Gradient Flow and Its Applications in Statistical Learning

Rui-Yang Zhang

Contents

Preface

Gradient flow has been an emerging topic in the field of computational statistics and machine learning. The topic of gradient flow initially started as a tool in optimal transport to study certain PDEs and the movements of distributions but has been adapted to study topics in statistical learning in recent years. Gradient flow has many applications in statistical learning, especially in designing and understanding optimisation and sampling algorithms such as the Langevin Monte Carlo and Stein Variational Gradient Descent. Albeit a large volume of research output in this area, the existing pedagogical material for gradient flow for statistical learning alternates between (very good!) tutorial slides and videos for machine learning conferences and chapters from a rather dense textbook on sampling or optimal transport. A middle ground is missing, and this note aims to fill that gap by providing a sufficiently technical expository (with few prerequisites) to gradient flow in statistical learning that should hopefully provide enough background material to understand the most recent advances in this growing field.

This note introduces the idea of gradient flow, in both the Euclidean space and the Wasserstein space. Gradient flow is an ODE that continuously minimises some function of interest, and it has been employed increasingly in statistical learning as a theoretical tool to understand many sampling and optimisation algorithms such as gradient descent and Langevin Monte Carlo. In Chapter 1, we introduce gradient flow in the Euclidean space setting and draw connections to common optimisation algorithms such as gradient descent and proximal point algorithm. Theoretical analyses of gradient flow are included as well, although they shed limited light on its discrete-time counterparts. In Chapter 2, the topic of optimal transport is described to motivate things such as the Kantorovich problem, Wasserstein distance, and the Wasserstein space. This gives us a suitable space and geometry to establish the gradient flow of probability distributions. We also draw links between the Langevin diffusion and the gradient flow in Wasserstein space, which is a key motivation for people to pay attention to this topic of gradient flow in Wasserstein space. Finally, in Chapter 3, we look at various popular sampling algorithms and indicate that they are indeed gradient flow in disguise, and how this revelation could help us understand these algorithms better theoretically.

Lastly, I would like to thank Chris Nemeth for the guidance and help he provided with writing this note.

Lancaster, UK
March, 2024

Chapter 1

Gradient Flow in Euclidean Space

In this chapter, we will look at the concept of gradient flow in Euclidean space, and how some common (convex) optimisation algorithms can be viewed as time-discretisations of them. Some of the material in this chapter is based on Santambrogio (2017).

1.1 Introducing Gradient Flow

Consider an objective function $F : \mathbb{R}^n \to \mathbb{R}$ that is sufficiently smooth (e.g. ∇F is Lipschitz continuous), and we wish to minimise F (and find the minimiser). This is one of the most fundamental questions in optimisation and Statistics. Mathematically, the problem can be written as $\min_x F(x)$.

One could consider the following ODE

$$\begin{cases} \dot{x}(t) & = -\nabla F(x(t)) \\ x(0) & = x_0 \end{cases} \tag{1}$$

where $t > 0$ and $x_0 \in \mathbb{R}^n$ would be an arbitrary initial position. This ODE is called the **gradient flow**, as we have $x(t)$ flowing smoothly towards the bottom of F using its gradient information. This is a well-defined difference equation under some smoothness condition of F, like when ∇F is Lipschitz continuous, and the existence and uniqueness of a solution are provided by the Cauchy-Lipschitz theorem (Arnold, 1992).

Intuitively speaking, the trajectory of $F(x(t))$ will always be towards the steepest decrease in the value of F, and we would assume (and it is true) that we would eventually reach the global minimum of F under some regularity condition of F. For the rest of this section, we will formally establish these results.

First, we will show that the gradient flow is decreasing the objective function. We have, using equation (1),

$$\frac{d}{dt}F(x(t)) = \nabla F(x(t))^T \frac{dx(t)}{dt} = -\nabla F(x(t))^T \nabla F(x(t)) = -\|\nabla F(x(t))\|_2^2 \leq 0, \tag{2}$$

as desired.

Next, we will establish the convergence of the gradient flow solution assuming certain conditions on F. Naturally, $F(x(t))$ could become periodic and never converge, so convergence does not always hold. We will show the convergence of $F(x(t))$ for m-strongly convex[1] F and for convex[2] F. As the name might suggest, m-strongly convex is a stronger condition than merely convex. For example, $f(x) = e^{-x}$ is convex, yet it is decaying too slowly for $x \to -\infty$ for it to be strongly convex.

Proposition 1.1. *The gradient flow with F being m-strongly convex converges, in the sense that for solution $x(t)$, $\lim_{t\to\infty} F(x(t))$ exists.*

Proof. For m-strongly convex F, we have

$$\min_y F(y) - F(x) \geq \min_y \nabla F(x)^T (y-x) + m/2\|y-x\|_2^2$$

where $F^* = \min_y F(y)$ and $x^* = \arg\min_y F(y)$. Since F is convex, x^* is a global and local minima of F although it may not be unique. To minimise the RHS of the above inequality, we take the partial derivation with respect to y and have

$$\frac{\partial}{\partial y} \nabla F(x)^T (y-x) + m/2\|y-x\|_2^2 = \nabla F(x) + m(y-x) = 0$$

so we should set $y = x - \nabla F(x)/m$. Substituting this value gives us

$$F^* - F(x) \geq -\nabla F(x)^T \nabla F(x)/m + m/2\|\nabla F(x)/m\|_2^2 = -\|\nabla F(x)\|_2^2/(2m)$$

which is called the Lojasiewicz inequality (Karimi et al., 2016). Next, recalling equation (2), and we have

$$\begin{aligned} \frac{d}{dt} F(x(t)) &= -\|\nabla F(x(t))\|_2^2 \\ -\|\nabla F(x(t))\|_2^2 &\leq -2m[F(x(t)) - F^*]. \end{aligned}$$

Combining them gives us

$$\begin{aligned} \frac{d}{dt} F(x(t)) &\leq -2m[F(x(t)) - F^*] \\ \frac{d}{dt}[F(x(t)) - F^*] &\leq -2m[F(x(t)) - F^*] \\ \frac{d}{dt} \log[F(x(t)) - F^*] &\leq -2m \\ F(x(t)) - F^* &\leq e^{-2mt}[F(x(0)) - F^*] \end{aligned}$$

where the final step is derived from integrating both sides of the inequality from 0 to t. So, we have established the (geometric) convergence of the gradient flow when F is m-strongly convex. □

Proposition 1.2. *The gradient flow with F being convex converges, in the sense that for solution $x(t)$, $\lim_{t\to\infty} F(x(t))$ exists.*

[1] F is m-strongly convex if $F(y) - F(x) \geq \nabla F(x)^T (y-x) + m/2\|y-x\|_2^2$ for any x, y.

[2] F is convex if $F(y) - F(x) \geq \nabla F(x)^T (y-x)$ for any x, y.

Proof. For convex F, we have $F(x^*) - F(x(t)) \geq \langle -\nabla F(x(t)), x(t) - x^* \rangle$. We then have

$$\begin{aligned}
\frac{d}{dt}\|x(t) - x^*\|_2^2 &= \langle 2(x(t) - x^*), \frac{d}{dt}x(t) \rangle \\
&= 2\langle x(t) - x^*, -\nabla F(x(t)) \rangle \\
&\leq 2(F^* - F(x(t))) = -2(F(x(t)) - F^*).
\end{aligned}$$

Reorganising the inequality gives us

$$\begin{aligned}
F(x(t)) - F^* &\leq -\frac{1}{2}\frac{d}{dt}\|x(t) - x^*\|_2^2 \\
\int_0^t F(x(u)) - F^* du &\leq -\frac{1}{2}\int_0^t \frac{d}{du}\|x(u) - x^*\|_2^2 du \\
\int_0^t F(x(u))du - tF^* &\leq -\|x(t) - x^*\|_2^2/2 + \|x(0) - x^*\|_2^2/2 \leq \|x(0) - x^*\|_2^2/2 \\
\frac{1}{t}\int_0^t F(x(u))du - F^* &\leq \frac{1}{2t}\|x(0) - x^*\|_2^2.
\end{aligned}$$

Next, as we have shown from earlier that $F(x(t))$ is decreasing in t, we have

$$F(x(t)) - F^* \leq \frac{1}{t}\int_0^t F(x(u))du - F^* \leq \frac{1}{2t}\|x(0) - x^*\|_2^2$$

and therefore we have established the convergence of gradient flow for convex F. □

Note that in the first part of the above proof we have established the result

$$\frac{d}{dt}\|x(t) - x^*\|_2^2 \leq -2[F(x(t)) - F^*].$$

This result, which assumes the function F of the gradient flow is convex, is known as the **evolution variational inequality** (EVI). This is a fundamental inequality, as one could in fact derive the gradient flow ODE with only this result, albeit it does not involve ∇F.

1.2 Discretisations of Gradient Flow

In the previous section, we looked at gradient flows in the Euclidean space. Gradient flow is an ODE that moves towards the minimum of an objective function F in continuous time. As minimisation (and equivalently maximisation) is often a task that we wish to conduct numerically, we could not (as of now) do so continuously using computers without some discretisation, except for the rare occasions when a closed form expression exists. In this section, we will look at two ways one could discretise the gradient flow in time to obtain implementable discretisation schemes.

Given an ODE, the easiest time-discretisation is the **Euler method** in Numerical Analysis (Burden and Faires, 2011), and the two optimisation algorithms below are essentially the explicit and implicit Euler method discretisations of the gradient flow.

The explicit Euler method applied to the gradient flow (1) with step size h will give the following output $\{x_k\}_{k\in\mathbb{N}}$:

$$x_0 = x_0, \qquad x_k = x_{k-1} - h\nabla F(x_{k-1}) \text{ for } k = 1, 2, \ldots.$$

This is the standard version of **gradient descent** that minimises a (differentiable) function F. There are variants of gradient descent that either use varying step size rather than fix (e.g. adaptive gradient descent), or use an unbiased estimate of ∇F at each iteration (e.g. stochastic gradient descent).

It is not hard to notice that the gradient descent will not guarantee to converge to the minimum of F, unlike the gradient flow. However, it is still reasonably good for small step sizes, and its simple formulation makes it easy to implement, which is why gradient descent is one of the most commonly used algorithms in optimisation. The theory behind gradient descent, such as its convergence with sufficiently small step size and the rates of convergence, will be omitted here, and they could not be derived from results about gradient flow.

The implicit Euler method applied to the gradient flow

$$\begin{cases} \dot{x}(t) & = -\nabla F(x(t)) \\ x(0) & = x_0 \end{cases}$$

with step size h will give the following output $\{x_k\}_{k\in\mathbb{N}}$:

$$x_0 = x_0, \qquad x_k = x_{k-1} - h\nabla F(x_k) \ \text{ for } k = 1, 2, \ldots.$$

This version is not exactly helpful, as if we are trying to run the optimisation, we would not be able to update x_k as it is defined using $\nabla F(x_k)$. We can instead transform it and get

$$\begin{aligned} \frac{x_k - x_{k-1}}{h} &= -\nabla F(x_k) \\ \frac{x_k - x_{k-1}}{h} + \nabla F(x_k) &= 0 \\ \frac{d}{dx_k}\left[\frac{\|x_k - x_{k-1}\|_2^2}{2h} + F(x_k)\right] &= \frac{x_k - x_{k-1}}{h} + \nabla F(x_k) = 0. \end{aligned}$$

So, as F is assumed to be convex, x_k would be the minimiser of $F(x) + \|x - x_{k+1}\|_2^2/(2h)$, and this becomes more feasible. Thus, the discretisation will yield the following output $\{x_k\}_{k\in\mathbb{N}}$:

$$x_0 = x_0, \qquad x_k \in \arg\min_x F(x) + \frac{\|x - x_{k-1}\|_2^2}{2h} = \operatorname{prox}_{hF}(x_{k-1}) \ \text{ for } k = 1, 2, \ldots$$

and this gives us the **proximal point algorithm**.

The above derivation assumes F to be differentiable, which is not exactly needed for the proximal point algorithm using the $\arg\min$ formulation. This algorithm uses the implicit Euler method, which is a more accurate discretisation scheme of the ODE, and thus would yield a better convergence. The exact theoretical justifications would be omitted here as well.

Currently, we cannot see much power of gradient flow, as for now, it is merely a neat way to unify various optimisation schemes while not providing much help with the theories. In the next chapter, we will look at gradient flows in Wasserstein space (instead of Euclidean space) which are far more powerful.

Chapter 2

Gradient Flow in Wasserstein Space

In this chapter, we will introduce and describe some basic results of optimal transport that allow us to formalise the notion of the Wasserstein distance, which is essential for our construction of the Wasserstein space. Next, we study a few key properties and structures of the Wasserstein space, such as its metric structure and its Riemannian structure. They will help us establish the notion of gradient flow in the Wasserstein space. Then, we will look at how one could view sampling as optimisation by minimising the KL divergence, and we derive the Wasserstein gradient flow of KL divergence, which has an amazing connection with the Langevin diffusion. Finally, we will study the discretisations of the Wasserstein gradient flow, as well as some convergence properties related to them. This chapter is based on Chapters 1.3 and 1.4 of Chewi (2023).

2.1 A Brief Introduction to Optimal Transport

The study of optimal transport (OT) started with finding the optimal way of transporting and allocating resources first proposed in Monge (1781), which is commonly known as the **Monge problem**. The Monge problem considered two probability densities f, g defined on $\mathbb{R}^d$ and aims to look for a map $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes f to g in the sense that

$$\int_A g(x)dx = \int_{T^{-1}(A)} f(y)dy$$

for all Borel subsets $A \subseteq \mathbb{R}^d$ and T minimises the cost

$$\int_{\mathbb{R}^d} \|T(x) - x\|_2 f(x)dx = \mathbb{E}_{X\sim f}\left[\|T(X) - X\|_2\right].$$

For the special case of $d = 1$ and f, g both being histograms, the problem can be intuitively thought of as transporting one pile of books (with relative numbers of books at each location following f) to a different configuration of arranging the books in a pile following g. For example, f could be putting half of the books at location 1 and the other half at location $2(f : 1 \mapsto 1/2, 2 \mapsto 1/2)$, while we would want to transform that into putting all the books at location 2 ($g : 2 \mapsto 1$), which could be solved by simply shifting all the books at location 1 to 2. However, if the target configuration is putting half of the books at location 1, but a quarter of the books at location 2 and new location 3, a transportation mapping would not exist as we are not allowed to half (or divide a whole into any proportions) the books at location 2. Even in this simple setting, it is not hard to notice that the Monge problem *does not* always have a solution.

At this stage, we would define the notion of a push forward. Consider measures μ on X and a measurable map $T : X \to Y$. We have the **push forward** of μ by T, which we denote by $T_{\#\mu}$, as a measure on Y defined by

$$T_{\#\mu}(A) = \mu(T^{-1}(A))$$

for all measurable subsets A of Y, and

$$\int_Y \phi \ d(T_{\#\mu}) = \int_X \phi \circ T d\mu$$

for all measurable functions ϕ on Y.

As we have mentioned using the book re-configuration example, the Monge problem does not always have a solution. One of the main problems is that we are not allowed to break a whole into portions. This restriction is relaxed by Kantorovich, and this relaxation of the Monge problem is known as the **Kantorovich problem** (Kantorovich, 2006) which we state below.

Consider the cost function $c : X \times X \to [0, +\infty]$ that is continuous and symmetric, denoted by the **optimal transport cost**. Given two probability measures $\mu, \nu \in \mathcal{P}(X)$[1], the Kantorovich problem is

$$\inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times X} c(x, x') \ d\gamma(x, x')$$

where $\Pi(\mu, \nu) = \{\gamma \in \mathcal{P}(X \times X) \mid (\pi_1)_{\#\gamma} = \mu, (\pi_2)_{\#\gamma} = \nu\}$ where π_1, π_2 are projections onto the first and second coordinate. The set Π is the collection of couplings between μ and ν that takes μ and ν marginally. The minimiser of the Kantorovich problem is called the **optimal transport plan**.

It is not hard to notice that if we set c to be the Euclidean distance, then the Monge problem considers the transportations that are couplings of $\mu, \nu = T_{\#\mu}$, which are contained in Π. So, the Kantorovich problem is a weaker problem than the Monge problem. The Monge problem does not always have a solution, but it can be shown that when the optimal transport cost c is lower semicontinous[2], there always exists an optimal transport plan (Ambrosio et al., 2005).

2.1.1 The p-Wasserstein Distance

The objective function in the Kantorovich problem is a way to measure the distance between two probability densities μ, ν via some cost function c. In the case where c is the l_p distance, the objective function is called the **p-Wasserstein distance**, denoted by $\mathcal{W}_p$ defined by

$$\mathcal{W}_p^p(\mu, \nu) := \inf_{\gamma \in \Pi(\mu,\nu)} \int_{X \times X} \|x - y\|_p^p \ d\gamma(x, y).$$

and we will define the p-Wasserstein distance on the space $\mathcal{P}_p(\mathbb{R}^d)$ defined by

$$\mathcal{P}_p(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \ \middle| \ \int |x|_2^p d\mu(x) < \infty \right\}.$$

[1] $\mathcal{P}(X)$ denotes the set of probability measures defined on X.

[2] f is lower semicontinuous at x' if for any ε, there exists $\delta > 0$ such that $f(x') - \varepsilon < f(x)$ for all δ ball of x'. If f is lower semicontinuous everywhere, the function is lower semicontinuous.

Another way to measure the distances between two densities (and functions) is via the L_p distance, defined by

$$\|f-g\|_p := \int_X \|f(x)-g(x)\|_p dx.$$

They are, obviously, not the same, as illustrated roughly in Figure 1.

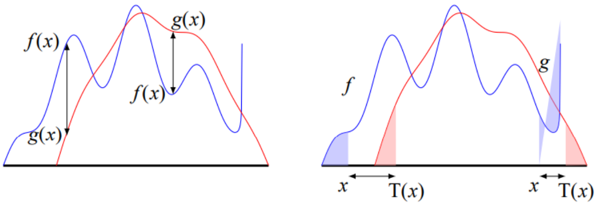

Figure 1: Comparison of L_p distance and $\mathcal{W}_p$ distance via transport map T, adapted from Santambrogio (2017).

In the rest of the notes, we will focus on the special case where $p = 2$ due to the various nice properties of $\mathcal{P}_2$ and $\mathcal{W}_2$ that we will explore next.

2.2 Dual and Solution to the Kantorovich Problem

So far, all we have considered is the solution to the Kantorovich problem and that its solution exists when the cost function is lower semicontinuous (and l_p cost functions are all lower semicontinuous). We have not investigated how one could *compute* a solution to the problem, which is often of more practical interest than knowing merely the existence of a solution. We will look more into the solution of the Kantorovich problem with $\mathcal{W}_2$ cost using the duality of the Kantorovich problem.

We will focus on the objective function $\mathcal{W}_2^2/2$ as it is nicer to work with. First, the marginal condition for the coupling could be rewritten as follows:

$$(\pi_1)_{\#\gamma} = \mu \iff \int f(x)d\gamma(x,y) = \int f(x)d\mu(x)\ \forall f \in L^1(\mu) = \left\{h \in L(\mathbb{R}^d) \mid \int |h|d\mu < \infty\right\}$$ [3]

and

$$(\pi_2)_{\#\gamma} = \nu \iff \int g(y)d\gamma(x,y) = \int g(y)d\nu(y)\ \forall g \in L^1(\nu) = \left\{h \in L(\mathbb{R}^d) \mid \int |h|d\nu < \infty\right\}.$$

[3] $L(\mathbb{R}^d)$ is the set of Lebesgue measurable functions on $\mathbb{R}^d$.

This allows up to rewrite $\mathcal{W}_2^2/2$ using sup:

$$\begin{aligned}\frac{1}{2}\mathcal{W}_2^2(\mu,\nu) &= \frac{1}{2}\inf_{\gamma\in\Pi(\mu,\nu)}\int \|x-y\|_2^2 d\gamma(x,y)\\ &= \frac{1}{2}\inf_{\gamma\in M_+(\mathbb{R}^d\times\mathbb{R}^d)}\sup_{f\in L^1(\mu),g\in L^1(\nu)}\Bigg[\int\frac{1}{2}\|x-y\|_2^2 d\gamma(x,y)\\ &\quad+\int fd\mu-\int f(x)d\gamma(x,y)+\int gd\nu-\int g(y)d\gamma(x,y)\Bigg]\\ &= \frac{1}{2}\sup_{f\in L^1(\mu),g\in L^1(\nu)}\inf_{\gamma\in M_+(\mathbb{R}^d\times\mathbb{R}^d)}\Bigg[\int\frac{1}{2}\|x-y\|_2^2-f(x)-g(y)d\gamma(x,y)\\ &\quad+\int fd\mu+\int gd\nu\Bigg]\\ &= \frac{1}{2}\sup_{(f,g)\in D(\mu,\nu)}\left[\int fd\mu+\int gd\nu\right]\end{aligned}$$

where $M_+(\mathbb{R}^d\times\mathbb{R}^d)$ is the space of non-negative finite measures on $\mathbb{R}^d\times\mathbb{R}^d$, and $D(\mu,\nu)$ is defined as

$$D(\mu,\nu) := \left\{(f,g)\in L^1(\mu)\times L^1(\nu)\mid f(x)+g(y)\leq\frac{1}{2}\|x-y\|_2^2\ \forall x,y\in\mathbb{R}^d\right\} \tag{3}$$

which is the set of f,g pairs that make the infimum of the integral zero rather than $-\infty$, i.e.

$$\inf_{\gamma\in M_+(\mathbb{R}^d\times\mathbb{R}^d)}\int\frac{1}{2}\|x-y\|_2^2-f(x)-g(y)d\gamma(x,y)=\begin{cases}0 & (f,g)\in D(\mu,\nu),\\ -\infty & \text{otherwise.}\end{cases}$$

The maximisers of the dual problem f,g are known as the **dual potential**. Therefore, we have turned the Kantorovich from an infimum problem to a supremum problem, and thus we have obtained the dual problem: Let $\mu,\nu\in\mathcal{P}^2(\mathbb{R}^d)$. the **dual Kantorovich problem** from μ to ν is the optimisation problem

$$\sup_{(f,g)\in D(\mu,\nu)}\left[\int fd\mu+\int gd\nu\right]$$

where D is the same as (3).

The following result summarises the key properties of the dual (and primal) problem. The proofs are omitted due to length constraints, and interested readers can find them at Ambrosio et al. (2005).

Theorem 2.1 (Fundamental Theorem of Optimal Transport). *Consider two densities $\mu,\nu\in\mathcal{P}_2(\mathbb{R}^d)$. Then, we have:*

1. *(strong duality) The value of the dual Kantorovich problem from μ to ν equals to $\mathcal{W}_2^2(\mu,\nu)/2$.*
2. *(existence of optimal dual potential) There exists an optimal pair (f^*,g^*) for the dual Kantorovich problem.*
3. *(optimal dual potential characterisation) The optimal pair is of the form*

$$f^*(x)=\|x\|_2^2/2-\varphi(x),\qquad g^*(y)=\|y\|_2^2/2-\varphi^*(y)$$

*where $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, called the **Kantorovich potential**, is a proper, convex, lower semicontinuous function and φ^* is its convex conjugate*[4]*.*

4. *(Brenier Theorem) If μ is also absolutely continuous w.r.t. the Lebesgue measure*[5]*, then the optimal transport plan is unique and it is induced by an **optimal transport map** T, i.e. for the optimal transport plan (μ^*, ν^*), we have $\nu^* = T_{\#\mu^*}$. Furthermore, the mapping T is characterised by the unique gradient of a proper, convex, lower semicontinuous function φ, called the **Brenier potential**, such that $T = \nabla\varphi$ and $(\nabla\varphi)_{\#\mu^*} = \nu^*$.*

The Brenier theorem indicates that under certain conditions on μ, ν, the optimal solution to the Kantorovich problem is the same as that of the Monge problem. Furthermore, in those cases, the solution can be captured by the optimal transport map, which is then characterised by the Brenier potential. In addition, using convex duality, $\nabla\varphi^* = (\nabla\varphi)^{-1}$. So, if ν is also absolutely continuous, then the optimal transport map from ν to μ would be $\nabla\varphi^*$. We would commonly denote the optimal transport map from μ to ν as $T_{\mu\to\nu}$ for simplicity. Also, as absolutely continuous measures are very convenient to work with, we will restrict our attention to those only, meaning that we will consider the space of absolutely continuous probability measures $\mathcal{P}^2_{ac}(\mathbb{R}^d) \subset \mathcal{P}^2(\mathbb{R}^d)$ instead.

2.3 2-Wasserstein Space and its Metric and Riemannian Structure

In the previous section, we have established some fundamental results of optimal transport and the Kantorovich problem when the cost function is chosen to be $\mathcal{W}_2$. This cost function, as it turns out, can be used as a metric on the space of probability measures $\mathcal{P}_2(\mathbb{R}^d)$. This space is commonly called the **2-Wasserstein space**. Furthermore, it also possesses a Riemannian structure. These two key properties allow us to study the motion of probability measures and provide a sensible notion of gradient in this space, which would be fundamental in the next section when we establish the Wasserstein gradient flow.

2.3.1 2-Wasserstein Space is a Metric Space

First, we will establish that the 2-Wasserstein space is indeed a metric space.

Proposition 2.2. *$(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is a metric space.*

Proof. We would like to show that $\mathcal{W}_2$ is indeed a metric on $\mathcal{P}_2(\mathbb{R}^d)$. The first two properties of a metric are trivial to show. Firstly, $\mathcal{W}_2$ is non-negative by definition as the l_2 norm is non-negative, and is symmetric for the same reason. Also, when $\mu = \nu$ a.e., we would have $\mathcal{W}_2(\mu, \mu) = 0$ by using the trivial coupling of product measure $\mu \otimes \mu$. Conversely, if $\mathcal{W}_2(\mu, \nu) = 0$, we would have $\|X - Y\| = 0$ a.s. by coupling γ, so $X = Y$ a.s. and thus $\mu = \nu$. We are now left with establishing the triangle inequality of $\mathcal{W}_2$.

An auxiliary lemma is needed here: for $\gamma_{1,2} \in \mathcal{P}_2(X \times Y)$ and $\gamma_{2,3} \in \mathcal{P}_2(Y \times Z)$ with the same marginal distributions on Y, there exists $\gamma_{1,2,3} \in \mathcal{P}_2(X \times Y \times Z)$ such that its marginal on $X \times Y$ is $\gamma_{1,2}$ and its marginal on $Y \times Z$ is $\gamma_{2,3}$. A proof of this can be found in Berti et al. (2015).

[4] φ^* is the convex conjugate of φ if we have $\varphi^*(y) = \sup_{x\in\mathbb{R}^d}[\langle x, y\rangle - \varphi(x)]$.

[5] A measure μ is absolutely continuous w.r.t. the Lebesgue measure λ if for any measurable subset A, we have $\lambda(A) = 0 \implies \mu(A) = 0$.

Consider $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$, we let $\gamma_{1,2}$ be the optimal coupling of μ_1, μ_2 in terms of $\mathcal{W}_2$ and $\gamma_{2,3}$ be the optimal coupling of μ_2, μ_3 in terms of $\mathcal{W}_2$. By the lemma above, there exists γ that glues up $\gamma_{1,2}$ and $\gamma_{2,3}$, and we let $\gamma_{1,3}$ denote the marginal of γ on the first and third coordinates. Then, we have

$$\begin{aligned}
\mathcal{W}_2(\mu_1, \mu_3) &= \sqrt{\int \|x_1 - x_3\|_2^2 d\gamma_{1,3}(x_1, x_3)} \\
&\leq \sqrt{\int \|x_1 - x_2\|_2^2 + \|x_2 - x_3\|_2^2 d\gamma(x_1, x_2, x_3)} \\
&\leq \sqrt{\int \|x_1 - x_2\|_2^2 d\gamma(x_1, x_2, x_3)} + \sqrt{\int \|x_2 - x_3\|_2^2 d\gamma(x_1, x_2, x_3)} \\
&= \sqrt{\int \|x_1 - x_2\|_2^2 d\gamma(x_1, x_2)} + \sqrt{\int \|x_2 - x_3\|_2^2 d\gamma(x_2, x_3)} \\
&= \mathcal{W}_2(\mu_1, \mu_2) + \mathcal{W}_2(\mu_2, \mu_3)
\end{aligned}$$

as desired. □

Knowing the metric space structure, we could study the dynamics of elements in the space, which are probability measures in this case. This would serve as a key component when we study gradient flow in Wasserstein space in the next section.

Consider a curve $t \mapsto \mu_t \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, we say the curve is **absolutely continuous** if for all t,

$$|\dot{\mu}|(t) := \lim_{s \to t} \frac{\mathcal{W}_2(\mu_s, \mu_t)}{|s - t|} < \infty.$$

We would also call $|\dot{\mu}|$ the **metric derivative** of the curve. This notion of derivative would help us to view the movement of measures as flows of fluid, which then would allow us to establish the continuity equation. The flows of fluid can be interpreted via the Lagrangian lens and via the Newtonian lens, where the Lagrangian viewpoint focuses on the trajectory of the particles and the Newtonian viewpoint focuses on the evolution of fluid density.

Suppose that $X_0 \sim \mu_0$ and $t \mapsto X_t$ evolves according to some ODE $\dot{X}_t = v_t(X_t)$ for vector field family (v_t). The ODE describes the motion of particle trajectories and thus is the Lagrangian viewpoint of the motion.

Correspondingly, we could have the following evolution equation to represent the Newtonian perspective.

Theorem 2.3. *Let $t \mapsto v_t$ be a family of vector fields such that the random variables $t \mapsto X_t$ evolve according to $\dot{X}_t = v_t(X_t)$. Then, the law μ_t of X_t evolves according to the* ***continuity equation***

$$\partial_t \mu_t + div(\mu_t v_t) = 0.$$

Proof. Given any test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have

$$\begin{aligned}\int \varphi \partial_t \mu_t = \partial_t \int \varphi d\mu_t &= \partial_t \mathbb{E}[\varphi(X_t)] \\ &= \mathbb{E}[\langle \nabla \varphi(X_t), \dot{X}_t \rangle] = \mathbb{E}[\langle \nabla \varphi(X_t), v_t(X_t) \rangle] \\ &= \int \langle \nabla \varphi, v_t \rangle d\mu_t = -\int \varphi \mathrm{div}(\mu_t v_t)\end{aligned}$$

and that gives us the desired continuity equation as it holds for any φ. □

The interpretation of this result is that, for any nice curve of measures $t \mapsto \mu_t$, we can interpret it as a fluid flow along a family of vector fields, although this family of fields is not necessarily unique as we can add an additional vector field (w_t) to any feasible (v_t) as long as $\mathrm{div}(v_t w_t) = 0$. We will then look for the optimal choice of family.

The first thing we want our vector field (v_t) to have is that we want it to minimise $\int \|v_t\|_2^2 d\mu_t$ which can be interpreted as the kinetic energy. Next, we would like v_t to be the gradient of some function, as that would make it more natural to characterise optimal transport maps. The following result summarises them and establishes the choice of the optimal family of vector fields. The proof is omitted here but could be found in Chewi (2023) as Theorem 1.3.19.

Theorem 2.4 (curves of measures as fluid flows)**.** *Let $t \mapsto \mu_t$ be an absolutely continuous curve of measures. We have*

1. *For any family of vector fields $t \mapsto \tilde{v}_t$ satisfying the continuity equation, we have $|\dot{\mu}|(t) \leq \|\tilde{v}_t\|_{L^2(\mu(t))}$ for all t.*
2. *Conversely, there exists a unique choice of vector fields $t \mapsto v_t$ such that the continuity equation holds and $\|v_t\|_{L^2(\mu(t))} \leq |\dot{\mu}|(t)$ for all t. The choice of vector fields is also characterised by the fact that the continuity equation holds for some function $\psi_t : \mathbb{R}^d \to \mathbb{R}$ with $v_t = \nabla \psi_t$ for all t.*

Furthermore, the optimal vector field (v_t) produced by the two results above satisfies

$$v_t = \lim_{\delta \searrow 0} \frac{T_{\mu_t \to \mu_{t+\delta}} - id}{\delta}$$

where $T_{\mu_t \to \mu_{t+\delta}}$ is the optimal transport from μ_t to $\mu_{t+\delta}$, and id is the identity map where $id_{\#\mu} = \mu$ for any μ.

As a consequence of this result, the optimal vector field v_t would satisfy $\|\tilde{v}_t\|_{L^2(\mu(t))} \leq |\dot{\mu}|(t)$. The metric derivative is supposed to be the "magnitude of the velocity".

2.3.2 2-Wasserstein Space and Riemannian Structure

So far, we have established the metric space structure of the 2-Wasserstein space. Next, we will move on to study the Riemannian structure of the same space. One major motivation for introducing some geometry into the space is that, since we have established some nice forms of curves, we would like to know about geodesics (shortest distance paths).

Before indicating the Riemannian structure of the Wasserstein space, we will first give a quick and informal description of Riemannian geometry and describe what a **Riemannian manifold** is.

A (smooth) **manifold** $\mathcal{M}$ can be defined as a space that is locally homeomorphic[6] to some Euclidean space $\mathbb{R}^n$, and such a manifold would then be of dimension n. For example, a circle S^1 embedded in $\mathbb{R}^2$ with only the circumference and no filling is locally homeomorphic to $\mathbb{R}^1$. If we zoom into one point on the circumference of the circle and enlarge everything, we would get something that is very close to a straight line. Therefore S^1 is of dimension 1, and thus the superscript 1. Using a similar logic, the sphere S^2 embedded in $\mathbb{R}^3$ (e.g. a round version of the Earth) is locally homeomorphic to $\mathbb{R}^2$, and walking on Earth would feel like walking on flat land although the Earth itself is round[7]. Given a manifold $\mathcal{M}$, we can associate any point $p \in \mathcal{M}$ to a vector space $T_p\mathcal{M}$ of all possible velocities of curves passing through p, and we call such a space the **tangent space** at p.

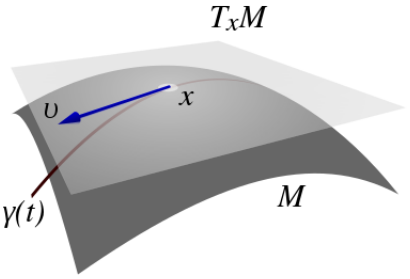

Figure 2: Tangent Space T_xM of manifold M at point x. Taken from Wikipedia.

A **Riemannian metric** is a smoothly varying choice of inner products $p \mapsto \langle \cdot, \cdot \rangle_p$ on the tangent space. This metric allows us to locally measure things such as the angles between two intersecting curves. In the context of gradient flow, this notion allows us to define the steepest descent direction for an objective function. Given a Riemannian metric, we can induce a distance function as

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

where $\dot{\gamma}(t)$ is the tangent vector to the curve γ at time t (note that this vector does not live on the manifold $\mathcal{M}$ but the tangent space $T_{\gamma(t)}\mathcal{M}$). The norm is w.r.t. the inner product on the tangent space $T_{\gamma(t)}\mathcal{M}$. The minimiser of the above infimum, if it exists, would be the **geodesic** curve γ. If in addition, speed of the curve $\|\dot{\gamma}(t)\|_{\gamma(t)}$ is constant for all t, the geodesic is called a **constant-speed geodesic**, and this will be the only type of geodesics we care about for the rest of the notes, and we will drop the prefix and call them only geodesics for simplicity.

Given a functional $F : \mathcal{M} \to \mathbb{R}$, the gradient of F at p is defined to be the unique element $\nabla F(p) \in T_p\mathcal{M}$ such that all curves (p_t) passing through p at time 0 with speed $v \in T_p\mathcal{M}$ satisfy $\partial_t F(p_t)|_{t=0} = \langle \nabla F(p), v \rangle_p$. This will be key to the construction of gradient flow in Wasserstein space.

A manifold $\mathcal{M}$ equipped with a Riemannian metric is thus known as a **Riemannian manifold**. We will show next that the 2-Wasserstein space is a Riemannian manifold.

[6] A function $f : X \to Y$ is a homeomorphism if f is a continuous bijection and the inverse is also continuous. If there exists a homeomorphism between spaces X and Y, then we say the two spaces are homeomorphic. A space X is locally homeomorphic to space Y if every point of X has a neighbourhood that is homeomorphic to an open subset of Y.

[7] at least to the non-believers of a flat Earth ...

The space of interest is the space of absolutely continuous probability measures with finite variances defined on $\mathbb{R}^d$, i.e. the space $\mathcal{P}_{2,ac}(\mathbb{R}^d)$. To define the tangent space for any $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, we have

$$T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) := \overline{\{\nabla\psi \mid \psi \in C_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)}$$

where the overline denotes the $L^2(\mu)$ closure[8] and $C_c^\infty(\mathbb{R}^d)$ is the space of compactly supported[9], smooth functions defined on $\mathbb{R}^d$. Using Theorem 2.4 about the optimal transport map, we can rewrite the tangent space as

$$T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d) := \overline{\{\lambda(T - \mathrm{id}) \mid \lambda > 0, T \text{ is an optimal transport map}\}}^{L^2(\mu)}.$$

We can equip this Riemannian metric with the L^2 norm to the space $\mathcal{P}_{2,ac}(\mathbb{R}^d)$. For this space to be a Riemannian manifold, we need to show that $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ is a manifold, which it **is not**. Nevertheless, we can still make use of the Riemannian structure of this space. To make a better connection with the 2-Wasserstein distance, we would want to show that the distance induced by the metric coincides with $\mathcal{W}_2$, and we would also want to know what a geodesic would look like in this case. These two points are illustrated in the following theorem. The proof is omitted as usual, and a detailed proof can be found in Chewi (2023) as Theorem 1.3.22.

Theorem 2.5 (Wasserstein geodesics). *Let $\mu_0, \mu_1 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. Then,*

$$W_2(\mu_0, \mu_1) := \inf \left\{ \int_0^1 \|v_t\|_{L^2(\mu_t)} dt \,\middle|\, \partial_t \mu_t + div(\mu_t v_t) = 0 \right\}.$$

The minimiser of the above infimum can be achieved as follows: let $X_0 \sim \mu_0$ and $X_1 \sim \mu_1$ be optimally coupled and let $X_t := (1-t)X_0 + tX_1$, and let $\mu_t := L(X_t)$ be the law of the random variable at time t. Then, $t \mapsto \mu_t$ is the unique constant-speed geodesic joining μ_0 to μ_1.

The minimising curve is called the **Wasserstein geodesic** joining μ_0 to μ_1, and it is also called the **displacement interpolation** or **McCann interpolation**. If there exists an optimal transport map T between μ_0 and μ_1, then the geodesic would be of the form

$$\mu_t = ((1-t)\mathrm{id} + tT)_{\#\mu_0}$$

for $t \in [0, 1]$. One should note that this is certainly different from the linear/mixture interpolation between μ_0 and μ_1, which is of the form

$$\mu_t := (1-t)\mu_0 + t\mu_1$$

for $t \in [0, 1]$. The two forms of interpolations are graphically compared below.

2.4 Wasserstein Gradient Flow and Sampling

In this section, we will finally show how one could define a gradient flow in the Wasserstein space, after an extended study of the basic properties of this space. The close connection between gradient flow in Wasserstein space and the Langevin diffusion would be illustrated, in the sense that the Langevin diffusion could be viewed as a gradient flow in the Wasserstein space,

[8] The $L^2(\mu)$ closure of a set includes all the elements of $L^2(\mu)$ such that the (possibly) enlarged set would be closed in $L^2(\mu)$ w.r.t. to its topology.

[9] A function f defined on X is compactly supported if the subset of domain of f of which the image is non-zero is compact, i.e. the set $\{u \in X \mid f(u) \neq 0\}$ is compact in X.

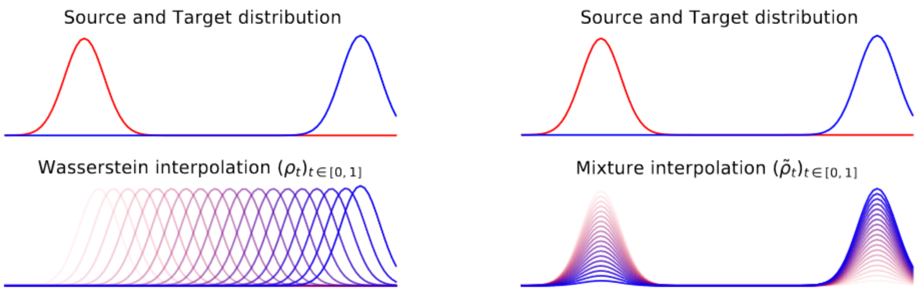

Figure 3: Comparison of Wasserstein Interpolation and Mixture Interpolation (Korba and Salim, 2022)

first introduced in Jordan et al. (1998). Additionally, we would also draw connections between sampling and optimisation, i.e. how certain sampling (from a distribution) techniques can be viewed as optimisation of the right objective function. Further examples of using gradient flow to study sampling algorithms will be introduced in the next chapter.

2.4.1 Wasserstein Gradient Flow

In the previous chapter, we looked at how one could define a gradient flow in the Euclidean space. Here, we will look at the Wasserstein gradient flow - gradient flows defined in the Wasserstein space $(\mathcal{P}_{2,ac}(\mathbb{R}^d), \mathcal{W}_2)$. This uses both the metric space structure and the Riemannian structure of the Wasserstein space, as they help us understand how one could study the motion of objects in the space (via the continuity equation) and how one could differentiate measures in the space (via tangent spaces). Recall that we have the gradient flow in Euclidean space

$$\dot{x}_t = -\nabla F(x_t)$$

for some (convex) function F and $x_t \in \mathbb{R}^d$. Trying to move things into the Wasserstein space where elements are probability measures μ_t instead of vectors x_t, we would want something like

$$\dot{\mu}_t \overset{?}{=} -\nabla_{\mathcal{W}_2} F(\mu_t) \tag{4}$$

for some functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$. The added subscript under the gradient operator above is because the geometry of the Wasserstein space is different from that of the Euclidean space, and the notion of gradient would require a slightly different form w.r.t. the right notion of distance, which is $\mathcal{W}_2$ in our case.

The left-hand side of the equation (4) is relatively well-defined at this stage since we have already studied the continuity equation for curves $t \mapsto \mu_t$. Therefore, we have

$$\dot{\mu}_t = \partial_t \mu_t = -\text{div}(\mu_t v_t)$$

where v_t is a family of vector fields. The choice of the family will become obvious soon, as we would want to make it match with the right-hand side of the gradient flow.

Next, we will look at the right-hand side of the equation (4). Consider some functional F : $\mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$, we would like to compute its gradient in the Wasserstein space at point

μ, which really means that we want to find the element $\nabla F(\mu) \in T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$ such that for any curve $t \mapsto \mu_t$ with $\mu_0 = \mu$, we have

$$\partial_t F(\mu_t)|_{t=0} = \langle \nabla_{\mathcal{W}_2} F(\mu), v_0 \rangle_\mu$$

for the tangent vector v_0 of the curve at time 0 with $\langle \cdot, \cdot \rangle_\mu$ being the inner product of the tangent space $T_\mu \mathcal{P}_{2,ac}(\mathbb{R}^d)$. We will give a formula for this quantity using the **first variation** of F at μ, which we will denote as $\delta F(\mu)$ and is defined to satisfy

$$\partial_t F(\mu_t)|_{t=0} = \int \delta F(\mu) \partial_t \mu_t|_{t=0}.$$

Using the continuity equation, we then have

$$\partial_t F(\mu_t)|_{t=0} = \int \delta F(\mu) \partial_t \mu_t|_{t=0} = -\int \delta F(\mu) \operatorname{div}(\mu v_0) = \int \langle \nabla \delta F(\mu), v_0 \rangle d\mu$$

for ∇ being the Euclidean gradient. Therefore, we have the **Wasserstein gradient** of the functional F at μ to be defined as

$$\nabla_{\mathcal{W}_2} F(\mu) = \nabla \delta F(\mu).$$

Now, combining what we have derived so far, we would have the **Wasserstein gradient flow** of F is by definition a curve of measures $t \mapsto \mu_t$ such that the tangent vector v_t at time t is $v_t = -\nabla_{\mathcal{W}_2} F(\mu_t)$, which will then give us the gradient flow equation

$$\partial_t \mu_t = \operatorname{div}(\mu_t \nabla_{\mathcal{W}_2} F(\mu_t) = \operatorname{div}(\mu_t \nabla \delta F(\mu_t)).$$

2.4.2 Langevin Diffusion as Gradient Flow

Consider the functional $F = \mathrm{KL}(\cdot|\pi)$ for some probability measure π and we wish to minimise this quantity using Wasserstein gradient flow. Before going into setting up the gradient flow, we first explore slightly the notion of KL divergence and why we would wish to minimise this quantity.

The KL divergence $\mathrm{KL}(\cdot|\cdot)$ is a commonly used measure of distance between probability distributions in Statistics and machine learning, defined as

$$\mathrm{KL}(\mu|\nu) := \int \log \frac{\mu}{\nu} \, d\mu$$

where μ, ν are probability measures and μ is absolutely continuous w.r.t. ν. One key property of KL divergence is that

$$\mathrm{KL}(\mu|\nu) \geq 0$$

for any μ, ν, and $\mathrm{KL}(\mu|\nu) = 0$ if and only if $\mu = \nu$ as measures. This means

$$\pi = \arg\min_{q \in Q} \mathrm{KL}(q|\pi)$$

as long as $\pi \in Q$. This property leads to the study of variational inference (Blei et al., 2017), where we try to learn about a complicated distribution π by optimising the KL divergence between π and elements of a sufficiently large class of distribution Q. In practice, the class of distribution is usually assumed to be of some parametric form, such as the class of Gaussian distributions with varying mean and variance parameters.

Another nice thing about KL divergence and the optimisation approach, especially in the case of Bayesian computation, is that it will still work even if the probability measure could be known up to a multiplicative constant, as we have, for constant $C > 0$,

$$\mathrm{KL}(\mu|C\nu) = \int \log \frac{\mu}{C\nu} \; d\mu = \int \left[\log \frac{\mu}{\nu} - \log C\right] \; d\mu = \mathrm{KL}(\mu|\nu) - \log C$$

so doing the optimisation using $\mathrm{KL}(\mu|C\nu)$ with varying μ would yield the same result as optimising $\mathrm{KL}(\mu|\nu)$.

KL divergence, in addition, can be used to do sampling. One (somewhat restrictive) way of thinking about sampling from a target distribution π is that we wish to construct a distribution of the form

$$\pi_N := \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$$

where $x_1, x_2, \ldots, x_N$ are samples from π and δ_{x_i} is the Dirac point mass at x_i. The π_N distribution is essentially a histogram made by N samples from π. We wish to approximate π using π_N for a fixed N, and the closeness of the approximation can be measured by none other than the KL divergence between them, i.e. $\mathrm{KL}(\pi_N|\pi)$. Different choices of samples $x_1, x_2, \ldots, x_N$ would then influence the KL divergence, and the minimising set of samples would be the desired output of the optimisation.[10] Therefore, we have established one link between sampling and optimisation, via minimising the KL divergence. This should provide sufficient reasons for constructing the Wasserstein gradient flow using KL divergence.

Recall from the start of this subsection that we wish to minimise the function $F = \mathrm{KL}(\cdot|\pi)$, and we assume here that $\pi \propto \exp(-V)$ for some function V, often viewed as the potential energy due to its connection with the Boltzmann distribution from statistical Physics (Faulkner and Livingstone, 2022). Using this definition, we have

$$F(\mu) = \int \log \frac{\mu}{\pi} d\mu = \int \log \mu d\mu - \int (-V) d\mu = \int V d\mu + \int \log \mu d\mu$$

and the first term can be interpreted as the *energy* part, and the second can be viewed as the (negative) *entropy* part. Using the definition of the first variation, we could have

$$\delta F(\mu) = V + \log \mu + \text{constant}$$

so the Wasserstein gradient of F becomes

$$\nabla_{\mathcal{W}_2} F(\mu) = \nabla V + \nabla \log \mu = \nabla \log \frac{\mu}{\pi}.$$

The Wasserstein gradient flow of F that we wished for in equation (4), therefore, is

$$\partial_t \mu_t = \mathrm{div}\left(\mu_t \nabla \log \frac{\mu_t}{\pi}\right). \tag{5}$$

[10] One should compare this approach to the more classical sampling approach via importance sampling, MCMC, etc. For importance sampling and MCMC, for example, the idea behind them is that we first draw samples from an easy-to-sample-from alternative distribution, and then correct the samples so that they become exact samples from the target distribution. It should not be too hard to notice that the variational approach to sampling is fundamentally different.

Note that for the Langevin diffusion

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$$

with $\{B_t\}_t$ being a standard Brownian motion has the stationary distribution $\pi \propto \exp(-V)$, and its Fokker-Planck equation is

$$\partial_t \pi_t = \mathcal{L}^* \pi_t = \mathrm{div}(\pi_t \nabla \log \pi_t/\pi) \tag{6}$$

where π_t is the law of X_t and $\mathcal{L}^*$ is the adjoint of the infinitesimal generator of the Langevin diffusion. More detail on the derivations of these results can be found in Oksendal (2013).

Comparing the Wasserstein gradient flow and the Fokker-Planck equation of the Langevin diffusion, we can see that **the law $t \mapsto \pi_t$ of $\{X_t\}$ from the Langevin diffusion with stationary distribution $\pi \propto \exp(-V)$, provided by equation (6), coincides with the Wasserstein gradient flow of KL$(\cdot|\pi)$ of equation (5).** This amazing connection between gradient flow and the Langevin diffusion was first drawn in Jordan et al. (1998).

2.5 Discretisations and Convergence Analysis of Wasserstein Gradient Flow

In Chapter 1, after introducing the gradient flow in the Euclidean space, we studied two ways of discretizing the ODE (explicitly and implicitly) and studied some convergence rate results of the gradient flow under varying convexity results of the objective function F. We will do the same in this section with the Wasserstein gradient flow.

2.5.1 Time-Discretisation of Wasserstein Gradient Flow

Consider a functional $F : \mathcal{P}_{2,ac}(\mathbb{R}^d) \to \mathbb{R}$ and the Wasserstein gradient flow would be of the form

$$\partial_t \mu_t = \mathrm{div}(\mu_t \nabla_{\mathcal{W}_2} F(\mu_t)) = \mathrm{div}(\mu_t \nabla \delta F(\mu_t)).$$

Using the explicit Euler discretisation of the above ODE, we would get the update scheme of the form

$$\mu_{m+1} = \mu_m - \gamma \nabla_{\mathcal{W}_2} F(\mu_m)_{\#\mu_m} = (\mathrm{id} - \gamma \nabla_{\mathcal{W}_2} F(\mu_m))_{\#\mu_m}$$

which is just pushing forward the current state μ_m along the Wasserstein geodesic with stepsize $\gamma > 0$. In the case where $F(\mu) = \mathrm{KL}(\mu|\pi)$, we would have

$$\nabla_{\mathcal{W}_2} F(\mu_m) = \nabla \log(\mu_m/\pi)$$

in the update step, which involves the density of μ_m and this density is not always attainable.

Using the implicit Euler discretisation, which is the same discretisation used in Jordan et al. (1998), we would get the update scheme of the form

$$\mu_{m+1} \in \arg \min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left[\gamma F(\mu) + \frac{1}{2} \mathcal{W}_2^2(\mu, \mu_m) \right]$$

using the proximal point ideas as mentioned in Chapter 1. Note that this involves minimising over $\mathcal{W}_2$, which is very hard to compute in practice when the probability measures we use to compute are not of some very nice form. This makes this scheme, often called the JKO scheme, impractical.

The direct approach of converting a continuous-time gradient flow to a discrete-time optimisation using the Euler method is not feasible here. One potential way to work around it is via splitting schemes, where at each full iteration, we do half a step of one update and half a step of another update. This may sometimes bypass some implementation problems. Some examples of such splitting schemes will be shown in Chapter 3.

2.5.2 Convergence Analysis of Wasserstein Gradient Flow

Now we shall switch focus back to the continuous time setting, and look at the various convexity conditions of $F = \mathrm{KL}(\cdot|\pi)$ under which the Wasserstein gradient flow would converge at certain rates.

Consider the curves $t \mapsto \mu_t$ in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ moving along the Wasserstein geodesic so there exists an optimal transport map T such that $\mu_t = [(1-t)\mathrm{id} + tT]_{\#\mu_0}$. We have

$$\partial_t F(\mu_t) = \langle \nabla_{\mathcal{W}_2} F(\mu_t), T - \mathrm{id}\rangle_{\mu_t}$$

when we pick the optimal vector field $T - \mathrm{id} \in T_{\mu_0}\mathcal{P}_{2,ac}(\mathbb{R}^d)$ as outlined in Theorem 2.4. Then, by taking the second derivative, we have

$$\partial_t^2 F(\mu_t)|_{t=0} = \langle \nabla_{\mathcal{W}_2} F(\mu_t)(T - \mathrm{id}), T - \mathrm{id}\rangle_{\mu_0}$$

since the acceleration term is zero due to the geodesic being of constant speed. If we can further show that the above quantity has a lower bound $\alpha\|T - \mathrm{id}\|_{\mu_0}^2$ for any μ_0 and T, then we have established that F is α-strongly convex.

Theorem 2.6. *For $\pi \propto \exp(-V)$ with α-strongly convex V, $KL(\cdot|\pi)$ along the Wasserstein geodesic is α-strongly convex.*

Proof. Consider $\pi \propto \exp(-V)$. We have

$$\mathrm{KL}(\mu|\pi) = \int \log \frac{\mu}{\pi}\mu = \underbrace{\int \log \mu d\mu}_{=:H(\mu)} + \underbrace{\int V d\mu}_{=:\mathcal{E}(\mu)}$$

where $H(\mu)$ is the negative entropy and $\mathcal{E}(\mu)$ is the potential energy. Along the Wasserstein geodesic, we have $X_t = (1-t)X_0 + tT(X_0)$ where $X_0 \sim \mu$ and T is the optimal transport map.

Studying the second derivative of F can be broken down into studying the second derivation of H and $\mathcal{E}$, which is what we are going to do here. First, we look at the derivatives of $\mathcal{E}$. We have

$$\begin{aligned}\partial_t \mathcal{E}(\mu_t) &= \partial_t \mathbb{E}[V(X_t)] = \mathbb{E}[\langle \nabla V(X_t), \dot{X}_t\rangle] = \mathbb{E}[\langle \nabla V(X_t), T(X_0) - X_0\rangle]\\ \partial_t^2 \mathcal{E}(\mu_t) &= \mathbb{E}[\langle \nabla V(X_t)(T(X_0) - X_0), (T(X_0) - X_0)\rangle].\end{aligned}$$

If V is α-strongly convex, the second derivative of $\mathcal{E}$ would be lower bounded by $\alpha\|T - \mathrm{id}\|_{\mu_0}^2$, meaning that $\mathcal{E}$ is α-strongly convex too. Next, we will look at H, which is not as straightforward as $\mathcal{E}$. We define $T_t := (1-t)\mathrm{id} + tT$ so that $(T_t)_{\#\mu_0} = \mu_t$. We can apply the change of variable formula of push forward and get $\det\nabla T_t = \mu_0/(\mu_t \circ T_t)$. So,

$$H(\mu_t) = \int \log \mu_t d\mu_t = \int \log(\mu_t \circ T_t) d\mu_0 = \int \log \frac{\mu_0}{\det\nabla T_t} d\mu_0 = H(\mu_0) - \int \log \det\nabla T_t d\mu_0.$$

Taking the second derivative and after some careful derivations, we could obtain

$$\partial_t^2 H(\mu_t) = \int \|\nabla T - \mathrm{id}\|_{HS}^2 d\mu_0 \geq 0$$

where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm. So, the α-strongly convex of H is established too. Combining the two results gives us the theorem.

□

The following result follows from the strong convexity under a Riemannian structure.

Corollary 2.7. *Consider $\pi \propto \exp(-V)$ with α-strongly convex V, we have*

$$KL(\nu|\pi) \geq KL(\mu|\pi) + \langle \nabla \log \frac{\mu}{\pi}, T_{\mu\to\nu} - id\rangle_\mu + \frac{\alpha}{2}\mathcal{W}_2^2(\mu,\nu)$$

for any $\mu, \nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$.

We will now exploit these results to understand the convergence of Wasserstein gradient flow. First, we want to get the basic results that say the Wasserstein gradient flow is indeed decreasing, and the conditions needed for it to converge quickly.

Proposition 2.8. *The Wasserstein gradient flow is always decreasing.*

Proof. Let the Wasserstein gradient flow be $t \mapsto F(\mu_t)$ for some functional F with $\inf F = 0$, we have

$$\partial_t F(\mu_t) = \langle \nabla_{\mathcal{W}_2} F(\mu_t), \dot{\mu}_t\rangle_\mu = \langle \nabla_{\mathcal{W}_2} F(\mu_t), -\nabla_{\mathcal{W}_2} F(\mu_t)\rangle_\mu = -\|F(\mu_t)\|_\mu \leq 0$$

as we chose the optimal vector field for the continuity equation. □

Proposition 2.9. *Under the* ***gradient domination*** *condition, or the* ***Polyak–Lojasiewicz (PL) inequality****, i.e.*

$$\|\nabla_{\mathcal{W}_2} F(\mu)\|_\mu \geq 2\alpha F(\mu)$$

for some $\alpha > 0$ and all $\mu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, the Wasserstein gradient flow converges exponentially fast.

Proof. The result follows directly from the previous proposition and the Gronwall inequality (Boyd and Vandenberghe, 2004). □

Proposition 2.10. *The α-strongly convex property of V implies the Polyak-Lojasiewicz inequality.*

Proof. The α-strongly convex property of $F = \mathrm{KL}(\cdot|\pi)$ gives us

$$F(\nu) \geq F(\mu) + \langle \nabla_{\mathcal{W}_2} F(\mu), T_{\mu\to\nu} - \mathrm{id}\rangle_\mu + \frac{\alpha}{2}\mathcal{W}_2^2(\mu,\nu).$$

If we pick $\nu := \arg\min F(\cdot)$ when we assume (WLOG) $\inf F = 0$, we would get

$$\begin{aligned} F(\mu) &\leq -\langle \nabla_{\mathcal{W}_2} F(\mu), T_{\mu\to\nu} - \mathrm{id}\rangle_\mu - \frac{\alpha}{2}\mathcal{W}_2^2(\mu,\nu). \\ &\leq \frac{1}{2\alpha}\|\nabla_{\mathcal{W}_2} F(\mu)\|^2 + \frac{\alpha}{2}\|T_{\mu\to\nu} - \mathrm{id}\|_\mu^2 - \frac{\alpha}{2}\mathcal{W}_2^2(\mu,\nu) = \frac{1}{2\alpha}\|\nabla_{\mathcal{W}_2} F(\mu)\|^2 \end{aligned}$$

using the fact that $\|T_{\mu\to\nu} - \mathrm{id}\|_\mu = \mathcal{W}_2(\mu,\nu)$ and the last inequality is due to Young's inequality.

□

Combining the previous two propositions gives us the fact that under α-strong convexity of V, the Wasserstein gradient flow produces an exponentially fast converging $F(\mu_t)$ in terms of KL divergence. We could also recover the log Sobolev inequality when we apply the above result to the Langevin diffusion. We will finish this section by stating without proof the following result that will re-appear in Section 3.1. A proof can be found as Theorem 23.9 of Villani et al. (2009).

Theorem 2.11. *For $\nu \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, the Wasserstein gradient of $\mu \mapsto \mathcal{W}_2(\mu, \nu)$ at μ is $-2(T_{\mu\to\nu} - id)$.*

Chapter 3

Sampling Algorithms as Gradient Flows

In this chapter, we will look at various sampling algorithms in Statistics and machine learning, and recognise their underlying gradient flow structure. This would then help us establish various convergence results about the sampling algorithms, and could also help us design variants to the standard algorithms. The algorithms that we will look at are the Langevin Monte Carlo / unadjusted Langevin algorithm, the Stein variational gradient descent, and the denoising diffusion model.

3.1 Langevin Monte Carlo as Gradient Flow

The Langevin diffusion is an SDE of the form

$$dX_t = -\nabla V(X_t)dt + \sqrt{2}dB_t$$

where $\{B_t\}$ is a standard Brownian motion. The stationary distribution of X_t is $\pi \propto \exp(-V)$ under regularity conditions of V, as outlined in Roberts and Tweedie (1996). So, if we wish to draw samples from the target distribution π, we could start from a starting point X_0 and let it follow the Langevin diffusion for a sufficient amount of time, wait until it converges to equilibrium, and the trajectory of X_t afterwards would all be samples from π. The practical problem with this idea, however, is that it is not an easy task to sample directly from a continuous SDE. One approach to (partially) resolve this issue is by taking an Euler discretisation of the Langevin diffusion, and we would get

$$X_{h(n+1)} = X_{hn} - h\nabla V(X_{hn}) + \sqrt{2h}\varepsilon$$

where $\varepsilon \sim N(0, 1)$. This is known as the **Langevin Monte Carlo** (LMC) in machine learning literature and **unadjusted Langevin algorithm** (ULA) in computational statistics literature Roberts and Tweedie (1996). The reason why this is called ULA is that it can be viewed as a Metropolis-adjusted Langevin algorithm (MALA) in the MCMC literature without the Metropolis-adjustment step at each iteration (Xifara et al., 2014). The output of LMC would not be exact samples from the target distribution π, while the output of MALA would be exact due to the added Metropolis adjustment step. One might think the bias of LMC output is due to the time discretisation of the Langevin diffusion. However, gradient descent, as time discretisation of gradient flow in the Euclidean space, converges under certain conditions on the objective function F (Boyd and Vandenberghe, 2004). Using similar logic, we would expect LMC, as time discretisation of gradient flow in the Wasserstein space, to converge under certain conditions on the target distribution π, but that is not the case (Wibisono, 2018). We will explore this point later in this section.

LMC is being used and analysed heavily despite its lack of exactness as it is easy to implement. The analysis of LMC started with Roberts and Tweedie (1996), and a major improvement was then made by Dalalyan (2017) where optimisation tools (for convergence analysis of gradient descent) were introduced into the study of LMC. Essentially, they break down the problem into studying the convergence of the Langevin diffusion and analysing the discretisation error by the LMC. A recent breakthrough in the theories of LMC is due to Durmus et al. (2019), where they built on the realisation made by Jordan et al. (1998) on the connection between Langevin diffusion and Wasserstein gradient flow, and exploited convex optimisation techniques for convergence rate analysis of gradient flow to study LMC. In this section, we will look into the theories developed in Durmus et al. (2019).

Following the observation in Wibisono (2018), LMC uses a **forward-flow splitting scheme** to discretise the Langevin diffusion. The following table lists how LMC updates the position and the distribution with stepsize h. We denote the law of position X_m by μ_m, and $*$ denotes convolution.

Forward Method	$\tilde{X}_{hn} = X_{hn} - h\nabla V(X_{kn})$	$\tilde{\mu}_{hn} = (\mathrm{id} - h\nabla V)_{\#\mu_{hn}}$
Flow Method	$X_{h(n+1)} = \tilde{X}_{hn} + \sqrt{2h}\varepsilon$	$\mu_{h(n+1)} = \tilde{\mu}_{hn} * N(0, 2hI)$

Recall from Section 2.4.2 that the Langevin diffusion and the Wasserstein gradient flow on KL divergence are the same. For target $\pi = \exp(-V)$ (here we use $=$ instead of $\propto$ only for simplicity), we have

$$\mathrm{KL}(\mu|\pi) = \int \log \frac{\mu}{\pi}\mu = \underbrace{\int \log \mu d\mu}_{=:H(\mu)} + \underbrace{\int V d\mu}_{=:\mathcal{E}(\mu)}$$

where $H(\mu)$ is the negative entropy and $\mathcal{E}(\mu)$ is the potential energy. The forward/explicit method of LMC can then also be seen as a gradient descent of $\mathcal{E}$ while the flow method of LMC can be viewed as a gradient flow of $-H$. This decomposition helps us to correctly locate the bias - if the flow method is replaced by the adjoint of the forward method (which is the backward/implicit method), then the overall update would be unbiased.

Now that we have correctly located the bias of LMC using Wasserstein gradient descent, we will look at how the same perspective can be used to show convergence results of LMC, following Durmus et al. (2019).

Consider we have the LMC algorithm targeting a d-dimensional distribution π with stepsize h.

Assumption 1. *$\pi = \exp(-V)$ with $0 \lesssim \alpha I_d \lesssim \nabla^2 V \lesssim \beta I_d$*[1].

We first need the following auxiliary lemma.

Lemma 3.1. *Under Assumption 1, if we let $(\mu_{kh})_k$ be the laws of the LMC output and we use stepsize $h \in [0, \sqrt{d}]$, then we have*

$$2hKL(\mu_{(k+1)h}|\pi) \leq (1 - \alpha h)\mathcal{W}_2^2(\mu_{kh}|\pi) - \mathcal{W}_2^2(\mu_{(k+1)h}|\pi) + 2\beta d h^2.$$

[1] $a \lesssim b$ means that $a = O(b)$, i.e. there exists constant $C > 0$ such that $a < Cb$.

The proof of this lemma will be delayed to the end of this section. Next, using this lemma, we can establish the following result.

Theorem 3.2. *Under Assumption 1, we have*

- *(weakly convex) When $\alpha = 0$, for $\varepsilon \in [0, \sqrt{d}]$, if we pick $h \asymp \varepsilon^2/(\beta d)$*[2]*, then the average law $\bar{\mu}_{Nh} := N^{-1} \sum_{k=1}^{N} \mu_{kh}$ satisfies*
$$\sqrt{KL(\bar{\mu}_{Nh}|\pi)} \leq \varepsilon$$
after $N = O(\beta d \mathcal{W}_2^2(\mu_0, \pi)/\varepsilon^4)$ steps.
- *(strongly convex) When $\alpha > 0$, denote the condition number of V by $\kappa := \beta/\alpha$, for $\varepsilon \in [0, \sqrt{d}]$, if we pick $h \asymp \varepsilon^2/(\beta d)$, then we have*
$$\sqrt{\alpha}\mathcal{W}_2(\mu_{Nh}, \pi) \leq \varepsilon, \qquad \sqrt{KL(\bar{\mu}_{Nh,2Nh}|\pi)} \leq \varepsilon$$
after $N = O(\kappa d/\varepsilon^2 \cdot \log[\sqrt{\alpha}\mathcal{W}_2(\mu_0, \pi)/\varepsilon])$ steps, where $\bar{\mu}_{Nh,2Nh} := N^{-1} \sum_{k=N+1}^{2N} \mu_{kh}$.

Proof. First, we will show the result under weakly convex V. Recall that if a function f is convex, then we have $f(N^{-1} \sum_{i=1}^{N} x_i) \leq N^{-1} \sum_{i=1}^{N} f(x_i)$ using induction and the definition of convexity. Rewriting Lemma 3.1, we have

$$\begin{aligned} 2h\mathrm{KL}(\mu_{(k+1)h}|\pi) &\leq \mathcal{W}_2^2(\mu_{kh}|\pi) - \mathcal{W}_2^2(\mu_{(k+1)h}|\pi) + 2\beta d h^2 \\ \mathrm{KL}(\mu_{(k+1)h}|\pi) &\leq [\mathcal{W}_2^2(\mu_{kh}|\pi) - \mathcal{W}_2^2(\mu_{(k+1)h}|\pi)]/(2h) + \beta d h, \end{aligned}$$

and using the convexity of the KL divergence (followed by definition and the log sum inequality), we have

$$\begin{aligned} \mathrm{KL}(\bar{\mu}_{Nh}|\pi) &\leq \frac{1}{N} \sum_{k=1}^{N} \mathrm{KL}(\mu_{kh}|\pi) \\ &\leq [\mathcal{W}_2^2(\mu_0|\pi) - \mathcal{W}_2^2(\mu_{Nh}|\pi)]/(2h) + N\beta d h \\ &\leq \mathcal{W}_2^2(\mu_0|\pi)/(2h) + N\beta d h \end{aligned}$$

which gives us the desired result under weakly convex V with the choice of ε and N specified in the theorem.

Next, we will show the result under strongly convex V. The idea is that we first run LMC for N steps so that the $\mathcal{W}_2$ distance is sufficiently small for us to ignore the effect of α in Lemma 3.1, then we can recover the weakly convex case and directly apply the previously established result.

Since the KL divergence is non-negative, we can rewrite Lemma 3.1 as

$$\mathcal{W}_2^2(\mu_{(k+1)h}|\pi) \leq (1 - \alpha h)\mathcal{W}_2^2(\mu_{kh}|\pi) + 2\beta d h^2$$

which we then apply recursively and yield

$$\mathcal{W}_2^2(\mu_{Nh}|\pi) \leq (1 - \alpha h)^N \mathcal{W}_2^2(\mu_0|\pi) + 2\beta d h^2 \sum_{k=0}^{N-1} (1 - \alpha h)^k \leq \exp(-\alpha h N)\mathcal{W}_2^2(\mu_0|\pi) + O(\kappa d h)$$

which would give us the desired $\mathcal{W}_2$ bound using the specifications of h and N. This allows us to ignore the α term in the statement of Lemma 3.1 after N steps, so the KL divergence bound can be established using the result form weakly convex V. □

[2] $a \asymp b$ if we have both $a \lesssim b$ and $b \lesssim a$.

Finally, we will prove Lemma 3.1.

Proof. (of Lemma 3.1) Let $Z \sim \pi$ be optimally coupled to X_{kh} and $\tilde{X}_{kh}$. We will break down the full KL bounds into smaller parts, by considering $\mathcal{E}$ and H separately.

Firstly, the $\mathcal{E}$ term. We have $\mathcal{E}(\tilde{\mu}_{kh}) - \mathcal{E}(\pi) = \mathbb{E}[V(\tilde{X}_{kh}) - Z]$. Using the evolution variational inequality of gradient descent (Boyd and Vandenberghe, 2004), we have

$$\begin{aligned}\mathcal{E}(\tilde{\mu}_{kh}) - \mathcal{E}(\pi) &= \mathbb{E}[V(\tilde{X}_{kh}) - Z] \\ &\leq \frac{1}{2h}\mathbb{E}\left[(1-\alpha h)\|X_{kh} - Z\|^2 - \|\tilde{X}_{kh} - Z\|^2\right] \\ &\leq \frac{1}{2h}\mathbb{E}\left[(1-\alpha h)\mathcal{W}_2^2(\mu_{kh}, \pi) - \mathcal{W}_2^2(\tilde{\mu}_{kh}, \pi)\right].\end{aligned}$$

Next, using the β-smoothnes of V, we have

$$\begin{aligned}\mathcal{E}(\mu_{(k+1)h}) - \mathcal{E}(\tilde{\mu}_{kh}) &= \mathbb{E}[V(X_{(k+1)h}) - V(\tilde{X}_{kh})] \\ &\leq \mathbb{E}\left[\langle \nabla V(\tilde{X}_{kh}, X_{(k+1)h} - \tilde{X}_{kh}\rangle + \frac{\beta}{2}\|X_{(k+1)h} - \tilde{X}_{kh}\|^2\right] \\ &= \mathbb{E}\left[\langle \nabla V(\tilde{X}_{kh}, B_{(k+1)h} - B_{kh}\rangle + \frac{\beta}{2}\|B_{(k+1)h} - B_{kh}\|^2\right] = \beta d h.\end{aligned}$$

Now we will look at the H term. Let $(Q_t)_t$ denote the heat semigroup[3], i.e. we have $Q_t f(x) := \mathbb{E}[f(x + \sqrt{2}B_t)]$, so that $\mu_{(k+1)h} = \tilde{\mu}_{kh}Q_h$. Recall that the heat flow is the Wasserstein gradient flow of H while the Wasserstein gradient of H is $\nabla_{\mathcal{W}_2}H(\mu) = \nabla \log \mu$, we would have, using Theorem 2.11,

$$\partial_t \mathcal{W}_2^2(\tilde{\mu}_{kh}Q_t, \pi) \leq 2\mathbb{E}\left[\langle \nabla \log \mu(\tilde{X}_{kh+t}, Z - \tilde{X}_{kh+t}\rangle\right]$$

where $\tilde{X}_{kh+t} \sim \tilde{\mu}_{kh}Q_t$. Also, using the convexity of H established in Section 2.5.2, we have

$$H(\pi) - H(\tilde{\mu}_{kh}Q_t) \geq \mathbb{E}\left[\langle \nabla \log \mu(\tilde{X}_{kh+t}, Z - \tilde{X}_{kh+t}\rangle\right].$$

Combining the two results, as well as the fact that $t \mapsto H(\tilde{\mu}_{kh}Q_t)$ is decreasing as it is the trajectory of the gradient flow of H, we would have

$$\mathcal{W}_2^2(\mu_{(k+1)h}, \pi) - \mathcal{W}_2^2(\tilde{\mu}_{kh}, \pi) \leq 2h[H(\pi) - H(\mu_{(k+1)h})].$$

Compiling everything we have derived so far would give us the desired result of the lemma. □

3.2 Stein Variational Gradient Descent as Gradient Flow

Liu and Wang (2016), Liu (2017), Duncan et al. (2019)

to be finished.

[3] for more information on semigroup theory, refer to Bakry et al. (2014)

Bibliography

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* Springer Science & Business Media, 2005.

Vladimir I Arnold. *Ordinary differential equations.* Springer Science & Business Media, 1992.

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Patrizia Berti, Luca Pratelli, and Pietro Rigo. Gluing lemmas and skorohod representations. 2015.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

Richard L. Burden and J. Douglas Faires. *Numerical analysis.* Brooks/Cole, Cengage Learning, 9th edition, 2011.

Sinho Chewi. *Log-Concave Sampling.* Unfinished Draft, 2023. https://chewisinho.github.io/main.pdf.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):651–676, 2017.

Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of langevin monte carlo via convex optimization. *The Journal of Machine Learning Research*, 20(1):2666–2711, 2019.

Michael F Faulkner and Samuel Livingstone. Sampling algorithms in statistical physics: a guide for statistics and machine learning. *arXiv preprint arXiv:2208.04751*, 2022.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Leonid V Kantorovich. On the translocation of masses. *Journal of mathematical sciences*, 133 (4):1381–1382, 2006.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge*

Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16, pages 795–811. Springer, 2016.

Anna Korba and Adil Salim. Sampling as first-order optimization over a space of probability measures. https://akorba.github.io/resources/Baltimore_July2022_ICMLtutorial.pdf, 2022. Accessed: 2024-02-06.

Qiang Liu. Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

Tatiana Xifara, Chris Sherlock, Samuel Livingstone, Simon Byrne, and Mark Girolami. Langevin diffusions and the metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91: 14–19, 2014.