distribution is a **coupling** $\gamma \in \Pi(\mu, \nu)$ of all possible joints.
- We can find an ‘optimal’ coupling which minimises a loss, e.g.

$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x_1 - x_2\|^2 \gamma(dx_1, dx_2).$$

- This optimal coupling gives a transport map T_μ^ν such that $\nu = (T_\mu^\nu)_\# \mu$ where $\nu(B) = [(T_\mu^\nu)_\# \mu](B) = \mu[(T_\mu^\nu)^{-1}(B)]$.

Wasserstein Gradient Flow

Coupling

- Consider two distributions $X_1 \sim \mu$ and $X_2 \sim \nu$, we can construct a joint distribution $p(X_1, X_2)$ with the right marginals in many ways.
- Such a joint distribution is a **coupling** $\gamma \in \Pi(\mu, \nu)$ of all possible joints.
- We can find an 'optimal' coupling which minimises a loss, e.g.

$$W_2(\mu, \nu)^2 = \inf_{\gamma \in \Pi(\mu,\nu)} \int \|x_1 - x_2\|^2 \gamma(dx_1, dx_2).$$

- This optimal coupling gives a transport map T_μ^ν such that $\nu = (T_\mu^\nu)_\# \mu$ where $\nu(B) = [(T_\mu^\nu)_\# \mu](B) = \mu[(T_\mu^\nu)^{-1}(B)]$.
- Turns out, this *Wasserstein distance* is a metric for the probability measure space with finite second moment, i.e. $\mathscr{P}_2(\mathbb{R}^d)$.

Wasserstein Gradient Flow

Dynamic Formulation

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

- And this vector field should be ‘optimal’, e.g. uses least kinetic energy in total …

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

- And this vector field should be ‘optimal’, e.g. uses least kinetic energy in total …

$$\int_0^1 \left[\int_x \|v_t(x)\|^2 d\mu_t(x) \right] dt = \int_0^1 \|v_t\|^2_{\mu_t} dt$$

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

- And this vector field should be ‘optimal’, e.g. uses least kinetic energy in total …

$$\int_0^1 \left[\int_x \|v_t(x)\|^2 d\mu_t(x) \right] dt = \int_0^1 \|v_t\|^2_{\mu_t} dt$$

- Thus,

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

- And this vector field should be ‘optimal’, e.g. uses least kinetic energy in total …

$$\int_0^1 \left[\int_x \|v_t(x)\|^2 d\mu_t(x) \right] dt = \int_0^1 \|v_t\|_{\mu_t}^2 dt$$

- Thus,

$$W_2(\mu,\nu)^2 = \inf \left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

Wasserstein Gradient Flow

Dynamic Formulation

- Due to Benamou and Brenier (2000), the Wasserstein distance can be formulated dynamically. [Most efficient flow from one point to another]
- Roughly speaking, we want to construct a vector field $\{v_t\}_{t\in[0,1]}$ that continuously transforms μ to ν as t goes from 0 to 1, i.e. …

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu$$

- And this vector field should be ‘optimal’, e.g. uses least kinetic energy in total …

$$\int_0^1 \left[\int_x \|v_t(x)\|^2 d\mu_t(x) \right] dt = \int_0^1 \|v_t\|_{\mu_t}^2 dt$$

- Thus,

$$W_2(\mu,\nu)^2 = \inf \left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

- The optimal transport map T_μ^ν also gives a (McCann) interpolation between μ and ν: $\mu_t = [(1-t)\mathrm{id} + tT_\mu^\nu]_\# \mu$.

Wasserstein Gradient Flow

Wasserstein Gradient

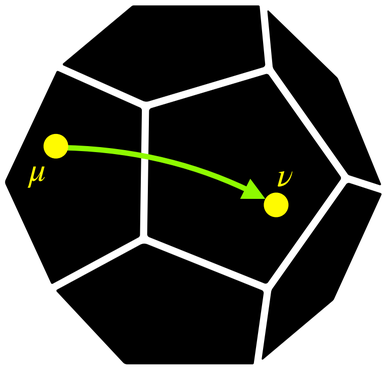
$$W_2(\mu,\nu)^2 = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla\cdot(v_t\mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

Wasserstein Gradient Flow

Wasserstein Gradient

$$W_2(\mu,\nu)^2 = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla\cdot(v_t\mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

Wasserstein distance offers a way to move from one point to another optimally.

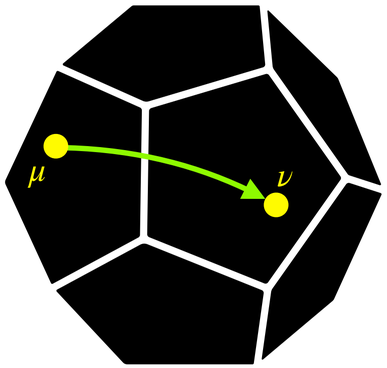


Wasserstein Gradient Flow

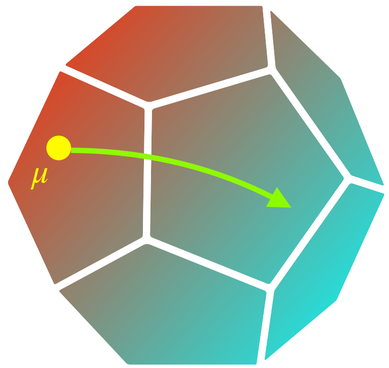
Wasserstein Gradient

$$W_2(\mu,\nu)^2 = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla\cdot(v_t\mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

Wasserstein distance offers a way to move from one point to another optimally.



But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?

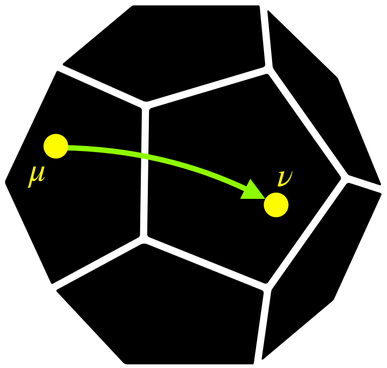


Wasserstein Gradient Flow

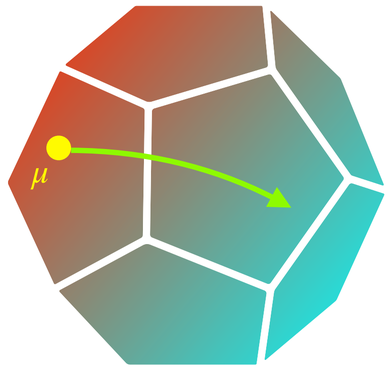
Wasserstein Gradient

$$W_2(\mu,\nu)^2 = \inf\left\{ \int_0^1 \|v_t\|_{\mu_t}^2 dt \,\middle|\, \partial_t \mu_t + \nabla\cdot(v_t\mu_t) = 0,\ \mu_0 = \mu,\ \mu_1 = \nu \right\}$$

Wasserstein distance offers a way to move from one point to another optimally.



But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



GOAL: μ_{t+1}"="$\mu_t - h\nabla_W F(\mu_t)$ and $\mu_\infty \approx \pi := \underset{\mu\in\mathscr{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} F(\mu)$

Wasserstein Gradient Flow

Wasserstein Gradient

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at* μ is given by

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at μ* is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at* μ is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

for measure χ ensuring $\mu + \varepsilon\chi \in \mathscr{P}_2(\mathbb{R}^d)$ for small ε.

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at* μ is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

  for measure χ ensuring $\mu + \varepsilon\chi \in \mathscr{P}_2(\mathbb{R}^d)$ for small ε.

- [Example 1] For potential energy $F(\mu) = \int V(x) d\mu(x)$, we have $\delta F(\mu) = V$.

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at* μ is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

  for measure χ ensuring $\mu + \varepsilon\chi \in \mathscr{P}_2(\mathbb{R}^d)$ for small ε.

- [Example 1] For potential energy $F(\mu) = \int V(x)d\mu(x)$, we have $\delta F(\mu) = V$.

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \lim_{\varepsilon \to 0^+} \varepsilon^{-1} \left[\int V d\mu + \varepsilon \int V d\chi - \int V d\mu \right] = \int V d\chi = \int \delta F(\mu) d\chi$$

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at μ* is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

  for measure χ ensuring $\mu + \varepsilon\chi \in \mathscr{P}_2(\mathbb{R}^d)$ for small ε.

- [Example 1] For potential energy $F(\mu) = \int V(x)d\mu(x)$, we have $\delta F(\mu) = V$.

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \lim_{\varepsilon \to 0^+} \varepsilon^{-1} \left[\int V d\mu + \varepsilon \int V d\chi - \int V d\mu \right] = \int V d\chi = \int \delta F(\mu) d\chi$$

- [Example 2] For internal energy $F(\mu) = \int U(\mu(x))dx$, we have $\delta F(\mu) = U'(\mu)$.

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$, [and further restrict the domain to be a.c. to Lebesgue to $\mu(dx) = \mu(x)dx$], its *first variation at* μ is given by

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \int \delta F(\mu) d\chi$$

  for measure χ ensuring $\mu + \varepsilon\chi \in \mathscr{P}_2(\mathbb{R}^d)$ for small ε.

- [Example 1] For potential energy $F(\mu) = \int V(x)d\mu(x)$, we have $\delta F(\mu) = V$.

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \lim_{\varepsilon \to 0^+} \varepsilon^{-1} \left[\int V d\mu + \varepsilon \int V d\chi - \int V d\mu \right] = \int V d\chi = \int \delta F(\mu) d\chi$$

- [Example 2] For internal energy $F(\mu) = \int U(\mu(x))dx$, we have $\delta F(\mu) = U'(\mu)$.

$$\lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu + \varepsilon\chi) - F(\mu)] = \lim_{\varepsilon \to 0^+} \varepsilon^{-1} \left[\int U(\mu(x) + \varepsilon\chi(x))dx - \int U(\mu(x))dx \right] = \int \lim_{\varepsilon \to 0^+} \varepsilon^{-1} \left[U(\mu + \varepsilon\chi)dx - U(\mu) \right] dx = \int U'(\mu) d\chi = \int \delta F(\mu) d\chi$$

Wasserstein Gradient Flow

Wasserstein Gradient

Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.

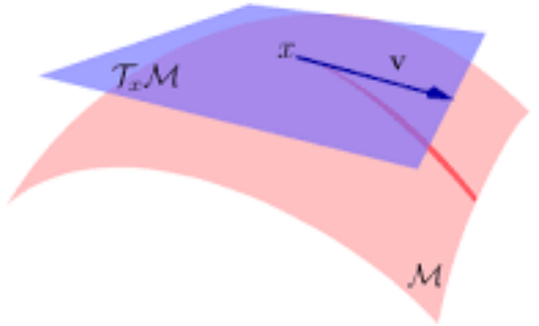
Wasserstein Gradient Flow

Wasserstein Gradient

- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.

Wasserstein Gradient Flow

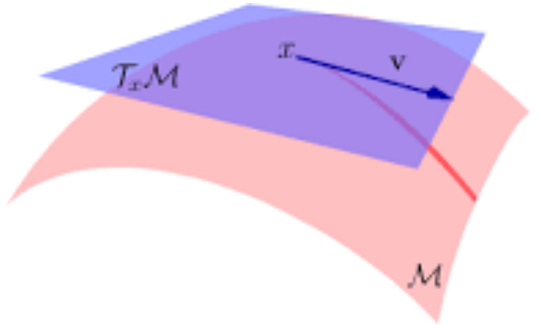
Wasserstein Gradient



- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.

Wasserstein Gradient Flow

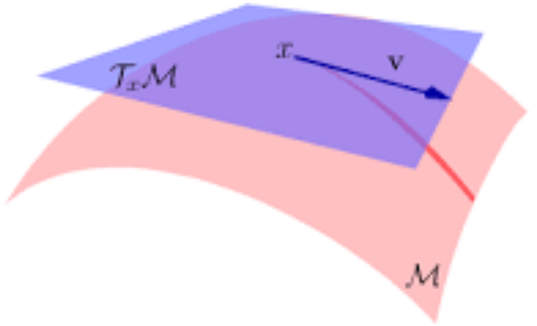
Wasserstein Gradient



- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.
- [Proposition] The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

Wasserstein Gradient Flow

Wasserstein Gradient

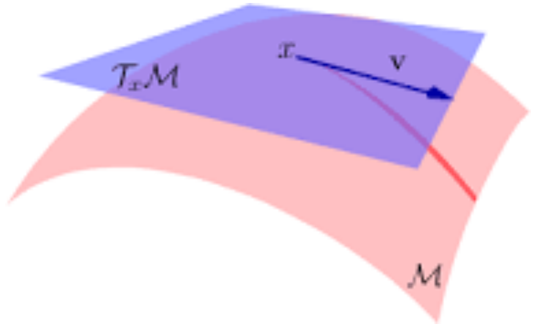


- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.
- [Proposition] The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

[Proof Sketch]

Wasserstein Gradient Flow

Wasserstein Gradient



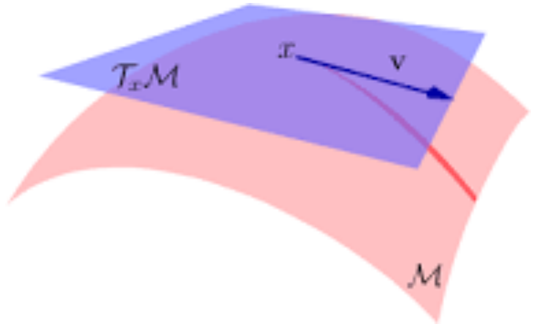
- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.
- [Proposition] The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

[Proof Sketch]

- Using the first variation, by picking $\chi = \partial_t \mu_t$, we have $\partial_t F(\mu_t) = \lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu_t + \varepsilon \partial_t \mu_t) - F(\mu_t)] = \int \delta F(\mu_t) d(\partial_t \mu_t)$.

Wasserstein Gradient Flow

Wasserstein Gradient



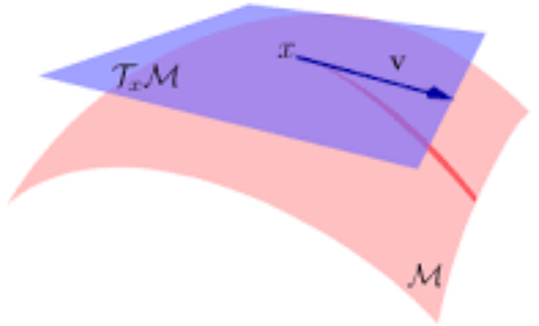
- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.
- [Proposition] The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

[Proof Sketch]

- Using the first variation, by picking $\chi = \partial_t \mu_t$, we have $\partial_t F(\mu_t) = \lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu_t + \varepsilon \partial_t \mu_t) - F(\mu_t)] = \int \delta F(\mu_t) d(\partial_t \mu_t)$.
- Using the continuity equation of $\partial_t \mu_t = -\nabla \cdot (\mu_t v_t)$, we have $\int \delta F(\mu_t) d(\partial_t \mu_t) = -\int \delta F(\mu_t) \nabla \cdot (\mu_t v_t) dx$.

Wasserstein Gradient Flow

Wasserstein Gradient



- Consider a curve of measures $\{\mu_t\}_t$ and corresponding vector field $\{v_t\}_t$.
- The Wasserstein gradient should satisfies $\partial_t F(\mu_t) = \langle \nabla_W F(\mu_t), v_t \rangle_{\mu_t} = \int \langle \nabla_W F(\mu_t), v_t \rangle d\mu_t$ by “chain rule”.
- [Proposition] The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

[Proof Sketch]

- Using the first variation, by picking $\chi = \partial_t \mu_t$, we have $\partial_t F(\mu_t) = \lim_{\varepsilon \to 0^+} \varepsilon^{-1}[F(\mu_t + \varepsilon \partial_t \mu_t) - F(\mu_t)] = \int \delta F(\mu_t) d(\partial_t \mu_t)$.
- Using the continuity equation of $\partial_t \mu_t = -\nabla \cdot (\mu_t v_t)$, we have $\int \delta F(\mu_t) d(\partial_t \mu_t) = -\int \delta F(\mu_t) \nabla \cdot (\mu_t v_t) dx$.
- Using integration by parts and vanishing boundaries, we have $-\int \delta F(\mu_t) \nabla \cdot (\mu_t v_t) dx = \int \nabla \delta F(\mu_t) v_t \mu_t dx = \int \langle \nabla \delta F(\mu_t), v_t \rangle d\mu_t$.

Wasserstein Gradient Flow

Wasserstein Gradient

- The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.

Wasserstein Gradient Flow

Wasserstein Gradient

- The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.
- Actually, this notion of gradient is the one possible gradient option [all Frechet subdifferentials] with the least norm [$L^2(\mu)$-norm].

Wasserstein Gradient Flow

Wasserstein Gradient

- The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.
- Actually, this notion of gradient is the one possible gradient option [all Frechet subdifferentials] with the least norm [$L^2(\mu)$-norm].
- [Example 1] For potential energy $F(\mu) = \int V(x) d\mu(x)$, we have $\delta F(\mu) = V$ and $\nabla_W F(\mu) = \nabla V$.

Wasserstein Gradient Flow

Wasserstein Gradient

- The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.
- Actually, this notion of gradient is the one possible gradient option [all Frechet subdifferentials] with the least norm [$L^2(\mu)$-norm].
- [Example 1] For potential energy $F(\mu) = \int V(x) d\mu(x)$, we have $\delta F(\mu) = V$ and $\nabla_W F(\mu) = \nabla V$.
- [Example 2] For internal energy $F(\mu) = \int U(\mu(x)) dx$, we have $\delta F(\mu) = U'(\mu)$ and $\nabla_W F(\mu) = \nabla U'(\mu)$.

Wasserstein Gradient Flow

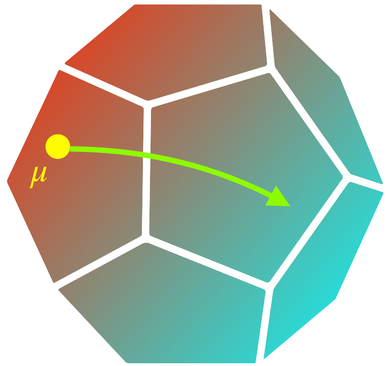
Wasserstein Gradient

- The *Wasserstein gradient* of function F at μ is given by $\nabla_W F(\mu) = \nabla(\delta F(\mu))$.
- Actually, this notion of gradient is the one possible gradient option [all Frechet subdifferentials] with the least norm [$L^2(\mu)$-norm].
- [Example 1] For potential energy $F(\mu) = \int V(x) d\mu(x)$, we have $\delta F(\mu) = V$ and $\nabla_W F(\mu) = \nabla V$.
- [Example 2] For internal energy $F(\mu) = \int U(\mu(x)) dx$, we have $\delta F(\mu) = U'(\mu)$ and $\nabla_W F(\mu) = \nabla U'(\mu)$.
- [Example 3] For KL divergence w.r.t. $\pi \propto \exp[-U]$, $F(\mu) = \mathrm{KL}(\mu \| \pi) = \int U d\mu + \int \mu \log \mu dx + \mathrm{const}$, so $\nabla_W F(\mu) = \nabla U + \nabla \log \mu = \nabla(-\log \pi) + \nabla \log \mu = \nabla \log(\mu/\pi)$.

Wasserstein Gradient Flow

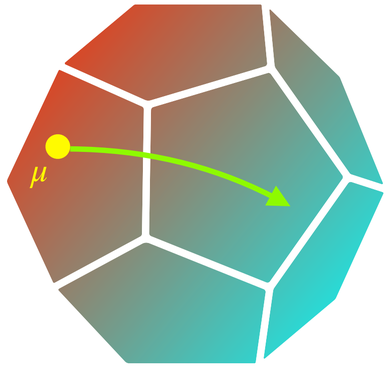
Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

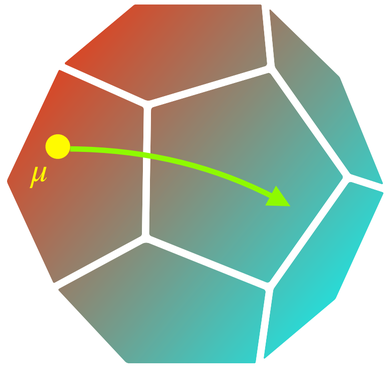
But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?

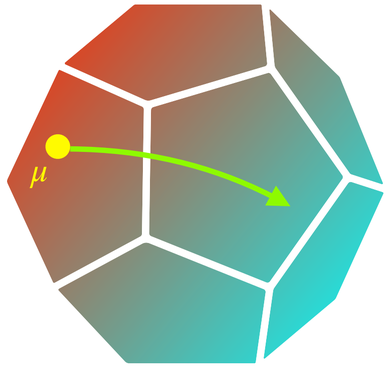


Wasserstein Gradient Flow

[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



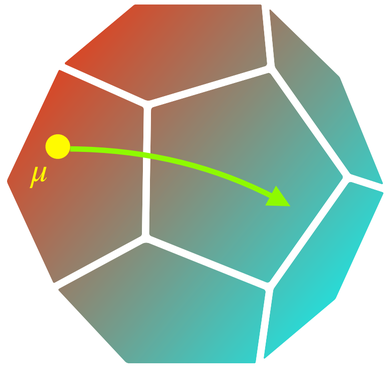
Wasserstein Gradient Flow

continuity equation

[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

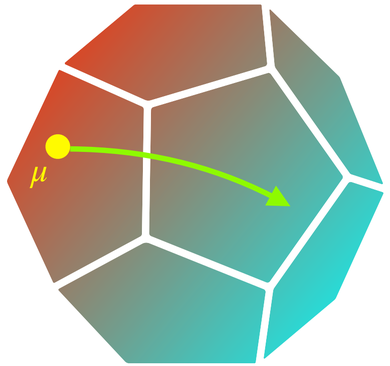
continuity equation

[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

[Lagrangian] $\frac{d}{dt} x_t = -\nabla_W F(\mu_t)(x_t)$

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

continuity equation

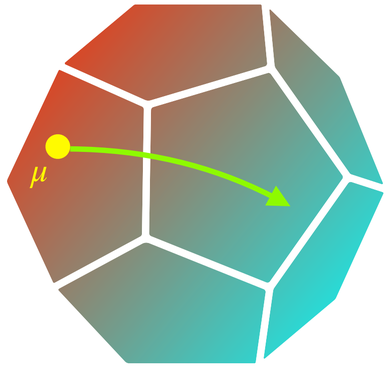
[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

[Lagrangian] $\frac{d}{dt} x_t = -\nabla_W F(\mu_t)(x_t)$

- F is *m-geodesic convex* if for any μ, ν with optimal transport interpolant μ_t we have $F(\mu_t) \leq (1-t)F(\mu) + tF(\nu) - t(1-t)m/2W_2^2(\mu, \nu)$. It is *geodesic convex* for $m = 0$.

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

continuity equation

[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

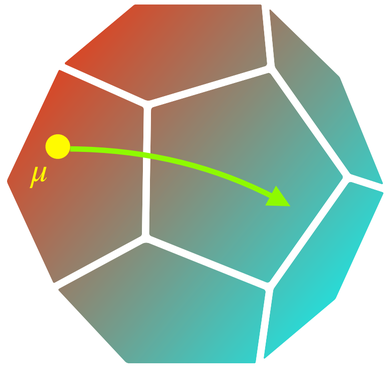
[Lagrangian] $\frac{d}{dt} x_t = -\nabla_W F(\mu_t)(x_t)$

- F is *m-geodesic convex* if for any μ, ν with optimal transport interpolant μ_t we have $F(\mu_t) \leq (1-t)F(\mu) + tF(\nu) - t(1-t)m/2W_2^2(\mu, \nu)$. It is *geodesic convex* for $m = 0$.

- If the functional is *geodesic convex*, the WGF converges to the global (!) optimum. (e.g. KL divergence)

Wasserstein Gradient Flow

But how could we move downwards following a functional $F : \mathscr{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$?



Wasserstein Gradient Flow

continuity equation

[Eulerian] $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$

[Lagrangian] $\frac{d}{dt} x_t = -\nabla_W F(\mu_t)(x_t)$

- F is *m-geodesic convex* if for any μ, ν with optimal transport interpolant μ_t we have $F(\mu_t) \leq (1-t)F(\mu) + tF(\nu) - t(1-t)m/2W_2^2(\mu, \nu)$. It is *geodesic convex* for $m = 0$.
- If the functional is *geodesic convex*, the WGF converges to the global (!) optimum. (e.g. KL divergence)
- If the functional is λ*-geodesic convex*, the WGF converges to the global optimum exponentially (!) quickly. (e.g. KL divergence with log-concave target)

WGF for Posterior Sampling Bayes or Otherwise

Bayes Posterior Sampling

JKO (1998)

Bayes Posterior Sampling

JKO (1998)

- Bayes posterior π can be viewed as the stationary point of a WGF down $F(\,\cdot\,) = \mathrm{KL}(\,\cdot\,\|\pi)$.

Bayes Posterior Sampling

JKO (1998)

- Bayes posterior π can be viewed as the stationary point of a WGF down $F(\,\cdot\,) = \mathrm{KL}(\,\cdot\,\|\pi)$.
- Recall $\nabla_W F(\mu) = \nabla(-\log \pi) + \nabla \log \mu$.

Bayes Posterior Sampling

JKO (1998)

- Bayes posterior π can be viewed as the stationary point of a WGF down $F(\,\cdot\,) = \mathrm{KL}(\,\cdot\,\|\pi)$.
- Recall $\nabla_W F(\mu) = \nabla(-\log \pi) + \nabla \log \mu$.
- The gradient flow is thus

Bayes Posterior Sampling

JKO (1998)

- Bayes posterior π can be viewed as the stationary point of a WGF down $F(\cdot) = \mathrm{KL}(\cdot \| \pi)$.
- Recall $\nabla_W F(\mu) = \nabla(-\log \pi) + \nabla \log \mu$.
- The gradient flow is thus

$$\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t)\mu_t) = \nabla \cdot [(\nabla(-\log \pi) + \nabla \log \mu_t)\mu_t]$$
$$= -\nabla \cdot [\nabla \log \pi \ \mu_t] + \nabla[\nabla \mu_t] = -\nabla \cdot [\nabla \log \pi \ \mu_t] + \nabla^2 \mu_t$$

Bayes Posterior Sampling

JKO (1998)

- Bayes posterior π can be viewed as the stationary point of a WGF down $F(\,\cdot\,) = \mathrm{KL}(\,\cdot\|\pi)$.
- Recall $\nabla_W F(\mu) = \nabla(-\log\pi) + \nabla\log\mu$.
- The gradient flow is thus

$$\partial_t \mu_t = -\nabla\cdot(-\nabla_W F(\mu_t)\mu_t) = \nabla\cdot[(\nabla(-\log\pi) + \nabla\log\mu_t)\mu_t]$$
$$= -\nabla\cdot[\nabla\log\pi\ \mu_t] + \nabla[\nabla\mu_t] = -\nabla\cdot[\nabla\log\pi\ \mu_t] + \nabla^2\mu_t$$

- which coincides with the Fokker-Planck of $dX_t = \nabla\log\pi(X_t)dt + \sqrt{2}dW_t$.

Bayes Posterior Sampling

Wibisono (2018)

Bayes Posterior Sampling

Wibisono (2018)

- Consider Euler-Maruyama discretisation of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$

Bayes Posterior Sampling

Wibisono (2018)

- Consider Euler-Maruyama discretisation of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$

$$X_{t+h/2} = X_t + h\nabla \log \pi(X_t)$$

$$X_{t+h} = X_{t+h/2} + \sqrt{2h}N(0,1)$$

Bayes Posterior Sampling

Wibisono (2018)

- Consider Euler-Maruyama discretisation of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$

$$X_{t+h/2} = X_t + h\nabla \log \pi(X_t)$$

$$X_{t+h} = X_{t+h/2} + \sqrt{2h}N(0,1)$$

- This is a splitting of WGF $\partial_t \mu_t = -\nabla \cdot [\nabla \log \pi \ \mu_t] + \nabla^2 \mu_t$ by first updating half-step with $-\nabla \cdot [\nabla \log \pi \ \mu_t]$ then updating another half-step with $\nabla^2 \mu_t$.

Bayes Posterior Sampling

Wibisono (2018)

- Consider Euler-Maruyama discretisation of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$

$$X_{t+h/2} = X_t + h\nabla \log \pi(X_t)$$

$$X_{t+h} = X_{t+h/2} + \sqrt{2h}N(0,1)$$

- This is a splitting of WGF $\partial_t \mu_t = -\nabla \cdot [\nabla \log \pi \ \mu_t] + \nabla^2 \mu_t$ by first updating half-step with $-\nabla \cdot [\nabla \log \pi \ \mu_t]$ then updating another half-step with $\nabla^2 \mu_t$.
- So ULA is a discretised [Lagrangian formulation of] WGF!

Bayes Posterior Sampling

Wibisono (2018)

- Consider Euler-Maruyama discretisation of $dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t$

$$X_{t+h/2} = X_t + h\,\nabla \log \pi(X_t)$$

$$X_{t+h} = X_{t+h/2} + \sqrt{2h}N(0,1)$$

- This is a splitting of WGF $\partial_t \mu_t = -\nabla \cdot [\nabla \log \pi\ \mu_t] + \nabla^2 \mu_t$ by first updating half-step with $-\nabla \cdot [\nabla \log \pi\ \mu_t]$ then updating another half-step with $\nabla^2 \mu_t$.
- So ULA is a discretised [Lagrangian formulation of] WGF!
- Aggregating many independent copies of the process $\mu_t^N := N^{-1} \sum_{i=1}^{N} \delta_{X_t^{(i)}}$ yields a *mean-field* approximation where $\mu_t^N \to \mu_t$ weakly as $N \to \infty$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

To see this, we have $\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) = \int \left[l_n + \log(q/q_0)\right] q(\theta) d\theta = \int \left[-\log(\exp(-l_n)) + \log(q/q_0)\right] q(\theta) d\theta$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

To see this, we have $\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) = \int \left[l_n + \log(q/q_0)\right] q(\theta) d\theta = \int \left[-\log(\exp(-l_n)) + \log(q/q_0)\right] q(\theta) d\theta$.

$$q^*(\theta) = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n)])] q d\theta = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n) Z^{-1}])] q d\theta = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}(q \| \pi(\theta \mid y))$$

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

To see this, we have $\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) = \int \left[l_n + \log(q/q_0)\right] q(\theta) d\theta = \int \left[-\log(\exp(-l_n)) + \log(q/q_0)\right] q(\theta) d\theta$.

$$q^*(\theta) = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n)])] q d\theta = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n) Z^{-1}])] q d\theta = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}(q \| \pi(\theta \mid y))$$

Since the minimisation is unconstrained, KL can reach zero and thus $q^*(\theta) = \pi(\theta \mid y)$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

To see this, we have $\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) = \int \left[l_n + \log(q/q_0)\right] q(\theta) d\theta = \int \left[-\log(\exp(-l_n)) + \log(q/q_0)\right] q(\theta) d\theta.$

$$q^*(\theta) = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n)])] q d\theta = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n) Z^{-1}])] q d\theta = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}(q \| \pi(\theta \mid y))$$

Since the minimisation is unconstrained, KL can reach zero and thus $q^*(\theta) = \pi(\theta \mid y)$.

Variational inference can be viewed as the above minimisation constrained over a parametric distribution family $\mathscr{P}_\kappa(\theta) \subset \mathscr{P}(\theta)$.

Post-Bayes Posterior Sampling

Variational Reformulation of Bayes Posteriors

Consider prior $q_0(\theta)$, and negative log-likelihood $l(\theta, y_i) = -\log p(y_i \mid \theta)$ and $l_n(\theta, y) = \sum_{i=1}^{n} l(y_i \mid \theta)$.

The corresponding Bayesian posterior is $\pi(\theta \mid y) = q_0(\theta)\exp[-l_n(\theta, y)]/Z$.

This posterior can be obtained variationally as $q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0)\right]$.

To see this, we have $\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) = \int \left[l_n + \log(q/q_0)\right] q(\theta) d\theta = \int \left[-\log(\exp(-l_n)) + \log(q/q_0)\right] q(\theta) d\theta$.

$$q^*(\theta) = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n)])] q d\theta = \underset{q}{\operatorname{argmin}} \int [\log(q/[q_0 \exp(-l_n) Z^{-1}])] q d\theta = \underset{q}{\operatorname{argmin}}\ \mathrm{KL}(q \| \pi(\theta \mid y))$$

Since the minimisation is unconstrained, KL can reach zero and thus $q^*(\theta) = \pi(\theta \mid y)$.

Variational inference can be viewed as the above minimisation constrained over a parametric distribution family $\mathscr{P}_\kappa(\theta) \subset \mathscr{P}(\theta)$.

Nowhere in the derivation do we use the explicit form of $l_n(\theta, y) = \sum_i -\log p(y_i \mid \theta)$, in fact, any [reasonable] *loss* l_n could work …

Post-Bayes Posterior Sampling

Post-Bayes Posterior Sampling

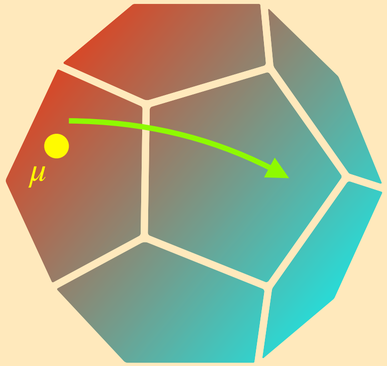
$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) \right] \qquad \pi(\theta \,|\, y) \propto q_0(\theta) \exp[-l_n(\theta, y)]$$

Post-Bayes Posterior Sampling

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) \right] \qquad \pi(\theta \,|\, y) \propto q_0(\theta) \exp[-l_n(\theta, y)]$$

Wasserstein Gradient Flow

$$\pi := \underset{\mu \in \mathscr{P}(\theta)}{\operatorname{argmin}} \; F(\mu) \qquad \partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$$

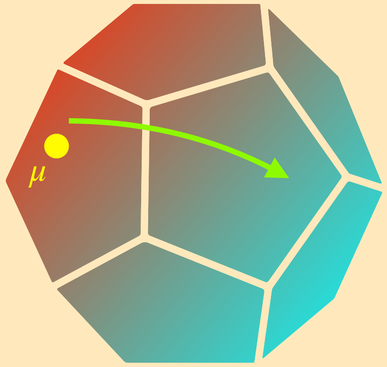


Post-Bayes Posterior Sampling

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) \right] \qquad \pi(\theta \,|\, y) \propto q_0(\theta) \exp[-l_n(\theta, y)]$$

Wasserstein Gradient Flow

$$\pi := \underset{\mu \in \mathscr{P}(\theta)}{\operatorname{argmin}} \; F(\mu) \qquad \partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$$



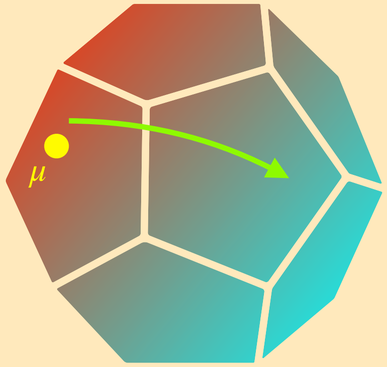
Sure, but it is not necessary

Post-Bayes Posterior Sampling

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathbb{E}_{q(\theta)}[l_n(\theta, y)] + \mathrm{KL}(q \| q_0) \right] \qquad \pi(\theta \,|\, y) \propto q_0(\theta) \exp[-l_n(\theta, y)]$$

Wasserstein Gradient Flow

$$\pi := \underset{\mu \in \mathscr{P}(\theta)}{\operatorname{argmin}} \; F(\mu) \qquad \partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t) \cdot \mu_t)$$



Sure, but it is not necessary

Unless … ?

Post-Bayes Posterior Sampling

Shen et al. (2025)

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

$$\mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

$$\mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

P_n is empirical DGP ; P_q is posterior predictive

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

$$\mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

nonlinear loss

P_n is empirical DGP ; P_q is posterior predictive

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathcal{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

$$\mathcal{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

nonlinear loss

P_n is empirical DGP ; P_q is posterior predictive

$$\mathrm{MMD}(P \| Q) = \sup_{f \in F} \left[\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)] \right]$$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathcal{P}(\theta)}{\operatorname{argmin}} \left[\mathcal{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

$$\mathcal{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

nonlinear loss

P_n is empirical DGP ; P_q is posterior predictive

$$\pi(\theta \mid y) \propto ???$$

$$\mathrm{MMD}(P \| Q) = \sup_{f \in F} \left[\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)] \right]$$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

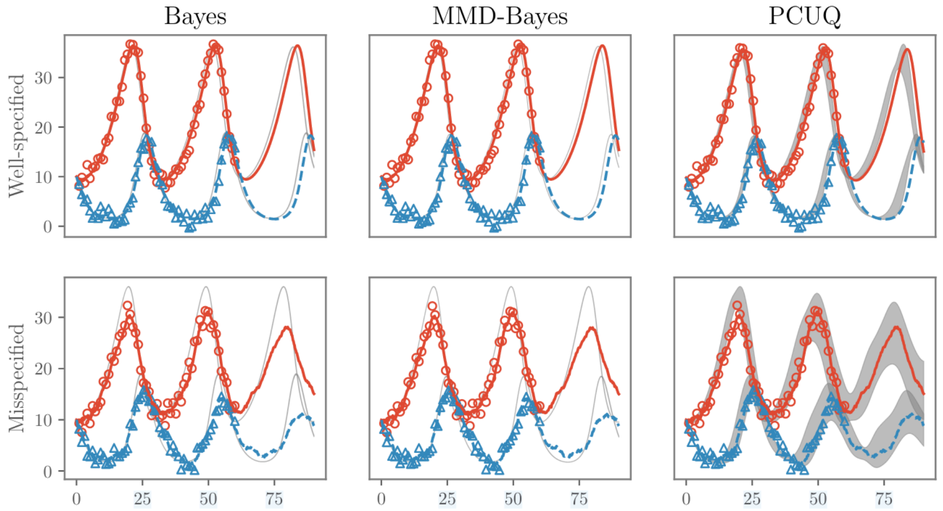
$$\mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

nonlinear loss

P_n is empirical DGP ; P_q is posterior predictive

$$\pi(\theta \mid y) \propto ???$$

$$\mathrm{MMD}(P \| Q) = \sup_{f \in F} \left[\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)] \right]$$



Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right]$$

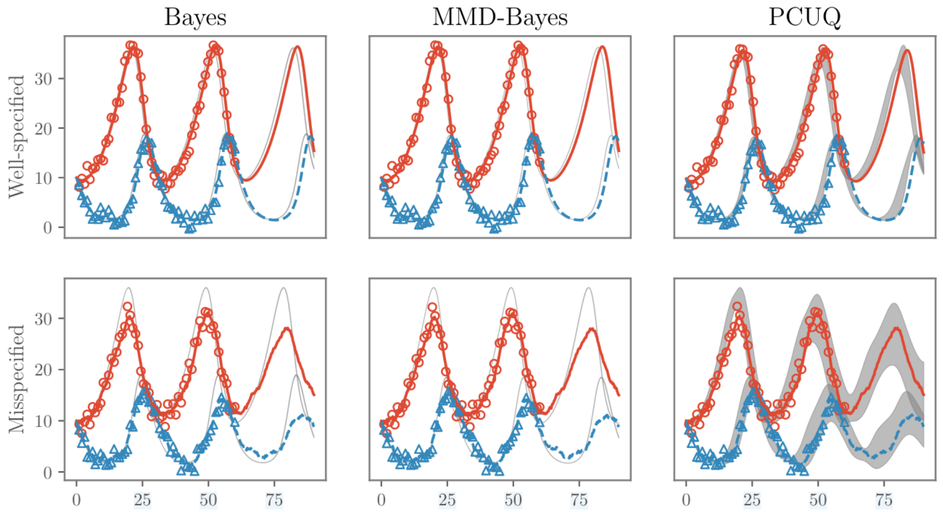
$$\mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

nonlinear loss

P_n is empirical DGP ; P_q is posterior predictive

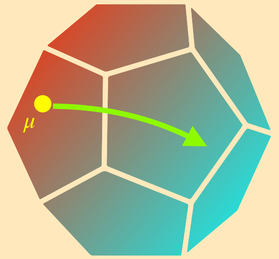
$$\pi(\theta \,|\, y) \propto ???$$

$$\mathrm{MMD}(P \| Q) = \sup_{f \in F} \left[\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)] \right]$$



Wasserstein Gradient Flow

$$\pi := \underset{\mu \in \mathscr{P}(\theta)}{\operatorname{argmin}} \; F(\mu)$$



Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Then, the Wasserstein gradient yields $\nabla_W F(q) = \nabla_W \mathscr{E}(q) + \nabla \log q.$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Then, the Wasserstein gradient yields $\nabla_W F(q) = \nabla_W \mathscr{E}(q) + \nabla \log q.$

So, the gradient flow of the above functional yields $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t)\mu_t) = -\nabla \cdot (-\nabla_W \mathscr{E}(\mu_t)\mu_t) + \nabla^2 \mu_t,$

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Then, the Wasserstein gradient yields $\nabla_W F(q) = \nabla_W \mathscr{E}(q) + \nabla \log q.$

So, the gradient flow of the above functional yields $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t)\mu_t) = -\nabla \cdot (-\nabla_W \mathscr{E}(\mu_t)\mu_t) + \nabla^2 \mu_t,$

which corresponds to, via Fokker-Planck, $d\theta_t = -\nabla_W \mathscr{E}(\mu_t)(\theta_t) + \sqrt{2} dW_t$ for μ_t being the law of θ_t.

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Then, the Wasserstein gradient yields $\nabla_W F(q) = \nabla_W \mathscr{E}(q) + \nabla \log q.$

So, the gradient flow of the above functional yields $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t)\mu_t) = -\nabla \cdot (-\nabla_W \mathscr{E}(\mu_t)\mu_t) + \nabla^2 \mu_t,$

which corresponds to, via Fokker-Planck, $d\theta_t = -\nabla_W \mathscr{E}(\mu_t)(\theta_t) + \sqrt{2} dW_t$ for μ_t being the law of θ_t.

Further calculations and particle approximations for μ_t gives a computable *interacting particle system* SDE.

Post-Bayes Posterior Sampling

Shen et al. (2025)

$$q^*(\theta) = \underset{q \in \mathscr{P}(\theta)}{\operatorname{argmin}} \left[\mathscr{L}(q) + \mathrm{KL}(q \| q_0) \right] \qquad \mathscr{L}(q) = \frac{1}{2} \mathrm{MMD} \left(P_q \parallel P_n \right)^2$$

Rewrite the objective functional as $F(q) = \mathscr{L}(q) + \mathrm{KL}(q \| q_0) = \mathscr{L}(q) - \int \log q_0 dq + \int \log q dq =: \mathscr{E}(q) + \int \log q dq.$

Then, the Wasserstein gradient yields $\nabla_W F(q) = \nabla_W \mathscr{E}(q) + \nabla \log q.$

So, the gradient flow of the above functional yields $\partial_t \mu_t = -\nabla \cdot (-\nabla_W F(\mu_t)\mu_t) = -\nabla \cdot (-\nabla_W \mathscr{E}(\mu_t)\mu_t) + \nabla^2 \mu_t,$

which corresponds to, via Fokker-Planck, $d\theta_t = -\nabla_W \mathscr{E}(\mu_t)(\theta_t) + \sqrt{2} dW_t$ for μ_t being the law of θ_t.

Further calculations and particle approximations for μ_t gives a computable *interacting particle system* SDE.

Thus, WGF offers a way to sample from this post-Bayes posterior that standard methods cannot sample from.

References

References

- Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the Fokker--Planck equation. SIAM journal on mathematical analysis, 29(1), 1-17.
- Benamou, J. D., & Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. Numerische Mathematik, 84(3), 375-393.
- Chewi, S., Niles-Weed, J., & Rigollet, P. (2025). Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX–2019. Springer Nature.
- Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In Conference on learning theory (pp. 2093-3027). PMLR.
- Shen, Z., Knoblauch, J., Power, S., & Oates, C. J. (2025). Prediction-Centric Uncertainty Quantification via MMD. In Proceedings of The 28th International Conference on Artificial Intelligence and Statistics.
- Sharrock, L., & Nemeth, C. (2025). Tuning-Free Sampling via Optimization on the Space of Probability Measures. arXiv preprint arXiv:2510.25315.
- Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. Journal of Machine Learning Research, 23(132), 1-109.

RKHS and Stein Stuff

Maximum Mean Discrepancy

- MMD is an integral probability metric that measures the distance between two probability measures.
- It can be directly computed using samples from the measures, via a *reproducing kernel Hilbert space* (RKHS).
- By definition, we have $\mathrm{MMD}(P\|Q) = \sup_{f \in F} \left[\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{y\sim Q}[f(y)] \right]$ for any function class F.
- We pick F to be a unit ball in the RKHS with kernel k, i.e. H_k.
- The *reproducing property* of H_k yields: $f(x) = \langle f, k(x, \cdot) \rangle_{H_k}$ for $f \in H_k$. Also, define *mean embedding* $\mu_P = \mathbb{E}_{x\sim P}[k(x, \cdot)]$.
- So, we have $\mathbb{E}_{x\sim P}[f(x)] = \mathbb{E}_{x\sim P}[\langle f, k(x, \cdot) \rangle_{H_k}] = \langle f, \mathbb{E}_{x\sim P}[k(x, \cdot)] \rangle_{H_k} = \langle f, \mu_P \rangle_{H_k}$.
- This way, $\mathrm{MMD}(P\|Q) = \sup_{f \in H_k, \|f\| \leq 1} \langle f, \mu_P - \mu_Q \rangle_{H_k}$.
- By Cauchy-Schartz, $\langle f, \mu_P - \mu_Q \rangle_{H_k} \leq \|f\|_{H_k} \|\mu_P - \mu_Q\|_{H_k}$ where equality holds for $f = (\mu_P - \mu_Q)/\|\mu_P - \mu_Q\|_{H_k}$, thus this is the maximising function for MMD.
- Therefore, $\mathrm{MMD}(P\|Q) = \|\mu_P - \mu_Q\|_{H_k} = \sqrt{\mathbb{E}_{x,x'\sim P}[k(x, x')] - 2\mathbb{E}_{x\sim P, y\sim Q}[k(x, y)] + \mathbb{E}_{y,y'\sim Q}[k(y, y')]}$.