CSC/DSCC 574 Continuous Algorithms for Optimization and Sampling Lecture 10

Wasserstein Space

Lecturer: Jiaming Liang *April 4, 2024*

1 Wasserstein space

Recall from Lecture 8 that the Fokker-Planck equation is

$$\frac{\partial \rho_t}{\partial t} = \nabla \cdot \left(\rho_t \nabla \log \frac{\rho_t}{\nu} \right) = \nabla \cdot (\rho_t \nabla f) + \Delta \rho_t. \tag{1}$$

We also claimed that a variational perspective on (1) is that it is the gradient flow for minimizing the relative entropy $\mathsf{KL}(\cdot \,\|\, \nu)$ in the space of probability distributions with the Wasserstein-2 metric. The objective of this lecture is to develop necessary calculus to substantiate this claim, understand the Wasserstein space, and analyze convergence along the Wasserstein gradient flow.

Background on Riemannian geometry A manifold $\mathcal{M}$ is a space that is locally homeomorphic to a Euclidean space. The tangent space $T_p\mathcal{M}$ at $p \in \mathcal{M}$ is an associated vector space containing all possible velocities (tangent bundle) of curves $\gamma(t)$ passing through p. A Riemannian metric is a choice of inner products $p \mapsto \langle \cdot, \cdot \rangle_p$ on the tangent spaces, denpending smoothly on p.

The Riemannian metric induces a distance function (in the sense of metric spaces) via

$$\mathrm{d}(p, q) := \inf \left\{ \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)} \mathrm{d}t \mid \gamma : [0,1] \to \mathcal{M}, \gamma(0) = p, \gamma(1) = q \right\}, \tag{2}$$

where $\dot\gamma(t) \in T_{\gamma(t)}\mathcal{M}$ denotes the tangent vector and $\|\cdot\|_{\gamma(t)}$ is induced by $\langle \cdot, \cdot \rangle_{\gamma(t)}$. If the infimum is achieved by a curve γ, then γ is referred to as a geodesic (a shortest path); if $t \mapsto \|\dot\gamma(t)\|_{\gamma(t)}$ is constant, then it is called a constant-speed geodesic. We say $(\mathcal{M}, \mathrm{d})$ is a Riemannian manifold.

Given a functional $\mathcal{F} : \mathcal{M} \to \mathbb{R}$, the gradient of $\mathcal{F}$ at p is defined to be the unique element $\nabla \mathcal{F}(p) \in T_p\mathcal{M}$ such that for all curves $(\rho(t))_{t \in \mathbb{R}}$ passing through p at time 0 with velocity $v \in T_p\mathcal{M}$ (i.e., $\rho(0) = p$ and $\rho'(0) = v$), it holds that $\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{F}(\rho(t)) \,|_{t=0} = \langle \nabla \mathcal{F}(p), v \rangle_p$.

In the rest of the lecture, we consider a specific Riemannian manifold, namely the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$, and develop the differential calculus in the Wasserstein space.

To develop the analogy of $\mathcal{W}_2$ and metric, we would like to define a metric structure $\langle \cdot, \cdot \rangle_\rho$ on each tangent space $T_\rho\mathcal{P}_2$, depending smoothly on ρ. This metric should define a norm $\|\cdot\|_\rho$ on each $T_\rho\mathcal{P}_2$ such that

$$\mathcal{W}_2(\rho_0, \rho_1)^2 = \inf \left\{ \int_0^1 \left\| \frac{\partial \rho}{\partial t} \right\|_{\rho(t)}^2 \mathrm{d}t \mid \rho : [0,1] \to \mathcal{P}_2(\mathbb{R}^d), \rho(0) = \rho_0, \rho(1) = \rho_1 \right\},$$

where the infimum is taken over all paths connecting ρ_0 and ρ_1.

A tangent vector $v \in T_\rho \mathcal{P}_2$ is $\frac{\mathrm{d}\rho}{\mathrm{d}t}|_{t=0}$ where $\rho(t)$ satisfies $\rho(0) = \rho$. From a fluid dynamics perspective, we want to see the path $\rho(t)$ as the time-evolving density of a set of particles moving continuously from ρ_0 to ρ_1 with a velocity field $c_t(x)$, that is the velocity of a particle is uniquely determined by its position. Now, we have the dynamics of the paticles

$$\dot{x}_t = c_t(x_t). \tag{3}$$

We want to study the evolution of their law $\rho_t(x) = \rho(t, x)$.

Lemma 1. *The continuity equation of (3) is*

$$\frac{\partial \rho_t(x)}{\partial t} + \nabla \cdot (\rho_t(x) \nabla \lambda(x)) = 0$$

for some smooth function λ. The length of the tangent vector is given by

$$\left\| \frac{\partial_t \rho}{\partial t} \right\|^2_{\rho(t)} = \mathbb{E}_{X \sim \rho_t}[\|\nabla \lambda(X)\|^2].$$

Proof. It follows from Theorem 1 of Lecture 8 that the continuity equaiton is

$$\frac{\partial \rho_t}{\partial t} + \nabla \cdot (\rho_t c_t) = 0.$$

The velocity field c_t is not unique: e.g., let w be a vector field with zero divergence (i.e., $\nabla \cdot w = 0$), then $c_t' = c_t + \varepsilon w / \rho_t$ for $\varepsilon \neq 0$ is also admissible since

$$\frac{\partial \rho_t(x)}{\partial t} + \nabla \cdot \left[\rho_t \left(c_t + \varepsilon \frac{w}{\rho_t} \right) \right] = 0. \tag{4}$$

Among all possible vector fields, we want to select the one with lowest kinectic energy, if possible. That is

$$\begin{aligned} &\inf_{c_t} \frac{1}{2} \int \rho_t(x) \|c_t(x)\|^2 \mathrm{d}x \\ &\text{s.t.} \quad -\nabla \cdot (\rho_t(x) c_t(x)) = \frac{\partial \rho_t(x)}{\partial t} \end{aligned}$$

Approach 1: Consider the Lagrange function with multiplier $\lambda(x)$

$$\begin{aligned} &\inf_{c_t} \frac{1}{2} \int \rho_t(x) \|c_t(x)\|^2 \mathrm{d}x + \int \lambda(x) \nabla \cdot (\rho_t(x) c_t(x)) \mathrm{d}x \\ =&\inf_{c_t} \frac{1}{2} \int \rho_t(x) \|c_t(x)\|^2 \mathrm{d}x - \int \langle \nabla \lambda(x), c_t(x) \rangle \rho_t(x) \mathrm{d}x. \end{aligned}$$

The problem boils down to a pointwise minimization and gives

$$c_t(x) = \nabla\lambda(x). \tag{5}$$

Approach 2: As in (4), $c_t + \varepsilon w/\rho_t$ is a feasible solution. Assume that c_t is a minimizing vector field, then we have

$$\int \rho_t \|c_t\|^2 \leq \int \rho_t \left\| c_t + \varepsilon \frac{w}{\rho_t} \right\|^2$$

Using an limiting argument (i.e., $\varepsilon \to 0$), we can show

$$0 = \int \langle c_t, w \rangle,$$

that is c_t is orthoganal to the set of divergence-free vector fields. This means c_t should be a gradient $\nabla\lambda$, so that

$$\int \langle c_t, w \rangle = \int \langle \nabla\lambda, w \rangle = \int \lambda \nabla \cdot w = 0.$$

In both approaches, we prove that (5) holds for some smooth function λ. Moreover, the norm $\|\cdot\|_\rho$ is given by the minimum kinetic engergy

$$\left\| \frac{\partial \rho_t}{\partial t} \right\|^2_{\rho(t)} = \int \rho_t(x) \|c_t(x)\|^2 \mathrm{d}x = \mathbb{E}_{X \sim \rho_t}[\|\nabla\lambda(X)\|^2].$$

□

A sufficient condition for uniqueness is that ρ_t satisfies the Poincaré inequality. Standard results on elliptic PDE assert the existence and uniqueness.

To sum up, one can write

$$\begin{cases} \mathcal{W}_2^2(\rho_0, \rho_1) = \inf \left\{ \int_0^1 \left\| \frac{\partial \rho}{\partial t} \right\|^2_{\rho(t)} \mathrm{d}t \mid \rho : \rho(0) = \rho_0, \quad \rho(1) = \rho_1 \right\} \\ \left\| \frac{\partial \rho}{\partial t} \right\|^2_\rho = \int \rho \|\nabla\lambda\|^2, \quad -\nabla \cdot (\rho \nabla\lambda) = \frac{\partial \rho}{\partial t}. \end{cases} \tag{6}$$

This definition formally endows $\mathcal{P}_2(\mathbb{R}^d)$ a Riemannian metric structure. It comes from a differential formulation of optimal trasport.

2 Otto calculus

By polarization and the norm $\|\cdot\|_\rho$, we can define the inner product of two tangent vectors $\frac{\partial \rho}{\partial t_1} \in T_p\mathcal{P}_2$ and $\frac{\partial \rho}{\partial t_2} \in T_p\mathcal{P}_2$: first, we know there exist smooth functions λ_1 and λ_2 satisfying

$$\frac{\partial \rho}{\partial t_1} = -\nabla \cdot (\rho \nabla\lambda_1), \quad \frac{\partial \rho}{\partial t_2} = -\nabla \cdot (\rho \nabla\lambda_2).$$

Thus, we define $\langle \cdot, \cdot \rangle_p$ as follows

$$\begin{aligned}\left\langle \frac{\partial \rho}{\partial t_1}, \frac{\partial \rho}{\partial t_2} \right\rangle_\rho &= \frac{1}{4}\left(\left\| \frac{\partial \rho}{\partial t_1} + \frac{\partial \rho}{\partial t_2} \right\|_\rho^2 - \left\| \frac{\partial \rho}{\partial t_1} - \frac{\partial \rho}{\partial t_2} \right\|_\rho^2 \right) \\ &= \frac{1}{4}\left(\int \rho \|\nabla \lambda_1 + \nabla \lambda_2\|^2 \mathrm{d}x - \int \rho \|\nabla \lambda_1 - \nabla \lambda_2\|^2 \mathrm{d}x \right) \\ &= \int \rho \langle \nabla \lambda_1, \nabla \lambda_2 \rangle \mathrm{d}x.\end{aligned}$$

Given a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the Wasserstein gradient of $\mathcal{F}$ at p is defined to be the unique element $\mathsf{grad}_W \mathcal{F}(p) \in T_p\mathcal{P}_2$ such that for all curves $(\rho(t))_{t\in\mathbb{R}}$ passing through p at time 0 with velocity $v \in T_p\mathcal{P}_2$ (i.e., $\rho(0) = p$ and $\rho'(0) = v$), it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}\left(\rho(t)\right)|_{t=0} = \langle \mathsf{grad}_W \mathcal{F}(p), v \rangle_p. \tag{7}$$

Definition 1. *The first variation of $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ evaluated at $\rho \in \mathcal{P}_2(\mathbb{R}^d)$ is a function $\frac{\delta \mathcal{F}}{\delta \rho}\colon \mathbb{R}^d \to \mathbb{R}$ that satisfies the following for all $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\lim_{\varepsilon \to 0} \frac{\mathcal{F}(\rho + \varepsilon(\nu - \rho)) - \mathcal{F}(\rho)}{\epsilon} = \left\langle \frac{\delta \mathcal{F}}{\delta \rho}, \nu - \rho \right\rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \frac{\delta \mathcal{F}}{\delta \rho}(x)\left(\nu(x) - \rho(x)\right)\mathrm{d}x\,.$$

Lemma 2. *The Wasserstein gradient of $\mathcal{F}$ is*

$$\mathsf{grad}_W \mathcal{F}(p) = -\nabla \cdot \left(p \nabla \frac{\delta \mathcal{F}}{\delta p} \right). \tag{8}$$

Proof. Let $p \in \mathcal{P}_2(\mathbb{R}^d)$ be given and consider a smooth curve $\rho(t)$ satisfying $\rho(0) = p$ and $\rho'(0) = v$ for some $v \in T_p\mathcal{P}_2$. Since $\mathsf{grad}_W \mathcal{F}(p) \in T_p\mathcal{P}_2$, we know there exist smooth functions f and λ solving the following continuity equaitons

$$\mathsf{grad}_W \mathcal{F}(p) = -\nabla \cdot (p \nabla f), \quad v = -\nabla \cdot (p \nabla \lambda). \tag{9}$$

On the one hand, by the definition of inner production $\langle \cdot, \cdot \rangle_p$, we have

$$\langle \mathsf{grad}_W \mathcal{F}(p), v \rangle_p = \int \langle \nabla f, \nabla \lambda \rangle p \mathrm{d}x.$$

On the other hand, by definitions and integration by parts, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}(\rho(t))|_{t=0} &= \lim_{t\to 0} \frac{\mathcal{F}(p + tv) - \mathcal{F}(p)}{t} = \left\langle \frac{\delta \mathcal{F}}{\delta p}, v \right\rangle_{L^2} \\ &= \int \frac{\delta \mathcal{F}(p)}{\delta p}(x) v(x) \mathrm{d}x = \int \frac{\delta \mathcal{F}}{\delta p}[-\nabla \cdot (p \nabla \lambda)] \mathrm{d}x \\ &= \int \left\langle \nabla \frac{\delta \mathcal{F}}{\delta p}, \nabla \lambda \right\rangle p \mathrm{d}x.\end{aligned}$$

In view of the definition of Wasserstein gradient (7), comparing the above two equations and noting v (and $\nabla\lambda$) is arbitrary, we must have

$$\nabla f = \nabla \frac{\delta \mathcal{F}}{\delta p}.$$

Plugging the above identity into (9), we prove (8). □

Examples of Wasserstein gradients

1) Negative entropy $\mathcal{F}(\rho) = -\mathsf{H}(\rho) = \int \rho \log \rho$,

$$\frac{\delta \mathcal{F}}{\delta \rho} = 1 + \log \rho, \quad \mathsf{grad}_W \mathcal{F}(\rho) = -\nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right) = -\nabla \cdot (\rho \nabla \log \rho) = -\Delta \rho,$$

then the Wasserstein gradient flow of entropy is the heat equation

$$\frac{\partial \rho_t}{\partial t} = -\mathsf{grad}_W \mathcal{F}(\rho_t) = \Delta \rho_t.$$

2) KL divergence (relative entropy) $\mathcal{F}(\rho) = \mathsf{KL}(\rho \,\|\, \nu) = \mathsf{H}_\nu(\rho) = \int \rho \log \frac{\rho}{\nu}$,

$$\frac{\delta \mathcal{F}}{\delta \rho} = 1 + \log \rho - \log \nu, \quad \mathsf{grad}_W \mathcal{F}(\rho) = -\nabla \cdot \left(\rho \nabla \frac{\delta \mathcal{F}}{\delta \rho} \right) = -\nabla \cdot (\rho \nabla \log \rho - \rho \nabla \log \nu) = -\Delta \rho - \nabla \cdot (\rho \nabla f),$$

then the Wasserstein gradient flow of KL divergence is the Fokker-Planck equation

$$\frac{\partial \rho_t}{\partial t} = -\mathsf{grad}_W \mathcal{F}(\rho_t) = \Delta \rho_t + \nabla \cdot (\rho_t \nabla f).$$

3) In general, consider three basic kinds of energies, $\mathcal{F} = \mathcal{U} + \mathcal{V} + \mathcal{W}$, where

- internal energy $\mathcal{U}(\rho) = \int U(\rho(x)) \mathrm{d}x$;
- potential energy $\mathcal{V}(\rho) = \int V(x) \rho(x) \mathrm{d}x$;
- interaction energy $\mathcal{W}(\rho) = \frac{1}{2} \int W(x - y) \rho(x) \rho(y) \mathrm{d}x \mathrm{d}y$.

$$\frac{\delta \mathcal{W}}{\delta \rho} = \int W(\cdot - y) \rho(y) \mathrm{d}y = W * \rho, \quad -\mathsf{grad}_W \mathcal{W}(\rho) = \nabla W * \rho.$$

Then the Wasserstein gradient flow is

$$\frac{\partial \rho_t}{\partial t} = -\mathsf{grad}_W \mathcal{F}(\rho_t) = \nabla \cdot \left[\rho_t \nabla U'(\rho_t) + \rho_t \nabla V + \rho_t (\rho_t * \nabla W) \right].$$

3 Geodesic convexity

From our construction, $\mathcal{W}_2$ is the geodesic length associated to the Riemannian structure (6). The geodesic is the McCann's displacement interpolation: given ρ_0 and ρ_1, the geodesic joining them is

$$\rho_t = [(1-t)\mathrm{Id} + t\nabla\varphi]\#\rho_0$$

where φ is a convex function and $\nabla\varphi$ is the optimal map in the Monge-Kantorovich problem from ρ_0 to ρ_1, i.e., $\nabla\varphi\#\rho_0 = \rho_1$. The corresponding particles evolve as follows

$$X_t = (1-t)X_0 + tX_1.$$

Thus,

$$\left\|\frac{\partial\rho_t}{\partial t}\right\|^2_{\rho(t)} = \mathbb{E}[\|\nabla\lambda(X)\|^2] = \mathbb{E}[\|c_t(X)\|^2] = \mathbb{E}[\|\dot{X}_t\|^2] = \mathbb{E}[\|X_1 - X_0\|^2] = \mathcal{W}_2^2(\rho_0,\rho_1)$$

does not depend on time. So $t \mapsto \rho_t$ is a constant speed geodesic and $\mathcal{W}_2(\rho_0,\rho_t) = t\mathcal{W}_2(\rho_0,\rho_1)$. Moreover, we have

$$c_0(x) = \nabla\varphi(x) - x, \quad c_t = c_0 \circ T_t^{-1},$$

where $T_t = (1-t)\mathrm{Id} + t\nabla\varphi$. This can be understood from both the Eulerian and Lagrangian perspectices: $c_t(x)$ is Eulerian, can be visualized by sitting on the bank of a river and watching the water pass the fixed location; $[c_0 \circ T_t^{-1}](x)$ is Lagrangian, can be visualized as sitting in a boat and drifting down a river. The velocity at x at t is the same as that of the particle starting from some $x(0) = X_0$ passing $x(t) = x$.

Over a Riemannian manifold $\mathcal{M}$, the correct way to define convexity is as follows. Let $\mathcal{M}$ be a Riemannian manifold and let $\mathcal{F}$ be a smooth functional over $\mathcal{M}$. For $\alpha \in \mathbb{R}$, we say that $\mathcal{F}$ is α-geodesically convex if one of the following equivalent conditions hold:

1. For all geodesics $(p_t)_{t\in[0,1]}$ and $t \in [0,1]$,

$$\mathcal{F}(p_t) \le (1-t)\mathcal{F}(p_0) + t\mathcal{F}(p_1) - \frac{\alpha t(1-t)}{2}\,\mathrm{d}(p_0,p_1)^2,$$

   where $\mathrm{d}(\cdot,\cdot)$ is the induced Riemannian distance (2).

2. For all $p, q \in \mathcal{M}$,

$$\mathcal{F}(q) \ge \mathcal{F}(p) + \left\langle \mathsf{grad}\mathcal{F}(p), \log_p(q)\right\rangle_p + \frac{\alpha}{2}\,\mathrm{d}(p,q)^2. \tag{10}$$

   Here, grad denotes the Riemannian gradient and $\log_p(q)$ denotes the logarithmic map $\log_p : \mathcal{M} : T_p\mathcal{M}$ taking q to the tanget vector $v \in T_p\mathcal{M}$ along which the endpoint at time 1 of the constant-speed gedesic starting from p is q. In particular, we have

$$\log_{\rho_0}(\rho_1) = -\nabla\cdot(\rho_0(\nabla\varphi - \mathrm{Id})).$$

3. For all constant-speed geodesics $(p_t)_{t\in[0,1]}$ with tangent vector $v_0 \in T_{p_0}\mathcal{M}$ at time 0, it holds that

$$\partial_t^2 \mathcal{F}(p_t)\big|_{t=0} = \langle \mathsf{Hess}\mathcal{F}(p_0)v_0\,, v_0\rangle_{p_0} \geq \alpha \|v_0\|_{p_0}^2\,. \tag{11}$$

4 Transport inequalities

Theorem 1. *$\nu \propto \exp(-f)$ and f is α-strongly convex, then $\mathsf{KL}(\cdot|\nu)$ is α-geodesically convex along Wasserstein geodesics.*

Proof. This is a well-known result. See for example page 48 of Sinho's book for the proof. □

Theorem 2. *$\nu \propto \exp(-f)$ and f is α-strongly convex, then for all $\rho_0, \rho_1 \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\mathsf{KL}(\rho_1 \,\|\, \nu) \geq \mathsf{KL}(\rho_0 \,\|\, \nu) + \mathbb{E}_{\rho_0}\left[\left\langle \nabla \log \frac{\rho_0(x)}{\nu(x)},\, \nabla\varphi(x) - x \right\rangle\right] + \frac{\alpha}{2}\mathcal{W}_2^2(\rho_0,\rho_1) \tag{12}$$

where $\nabla\varphi\#\rho_0 = \rho_1$.

Proof. It follows from Theorem 1 the definition of geodesic convexity (10) that

$$\mathsf{KL}(\rho_1 \,\|\, \nu) \geq \mathsf{KL}(\rho_0 \,\|\, \nu) + \left\langle \mathsf{grad}_W \mathsf{KL}(\rho_0 \,\|\, \nu),\, \log_{\rho_0}(\rho_1)\right\rangle_{\rho_0} + \frac{\alpha}{2}\mathcal{W}_2^2(\rho_0,\rho_1),$$

which can be reformulated to (12). □

Taking $\rho_0 = \rho$ and $\rho_1 = \nu$ in (12), we have the *HWI inequality*

$$\mathsf{KL}(\rho \,\|\, \nu) \leq \mathcal{W}_2(\rho,\nu)\sqrt{\mathsf{FI}(\rho \,\|\, \nu)}) - \frac{\alpha}{2}\mathcal{W}_2^2(\rho,\nu).$$

Clearly, the above inequality and the Cauchy-Schwarz inequality imply that

$$\mathsf{KL}(\rho \,\|\, \nu) \leq \mathcal{W}_2(\rho,\nu)\sqrt{\mathsf{FI}(\rho \,\|\, \nu)}) - \frac{\alpha}{2}\mathcal{W}_2^2(\rho,\nu) \leq \frac{1}{2\alpha}\mathsf{FI}(\rho|\nu).$$

Hence, the HWI inequality implies LSI, so we prove that if ν is α-SLC then ν satisfies α-LSI. Actually, we have a more general inequality than HWI, i.e.,

$$\mathsf{KL}(\rho_0 \,\|\, \nu) \leq \mathsf{KL}(\rho_1 \,\|\, \nu) + \mathcal{W}_2(\rho_0,\rho_1)\sqrt{\mathsf{FI}(\rho_0|\nu)}) - \frac{\alpha}{2}\mathcal{W}_2^2(\rho_0,\rho_1).$$

Taking $\rho_1 = \nu$, we recover HWI. Taking $\rho_0 = \nu$, we recover the Talagrand inequality (α-TI)

$$\mathcal{W}_2^2(\rho,\nu) \leq \frac{2}{\alpha}\mathsf{KL}(\rho \,\|\, \nu).$$

If ν is log-concave and satisfies λ-TI, then ν also satisfies $(\lambda/4)$-LSI. Indeed, it follows from HWI with $\alpha = 0$

$$\mathsf{KL}(\rho \,\|\, \nu) \leq \mathcal{W}_2(\rho,\nu)\sqrt{\mathsf{FI}(\rho \,\|\, \nu)}) \leq \sqrt{\frac{2}{\lambda}\mathsf{KL}(\rho \,\|\, \nu)}\sqrt{\mathsf{FI}(\rho \,\|\, \nu)}),$$

so

$$\mathsf{KL}(\rho \,\|\, \nu) \le \frac{2}{\lambda}\mathsf{FI}(\rho \,\|\, \nu).$$

Optimization inequalities correspondence

Consider optimization in a smooth Riemannian manifold with an objective function $f : \mathcal{M} \to \mathbb{R}$, we have the following conditions guaranteeing exponential convergence:

(1) f is α-strongly convex: $\forall x \in \mathcal{M}$, $\mathrm{Hess}_x f \succeq \alpha I$, i.e., $(\mathrm{Hess}_x f)(v, v) \ge \alpha \|v\|_x^2$ for all $v \in T_x\mathcal{M}$;

(2) f is α-gradient dominant: $\forall x \in \mathcal{M}$, $\|\mathrm{grad}_x f\|_x^2 \ge 2\alpha(f(x) - \min f)$;

(3) f is α-sufficient growth: $\forall x \in \mathcal{M}$, $f(x) - \min f \ge \frac{\alpha}{2}\mathrm{d}(x, x_*)^2$.

The relationship among the three inequalities is (1) $\implies$ (2) $\implies$ (3).

In the setting of Wasserstein space $(\mathcal{M}, \mathrm{d}) = (\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ and the objective functional being the KL divergence $\mathsf{KL}(\cdot \,\|\, \nu)$, we have corresponding functional/transport inequalities

(a) ν is α-strongly logconcave, i.e., (11);

(b) ν satisfies α-LSI: $\forall \rho \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathsf{FI}(\rho \,\|\, \nu) \ge 2\alpha\mathsf{KL}(\rho \,\|\, \nu)$;

(c) ν satisfies α-TI: $\forall \rho \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathsf{KL}(\rho \,\|\, \nu) \ge \frac{\alpha}{2}\mathcal{W}_2^2(\rho, \nu)$.

The same relationship also holds for the three transport inequalities: (a) $\implies$ (b) $\implies$ (c).

5 New convergence analysis

Now we have develop enough mathematics for Wasserstein space. We would like to study the convergence along Langevin dynamics from the Wasserstein gradient flow perspective. The following result is the same as Lemma 3 and Theorem 3 of Lecture 8, while we present a simpler derivation of the de Bruijn's identity.

Theorem 3. *Along the Langevin dynamics, we have the de Bruijn's identity*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\rho_t \,\|\, \nu\right) = -\mathsf{FI}(\rho_t \,\|\, \nu).$$

If ν satisfies α-LSI, we have

$$\mathsf{KL}\left(\rho_t \,\|\, \nu\right) \le e^{-2\alpha t}\mathsf{KL}\left(\rho_0 \,\|\, \nu\right).$$

Proof. We know that the Fokker-Planck equation is the continuity equation of the Langevin dynamic. Moreover, we have shown that the Fokker-Planck equation is also the Wasserstein gradient flow of KL divergence

$$\frac{\partial \rho_t}{\partial t} = \Delta\rho_t + \nabla \cdot (\rho_t \nabla f) = -\mathsf{grad}_W \mathsf{KL}(\rho_t).$$

Hence, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}\left(\rho_t \,\|\, \nu\right) = \left\langle \mathsf{grad}_W \mathsf{KL}(\rho_t \,\|\, \nu), \frac{\partial \rho_t}{\partial t} \right\rangle_{\rho_t} = -\left\|\mathsf{grad}_W \mathsf{KL}(\rho_t \,\|\, \nu)\right\|_{\rho_t}^2$$
$$= -\mathbb{E}_{\rho_t}\left[\left\|\nabla \log \frac{\rho_t}{\nu}\right\|^2\right] = -\mathsf{FI}(\rho_t \,\|\, \nu).$$

The rest of the proof follows that of Theorem 3 of Lecture 8. □

Lemma 3. *Given $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Wasserstein gradient of $\rho \mapsto \mathcal{W}_2^2(\rho, \nu)$ is $-2\log_\rho(\nu)$. In general, on a Riemannian manifold, the gradient of $\mathrm{d}(\cdot, q)^2$ at p is $-2\log_p(q)$.*

Theorem 4. *If ν is log-concave, then along the Langevin dynamics, we have*

$$\mathsf{KL}\left(\rho_t \,\|\, \nu\right) \le \frac{\mathcal{W}_2^2(\rho_0, \nu)}{2t}. \tag{13}$$

Proof. Recall from Theorem 2 that

$$\mathsf{KL}(\rho_1 \,\|\, \nu) \ge \mathsf{KL}(\rho_0 \,\|\, \nu) + \left\langle \mathsf{grad}_W \mathsf{KL}(\rho_0 \,\|\, \nu), \log_{\rho_0}(\rho_1) \right\rangle_{\rho_0} + \frac{\alpha}{2}\mathcal{W}_2^2(\rho_0, \rho_1).$$

Taking $\rho_0 = \rho_t$, $\rho_1 = \nu$, and $\alpha = 0$, we have

$$\left\langle \mathsf{grad}_W \mathsf{KL}(\rho_t \,\|\, \nu), \log_{\rho_t}(\nu) \right\rangle_{\rho_t} \le -\mathsf{KL}\left(\rho_t \,\|\, \nu\right). \tag{14}$$

Consider a Lyapunov functional $\mathcal{L}_t := t\mathsf{KL}\left(\rho_t|\nu\right) + \frac{1}{2}\mathcal{W}_2^2\left(\rho_t, \nu\right)$. Taking time derivative gives

$$\begin{aligned}
\frac{\partial}{\partial t}\mathcal{L}_t &= \mathsf{KL}\left(\rho_t \,\|\, \nu\right) - t\mathsf{FI}\left(\rho_t \,\|\, \nu\right) + \left\langle \frac{\partial \rho_t}{\partial t}, -\log_{\rho_t}(\nu) \right\rangle_{\rho_t} \\
&= \mathsf{KL}\left(\rho_t \,\|\, \nu\right) - t\mathsf{FI}\left(\rho_t \,\|\, \nu\right) + \left\langle \mathsf{grad}_W \mathsf{KL}(\rho_t \,\|\, \nu), \log_{\rho_t}(\nu) \right\rangle_{\rho_t} \\
&\le -t\mathsf{FI}\left(\rho_t \,\|\, \nu\right) \le 0
\end{aligned}$$

where the first inequality is due to (14). Therefore, $\mathcal{L}_t \le \mathcal{L}_0$ and (13) holds. □

We provide a convergence guarantee of the proximal sampling algorithm in Lecture 9 for the log-concave case, which closely resembles the convergence of proximal point method in optimization for convex functions.

Theorem 5. *If ν is log-concave, then ρ_k^X of the proximal sampling algorithm satisfies*

$$\mathsf{KL}\left(\rho_k^X \,\|\, \nu\right) \le \frac{\mathcal{W}_2^2(\rho_0^X, \nu)^2}{k\eta}.$$

Proof. Since ν_t is log-concave (log-concavity is preserved by convolution), we have HWI inequality

$$\mathsf{KL}(\rho_t \parallel \nu_t) \le \mathcal{W}_2(\rho_t, \nu_t)\sqrt{\mathsf{FI}(\rho_t \parallel \nu_t)}.$$

It follows from Lemma 3 of Lecture 9 and the above HWI inequality that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t \parallel \nu_t) = -\frac{1}{2}\mathsf{FI}(\rho_t \parallel \nu_t) \le -\frac{1}{2}\frac{\mathsf{KL}(\rho_t \parallel \nu_t)^2}{\mathcal{W}_2^2(\rho_t, \nu_t)} \le -\frac{1}{2}\frac{\mathsf{KL}(\rho_t \parallel \nu_t)^2}{\mathcal{W}_2^2(\rho_0, \nu_0)} = -\frac{1}{2}\frac{\mathsf{KL}(\rho_t \parallel \nu_t)^2}{\mathcal{W}_2^2(\rho_k^X, \pi^X)},$$

where the last inequality is due to the fact that $\mathcal{W}_2$ contraction along heat flow (proof by coupling). Solving this differential inequality yields

$$\frac{1}{\mathsf{KL}(\rho_k^Y \parallel \pi^Y)} = \frac{1}{\mathsf{KL}(\rho_\eta \parallel \nu_\eta)} \ge \frac{1}{\mathsf{KL}(\rho_k^X \parallel \pi^X)} + \frac{\eta}{2\mathcal{W}_2^2(\rho_k^X, \pi^X)}. \tag{15}$$

For the backward process, we similarly have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{KL}(\rho_t^- \parallel \nu_t^-) = -\frac{1}{2}\mathsf{FI}(\rho_t^- \parallel \nu_t^-) \le -\frac{1}{2}\frac{\mathsf{KL}(\rho_t^- \parallel \nu_t^-)^2}{\mathcal{W}_2^2(\rho_t^-, \nu_t^-)}.$$

Recall from Lecture 9 that the backward channel can be modeled by the following SDE

$$\mathrm{d}Y_t = \nabla \log \nu_{\eta-t}(Y_t)\mathrm{d}t + \mathrm{d}W_t. \tag{16}$$

Since $\log \nu_{\eta-t}$ is log-concave, by a coupling argument, we can show $t \mapsto \mathcal{W}_2^2(\rho_t^-, \nu_t^-)$ is descreasing. Thus,

$$\mathcal{W}_2^2(\rho_t^-, \nu_t^-) \le \mathcal{W}_2^2(\rho_0^-, \nu_0^-) = \mathcal{W}_2^2(\rho_k^Y, \pi^Y) \le \mathcal{W}_2^2(\rho_k^X, \pi^X).$$

Therefore, we obtain

$$\frac{1}{\mathsf{KL}(\rho_{k+1}^X \parallel \pi^X)} = \frac{1}{\mathsf{KL}(\rho_\eta^- \parallel \nu_\eta^-)} \ge \frac{1}{\mathsf{KL}(\rho_k^Y \parallel \pi^Y)} + \frac{\eta}{2\mathcal{W}_2^2(\rho_k^X, \pi^X)}.$$

Combining the above inequality and (15), we have

$$\frac{1}{\mathsf{KL}(\rho_{k+1}^X \parallel \pi^X)} \ge \frac{1}{\mathsf{KL}(\rho_k^X \parallel \pi^X)} + \frac{\eta}{\mathcal{W}_2^2(\rho_k^X, \pi^X)} \ge \frac{1}{\mathsf{KL}(\rho_k^X \parallel \pi^X)} + \frac{\eta}{\mathcal{W}_2^2(\rho_0^X, \pi^X)},$$

i.e.,

$$\frac{1}{\mathsf{KL}(\rho_{k+1}^X \parallel \nu)} \ge \frac{1}{\mathsf{KL}(\rho_k^X \parallel \nu)} + \frac{\eta}{\mathcal{W}_2^2(\rho_0^X, \nu)}.$$

Summing the above inequality over iterations, we arrive at

$$\frac{1}{\mathsf{KL}(\rho_k^X \parallel \nu)} \ge \frac{1}{\mathsf{KL}(\rho_0^X \parallel \nu)} + \frac{k\eta}{\mathcal{W}_2^2(\rho_0^X, \nu)} \ge \frac{k\eta}{\mathcal{W}_2^2(\rho_0^X, \nu)}.$$

□

6 Revisiting proximal sampling

Recall that sampling from $\nu \propto \exp(-f)$ can be understood as minimizing $\mathsf{KL}(\cdot \parallel \nu)$ over $\mathbb{R}_2(\mathbb{R}^d)$. Because the Fokker-Planck equation is the Wasserstein gradient flow of KL divergence, in the language of numerical methods, the Langevin dynamics (and hence LMC) can be viewed as an explicit scheme. As we know explicit schemes are less stable than implicit schemes, and an example in optimization is gradient descent versus proximal point method. In sampling, the implicit scheme is known as the JKO scheme, which repeatedly invokes the proximal operator over the Wasserstein space

$$\rho_{k+1} = \mathrm{prox}_{\eta\mathcal{F}}(\rho_k) = \operatorname*{argmin}_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathcal{F}(\rho) + \frac{1}{2\eta}\mathcal{W}_2^2(\rho_k, \rho) \right\}.$$

Hence, the JKO scheme is a Wasserstein analogue of the proximal point method.

We revisit the proximal sampling algorithm in Lecture 9 from this variational perspective. The first result shows that RGO is a proximal operator on the Wasserstein space.

Lemma 4. *Recall RGO is a sampling oracle for given $y \in \mathbb{R}^d$ that returns $x \sim \pi^{X|Y}(x|y)$ where*

$$\pi^{X|Y}(x \mid y) \propto \exp\left(-f(x) - \frac{1}{2\eta}\|x-y\|^2\right).$$

Then, RGO is a proximal operator of KL divergence, i.e.,

$$\pi^{X|Y=y} = \mathrm{prox}_{\eta\mathsf{KL}(\cdot\|\pi^X)}(\delta_y).$$

Proof. Since $\pi^X(x) = \nu(x) \propto \exp(-f(x))$, we know

$$\pi^{X|Y}(x \mid y) \propto \exp\left(-\frac{1}{2\eta}\|x-y\|^2\right)\pi^X(x).$$

Also, note that

$$\begin{aligned}\mathsf{KL}(\rho^X \parallel \pi^{X|Y}) &= \int \rho^X \log\frac{\rho^X}{\pi^{X|Y}} = \int \rho^X \left(\log\frac{\rho^X}{\pi^X} + \frac{1}{2\eta}\|x-y\|^2\right) + C(y)\\ &= \mathsf{KL}(\rho^X \parallel \pi^X) + \int \rho^X \frac{1}{2\eta}\|x-y\|^2 + C(y),\end{aligned}$$

where $C(y)$ is a function of y. Then, RGO can be expressed as

$$\begin{aligned}\pi^{X|Y=y} &= \operatorname*{argmin}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \mathsf{KL}(\rho^X \parallel \pi^{X|Y}) = \operatorname*{argmin}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho^X \parallel \pi^X) + \frac{1}{2\eta}\int \|x-y\|^2 d\rho^X(x) \right\}\\ &= \operatorname*{argmin}_{\rho^X \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \mathsf{KL}(\rho^X \parallel \pi^X) + \frac{1}{2\eta}\mathcal{W}_2^2(\rho^X, \delta_y) \right\} = \mathrm{prox}_{\eta\mathsf{KL}(\cdot\|\pi^X)}(\delta_y).\end{aligned}$$

□

A variant of the JKO scheme introduces an extra entropic regularization term

$$\rho_{k+1} = \underset{\rho\in\mathcal{P}_2\left(\mathbb{R}^d\right)}{\operatorname{argmin}} \left\{\mathcal{F}(\rho) + \frac{1}{2\eta}\mathcal{W}_{2,\varepsilon}^2\left(\rho_k, \rho\right)\right\}$$

where $\mathcal{W}_{2,\epsilon}$ is the entropy-regularized Wasserstein-2 distance defined as

$$\mathcal{W}_{2,\epsilon}^2(\mu,\nu) := \min_{\gamma\in\Gamma(\mu,\nu)} \left\{\int \|x-y\|^2\gamma(x,y)\mathrm{d}x\mathrm{d}y - \varepsilon\mathsf{H}(\gamma)\right\} \tag{17}$$

and $\mathsf{H}(\gamma) = -\int \gamma\log\gamma$ denotes the entropy.

The next result shows that the proximal sampling algorithm can be viewed as the entropy-regularized JKO scheme.

Theorem 6. *Let $\rho_k^X, \rho_k^Y, \rho_{k+1}^X$ be the distributions of x_k, y_k, x_{k+1}, respectively, in one iteration of the proximal sampling algorithm. Then, they follow the entropy-regularized JKO scheme*

$$\rho_k^Y = \underset{\rho\in\mathcal{P}_2\left(\mathbb{R}^d\right)}{\operatorname{argmin}} \frac{1}{2\eta}\mathcal{W}_{2,2\eta}^2\left(\rho_k^X, \rho\right), \tag{18}$$

$$\rho_{k+1}^X = \underset{\rho\in\mathcal{P}_2\left(\mathbb{R}^d\right)}{\operatorname{argmin}} \left\{\int f\rho + \frac{1}{2\eta}\mathcal{W}_{2,2\eta}^2\left(\rho_k^Y, \rho\right)\right\}. \tag{19}$$

Proof. Plugging (17) into the right-hand side of (18) gives a solution γ^Y with γ being the solution to

$$\min_{\substack{\gamma\in\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)\\ \gamma^X=\rho_k^X}} \left\{\int \frac{1}{2\eta}\|x-y\|^2\gamma(x,y)\mathrm{d}x\mathrm{d}y - \mathsf{H}(\gamma)\right\}.$$

Consider the Lagrangian function

$$\begin{aligned}
&\int \frac{1}{2\eta}\|x-y\|^2\gamma(x,y)\mathrm{d}x\mathrm{d}y - \mathsf{H}(\gamma) + \int \lambda(x)[\gamma^X(x) - \rho_k^X(x)]\mathrm{d}x\\
=&\int \frac{1}{2\eta}\|x-y\|^2\gamma(x,y)\mathrm{d}x\mathrm{d}y - \mathsf{H}(\gamma) + \int \lambda(x)[\gamma(x,y) - \rho_k^X(x)]\mathrm{d}x\mathrm{d}y
\end{aligned}$$

Taking the first variation yields

$$\frac{1}{2\eta}\|x-y\|^2 + 1 + \log\gamma + \lambda(x) = 0.$$

So

$$\gamma(x,y) \propto \exp\left(-\frac{1}{2\eta}\|x-y\|^2 - \lambda(x)\right),$$

and the X-marginal is $\gamma^X \propto \exp(-\lambda(x))$. On the other hand, from the constraint, we know $\gamma^X = \rho_k^X$. So

$$\gamma(x,y) \propto \exp\left(-\frac{1}{2\eta}\|x-y\|^2\right)\rho_k^X(x)$$

and $\gamma^Y = \rho_k^X * \mathcal{N}(0, \eta I) = \rho_k^Y$. Thus, we verify (18).

Similarly, plugging (17) into the right-hand side of (19) gives a solution γ^X with γ being the solution to

$$\underset{\substack{\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \\ \gamma^Y = \rho_k^Y}}{\operatorname{argmin}} \left\{ \int \left[f(x) + \frac{1}{2\eta} \|x - y\|^2 \right] \gamma(x, y) \mathrm{d}x\mathrm{d}y - \mathsf{H}(\gamma) \right\}.$$

Clearly, $\gamma(x, y) \propto \rho_k^Y(y) \exp\left(-f(x) - \frac{1}{2\eta}\|x - y\|^2\right)$. Hence, the X-marginal γ^X has the same distribution as ρ_{k+1}^X. We thus complete the proof. □

The above interpretation of the proximal sampling algorithm provides some insights on its connections to optimization. Define a more general proximal sampling algorithm with a different level of entropy regularization

$$\rho_k^Y = \underset{\rho \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \frac{1}{2\eta} \mathcal{W}_{2, 2\eta\varepsilon}^2 \left(\rho_k^X, \rho\right), \tag{20}$$

$$\rho_{k+1}^X = \underset{\rho \in \mathcal{P}_2(\mathbb{R}^d)}{\operatorname{argmin}} \left\{ \int f\rho + \frac{1}{2\eta} \mathcal{W}_{2, 2\eta\varepsilon}^2 \left(\rho_k^Y, \rho\right) \right\}. \tag{21}$$

The next result reveals the well-known fact that optimization is the limit of sampling on the particular example of proximal sampling and optimization.

Theorem 7. *As $\varepsilon \to 0$, (20) and (21) reduce to the proximal point method in optimization.*

Proof. Following a similar argument as in the proof of Theorem 6, we can show that (20)-(21) correspond to the following proximal sampling

$$\begin{aligned} y_k &\sim \pi_\varepsilon^{Y|X=x_k} = \mathcal{N}\left(x_k, \varepsilon\eta I\right), \\ x_{k+1} &\sim \pi_\varepsilon^{X|Y=y_k} \propto \exp\left[-\frac{1}{\varepsilon}\left(f(x) + \frac{1}{2\eta}\|x - y_k\|^2\right)\right]. \end{aligned}$$

As $\varepsilon \to 0$, the above steps becomes to

$$\begin{aligned} y_k &= x_k, \\ x_{k+1} &= \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ f(x) + \frac{1}{2\eta}\|x - y_k\|^2 \right\}. \end{aligned}$$

Putting them together, we have $x_{k+1} = \operatorname{prox}_{\eta f}(x_k)$, i.e., the proximal point method. □

We finally note that the stationary distribution of the new proximal sampling algorithm (20)-(21) is $\pi_\varepsilon^X \propto \exp(-f/\varepsilon)$, which converges (as $\varepsilon \to 0$) to a Dirac distribution concentrating on the minimizer of f (or a uniform distribution over the minimizer set of f).