CPSC 661: Sampling Algorithms in ML

Andre Wibisono

March 31, 2021

Yale University

Last time

- Wasserstein metric
- Otto calculus (gradient rule)
- Gradient flow of potential energy

Today: Optimization of potential energy

References

- Villani, *Topics in Optimal Transportation*, Springer, 2003
- Villani, *Optimal Transport: Old and New*, Springer, 2008
- Ambrosio, Gigli & Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Springer, 2005
- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018

Dynamics and distributions

Continuity equation

Recall a dynamics in $\mathbb{R}^n$

$$\dot{X}_t = v_t(X_t)$$

induces a dynamics in $\mathcal{P}(\mathbb{R}^n)$ via the *continuity equation*:

$$\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$$

if $X_0 \sim \rho_0$

and $\dot{X}_t = v_t(X_t)$

then $X_t \sim \rho_t$ follows continuity equation.

Dynamics of distributions

Let $\mathcal{P}(\mathbb{R}^n)$ be the space of probability distributions on $\mathbb{R}^n$

A dynamics in $\mathcal{P}(\mathbb{R}^n)$ is a curve $(\rho_t)_{t\geq 0}$ following a vector field ξ

$$\dot{\rho}_t = \xi(\rho_t)$$

Dynamics of distributions

Let $\mathcal{P}(\mathbb{R}^n)$ be the space of probability distributions on $\mathbb{R}^n$

A dynamics in $\mathcal{P}(\mathbb{R}^n)$ is a curve $(\rho_t)_{t\geq 0}$ following a vector field ξ

$$\dot{\rho}_t = \xi(\rho_t)$$

Examples:

1. Continuity equation: $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$ for some $v_t : \mathbb{R}^n \to \mathbb{R}^n$
2. Gradient flow: $\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$ for some $f : \mathbb{R}^n \to \mathbb{R}$
3. Heat equation: $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$

Implementable dynamics

We say a dynamics in $\mathcal{P}(\mathbb{R}^n)$

$$\dot{\rho}_t = \xi(\rho_t)$$

is **implementable** if it arises as the continuity equation of some (possibly stochastic) dynamics in $\mathbb{R}^n$

$$\dot{X}_t = v_t(X_t)$$

Implementable dynamics

We say a dynamics in $\mathcal{P}(\mathbb{R}^n)$

$$\dot{\rho}_t = \xi(\rho_t)$$

is **implementable** if it arises as the continuity equation of some (possibly stochastic) dynamics in $\mathbb{R}^n$

$$\dot{X}_t = v_t(X_t)$$

$\Rightarrow$ Can simulate dynamics of ρ_t in $\mathcal{P}(\mathbb{R}^n)$ via a *sample* $X_t \sim \rho_t$ in $\mathbb{R}^n$

Implementable dynamics

Examples:

1. Continuity equation: $\frac{\partial \rho_t}{\partial t} = -\nabla \cdot (\rho_t v_t)$ is implemented by

$$\dot{X}_t = v_t(X_t)$$

2. Gradient flow: $\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$ is implemented by

$$\dot{X}_t = -\nabla f(X_t)$$

3. Heat equation: $\frac{\partial \rho_t}{\partial t} = \Delta \rho_t$ is implemented by Brownian motion

$$dX_t = \sqrt{2}\, dW_t$$

Optimization dynamics

Some dynamics in $\mathcal{P}(\mathbb{R}^n)$ optimize a functional $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

1. Gradient flow:

$$\dot{\rho}_t = -\mathrm{grad}\, F(\rho_t)$$

2. Gradient descent:

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta\, \mathrm{grad}\, F(\rho_k))$$

3. Proximal method:

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta\, \mathrm{grad}\, F(\rho_{k+1}))$$

Optimization dynamics

Example: For **potential energy**

$$F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$$

for some $f : \mathbb{R}^n \to \mathbb{R}$

the gradient flow is

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

- In Wasserstein W_2 metric with Otto calculus

Potential energy

Space:

$$\mathbb{R}^n$$

Objective function:

$$f : \mathbb{R}^n \to \mathbb{R}$$

Gradient flow:

$$\dot{X}_t = -\nabla f(X_t)$$

$\xrightarrow{\text{implements}}$

Space:

$$\mathcal{P}(\mathbb{R}^n)$$

Potential energy $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

$$F(\rho) = \mathbb{E}_\rho[f]$$

Gradient flow:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

Review: Wasserstein metric

Wasserstein metric

$\mathcal{P} \equiv \mathcal{P}(\mathbb{R}^n)$ is space of probability distributions ρ with $\mathbb{E}_\rho[\|X\|^2] < \infty$

Tangent vector $\phi \in \mathsf{T}_\rho\mathcal{P}$ is a function $\phi \colon \mathbb{R}^n \to \mathbb{R}$ of the form

$$\phi = -\nabla \cdot (\rho \nabla u)$$

for some $u \colon \mathbb{R}^n \to \mathbb{R}$

- Tangent space $\mathsf{T}_\rho\mathcal{P}$ can be parameterized by functions $u \colon \mathbb{R}^n \to \mathbb{R}$ via their gradients ∇u

Wasserstein metric

$\mathcal{P} \equiv \mathcal{P}(\mathbb{R}^n)$ is space of probability distributions ρ with $\mathbb{E}_\rho[\|X\|^2] < \infty$

Tangent vector $\phi \in \mathrm{T}_\rho\mathcal{P}$ is a function $\phi\colon \mathbb{R}^n \to \mathbb{R}$ of the form

$$\phi = -\nabla \cdot (\rho \nabla u)$$

for some $u\colon \mathbb{R}^n \to \mathbb{R}$

- Tangent space $\mathrm{T}_\rho\mathcal{P}$ can be parameterized by functions $u\colon \mathbb{R}^n \to \mathbb{R}$ via their gradients ∇u

Wasserstein metric:

$$\|\phi\|_\rho^2 = \mathbb{E}_\rho[\|\nabla u\|^2] = \int_{\mathbb{R}^n} \rho(x)\, \|\nabla u(x)\|^2\, dx$$

- Generates $W_2(\rho, \nu)^2 = \inf_{\pi \in \Pi(\rho,\nu)} \mathbb{E}[\|X - Y\|^2]$ as geodesic distance

Wasserstein inner product

For $\phi_1, \phi_2 \in \mathrm{T}_\rho\mathcal{P}$ with

$$\phi_1 = -\nabla \cdot (\rho \nabla u_1)$$
$$\phi_2 = -\nabla \cdot (\rho \nabla u_2)$$

Wasserstein inner product:

$$\langle \phi_1, \phi_2 \rangle_\rho = \mathbb{E}_\rho[\langle \nabla u_1, \nabla u_2 \rangle] = \int_{\mathbb{R}^n} \rho(x) \langle \nabla u_1(x), \nabla u_2(x) \rangle \, dx$$

- Follows from *polarization identity*:

$$\langle a, b \rangle = \frac{1}{4} \left(\|a + b\|^2 - \|a - b\|^2 \right)$$

Let $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$ be for some $u \colon \mathbb{R}^n \to \mathbb{R}$

$\Longleftrightarrow \nabla^2 u(x) \succeq -I$

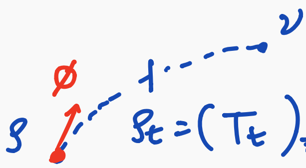
Lemma: Assume $\frac{1}{2}\|x\|^2 + u(x)$ is convex. The **geodesic** from $\rho_0 = \rho$ along direction $\dot{\rho}_0 = -\nabla \cdot (\rho \nabla u)$ is:

$$\rho_t = (T_t)_{\#}\rho$$

for $0 \leq t \leq 1$, where

$$T_t = I + t\nabla u$$

$$T_t(x) = x + t\nabla u(x)$$



$\rho_t = (T_t)_{\#}\rho$

$T_t(x) = (1-t)x + t\nabla\psi(x)$

where $(\nabla\psi)_{\#}\rho = \nu$

this corresponds to

$\phi = -\nabla\cdot(\rho\nabla u)$

where $u(x) = \psi(x) - \frac{1}{2}\|x\|^2$

Exponential map on $\mathcal{P}(\mathbb{R}^n)$

Let $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$ be for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Exponential map: If $\frac{1}{2}\|x\|^2 + u(x)$ is convex

$$\mathrm{Exp}_\rho(\phi) = (I + \nabla u)_{\#}\rho$$

Exponential map on $\mathcal{P}(\mathbb{R}^n)$

Let $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$ be for some $u \colon \mathbb{R}^n \to \mathbb{R}$

Exponential map: If $\frac{1}{2}\|x\|^2 + u(x)$ is convex

$$\mathrm{Exp}_\rho(\phi) = (I + \nabla u)_{\#} \rho$$

- Can *implement* via map $I + \nabla u$ in space:

$$\begin{aligned} \text{If} \quad & X \sim \rho \\ \text{then} \quad & Y = X + \nabla u(X) \sim \mathrm{Exp}_\rho(\phi) \end{aligned}$$

Logarithm map on $\mathcal{P}(\mathbb{R}^n)$

Let $\rho, \nu \in \mathcal{P}(\mathbb{R}^n)$

Let $\nabla\psi$ be optimal transport map from ρ to ν, for some ψ convex

Logarithm map:

$$\mathrm{Log}_\rho(\nu) = -\nabla \cdot (\rho \nabla u) \in T_\rho \mathcal{P}$$

where

$$u(x) = \psi(x) - \frac{1}{2}\|x\|^2$$

so $\nabla u = \nabla\psi - I$

Logarithm map on $\mathcal{P}(\mathbb{R}^n)$

Let $\rho, \nu \in \mathcal{P}(\mathbb{R}^n)$

Let $\nabla\psi$ be optimal transport map from ρ to ν, for some ψ convex

Logarithm map:

$$\mathrm{Log}_\rho(\nu) = -\nabla \cdot (\rho \nabla u)$$

where

$$u(x) = \psi(x) - \frac{1}{2}\|x\|^2$$

so $\nabla u = \nabla\psi - I$

$$\begin{aligned}
\mathrm{Exp}_\rho(\mathrm{Log}_\rho(\nu)) &= \mathrm{Exp}_\rho(-\nabla \cdot (\rho \nabla u)) \\
&= (I + \nabla u)_{\#}\rho \\
&= (\nabla\psi)_{\#}\rho = \nu
\end{aligned}$$

Gradient

The **gradient** of $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ at ρ is

$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \, \nabla \frac{\delta F}{\delta \rho}\right) \quad \in T_\rho \mathcal{P}$$

where $\frac{\delta F}{\delta \rho} \colon \mathbb{R}^n \to \mathbb{R}$ is the L^2 derivative

$$\frac{\delta F}{\delta \rho}(x) = \frac{\partial F(\rho)}{\partial \rho(x)}$$

Gradient

Lemma:

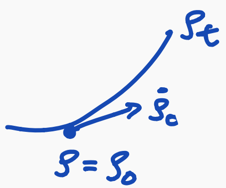
$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \, \nabla \frac{\delta F}{\delta \rho} \right)$$

Proof: For any geodesic ρ_t from $\rho_0 = \rho$ with $\dot\rho_0 = -\nabla \cdot (\rho \nabla u)$

$$\left.\frac{d}{dt}\right|_{t=0} F(\rho_t) = \int_{\mathbb{R}^n} \frac{\delta F(\rho)}{\delta \rho(x)} \, \dot\rho_0(x) \, dx$$

$$= -\int_{\mathbb{R}^n} \frac{\delta F(\rho)}{\delta \rho(x)} \, \nabla \cdot (\rho \nabla u)(x) \, dx$$

$$= \int_{\mathbb{R}^n} \rho(x) \left\langle \nabla \frac{\delta F(\rho)}{\delta \rho(x)}, \nabla u(x) \right\rangle dx$$



on $\mathbb{R}^n$:

if $\dot x_t = v(x_t)$

$$\frac{d}{dt} F(x_t) = \langle \nabla F(x_t), \dot x_t \rangle$$

$$= \langle \nabla F(x_t), v(x_t) \rangle$$

$$= \sum_{i=1}^{n} \frac{\partial F(x_t)}{\partial x_i} v_i(x_t)$$

Write $\operatorname{grad} F(\rho) = -\nabla \cdot (\rho \nabla \psi)$ for some $\psi \colon \mathbb{R}^n \to \mathbb{R}$. Then

$$\begin{aligned} \frac{d}{dt} F(\rho_t)\Big|_{t=0} &= \langle \operatorname{grad} F(\rho),\, \dot{\rho}_0 \rangle_\rho \\ &= \int_{\mathbb{R}^n} \rho(x) \, \langle \nabla \psi(x),\, \nabla u(x) \rangle \; dx \end{aligned}$$

Therefore,

$$\psi = \frac{\delta F}{\delta \rho}$$

$$\operatorname{grad} F(\rho) = -\nabla \cdot \left(\rho \nabla \frac{\delta F}{\delta \rho} \right)$$

$\square$

Potential energy

Potential energy

$$\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \quad F(\rho) = \mathbb{E}_\rho[f]$$

Potential energy $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is

$$F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$$

for some function $f\colon \mathbb{R}^n \to \mathbb{R}$

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x)$$

Potential energy

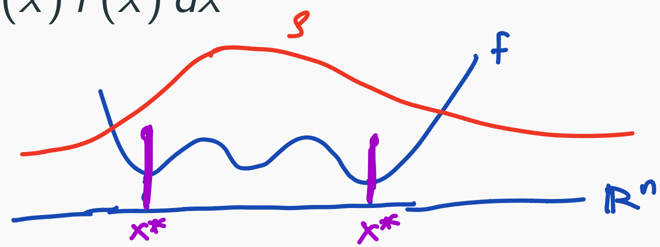
$\delta_{x^*}(x) = \begin{cases} \infty & \text{if } x = x^* \\ 0 & \text{else} \end{cases}$

with $\int_{\mathbb{R}^n} \delta_{x^*}(x)\,dx = 1$

Potential energy $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is

$$F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$$

for some function $f\colon \mathbb{R}^n \to \mathbb{R}$



- Minimized by any probability distribution ν supported on the minimizer set $x^*(f) = \arg\min_{x\in\mathbb{R}^n} f(x)$
- Minimum value is $\min_{\rho\in\mathcal{P}(\mathbb{R}^n)} F(\rho) = \min_{x\in\mathbb{R}^n} f(x)$
- F inherits convexity and smoothness of f

Gradient of potential energy

Potential energy: $F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$

L^2 derivative:

$$\frac{\delta F}{\delta \rho}(x) = f(x) = \frac{\partial F(\rho)}{\partial \rho(x)}$$

Gradient:

$$\operatorname{grad} F(\rho) = -\nabla \cdot (\rho \nabla f)$$

Norm of gradient:

$$\|\operatorname{grad} F(\rho)\|_\rho^2 = \mathbb{E}_\rho[\|\nabla f\|^2]$$

Gradient flow of potential energy

Potential energy: $F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$

Gradient flow:

$$\dot{\rho}_t = -\mathrm{grad}\, F(\rho_t) = \nabla \cdot (\rho_t \nabla f)$$

- Implemented by gradient flow of f:

$$\dot{X}_t = -\nabla f(X_t)$$

Gradient domination of potential energy

Lemma: Assume $f : \mathbb{R}^n \to \mathbb{R}$ is α-gradient dominated:

$$f(x) - \min f \geq \frac{\alpha}{2} \|\nabla f(x)\|^2 \qquad \forall\, x \in \mathbb{R}^n$$

Then potential energy $F(\rho) = \mathbb{E}_\rho[f]$ is also α-gradient dominated:

$$F(\rho) - \min F \geq \frac{\alpha}{2} \|\mathrm{grad}\, F(\rho)\|_\rho^2 \qquad \forall\, \rho \in \mathcal{P}(\mathbb{R}^n)$$

The converse also holds.

Gradient domination of potential energy

Lemma: Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ is α-gradient dominated:

$$f(x) - \min f \geq \frac{\alpha}{2}\|\nabla f(x)\|^2$$

Then potential energy $F(\rho) = \mathbb{E}_\rho[f]$ is also α-gradient dominated:

$F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

$$F(\rho) - \min F \geq \frac{\alpha}{2}\|\mathrm{grad}\, F(\rho)\|_\rho^2$$

The converse also holds.

Proof: Since

$F(\rho) - \min F = \mathbb{E}_\rho[f(X) - \min f] \geq \frac{\alpha}{2}\mathbb{E}_\rho[\|\nabla f(X)\|^2] = \frac{\alpha}{2}\|\mathrm{grad}\, F(\rho)\|_\rho^2.$

Conversely, can choose $\rho \to \delta_x$ for any $x \in \mathbb{R}^n$. $\square$

Convergence rate of potential energy

Theorem: Assume $f : \mathbb{R}^n \to \mathbb{R}$ is α-gradient dominated.

Along gradient flow:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

potential energy $F(\rho) = \mathbb{E}_\rho[f]$ converges exponentially fast:

$$F(\rho_t) - \min F \le e^{-2\alpha t}(F(\rho_0) - \min F)$$

$$\Leftrightarrow \qquad \mathbb{E}_{\rho_t}[f(X_t) - \min f] \le e^{-2\alpha t}\,\mathbb{E}_{\rho_0}[f(X_0) - \min f]$$

Furthermore, can implement via gradient flow of f: $\dot{X}_t = -\nabla f(X_t)$

Gradient descent of potential energy

Gradient descent of potential energy

Lemma: Assume f is L-smooth ($\nabla^2 f(x) \preceq LI$). For $0 < \eta \leq \frac{1}{L}$, the **gradient descent** of potential energy $F(\rho) = \mathbb{E}_\rho[f]$

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta \, \mathrm{grad}\, F(\rho_k))$$

is given by the pushforward map

$$\rho_{k+1} = (I - \eta \nabla f)_\# \rho_k$$

which can be implemented as *gradient descent* of $f(x)$

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Gradient descent of potential energy

Proof: Gradient of potential energy is $\operatorname{grad} F(\rho) = -\nabla \cdot (\rho \nabla f)$.

Since f is L-smooth and $\eta \leq \frac{1}{L}$, $\frac{1}{2}\|x\|^2 - \eta f(x)$ is convex.

Then gradient descent of F is

$$
\begin{aligned}
\rho_{k+1} &= \mathrm{Exp}_{\rho_k}(-\eta \operatorname{grad} F(\rho_k)) \\
&= \mathrm{Exp}_{\rho_k}(-\nabla \cdot (\rho_k(-\eta \nabla f))) \\
&= (I - \eta \nabla f)_{\#}\rho_k
\end{aligned}
$$

This is the pushforward map of the gradient descent of f

$$
x_{k+1} = x_k - \eta \nabla f(x_k)
$$

$$
= (I - \eta \nabla f)(x_k)
$$

$\square$

Proximal method of potential energy

Lemma: Assume f is L-smooth ($-LI \preceq \nabla^2 f(x) \preceq LI$).

For $0 < \eta \leq \frac{1}{L}$, the **proximal method** of potential energy $F(\rho) = \mathbb{E}_\rho[f]$

$$\rho_{k+1} = \arg \min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \left\{ F(\rho) + \frac{1}{2\eta} W_2(\rho, \rho_k)^2 \right\}$$

is implemented by the *proximal method* of $f(x)$

$$x_{k+1} = \arg \min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}$$

$$x_{k+1} = x_k - \eta \nabla f(x_{k+1})$$

$$\Leftrightarrow x_{k+1} = (I + \eta \nabla f)^{-1}(x_k)$$

- Wibisono, *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*, COLT 2018, Appendix E

Recap: Algorithms for potential energy

Objective function

$$f : \mathbb{R}^n \to \mathbb{R}$$

Gradient flow:

$$\dot{X}_t = -\nabla f(X_t)$$

Gradient descent:

$x_k \in \mathbb{R}^n$

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Proximal method:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}$$

Potential energy $F : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

$$F(\rho) = \mathbb{E}_\rho[f]$$

Gradient flow:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

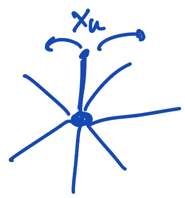
Gradient descent:

$\rho_{k+1} : \mathbb{R}^n \to \mathbb{R}$

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta \operatorname{grad} F(\rho_k))$$

Proximal method:

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \left\{ F(\rho) + \frac{1}{2\eta} W_2(\rho, \rho_k)^2 \right\}$$



$G = (V, E)$

$|V| = n$

$\rho_k \in \mathcal{P}(V)$

RW:

from $x_k \sim \rho_k$:

$x_{k+1} \mid x_k \sim P_{x_k}$

maintain sample on 1 vertex

$\rho_k = (\rho_{k,1}, \dots, \rho_{k,n})$, $\rho_{k,i} \geq 0$, $\sum_{i=1}^{n} \rho_{k,i} = 1$

$\rho_{k+1} = P \cdot \rho_k$

$\rho_k = P^k \cdot \rho_0 \longrightarrow \nu$ as $k \to \infty$

a vector of size n

Displacement convexity

Displacement convexity

$F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is **displacement convex** if it is convex along displacement interpolations:

$$t \mapsto F(\rho_t) \quad \text{is convex}$$

where

$$\rho_t = (T_t)_{\#}\rho_0$$
$$T_t = (1-t)x + t\nabla\phi(x)$$

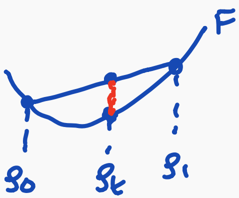
and $\nabla\phi$ is the optimal transport map from ρ_0 to ρ_1

- Displacement interpolation is geodesic in W_2 metric
- Displacement convexity is **geodesic convexity** in W_2 metric
- Similarly for F displacement strongly convex

Displacement convexity

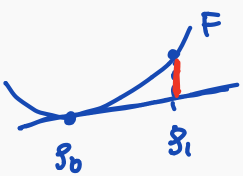
Let $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ be α-strongly displacement convex

1. F is α-strongly convex along displacement interpolation $(\rho_t)_{0 \le t \le 1}$:



$$tF(\rho_1) + (1-t)F(\rho_0) - F(\rho_t) \ge \frac{\alpha}{2} t(1-t) W_2(\rho_0, \rho_1)^2$$

2. If F is differentiable, then



$$F(\rho_1) \ge F(\rho_0) + \langle \mathrm{grad}\, F(\rho_0), \mathrm{Log}_{\rho_0}(\rho_1) \rangle_{\rho_0} + \frac{\alpha}{2} W_2(\rho_0, \rho_1)^2$$

3. If F is twice differentiable, then

$$\mathrm{Hess}\, F(\rho) \succeq \alpha I$$

Hessian

Hessian of $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ at $\rho \in \mathcal{P}(\mathbb{R}^n)$ is a bilinear form

$$\mathrm{Hess}\, F(\rho)\colon \mathrm{T}_\rho\mathcal{P} \times \mathrm{T}_\rho\mathcal{P} \to \mathbb{R}$$

that sends a tangent vector $\phi \in \mathrm{T}_\rho\mathcal{P}$ to the acceleration of F:

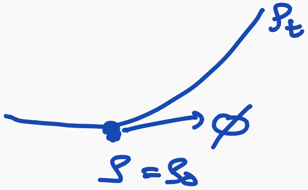
$$\frac{d^2}{dt^2} f(x+tv)$$

$$= \frac{d}{dt} \langle \nabla f(x+tv), v \rangle$$

$$= v^T \nabla^2 f(x+\cancel{t}^{\,0} v)\, v$$

$$(\mathrm{Hess}\, F(\rho))(\phi, \phi) = \left.\frac{d^2}{dt^2}\right|_{t=0} F(\rho_t)$$

where (ρ_t) is geodesic from $\rho_0 = \rho$ along direction $\dot{\rho}_0 = \phi$



Smoothness

$F: \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ is L-**displacement smooth** if

$$\mathrm{Hess}\, F(\rho) \preceq LI$$

- If F is both α-displacement strongly convex and L-displacement smooth, then define condition number

$$\kappa = \frac{L}{\alpha}$$

- Will drop the term "displacement" for convenience

Hessian of potential energy

Lemma: The Hessian of potential energy $F(\rho) = \mathbb{E}_\rho[f]$ sends

$$\phi = -\nabla \cdot (\rho \nabla u) \in \mathsf{T}_\rho \mathcal{P}$$

to

$$\begin{aligned}(\mathrm{Hess}\, F(\rho))(\phi, \phi) &= \mathbb{E}_\rho \left[\langle \nabla u, (\nabla^2 f)\, \nabla u \rangle \right] \\ &= \int_{\mathbb{R}^n} \rho(x)\, \nabla u(x)^\top \nabla^2 f(x) \nabla u(x)\, dx\end{aligned}$$

- See [Villani 2003, §9.1.2]

Convexity of potential energy

Potential energy: $F(\rho) = \mathbb{E}_\rho[f] = \int_{\mathbb{R}^n} \rho(x)\, f(x)\, dx$

Theorem:

1. f is α-strongly convex $\;\Leftrightarrow\;$ F is α-strongly convex
2. f is α-gradient dominated $\;\Leftrightarrow\;$ F is α-gradient dominated
3. f is L-smooth $\;\Leftrightarrow\;$ F is L-smooth

Convexity of potential energy

Proof:

1. Assume f is α-strongly convex: $\nabla^2 f(x) \succeq \alpha I$, which means

$$v^\top \nabla^2 f(x) v \geq \alpha \|v\|^2$$

for all $v \in \mathbb{R}^n$. Then for all $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$

$$\begin{aligned}
(\mathrm{Hess}\, F(\rho))(\phi, \phi) &= \int_{\mathbb{R}^n} \rho(x)\, \nabla u(x)^\top \nabla^2 f(x) \nabla u(x)\, dx \\
&\geq \alpha \int_{\mathbb{R}^n} \rho(x)\, \|\nabla u(x)\|^2\, dx \\
&= \alpha \|\phi\|_\rho^2
\end{aligned}$$

which means $\mathrm{Hess}\, F(\rho) \succeq \alpha I$, so F is α-strongly convex.

Conversely, can take $\rho \to \delta_x$ for any $x \in \mathbb{R}^n$.

Convexity of potential energy

3. Assume f is L-smooth: $\nabla^2 f(x) \preceq LI$, which means

$$v^\top \nabla^2 f(x) v \le L\|v\|^2$$

for all $v \in \mathbb{R}^n$. Then for all $\phi = -\nabla \cdot (\rho \nabla u) \in \mathrm{T}_\rho \mathcal{P}$

$$\begin{aligned}
(\mathrm{Hess}\, F(\rho))(\phi, \phi) &= \int_{\mathbb{R}^n} \rho(x)\, \nabla u(x)^\top \nabla^2 f(x) \nabla u(x)\, dx \\
&\le L \int_{\mathbb{R}^n} \rho(x)\, \|\nabla u(x)\|^2\, dx \\
&= L\|\phi\|_\rho^2
\end{aligned}$$

which means $\mathrm{Hess}\, F(\rho) \preceq LI$, so F is L-smooth.

Conversely, can take $\rho \to \delta_x$ for any $x \in \mathbb{R}^n$.

$\square$

Recap: Convexity of potential energy

Objective function

$$f\colon \mathbb{R}^n \to \mathbb{R}$$

Potential energy $F\colon \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$

$$F(\rho) = \mathbb{E}_\rho[f]$$

Strong convexity:

$$\nabla^2 f(x) \succeq \alpha I$$

$\Longleftrightarrow$

Strong convexity:

$$\operatorname{Hess} F(\rho) \succeq \alpha I$$

Gradient dominated:

$$f(x) - \min f \geq \tfrac{\alpha}{2}\|\nabla f(x)\|^2$$

$\Longleftrightarrow$

Gradient dominated:

$$F(\rho) - \min F \geq \tfrac{\alpha}{2}\|\operatorname{grad} F(\rho)\|_\rho^2$$

Smoothness:

$$\nabla^2 f(x) \preceq LI$$

$\Longleftrightarrow$

Smoothness:

$$\operatorname{Hess} F(\rho) \preceq LI$$

Recap: Algorithms for potential energy

$$f : \mathbb{R}^n \to \mathbb{R}$$

Gradient flow:

$$\dot{X}_t = -\nabla f(X_t)$$

Gradient descent:

$$x_{k+1} = x_k - \eta \nabla f(x_k)$$

Proximal method:

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x) + \frac{1}{2\eta} \|x - x_k\|^2 \right\}$$

$$F(\rho) = \mathbb{E}_\rho[f]$$

Gradient flow:

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot (\rho_t \nabla f)$$

Gradient descent:

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta \, \mathrm{grad}\, F(\rho_k))$$

Proximal method:

$$\rho_{k+1} = \arg\min_{\rho \in \mathcal{P}(\mathbb{R}^n)} \left\{ F(\rho) + \frac{1}{2\eta} W_2(\rho, \rho_k)^2 \right\}$$

Same rates of convergence

Convergence rate of gradient descent

Theorem: Assume $f : \mathbb{R}^n \to \mathbb{R}$ is α-gradient dominated and L-smooth,

Along gradient descent with $\eta = 1/L$:

$\kappa = \frac{L}{\alpha}$

$$\rho_{k+1} = \mathrm{Exp}_{\rho_k}(-\eta \, \mathrm{grad} \, F(\rho_k))$$

potential energy $F(\rho) = \mathbb{E}_\rho[f]$ converges exponentially fast:

$$F(\rho_k) - \min F \leq \left(1 - \tfrac{1}{\kappa}\right)^k (F(\rho_0) - \min F)$$

$$\mathbb{E}_{\rho_k}[f(x_k) - \min f] \leq \left(1 - \tfrac{1}{\kappa}\right)^k \mathbb{E}_{\rho_0}[f(x_0) - \min f]$$

Can implement via gradient descent of f: $x_{k+1} = x_k - \eta \nabla f(x_k)$