CS2910 - Optimization Summer 2023

Lecture B. Wasserstein Gradient Flow

Lecturer: Yunwei Ren *Scribed by Zhidan Li*

1 Overview

Now we focus on how to solve the optimal transport problem, with the cost function $c(x,y) = \frac{1}{2}\|x-y\|^2$.

Given a state space Ω (in this lecture, we will assume $\Omega = \mathbb{R}^d$), let $P_2(\Omega)$ be the collection of all probability measures over Ω with finite second moments, i.e.,

$$P_2(\Omega) \triangleq \left\{ \mu \middle| \int_{\Omega} |x|^2 \, d\mu(x) < \infty \right\}.$$

Then we want to solve the following optimization problem, for $\mu, \nu \in P_2(\mathbb{R}^d)$,

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2}\|x-y\|^2 \, d\gamma(x,y). \tag{1}$$

In this lecture, we will define a method of gradient flow over a kind of metric space called *Wasserstein-2 space*. This part might need a little of mathematical techniques to ensure the terms we introduce and use in the lecture are well-defined.

2 Gradient Flow in Wasserstein Space

To solve the optimization problem (1), we want to employ a 'gradient flow'-like algorithm to solve it. However, since the space is not $\mathbb{R}^d$, it is necessary to define the 'gradient' specifically in $P_2(\mathbb{R}^d)$.

2.1 Wasserstein-2 distance and Wasserstein-2 space

For two probability measures $\mu, \nu \in P_2(\mathbb{R}^d)$, we define the *Wasserstein-2 distance* between μ and ν as

$$W_2(\mu,\nu) = \left(\inf_{\gamma \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{1}{2}\|x-y\|_2^2 \, d\gamma(x,y) \right)^{1/2}.$$

Informally speaking, W_2 is some kind of 'distance' in the space $P_2(\mathbb{R}^d)$. Then it can shown that, $\mathcal{W}_2(\mathbb{R}^d) \triangleq \left(P_2(\mathbb{R}^d), W_2\right)$ is a metric space.

2.2 Gradient in Wasserstein-2 space

Now we construct the gradient in Wasserstein-2 space $\mathcal{W}_2(\mathbb{R}^d)$. We put our eyes on gradient flow over $\mathbb{R}^d$:

$$\dot{x_t} = -\nabla f(x_t).$$

This ODE means, at each time t, we look at the linear approximation of f, and we want to locally minimize it. That is to say, when x is around x_t, we know

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle .$$

To minimize the linear function, we choose the $-\nabla f(x_t)$ as the 'moving direction'. This interpretation of gradient flow intuitively inspires us how to define gradient flow over Wasserstein metric space $\mathcal{W}_2(\mathbb{R}^d)$:

(a) Firstly we will show how to locally minimize a 'linear functional' $F : \mathcal{W}_2(\mathbb{R}^d) \to \mathbb{R}$.

(b) Secondly we define the linear approximation of a non-linear functional $F : \mathcal{W}_2(\mathbb{R}^d) \to \mathbb{R}$.

(c) Finally, based on the two things above, we can immediately define the Wasserstein gradient flow.

2.2.1 Gradient for linear functionals

Before we discuss how to locally minimize a linear functional, it is of great necessity for us to answer the question: which kind of functionals are called 'linear'? To answer it, firstly we introduce some notations.

Let $\mathcal{M}_\pm(\mathbb{R}^d)$ be the collection of signed measures on $\mathbb{R}^d$. It is trivial that $\mathcal{M}_\pm(\mathbb{R}^d)$ can be made into a vector space equipped with operation $+ : \mathcal{M}_\pm(\mathbb{R}^d) \times \mathcal{M}_\pm(\mathbb{R}^d) \to \mathcal{M}_\pm(\mathbb{R}^d)$ as: for all $\mu, \nu \in \mathcal{M}_\pm(\mathbb{R}^d)$ and $a, b \in \mathbb{R}$, for all measurable $E \subseteq \mathbb{R}^d$,

$$(a\mu + b\nu)(E) = a\mu(E) + b\nu(E).$$

For a functional $F : \mathcal{M}_\pm(\mathbb{R}^d) \to \mathbb{R}$, F is said to be *linear* if for all $\mu, \nu \in \mathcal{M}_\pm(\mathbb{R}^d)$, $a, b \in \mathbb{R}$, it holds that

$$F(a\mu + b\nu) = aF(\mu) + bF(\nu).$$

Equivalent, if F is a linear functional, there exists a function $V : \mathbb{R}^d \to \mathbb{R}$ such that

$$F(\mu) = \int_{\mathbb{R}^d} V(x)\, d\mu(x), \forall \mu \in \mathcal{M}_\pm(\mathbb{R}^d).$$

Intuitively, we can view $V(x)$ as the cost of putting 1 unit of particles at x, μ as the distribution of the particles. Then $F(\mu)$ means the total cost of putting particles as μ.

Now we focus on how to locally minimize F. Note that, if we want to globally minimize F, we just need to put all particles at $x_* = \operatorname{argmin}_{x \in \mathbb{R}^d} \{f(x)\}$. This is no meaning.

Roughly speaking, to locally minimize F at μ_t, for small η, we need to consider the probability measure $\mu_{t+\eta}$ such that the distance $W_2(\mu_t, \mu_{t+\eta})$ is small subject to there exists $T : \mathbb{R}^d \to \mathbb{R}^d$, $T\#\mu = \mu_{t+\eta}$. We put our eyes on the position of each particle.

It is not surprising that, the movement of particles at each position $x \in \mathbb{R}^d$ corresponds to the movement of the probability measure, since the probability measure describes the distribution of particles. For linear functional $F(\mu) = \int V \, d\mu$, to locally minimize F, it suffices to locally minimize the 'cost' of the movement of each particle. For a particle positioned at x_t, the movement of it is exactly the gradient flow, i.e.,

$$\frac{d}{dt}x_t = -\nabla V(x_t).$$

Based on the discussion above, we formally define the flow of linear functionals on $\mathcal{W}_2(\mathbb{R}^d)$.

Definition 1. *Given a (time-dependent) velocity field $v_t : \mathbb{R}^d \to \mathbb{R}^d$, we define its associated flow $\Phi : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}$ as*

$$\Phi(x, t) \triangleq x_t$$

where x_t is the solution to the following ODE

$$\dot{x_t} = v_t(x_t), x_0 = x.$$

Proposition 2. *Given a (time-dependent) velocity field $v_t : \mathbb{R}^d \to \mathbb{R}^d$, and an initial configuration $\mu_0 \in P(\mathbb{R}^d)$, define $\mu_t \triangleq \Phi_t\#\mu_0$. Then μ_t satisfies the continuity equation*

$$\partial_t\mu_t + \nabla \cdot (\mu_t v_t) = 0.$$

Remark 1. The term $\partial_t\mu_t$ means the change of the density, and the second term can be viewed as the amount of out-flow particles.

Then we specify Wasserstein gradient flow.

Definition 3. *We say μ_t is the Wasserstein gradient flow with respect to $F = \int V \, d\mu$ if $\mu_0 = \mu$, it holds*

$$\forall t \geq 0, x \in \mathbb{R}^d, \frac{d}{dt}x_t = -\nabla V(x_t).$$

Or equivalently,

$$\partial_t\mu_t - \nabla \cdot (\mu_t\nabla V) = 0.$$

Indeed, the Wasserstein gradient flow is just the specification of Definition 1 when $v_t \equiv -\nabla V$.

2.2.2 First-order calculus in $\mathcal{W}_2(\mathbb{R}^d)$

Consider the space $\mathbb{R}^d$, to describe a point $u \in \mathbb{R}^d$, it is equivalent to describe the linear functional $v \mapsto \langle u, v\rangle$. Also, if we want to describe $\nabla f(x) \in \mathbb{R}^d$, it suffices to describe the linear functional $v \mapsto \langle\nabla f(x), v\rangle$.

For some small $\varepsilon > 0$, we consider the curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^d$ such that

$$x(0) = x, \frac{d}{dt}x(t)\Big|_{t=0} = v.$$

Then by the chain rule,

$$\frac{d}{dt}f(x(t))\Big|_{t=0} = \left\langle \nabla f(x(t)), \dot{x(t)} \right\rangle \Big|_{t=0} = \langle \nabla f(x), v \rangle .$$

Now we generalize the analogue things in the space $P_2(\mathbb{R}^d)$.

Definition 4 (first variation)**.** *Given a functional $F : P_2(\mathbb{R}^d) \to \mathbb{R}$ and $\mu \in P_2(\mathbb{R}^d)$, we say $G : \mathbb{R}^d \to \mathbb{R}$ is the first variation of F at μ if for all perturbation $\chi \in \mathcal{M}_\pm(\mathbb{R}^d)$ with $\mu \in \varepsilon\chi \in P_2(\mathbb{R}^d)$ for all small $\varepsilon > 0$, we have*

$$\frac{d}{d\varepsilon}F(\mu + \varepsilon\chi)\Big|_{\varepsilon=0} = \int G \, d\chi.$$

Note that G does not necessarily exist. If G exists, we denote the first variation by $\frac{\delta F}{\delta \mu}[\mu]$.

Examples:

- For a linear functional $F(\mu) = \int V \, d\mu$, we compute

$$\begin{aligned} \frac{d}{d\varepsilon}F(\mu + \varepsilon\chi) &= \frac{d}{d\varepsilon}\int_{\mathbb{R}^d} V(x) \, d(\mu(x) + \varepsilon\chi(x)) \\ &= \frac{d}{d\varepsilon}\varepsilon \int_{\mathbb{R}^d} V(x) \, d\chi(x) \\ &= \int_{\mathbb{R}^d} V(x) \, d\chi(x). \end{aligned}$$

  Thus,

$$\frac{\delta F}{\delta \mu}[\mu] = V.$$

- For $F(\mu) = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} W(x, y) \, d\mu(x) \, d\mu(y)$, by elementary calculation

$$\begin{aligned} \frac{d}{d\varepsilon}F(\mu + \varepsilon\chi) &= \frac{d}{d\varepsilon}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} W(x, y)(\, d\mu(x) + \varepsilon \, d\chi(x))(\, d\mu(y) + \varepsilon \, d\chi(y)) \\ &= \frac{d}{d\varepsilon}\left(\varepsilon \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (W(x, y) + W(y, x)) \, d\mu(y) \, d\chi(x) + O(\varepsilon^2)\right) \\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (W(x, y) + W(y, x)) \, d\mu(y) \, d\chi(x). \end{aligned}$$

  Then we show

$$\frac{\delta F}{\delta \mu}[\mu](x) = \int_{\mathbb{R}^d} (W(x, y) + W(y, x)) \, d\mu(y).$$

2.2.3 Wasserstein gradient flow

Now we define the Wasserstein gradient flow in $P_2(\mathbb{R}^d)$. The key step is to locally 'minimize' F at some $\mu \in P_2(\mathbb{R}^d)$. Based on the first variation we define above, since $\int \frac{\delta F}{\delta \mu}[\mu]\, d\mu + C$ is the linear approximation of F at μ, to achieve the local minimum of F, it suffices to locally minimize the linear functional $\int \frac{\delta F}{\delta \mu}[\mu]\, d\mu$.

Definition 5. *Given a functional $F : \mathcal{W}_2(\mathbb{R}^d) \to \mathbb{R}$, we say μ_t is the Wasserstein gradient flow with respect to F if it satisfies*

$$\frac{d}{dt} x_t = -\nabla \frac{\delta F}{\delta \mu_t}[\mu_t](x_t), \forall t \geq 0, x \in \mathbb{R}^d.$$

Or equivalently,

$$\partial_t \mu_t - \nabla \cdot \left(\mu_t \nabla \frac{\delta F}{\delta \mu}[\mu_t] \right) = 0.$$

Now we establish the decay of F during the Wasserstein gradient flow.

Proposition 6. *Let μ_t be the Wasserstein gradient flow with respect to $F : \mathcal{W}_2 \to \mathbb{R}^d$. Under some regularity conditions, we have*

$$\frac{d}{dt} F(\mu_t) = -\int \left\| \nabla \frac{\delta F}{\delta \mu}[\mu_t](x) \right\|_2^2 d\mu_t(x).$$

Proof assuming all regularity conditions. Now we prove the proposition assuming that all regularity conditions we need. By elementary calculation,

$$\begin{aligned}
\frac{d}{dt} F(\mu_t) &= \int \frac{\delta F}{\delta \mu}[\mu_t](x) \partial_t \mu_t(x)\, dx \\
&= \int \frac{\delta F}{\delta \mu}[\mu_t](x) \nabla \cdot \left(\mu_t(x) \nabla \frac{\delta F}{\delta \mu}[\mu_t](x) \right) dx \\
&= \sum_{k=1}^d \int \frac{\delta F}{\delta \mu}[\mu_t](x) \partial_k \left[\mu_t(x) \partial_k \frac{\delta F}{\delta \mu}[\mu_t](x) \right] dx \\
&= -\sum_{k=1}^d \int \partial_k \frac{\delta F}{\delta \mu}[\mu_t](x) \mu_t(x) \partial_k \frac{\delta F}{\delta \mu}[\mu_t](x)\, dx \\
&= -\int \left\| \nabla \frac{\delta F}{\delta \mu}[\mu_t](x) \right\|_2^2 d\mu_t(x)
\end{aligned}$$

where the last two equalities hold under the assumption that F has some good boundary conditions.

□