Optimal Transport and Applications to Stochastic Control

Anthony Salib

Supervisor: Gregoire Loeper

Honours Thesis submitted as part of the B.Sc. (Honours) degree in the School of Mathematical Sciences, Monash University.

Date of Submission: 20th August 2019

Abstract

The theory of optimal transportation has given rise to a framework which makes analysis of functions on the space of probability measures possible. In this thesis we study the optimal transportation problem, duality methods to its solution and the continuous in time version of the problem. This time dependent formalism motivates a Riemannian structure on the space of probability measures known as the Otto calculus. This formalism is studied in detail and we review the application of it to the theory of gradient flows. The deterministic optimal control problem is outlined along with the dynamic programming principle as well as their analogues in the stochastic control problem. We then investigate how the framework of the Wasserstein space can be applied to the stochastic optimal control problem of Mckean-Vlasov stochastic differential equations. We recast this into a deterministic problem on the underlying probability distribution of the process, and then the dynamic programming principle is used to construct an abstract form of a Hamiltonian which can be minimized to find the optimal control.

Introduction

This research project is summarized into three stages. The first stage is the study of the **optimal transport problem** whose solution motivates the Riemannian structure that is named the Otto calculus. The second stage is the study of both the deterministic and stochastic optimal control problems, whose treatments are standard. For this reason, we call this stage of the project the study of **classical optimal control theory**. The third and final stage was the union of the first two studies in the investigation of stochastic optimal control of **Mckean-Vlasov stochastic differential equations** (MKVSDEs).

The first stage of the project followed closely the course by Villani on optimal transport [45], specifically chapters 1, 2, 5, 7, and 8, as well as sections of [8] when necessary preliminary material on functional analysis was required. The content of the first two chapters and the appendices are mostly taken from these sources, however, in an effort to present a first hand account, many of the original papers on optimal transportation were consulted and referenced.

In Chapter 1 we introduce the optimal transportation problem in both it's strong and weak formulations, known respectively as the Monge and Kantorovich problems. The Kantorovich duality is discussed in detail and the existence results that were based of this duality are presented. The Wasserstein metrics and properties of the Wasserstein spaces are then studied as products of the solution to the optimal transport problem; their utility will be indispensable when we study the optimal control of MKVSDEs. We finally present the Lagrangian formulation of the time continuous optimal transport problem and discuss its solution.

Chapter 2 is dedicated to the study of the Riemannian structure on the space of probability measures and it's applications. We first present the Eulerian formalism of the time-dependent optimal transportation problem and use these results to prove the celebrated Benamou-Brenier formula for their reformulation of the optimal transport problem, which

was performed in the framework of fluid mechanics. Drawing on this formula, we formally present the Riemannian structure endowed on the space of probability measures $P_2(\mathbb{R}^d)$. In closing the chapter, we briefly outline various examples of applications of this calculus, showcasing the enormous success of this formalism.

The second stage of the thesis reviews the classical theory of optimal control problems, focusing on the use of the dynamic programming principle (DPP) in deterministic and stochastic settings. As this theory is already contained in numerous references in textbook form and is completely standard by now, we reference the textbooks as the sources used and no literature review was performed.

In Chapter 3, the deterministic optimal control problem is introduced and the DPP is presented. We give a formal derivation of the Hamilton-Jacobi-Bellman (HJB) equation for the value function and the notion of viscosity solution is discussed. We then move on to a brief treatment of the stochastic control problem focusing on the DPP method. As our own treatment of Mckean-Vlasov optimal control is based solely on the DPP, and since the purpose of this chapter is to introduce the framework in which we will be working, we make no attempt to study the Pontryagin maximum principle.

The final chapter deals with the optimal control of MKVSDEs in the framework of the dynamic programming principle. This is a new area of research and so after introducing the problem we move on to summarize the advances made in applying the DPP. Following this review, a minor variant of some of the ideas in the literature is presented as a new approach to applying the DPP to this optimal control problem.

We present most theorems without proofs as they can be found elsewhere and references will be given when required. However, some proofs are given in the case when it showcases necessary concepts and ideas.

Contents

Chapter 1

Optimal Transportation

In this chapter we will detail the Monge-Kantorovich problem and its solution. We then introduce the p-Wasserstein distances which are induced from the solution to the transportation problem, and study their properties when they are endowed upon the space of probability measures with bounded p-moments. Finally the time dependent optimal transportation problem is introduced from the Lagrangian point of view and its solution is presented. If no reference is given for a theorem or result, the corresponding result is taken from [45].

1.1 The Monge Problem

In 1781, Monge was interested in finding the most economic way of moving a pile of soil to a mound [36]. Given points of masses $\{m_1, m_2, ...\}$, at locations $\{x_1, x_2, ...\}$ in $\mathbb{R}^3$ that needed to be moved to locations $\{y_1, y_2, ...\}$, Monge was interested in finding a bijective map $T : \{x_1, x_2, ...\} \to \{y_1, y_2, ...\}$ that minimized the weighted distance cost

$$I[T] = \sum m_i |x_i - T(x_i)|. \tag{1.1}$$

This minimizing map is known as the optimal transport map. In his memoir, he used geometric arguments to deduce that if an optimal map does exist then it must be determined by a potential ϕ. As stated in [16], the precise contribution given by Monge was that if the optimal map T exists, then it must satisfy

$$\frac{T(x)-x}{|T(x)-x|} = -D\phi. \tag{1.2}$$

A formal proof of (1.2) was given by Appell [3] in 1884, one hundred years after the work of Monge.

A continuous, general version of the Monge problem can be restated as follows. Let X and Y be measure spaces with probability measures μ and ν respectively - these represent the pile of soil we wish to move and the mound we wish to create. We define the cost function c as a measurable function $c(x,y) : X \times Y \to \mathbb{R} \cup \{+\infty\}$. Note that in Monge's original consideration, this was simply the weighted distance in (1.1). A *transport map* $T : X \to Y$ is a measurable map such that ν is the push forward of μ by T. This means that for any measurable set $B \subset Y$, we should have that

$$\nu(B) = \mu(T^{-1}(B)), \tag{1.3}$$

and we write $\nu = T\#\mu$. With the above defined, the Monge Optimisation problem is to minimise

$$I[T] := \int_X c(x, T(x))d\mu \tag{1.4}$$

over all measurable maps T that satisfy $\nu = T\#\mu$. The solution to the Monge Problem (1.4) is called the optimal transport map and the cost assoociated with it is known as the optimal transportation cost. We will denote the optimal transportation cost between probability measures μ and ν by $\mathcal{T}_c(\mu,\nu)$; i.e.

$$\mathcal{T}_c(\mu,\nu) = \inf_{T\#\mu=\nu} I[T].$$

1.2 The Kantorovich Problem

In 1942, Kantorovich [23] introduced the following relaxed version of the Monge optimisation problem. We again consider two measure spaces X and Y with probability measures μ and ν respectively. We define an admissible transport plan π as a joint probability measure on $X \times Y$ that

has marginals μ and ν. This means that for all measurable sets $A \subset X$ and $B \subset Y$ we have

$$\pi[A \times Y] = \mu[A], \qquad \pi[X \times B] = \nu[B]. \tag{1.5}$$

The set of all admissible transport plans will be denoted as $\Pi(\mu, \nu)$. Kantorovich's minimisation problem is to minimise

$$I[\pi] = \int_{X \times Y} c(x, y) d\pi(x, y) \tag{1.6}$$

for all $\pi \in \Pi(\mu, \nu)$. He showed that a solution to his problem exists and that it is indeed determined by a potential, as was argued by Monge in [36]. In 1948, Kantorovich related his problem to the Monge optimization problem [22] and showed that his results from [23] could be applied. In fact, the Kantorovich problem is just a relaxed version of the Monge problem. The Monge problem is more stringent in that it does not allow a piece of mass at a point $x \in X$ to be split up and sent to multiple different locations in Y. We can write the transport plans for the Kantorovich problem in terms of the transport maps of the Monge problem as

$$d\pi(x, y) = d\mu(x)\delta[y = T(x)], \tag{1.7}$$

where $\delta[y = T(x)]$ is the Dirac measure defined to be 1 if $y = T(x)$ and 0 otherwise.

Kantorovich only showed that the optimal transport plan π exists [22], yet it was not shown if the solution could be represented as a map $T : X \to Y$. This was later done by Sudakov [43] in 1979. We also note that Gangbo and Evans took a PDE approach to proving the existence of optimal transport plans for the distance cost function (Monge's original problem) in [16]. Although these results are historically important as the distance cost function was the original problem considered by Monge, the case of the quadratic cost function, $c(x, y) = |x - y|^2$, is the one most widely studied and that is also found in many applications. The first existence results in the case of the quadratic cost function relied on studying the dual problem introduced by Kantorovich in the 1940's, which is the subject of our next section. Although the duality was first used in the study of the quadratic cost function, in [10], the authors used the dual problem to prove the existence of a minimiser for the Monge problem with distance cost function.

1.3 Kantorovich Duality

We begin this section by stating the Kantorovich duality theorem.

Theorem 1.3.1 (Kantorovich duality)**.** *Let X and Y be complete, separable metric spaces (Polish) and let $\mu \in P(X)$, $\nu \in P(Y)$. Take $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ to be a lower semi continuous cost function. Now, for $(\phi, \psi) \in L^1(d\mu) \times L^1(d\nu)$ define*

$$J(\phi, \psi) = \int_X \phi d\nu + \int_Y \psi d\nu, \tag{1.8}$$

and

$$\Phi_c = \{(\phi, \psi) \in L^1(d\mu) \times L^1(d\nu) : \phi(x) + \psi(y) \leq c(x, y)\},$$

where the inequality holds for $d\mu-$almost all $x \in X$ and $d\nu-$almost all $y \in Y$. It then follows that

$$\inf_{\pi \in \Pi(\mu,\nu)} I[\pi] = \sup_{(\phi,\psi)\in\Phi_c} J(\phi, \psi). \tag{1.9}$$

The above theorem can be interpreted in terms of economics.

The owner of a chocolate company needs to ship a fixed amount of chocolate to different confectionery stores across their town. If they do the shipping themselves, it will cost $c(x, y)$ to move one box of chocolate from point x to y. This problem can be translated into it's Monge-Kantorovich problem and be solved.

However a shipping company will only charge a set cost $\phi(x)$ to pick up the boxes at some point x and $\psi(y)$ to unload it at some point y, while ensuring that the cost of shipping, $\phi(x) + \psi(y)$, will be less than $c(x, y)$.

The Kantorovich duality theorem states that the shipping company can adjust the prices so that it will cost the chocolate company the same amount as if they had handled it themselves.

At least in the compact case, theorem 1.9 is a consequence of the Frenchel-Rockerfella duality theorem. For definitions see the Appendix.

Theorem 1.3.2 (Frenchel-Rockerfella)**.** *For two convex functions on a normed vector space E, $\alpha, \beta : E \to (-\infty, \infty]$, assume that there is some $x_0 \in Dom(\alpha) \cap Dom(\beta)$ such that α is continuous at x_0. Then*

$$\inf_{x\in E}\{\alpha(x)+\beta(x)\} = \sup_{f\in E^*}\{-\alpha^*(-f)-\beta^*(f)\} = \max_{f\in E^*}\{-\alpha^*(-f)-\beta^*(f)\}. \tag{1.10}$$

This theorem states that a minimization problem can be turned into a maximization problem. Indeed, theorems like this are bountiful and extremely useful. This particular type of statement is known as a **minimax** principle. The idea of a minimax principle is that it allows us to write "min max" problems as "max min" problems, where the latter is usually easier to solve. Indeed, this is the route taken to prove existence of optimal transport plans which is done via the Kantorovich Duality. The full proof of theorem 1.9 is too technical to present here. Rather, we prove the result in a restricted setting and then sketch how to extend this to the more general case by approximations. This proof is taken from [45] and is presented here to showcase the power of duality methods in optimal transportation.

Sketch of proof of Theorem 1.9. As a preliminary step, we show that we can change what we need to show in the definition of Φ_c to restrict to functions (ϕ, ψ) that are bounded and continuous. This restricted set we will denote as $\Phi_c \cap C_b$. Since $\Phi_c \cap C_b \subset \Phi_c$ we automatically have that $\sup_{\Phi_c \cap C_b} J(\phi, \psi) \leq \sup_{\Phi_c} J(\phi, \psi)$. Now, since $\phi(x) + \psi(y) \leq c(x, y)$ it follows that for $\pi \in \Pi(\mu, \nu)$

$$\begin{aligned} J(\phi, \psi) &:= \int_X \phi(x) d\mu + \int_Y \psi(y) d\nu \\ &= \int_{X \times Y} (\phi(x) + \psi(y)) d\pi(x, y) \\ &\leq \int_{X \times Y} c(x, y) d\pi(x, y) \\ &=: I[\pi]. \end{aligned}$$

Hence, we get that $\sup_{\Phi_c} J(\phi, \psi) \leq \inf_{\Pi(\mu,\nu)} I[\pi]$. This means that if we prove that $\sup_{\Phi_c \cap C_b} J(\phi, \psi) = \inf_{\Pi(\mu,\nu)} I[\pi]$, we get that it holds for Φ_c too.

We now show that the result is true when X and Y are compact and c is continuous. In this case, it is a direct application of theorem 1.3.2. Let $E = C_b(X \times Y)$ and note by Riesz Theorem that it's dual space E^* is the set of all Borel measures on $X \times Y$, which is denoted as $M(X \times Y)$. Now, let us define $\alpha : C_b(X \times Y) \to \mathbb{R}$ as the mapping

$$u(x, y) \mapsto \begin{cases} 0, & \text{if } u(x, y) \geq -c(x, y) \\ +\infty, & \text{else} \end{cases} \tag{1.11}$$

and $\beta : C_b(X \times Y) \to \mathbb{R}$ as the mapping

$$u(x,y) \mapsto \begin{cases} \int_X \phi d\mu + \int_Y \psi d\nu, & \text{if } u(x,y) = \phi(x) + \psi(y) \\ +\infty, & \text{else} \end{cases}. \tag{1.12}$$

Now, α is the indicator function for a convex set and so it must be convex. Furthermore, at $u = 1$, α is obviously continuous. On the other hand β is convex because it extends a convex function to infinity outside of it's domain. Hence, we can use the Frenchel-Rockerfella duality theorem. Computing the convex conjugates of α and β we get that for $\pi \in E^*$

$$\alpha^*(-\pi) = \begin{cases} \int c(x,y) d\pi(x,y), & \text{if } \pi \in M_+(X \times Y) \\ +\infty, & \text{else} \end{cases}$$

and

$$\beta^*(-\pi) = \begin{cases} 0, & \text{if } \int \phi(x) + \psi(y) d\pi(x,y) = \int \phi(x) d\mu + \int \psi(y) d\nu \\ +\infty, & \text{else} \end{cases}.$$

From 1.10 we can see that

$$\inf_{x \in E} \{\alpha(u) + \beta(u)\} = \sup_{\pi \in E^*} \{-\alpha^*(-\pi) - \beta^*(\pi)\}. \tag{1.13}$$

Now, by looking carefully at the piece-wise conditions in both α and β we see that

$$\inf_{x \in E} \{\alpha(u) + \beta(u)\} = - \sup_{(\phi,\psi) \in \Phi_c} \left\{ \int_X \phi d\mu + \int_Y \psi d\nu \right\}.$$

Furthermore, by looking at $-\alpha^*(-\pi) - \beta^*(\pi)$ it is clear that

$$\sup_{\pi \in E^*} \{-\alpha^*(-\pi) - \beta^*(\pi)\} = - \inf_{\Pi(\mu,\nu)} \left\{ \int_{X \times Y} c(x,y) d\pi(x,y) \right\},$$

and so we arrive at the announced

$$\inf_{\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)d\pi(x,y) = \sup_{\Phi_c\cap C_b} \int_X \phi d\mu + \int_Y \psi d\nu. \tag{1.14}$$

To complete the proof, we extend this result to the general case by approximation arguments. A measure on a Polish space is automatically inner regular [6] which means for $\delta > 0$ arbitrarily small we can find a compact set $K \subset X \times Y$ such that $\pi[(X \times Y)\backslash K] < \delta$. The idea is to use the fact that the duality theorem holds on K and then extend it to the rest of $X \times Y$, and this will do away with the compactness assumptions. Then, since any lower semi-continuous function can be approximated by a sequence of uniformly continuous functions, we can express $c = \sup c_n$, and by this approximation, the general result will follow. □

Remark 1.3.1. *We note that there are more abstract and general spaces in which the theorem holds, for example, the topological space is not needed as shown in* [41].

The first existence results for optimal transport maps were proven using theorem 1.9, and these are the subject of the next section.

1.4 Existence of Optimal Transport Plans

In this section we briefly outline the existence results of optimal transport plans, we first deal with the case of the quadratic cost and then we mention the results for general cost functions.

1.4.1 The Quadratic Cost

For the quadratic cost, $c(x,y) = |x-y|^2$, we have that the first existence result was due to Knott and Smith [26] and then completed by Rachev and Ruschendorff [42]. The important results from convex analysis used in the proof and statement of the theorem are presented in the appendix. The statements of the following theorems are taken from the text [45].

Theorem 1.4.1 (Knott-Smith)**.** *Let μ and ν be probability measures on $\mathbb{R}^n$ that satisfy*

$$\int_{\mathbb{R}^n} |x|^2 d\mu < +\infty, \quad \int_{\mathbb{R}^n} |x|^2 d\nu < +\infty. \tag{1.15}$$

Then in the Monge-Kantorovich problem with $c(x,y) = |x-y|^2$, $\pi \in \Pi(\mu,\nu)$ *is optimal iff there exists a convex lower semi-continuous function* ϕ *such that*

$$Supp(\pi) \subset Graph(\partial\phi). \tag{1.16}$$

Furthermore, (ϕ, ϕ^*) *will be a minimizer of*

$$\inf\left\{\int_{\mathbb{R}^n} \phi d\mu + \int_{\mathbb{R}^n} \psi d\nu; \quad \forall (x,y) \text{ such that } x\cdot y \leq \phi(x) + \psi(y)\right\}. \tag{1.17}$$

Brenier [7] achieved the following result for the case of the quadratic cost and when $X = Y = \mathbb{R}^d$.

Theorem 1.4.2 (Brenier)**.** *Let* μ *and* ν *be probability measures on* $\mathbb{R}^n$ *that satisfy* (1.15)*. If* μ *is absolutely continuous to the Lebesgue measure on* $\mathbb{R}^n$*, then there is a unique optimal* π *of the form*

$$\pi = (Id \times \nabla\phi)\#\mu, \tag{1.18}$$

where $\nabla\phi$ *is the gradient of a convex function, that is uniquely determined* $d\mu$*-almost everywhere, such that* $\nu = \nabla\phi\#\mu$ *and it holds that*

$$Supp(\nu) = \overline{\nabla\phi(Supp(\mu))}.$$

The above theorems were proved using the Kantorovich duality theorem. However, Brenier's theorem was later proved without relying on the duality theorem by Gangbo [18] and then later by McCann [34]. In the case of a smooth Riemannian Manifold, McCann [33] showed that theorems 1.4.1 and 1.4.2 hold.

Remark 1.4.1. *The condition that* μ *is absolutely continuous to the Lebesgue measure is rather strong and is not necessary, although this is the assumption contained in* [7]*. We can replace this with the weaker assumption that* μ *should not give mass to small sets meaning that it does not give mass to sets of Hausdorff dimension* $n-1$*, whose definition is given in the appendix.*

1.4.2 General cost functions

Other than the case of the quadratic cost function and the distance cost function, Gangbo and McCann showed the existence of optimal transport maps for strictly concave and strictly convex costs in [19]. Again, all necessary definitions are found in the appendix.

Theorem 1.4.3 (Strictly Convex Costs)**.** *Let c be a strictly convex, super-linear cost on $\mathbb{R}^n$ and let μ, ν be probability measures on $\mathbb{R}^n$ and we require that the total cost of transportation between them is not identically infinity and that μ is absolutely continuous to the Lebesgue measure. Then there exists a unique optimal transport plan for the Monge-Kantorovich problem and it is of the form $\pi = (Id \times T)\#\mu$ where T is uniquely determined $d\mu$ almost everywhere by $T\#\mu = \nu$ and*

$$T(x) = x - \nabla c^*(\nabla \phi(x))$$

for some c-concave function ϕ.

Theorem 1.4.4 (Strictly Concave Costs)**.** *Let c be a strictly concave cost on $\mathbb{R}^n$ and let μ, ν be probability measures on $\mathbb{R}^n$ and we require that the total cost of transportation between them is not identically infinity and that μ does not give mass to small sets. Furthermore, if μ and ν are singular to each other, then there exists a unique optimal transport plan for the Monge-Kantorovich problem and it is of the form $\pi = (Id \times T)\#\mu$ where T is uniquely determined $d\mu$ almost everywhere by $T\#\mu = \nu$ and*

$$T(x) = x - (\nabla c)^{*-1}(\nabla \phi(x))$$

for some c-concave function ϕ.

1.5 Wasserstein Distances

Using the optimal transportation cost, $\mathcal{T}_p(\mu, \nu)$ corresponding to the cost function $c(x, y) = |x - y|^p$, we can define the Wasserstein distances which are metrics on the space of probability measures with bounded moments of order p, $P_p(\mathbb{R}^n)$. This is the subject of the following theorem.

Theorem 1.5.1 (Wasserstein Distances)**.** *Let μ, ν be two probability measures on $\mathbb{R}^n$. For all $p \in [1, \infty)$, $W_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)^{1/p}$ is a metric on $P_p(\mathbb{R}^n)$. If $p \in [0, 1)$ then $W_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)$ is a metric on $P_p(\mathbb{R}^n)$.*

From here on, we call the metric W_p the p-Wasserstein metric and the space $P_2(\mathbb{R}^n)$ endowed with this metric the p-Wasserstein space. In particular, W_2 is called the quadratic Wasserstein metric and the the space $P_2(\mathbb{R}^n)$ is endowed with W_2 is called the quadratic Wasserstein space, or simply, the Wasserstein space.

As stated in the introduction, the main contribution which we take from Optimal Transportation is the structure it gives on the space of probability measures. The remainder of this section is devoted to examining the properties of the Wasserstein metrics and the Wasserstein space. The first property which we state is that the metrics are bounded by the total variation measure.

Proposition 1.5.2 (Wasserstein Distances are bounded)**.** *Let $\mu, \nu \in P_p(\mathbb{R}^n)$. For $p \geq 0$ and any $x_0 \in \mathbb{R}^n$ we have that*

$$\mathcal{T}_p(\mu, \nu) \leq \max(1, 2^{p-1}) \int |x_0 - x|^p d|\mu - \nu|(x) \tag{1.19}$$

What this tells us is that the optimal transportation cost is bounded by the total variation norm, that is, the Wasserstein distances are controlled by the total variation distance.

The Wasserstein distances also meterize weak convergence in the sense that convergence in the W_p metric is equivalent to the notion of weak convergence. We recall that a sequence of measures in $(\mu_n) \subset P(\mathbb{R}^n)$ converges weakly to μ if for all $\phi \in C_b(\mathbb{R}^n)$,

$$\lim_{n\to\infty} \int_{\mathbb{R}^n} \phi d\mu_n = \int_{\mathbb{R}^n} \phi d\mu.$$

Theorem 1.5.3 (Wasserstein Distances and Weak Convergence)**.** *Let $p \in (0, \infty)$ and $(\mu_k)_{k\in\mathbb{N}}$ be a sequence of probability measures in $\mathcal{P}_p(X)$. For some $\mu \in \mathcal{P}(X)$, $W_p(\mu_k, \mu) \to 0$ as $k \to \infty$ iff $\mu_k \to \mu$ in the weak sense and (μ_k) satisfies for any x_0*

$$\lim_{R\to\infty} \limsup_{k\to\infty} \int_{d(x_0,x)\geq R} d(x_0, x)^p d\mu_k(x) = 0. \tag{1.20}$$

Although we have introduced these distances for general cost, $|x - y|^p$, from here on we will only be interested in the quadratic cost, $p = 2$.

1.6 Time Dependent Formulation

The Monge-Kantorovich problem we introduced does not depend on time, it is only a function of the start and final locations, what happens in between is not important. For the time dependent formulation of the problem, we are interested in the trajectory of each particle at each point x. To this end, at each point $x \in X$ we will attach a trajectory $(T_t(x))_{0\leq t\leq 1}$ and then we define $C[(T_t(x))_{0\leq t\leq 1}]$ to be the cost of this transport. With this notation, the time dependent minimization problem becomes finding

$$\inf\Big\{ \int_X C[(T_t(x))_{0\leq t\leq 1}]d\mu(x) \mid \ T_0 = Id, \ \ T_1\#\mu = \nu\Big\}. \tag{1.21}$$

The time independent and time dependent problems are compatible if

$$c(x,y) = \inf\left\{C[(z_t)_{0\leq t\leq 1}];\ z_0 = x,\ z_1 = y\right\}. \tag{1.22}$$

This condition really asserts that up to trajectories defined on null sets, each trajectory $(T_t(x))_{0\leq t\leq 1}$ is optimal. Now using this compatibility condition we have that the quadratic cost function, $c(x,y) = |x-y|^2$ in the time dependent formulation is

$$C[(z_t)] = \int_0^1 |\dot{z}_t|^2 dt, \tag{1.23}$$

where $\dot{z}_t$ is the derivative of the trajectory z_t with respect to time t. This form of a cost function in the time dependent formulation is very common. In fact, for any cost that can be written as $c(x,y) =: c(x-y)$ we have that

$$C[(z_t)] = \int_0^1 c(\dot{z}_t)dt.$$

In these cases, $c(z)$ is called a differential cost and for any convex cost we have that the minimizing trajectory is a straight line. Indeed by Jensen's inequality,

$$\int_0^1 c(\dot{z}_t)dt \geq c\left(\int_0^1 \dot{z}_t dt\right) = c(y-x), \tag{1.24}$$

and this is the value of the cost when we choose the trajectory to be

$$z_t = (1-t)x + ty,$$

a straight line between x and y. This result is contained in the following proposition.

Proposition 1.6.1 (Convex costs admit straight lines as optimal trajectories)**.** *If c is a convex cost function of $\mathbb{R}^n$ then*

$$\inf\left\{\int_0^1 c(\dot{z}_t)dt;\ ;\ z_0 = x,\ z_1 = y\right\} = c(x-y). \tag{1.25}$$

We know that that at time $t = 0$, the trajectory is just the identity, that is, $T_0(x) = x$. At time $t = 1$, theorem 1.4.3 tells us (under suitable assumptions) that the map has to be $T_1(x) = x - \nabla c^*(\nabla\phi(x))$ for some c-concave function ϕ which satisfies $(Id - \nabla c^*(\nabla\phi(x)))\#\mu = \nu$. Then, from proposition 1.6.1, we get that the optimal trajectory for a convex cost will be

$$T_t(x) = x - t\nabla c^*(\nabla\phi(x)).$$

We make this precise in the following theorem.

Theorem 1.6.2. *Consider the cost function given by $c(x,y) = c(x-y)$ in $\mathbb{R}^n$ where c is strictly convex and $c(0) = 0$. Let μ and ν be two probability measures on $\mathbb{R}^n$ that are absolutely continuous to the Lebesgue measure, and let $C[(z_t)] = \int_0^1 c(\dot{z}_t)dt$. If $\nabla\phi$ is the gradient of a c-concave function such that $(Id - \nabla c^*(\nabla\phi(x)))\#\mu = \nu$, then*

$$T_t(x) = x - t\nabla c^*(\nabla\phi(x)) \tag{1.26}$$

is the solution to the time dependent optimal transportation problem.

Now we restrict our discussion to that of the quadratic cost $c(x,y) = |x-y|^2$ and what follows will be a short summary of McCann's work [32].

1.6.1 Displacement Interpolation and Displacement Convexity

In the case of the quadratic cost, the solution to the optimal transportation problem is

$$\rho_t = [\mu, \nu]_t := [(1-t)Id + t\nabla\phi]\#\mu, \tag{1.27}$$

where μ and ν do not give mass to small sets and so $\nabla\phi$ is given by Brenier's theorem 1.4.2. This is known as McCann's interpolation and defines a family of probability measures that interpolate between μ and ν such that

$$W_2(\mu, \rho_t) = tW_2(\mu, \nu). \tag{1.28}$$

Now we wish to study how a functional of ρ_t varies with time. To do this we introduce some definitions and notation.

Let $P_{ac}(\mathbb{R}^n)$ be the set of absolutely continuous probability measures on $\mathbb{R}^n$ and we identify each $\rho \in P_{ac}(\mathbb{R}^n)$ with it's Lebesgue density such that $d\rho(x) = \rho(x)dx$.

Definition 1.6.1 (Displacement Convexity)**.** *$P \subset P_{ac}(\mathbb{R}^n)$ is displacement convex if for all $\mu, \nu \in P$, for all $t \in [0,1]$ $\rho_t = [\mu, \nu]_t \in P$. Furthermore, given a displacement convex subset P, the functional $F : P \to \mathbb{R} \cup \{+\infty\}$, is (strictly) displacement convex on P if $t \to F(\rho_t)$ is (strictly) convex on $[0,1]$.*

An example of a typical functional is

$$F(\rho) = \int_{\mathbb{R}^n} U(\rho(x))dx, \tag{1.29}$$

and in [32], McCann studied this functional and others where he gave criteria for these functionals to be displacement convex. He then used his notion of strict displacement convexity to prove that there exists a unique ground state energy level of a physical system. This example showcases a successful application of optimal transport. However, our reason for presenting these results is that this notion of convexity is the one used to define convexity of functionals on the Wasserstein space. This will be related to the Riemannian structure we introduce in the next chapter.

Chapter 2

The Otto Calculus

We now introduce the differential view of optimal transport. We first begin with the reformulation by Benamou and Brenier and then introduce the Otto Calculus and finally discuss examples of its use.

2.1 Eulerian Perspective

In the last chapter, we concluded by introducing the time dependent optimal transportation problem where the problem was to find optimal trajectories which transported μ to ν. We remark that the view we considered before, in which we studied trajectories of the particles, was the Lagrangian point of view. In this section, we are interested in developing a field theory, that is, we want to characterize the optimal transportation in terms of a velocity field which dictates how the particles move. Obviously, the velocity field must be related to the trajectories by $v(t,x) = \frac{dT_t}{dt}$, in fact, this is how you switch between the Eulerian and Lagrangian perspectives. Assuming that we have a family of optimal trajectories $(T_t)_{0\leq t\leq 1}$, the intermediate configuration of the mass while it is being transported at time t is given by $\rho_t = T_t\#\mu$. We would like to answer the following questions:

- What is the evolution equation for ρ_t?
- Can we characterize the optimal trajectories?

We have already answered these questions in section 1.6; proposition 1.6.1 gives us the optimal trajectories and theorem 1.6.2 gives us how T_t

evolves, from which the evolution of ρ_t can be easily deduced. We stress that the answers we now seek must be in terms of the velocity field $v(t,x)$, giving us an Eulerian description of the optimal transportation problem. The reason for doing so is that this Eulerian view gives us a Riemannian structure on $P(\mathbb{R}^n)$ which has found utility in many applications.

We answer the first question in the following theorem.

Theorem 2.1.1 (Evolution equation for trajectories)**.** *Consider the time-dependent mass transportation on $\mathbb{R}^n$ and let $(T_t)_{0\leq t\leq 1}$ be a locally Lipschitz family of diffeomorphisms in $\mathbb{R}^n$. Let $v = v(t,x)$ be the velocity field associated with the trajectories, that is $\frac{dT_t}{dt} = v(t,x)$. Then, for $\mu \in P(\mathbb{R}^n)$, $\rho_t := T_t\#\mu$ is the unique solution to*

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho v) = 0 \quad \text{in } C([0,T]; P(\mathbb{R}^n)) \\ \rho_0 = \mu \end{cases} \tag{2.1}$$

in the sense of distributions, where $P(\mathbb{R}^n)$ is endowed with the weak topology.

The notion of weak solutions to PDEs is handled in the Appendix, refer to the definitions therein for the notion of "in the sense of distributions". Again this proof is taken from [45].

Proof. We will only show that ρ_t satisfies (2.1) in the sense of distributions. Uniqueness is handled by a duality argument and can be found in [45].

We first show that the map $t \to \int_{\mathbb{R}^n} \phi d\rho_t$ is continuous.

To do this we note that for $\phi \in D(\mathbb{R}^n)$, by the definition of the push forward measure,

$$\int \phi d\rho_t = \int \phi \circ T_t d\mu,$$

and we note that this map inside the integral on the right is Lipschitz since both ϕ and T_t is Lipschitz. Now, take a sequence $t_n \to t_*$ and note that since ϕ is uniformly bounded, so is $\phi \circ T_{t_n}$. By the Lebesgue dominated convergence theorem

$$\lim_{n\to\infty} \int \phi d\rho_{t_n} = \int \lim_{n\to\infty} \phi \circ T_{t_n} d\mu \tag{2.2}$$

$$= \int \phi \circ T_{t_*} d\mu \tag{2.3}$$

$$= \int \phi d\rho_{t_*}. \tag{2.4}$$

Then, using the definition of the derivative,

$$\lim_{h\to 0}\frac{1}{h}\left(\int \phi d\rho_{t+h} - \int \phi d\rho_t\right) = \lim_{h\to 0}\int \frac{\phi\circ T_{t+h}(x) - \phi\circ T_t(x)}{h} d\mu.$$

Now, since the map is Lipschitz, the quotient inside the integral on the right hand side is uniformly bounded, and so by the Lebesgue Dominated Convergence Theorem, we can pass the limit inside the integral and we get that

$$\lim_{h\to 0}\int \frac{\phi\circ T_{t+h}(x) - \phi\circ T_t(x)}{h} d\mu = \int \frac{d}{dt}\phi\circ T_t d\mu$$

applying the chain rule

$$\begin{aligned}
&= \int (\nabla\phi\circ T_t)\cdot \partial_t T_t d\mu \\
&= \int (\nabla\phi\circ T_t)\cdot v_t\circ T_t d\mu \\
&= \int \nabla\phi\cdot v_t d\rho_t \\
&= -\int \phi d[\nabla\cdot(v_t\rho_t)]
\end{aligned}$$

where we used the duality definition of the ∇ operator as discussed in the appendix. □

We now proceed to answer the second question and characterize the velocity fields which will give us optimal trajectories. First recalling that for convex costs, the optimal trajectories are straight lines, we can begin with the fact that the acceleration of the particles on the trajectories must be zero. Using the chain rule we can see that

$$0 = \frac{d^2}{dt^2}(T_t x) = \frac{d}{dt}(v(T_t(x))) = \partial_t v(T_t(x)) + \nabla v(T_t(x))\cdot v(T_t(x)).$$

Hence, we have that optimal velocity fields are solutions to the PDE

$$\partial_t v + v\cdot\nabla v = 0.$$

With this, we can fully specify the Eulerian system of time dependent mass transportation. The trajectories and corresponding velocity fields must solve

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho v) = 0 \quad \rho(0, \cdot) = \mu \\ \partial_t v + v \cdot \nabla v = 0. \end{cases} \tag{2.5}$$

Now the final task is to specify what role the cost function plays in this perspective. Within the light of theorem 1.6.2, we can see that we will have optimal transportation for the cost c if and only if

$$v(0, x) = -\nabla c^*(\nabla \psi)$$

for some c-concave function ψ.

We note that our entire justification of the Eulerian point of view was done using the far simpler Lagrangian perspective, and it is only presented here as to lay the foundations for the Otto calculus we are working towards. In the next section, we use this Eulerian perspective here to introduce a reformulation of the time dependent optimal transportation problem in the language of fluid mechanics.

2.2 The Benamou-Brenier Formulation

Consider that you have a collection of particles that have some density distribution ρ_0 at time $t = 0$ and ρ_1 at time $t = 1$. Let the position of the particles be modeled by the function $X = X(t)$ and suppose that there is some velocity field $v_t = v(t, X)$ in the region of the particles such that

$$\frac{dX}{dt} = v_t. \tag{2.6}$$

If 2.6 is uniformly Lipschitz, Cauchy-Lipschitz theory guarantees a well defined flow for $t \in [0, 1]$, i.e. unique solutions exist for each different initial condition x_0. So with this, we have a unique trajectory of a particle starting at a position x_0, call it X_{x_0}.

Since the map $(x_0, t) \mapsto X_{x_0}(t)$ is Lipschitz and one to one (by uniqueness of the solutions), we have that the density of the particles evolve as weak solutions of the continuity equation

$$\partial_t(\rho_t) + \nabla_x \cdot (\rho_t v_t) = 0. \tag{2.7}$$

The particles have kinetic energy

$$E(t) = \int_{\mathbb{R}^n} \rho_t(x)|v_t(x)|^2 dx, \tag{2.8}$$

and so the action of the velocity field is

$$A[\rho, v] = \int_0^1 (\int_{\mathbb{R}^n} \rho_t(x)|v_t(x)|^2 dx)dt. \tag{2.9}$$

The Benamou-Brenier minimization problem is to minimize the action $A[\rho, v]$ over all possible families of densities and velocities $(\rho, v) = (\rho_t, v)_{t\in[0,1]}$ that satisfy the following 5 conditions:

- $\rho \in C([0,1] : w * -P_{ac})$;
- $v \in L^2(d\rho_t(x)dt)$;
- $\cup_{t\in[0,1]}\text{supp}(\rho_t)$ is bounded;
- $\partial_t(\rho_t) + \nabla_x \cdot (\rho_t v_t) = 0$ in the distributional sense;
- $\rho(x, 0) = \rho_0(x)$ and $\rho(x, 1) = \rho_1(x)$.

In reality, these 5 conditions are very natural physical considerations. For example, the first and last conditions ensure that we have a continuous transformation of the particles that begin with the initial configuration and end with the final configuration. The second condition ensures that the kinetic energy is not infinite: a physically impossible scenario, while condition 3 ensures that the particles don't disperse out to infinity. The set of (ρ, v) that satisfy these 5 conditions is called $V(\rho_0, \rho_1)$. The following formula is due to the work of Benamou and Brenier at the end of the 90's [5].

Theorem 2.2.1 (Benamou-Brenier Formula)**.** *Let ρ_0 and ρ_1 be two compactly supported absolutely continuous probability measures on $\mathbb{R}^d$. Then*

$$\mathcal{T}_2(\rho_0, \rho_1) = \inf \{A[\rho, v]; (\rho, v) \in V(\rho_0, \rho_1)\}. \tag{2.10}$$

We first remark that the velocity fields and probability measures need not be smooth. Indeed, the proof presented in [45] first proves it in the case of smooth velocity fields and then uses a mollifying argument to reduce to the case of smooth velocity fields. We re-write the proof assuming that the velocity field is globally Lipschitz continuous, so as to bypass the use of mollifiers and to keep the main ideas of the proof unobscured. We present the proof here as it ties together all the considerations of the previous sections and leads us to the main subject of this chapter, the Otto calculus.

Proof of theorem 2.10. The idea of the proof is to first show that

$$\mathcal{T}_2(\rho_0, \rho_1) \leq \inf \{A[\rho, v]; (\rho, v) \in V(\rho_0, \rho_1)\},$$

and then construct a pair $(\rho, v) \in V(\rho_0, \rho_1)$ such that the minimum is attained, that is, $A[\rho, v] = \mathcal{T}_2(\rho_0, \rho_1)$.

We begin by noting that since v is Lipschitz continuous, the trajectories $T_t(x)$ can be defined as the solution to $\frac{d}{dt}T_t(x) = v_t(T_t(x))$, with $T_0(x) = x$. From our assumptions on $V(\rho_1, \rho_2)$ and theorem 2.1.1, we can write that $\rho_t = T_t \# \rho_0$. Then,

$$\begin{aligned}
A[\rho, v] &:= \int_0^1 \int_{\mathbb{R}^n} \rho_t(x)|v_t(x)|^2 dxdt \\
&= \int_0^1 \int_{\mathbb{R}^n} |v_t(x)|^2 d\rho_t(x)dt \\
&= \int_0^1 \int_{\mathbb{R}^n} |v_t(x)|^2 \circ T_t(x) d\rho_0(x)dt \\
&= \int_0^1 \int_{\mathbb{R}^n} \rho_0(x)|v_t(T_t(x))|^2 dxdt \\
&= \int_0^1 \int_{\mathbb{R}^n} \rho_0(x)\left|\frac{d}{dt}T_t(x)\right|^2 dxdt
\end{aligned}$$

Since $v \in L^2(d\rho_t(x)dt)$ we use Tonelli's theorem and hence Fubini's theorem to exchange the integrals

$$= \int_{\mathbb{R}^n} \rho_0(x) \int_0^1 \left|\frac{d}{dt}T_t(x)\right|^2 dtdx$$

By Jensen's inequality

$$\begin{aligned}
&\geq \int_{\mathbb{R}^n} \rho_0(x) \left| \int_0^1 \frac{d}{dt} T_t(x) dt \right|^2 dx \\
&= \int_{\mathbb{R}^n} \rho_0(x) |T_1(x) - x|^2 dx \\
&\geq \mathcal{T}_2(\rho_0, \rho_1).
\end{aligned}$$

We now construct a (ρ, v) such at $A[\rho, v] = \mathcal{T}_2(\rho_0, \rho_1)$. Given the time independent optimal transport problem between ρ_0 and ρ_1 for the quadratic cost, our assumption that these are absolutely continuous allows us to use Brenier's theorem to conclude that there exists an optimal transport map $T = \nabla\phi$, and since ϕ is convex we have that $(\nabla\phi)^{-1} = \nabla\phi^*$ (see appendix). We now construct

$$T_t(x) = (1-t)x + t\nabla\phi(x), \tag{2.11}$$

$$\rho_t = T_t \# \rho_0, \tag{2.12}$$

and finally

$$v_t(x) = (T - Id) \circ T_t^{-1}. \tag{2.13}$$

From the proof of theorem 2.1.1, it is easy to see that the above defined ρ_t and v_t solve the transport equation in the sense of distributions. So by definition of push forward measure we can write

$$\int \rho_t(x)|v_t(x)|^2 dx = \int \left|(T - Id) \circ T_t^{-1}(x)\right| \circ T_t(x) d\rho_0(x) = \int \rho_0(x)|T(x) - x|^2 dx.$$

Since this is true for all t, integrating over time doesn't change this and so we get that $A[\rho, v] = \mathcal{T}_2(\rho_0, \rho_1)$. □

This formula is nothing but the square of W_2. Otto [37] used this to develop what we now call the Otto Calculus, which was inspired by his study of dissipative equations [37].

2.3 Formal Presentation of Otto Calculus

The aim of this section is to define a metric structure on the tangent space $T_\rho P$ for each $\rho \in P(\mathbb{R}^n)$. We further require that the norm given by this metric structure recovers the squared Wasserstein distance

$$W_2^2(\rho_0,\rho_1) = \inf\left\{\int_0^1 \left\|\frac{\partial\rho}{\partial t}\right\|_{\rho(t)}^2 dt; \rho(0)=\rho_0, \rho(1)=\rho_1\right\}.$$

In fluid mechanics, we view $\rho(t)$ as the density of particles evolving with time under some velocity field $v(t)$, we will use this to define what elements of the tangent space $T_\rho P$ look like. Now, ρ must satisfy

$$\frac{\partial\rho}{\partial t} = -\nabla\cdot(\rho v),$$

and so we expect the tangent space to be the space of probability densities with the form $-\nabla\cdot(\rho v)$. Since we wish to only consider physically admissible velocity fields, i.e the ones that give finite kinetic energy,

$$\int |v|^2 d\rho < +\infty,$$

we will require that $v \in L^2(d\rho;\mathbb{R}^n)$. With all these considerations, we define the norm of the tangent vector at $\rho \in P(\mathbb{R}^n)$ to be

$$\left\|\frac{\partial\rho}{\partial t}\right\|^2 = \inf_{v\in L^2(d\rho;\mathbb{R}^n)}\left\{\int \rho|v|^2 dx; \quad \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho v) = 0\right\}. \tag{2.14}$$

From a physical perspective, this is a very natural definition. It says that the norm of the tangent vector to an evolving density of particles, is the lowest possible kinetic energy given by an admissible velocity field. The admissible conditions here are that you cannot have a velocity that gives you an infinite kinetic energy and the velocity must satisfy the continuity equation. When ρ is smooth and positive, the characterization of optimal velocity fields can be formally derived by considering a small perturbation to a minimizing velocity field, v_0. This perturbation is taken to be $\epsilon w/\rho$ where w is a divergence free vector field and $\epsilon \neq 0$. We first note that $v_0 + \epsilon w/\rho$ satisfies the continuity equation as

$$-\nabla\cdot[\rho(v_0+\epsilon w/\rho)] = -\nabla\cdot(\rho v_0) - \epsilon\nabla\cdot w = -\nabla\cdot(\rho v_0) = \partial_t\rho.$$

Now, since v_0 is the minimizing velocity field, it follows that

$$\int \rho|v_0|^2 \leq \int \rho|v_0+\epsilon w/\rho|^2,$$

and so we get that

$$\int \rho|v_0|^2 \leq \int \rho|v_0|^2 + \int 2\epsilon v_0 \cdot w + \int \epsilon^2|w|^2/\rho.$$

For $\epsilon > 0$, we re-arrange and divide by ϵ to get

$$0 \leq \int 2v_0 \cdot w + \int \epsilon|w|^2/\rho,$$

and finally letting $\epsilon \to 0$, we arrive at

$$0 \leq \int v_0 \cdot w.$$

Repeating the same argument for $\epsilon < 0$, we get that

$$\int v_0 \cdot w = 0.$$

Hence, v_0 should be orthogonal, in the L^2 inner product, to the set of divergence free vector fields. This means that v_0 has to be a gradient, i.e. $v_0 = \nabla u_0$.

2.4 Riemannian Structure

We now have a Riemannian structure on the space of probability measure $P(\mathbb{R}^n)$ with the norm defined in (2.14). Using (2.14) and the polarization identity, we can define the metric on the tangent space at ρ. Take two elements of the tangent space $\frac{\partial \rho}{\partial t_1}$ and $\frac{\partial \rho}{\partial t_2}$ that have optimal velocity fields ∇u_1 and ∇u_2.

$$\begin{aligned}
\langle \frac{\partial \rho}{\partial t_1}, \frac{\partial \rho}{\partial t_2} \rangle &= \frac{1}{4}\left(\left\| \frac{\partial \rho}{\partial t_1} + \frac{\partial \rho}{\partial t_2} \right\|^2 - \left\| \frac{\partial \rho}{\partial t_1} - \frac{\partial \rho}{\partial t_2} \right\|^2 \right) \\
&= \frac{1}{4}\int \rho \left(|\nabla u_1 + \nabla u_2|^2 - |\nabla u_1 - \nabla u_2|^2 \right) dx \\
&= \int \rho \langle \nabla u_1, \nabla u_2 \rangle dx.
\end{aligned}$$

With this metric, the geodesics have length given by W_2. Furthermore, geodesics are the McCann's displacement interpolation, and geodesic convexity in this Riemannian structure is McCann's displacement convexity.

Now that we have this metric, we have that the gradient of functions in this setting is

$$\text{grad}_W F(\rho) = -\nabla \cdot (\rho \nabla \frac{\delta F}{\delta \rho}), \tag{2.15}$$

where $\frac{\delta F}{\delta \rho}$ is $\text{grad}_{L^2} F$ which is defined by

$$\int \frac{\delta F}{\delta \rho} \phi(x) = \left[\frac{d}{d\epsilon} F(\rho + \epsilon \phi) \right] |_{\epsilon=0}. \tag{2.16}$$

The subscript W is to signify that the gradient and inner product are those associated with the quadratic Wasserstein space.

Since we identify measures with their (Lebesgue) densities, and these measures have bounded moments of order 2, we can take $F(\rho)$ to be a functional on $L^2(\mathbb{R}^n)$ and so in this case $DF = \frac{\delta F}{\delta \rho}$ as above.

As elements of $T_\rho P$, by the Otto Calculus we must have optimal velocities ∇u_1 and ∇u_2 such that

$$\text{grad}_W F(\rho) = -\nabla \cdot (\rho \nabla u_1),$$

and

$$\partial_t \rho = -\nabla \cdot (\rho \nabla u_2).$$

Now by definition of the gradient we require that

$$\begin{aligned} \langle \text{grad}_W F(\rho), \partial_t \rho \rangle &= \langle DF, \partial_t \rho \rangle_{L^2} \\ &= \int \frac{\delta F}{\delta \rho} \partial_t \rho dx \\ &= -\int \frac{\delta F}{\delta \rho} \nabla \cdot (\rho \nabla u_2) dx \\ &= \int \nabla \frac{\delta F}{\delta \rho} \nabla u_2 \rho(x) dx, \end{aligned}$$

where the last follows after integrating by parts. By definition of the metric on the Wasserstein space we get that

$$\langle \text{grad}_W F(\rho), \partial_t \rho \rangle = \int \nabla u_1 \nabla u_2 \rho(x) dx.$$

Comparing these two expressions we can see that $\nabla u_1 = \nabla \frac{\delta F}{\delta \rho}$ and so it follows that

$$\text{grad}_W F(\rho) = -\nabla \cdot (\rho \nabla u_1) = -\nabla \cdot (\rho \nabla \frac{\delta F}{\delta \rho}),$$

which, at least formally, justifies (2.15).

These formal considerations are made rigorous in [1] in which a geometric notion of derivative on the Wasserstein space is given in terms of sub/super differentials.

We now present some already well established applications of the Otto Calculus.

2.5 Applications to Gradient flows

There are many physical systems that arise as the gradient of the energy functional of the system. An important class are gradient flows

$$\frac{dX}{dt} = -\text{grad}E(X) \tag{2.17}$$

where $E(X)$ is the energy functional. A simple example of a gradient flow is the heat equation

$$\frac{\partial u}{\partial t} = \nabla^2 u,$$

which is the gradient flow of the energy functional $E(u) = \|\nabla u\|^2_{L^2}$ with respect to grad_{L^2}. However, the heat equation is also the gradient flow of the energy functional

$$E(\rho) = \int \rho \log(\rho), \tag{2.18}$$

with respect to the Riemannian structure given by optimal transportation, specifically the W_2 metric. Indeed, using (2.15),

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= -\text{grad} E(\rho) \\ &= \nabla \cdot (\rho \nabla (\frac{\delta F}{\delta \rho})) \\ &= \nabla \cdot (\rho \nabla (\log(\rho)) \\ &= \nabla \cdot (\rho \frac{\nabla \rho}{\rho}) \qquad (2.19) \\ &= \nabla^2 \rho. \qquad (2.20) \end{aligned}$$

There are many other important examples of gradient flows with respect to this Otto Calculus, including the linear Fokker-Planck equation

$$\frac{\partial \rho}{\partial t} = \nabla^2 \rho + \nabla \cdot (\rho \nabla V) \tag{2.21}$$

which is the gradient flow of the energy functional

$$E(\rho) = \int \rho \log \rho + \int \rho V. \tag{2.22}$$

This equation was studied via a time-discretization approach in [40] and does not rely on the notion of a gradient, or any underlying Riemannian structure. We present here a generalized scheme to approximate the gradient flow of an energy functional in any abstract metric space.

We first need to discretize the time domain, and for that we introduce the time discretization variable τ and then define a sequence

$$X_\tau^0 = X^0, \quad X_\tau^{n+1} = \text{argmin}[E(X) + \frac{d(X_\tau^n, X)^2}{2\tau}].$$

The Euler-Lagrange equation in the Euclidean setting for the equation of X_τ^{n+1} is

$$\frac{X_\tau^{n+1} - X_\tau^n}{\tau} = -\text{grad}E(X_\tau^{n+1}),$$

which is a time discretized version of the gradient flow. In general, Otto showed in [37] that the time discretization introduced is indeed that of the gradient flow.

We then define the piece wise constant function

$$X_\tau = \sum_n X_\tau^n \chi_{[n\tau,(n+1)\tau]}.$$

Passing to the limit $\tau \to 0$ of X_τ we recover what is known as the generalized gradient flow. Obviously, one has to check that we can actually pass the limit in each case, which requires various estimates, all of which are covered in [45].

Although there are many examples of gradient flows in the Wasserstein space, in this thesis, we are particularly interested in that of the linear Fokker Planck equation, which is also known as the Kolmogorov forward equation. For our study of stochastic optimal control, we note that the classical SDEs introduced in the following chapter admit a probability density that evolves under the Kolmogorov forward equation. In light of this, the probability measure of the state of the process evolves as a gradient flow in the Wasserstein space. This is however, not true for the Mckean Vlasov case, due to the co-efficients dependence upon the probability measure. In the classical case, this time discretization discussed gives insight into numerical methods that can be used to numerically solve for the underlying probability distribution as carried out in [29].

Chapter 3

Classical Control Theory

The content of this chapter is standard by now and is contained in many references. The sources used in constructing the following exposition are [17, 46] as well as the lectures by Andrzej Święch in Tohoku University [44].

3.1 Deterministic Control

In this section we introduce the deterministic optimal control problem and the dynamic programming principle.

3.1.1 The Control Problem

An ODE of the form

$$x'(t) = f(t, x(t)),$$

could describe the natural motion of a system, like a pendulum for example. Now let us consider the case where we don't want our system to evolve naturally but rather we would like to control how it evolves or what it's final state would be. With the pendulum example for instance, we may wish to control it's motion so that it's final state is upside down, i.e. it is at an angle of π from the vertical. In other instances, we may be given the initial position and some cost function to minimize. In this case, we would like to control the trajectory so that the total cost is

a minimum. In cases like this, we introduce the controlled differential equation

$$\begin{cases} x'(s) = f(s, x(s), \alpha(s)) & s \in (0, T) \\ x(0) = x_0 \end{cases} \tag{3.1}$$

where $f : \mathbb{R}^n \times A \to \mathbb{R}^n$ is bounded and Lipschitz continuous, and A is a compact subset of $\mathbb{R}^m$. We note that the function $\alpha(\cdot)$ that f depends on is defined on the interval $[0, T]$ and is called the **control** and takes values in A. This control selects parameters from A that adjusts how the state of the particle evolves over time. In practice, we would only like to deal with certain types of controls and so we must introduce restrictions on which types of controls we will consider. In this particular case, we will be content with considering controls which are measurable. With this restriction, we obtain a set of controls which we call the *admissible controls*, which we denote by $\mathcal{A}$.

In light of Cauchy-Lipschitz theory (Appendix), the assumptions on f ensure the existence and uniqueness of solutions to 3.1 for a fixed control $\alpha \in \mathcal{A}$, we will denote this solution as $x(\cdot) := x(\cdot; \alpha(\cdot))$, and we will call it the trajectory of the system. The notation $x(t; \alpha())$ asserts we fix some α and that x is a function of t. The aim of the control problem is, given some initial position x, find the control $\alpha^*(\cdot)$ that adjusts the dynamics of the system so that the cost functional

$$J(\alpha; x_0, t_0) = \int_t^T r(x(s), \alpha(s))ds + g(x(T)), \tag{3.2}$$

is minimized, where $r : \mathbb{R}^n \times A \to \mathbb{R}$ is the running cost and $g : \mathbb{R}^n \to \mathbb{R}$ is the terminal cost, are bounded as well as Lipschitz continuous in x. In (3.2), we note that $x(\cdot)$ is the solution to (3.1) with the control $\alpha()$ and x_0 is the initial state of the trajectory at time t_0; although this notation seems redundant it will be useful to specify which initial value problem we are considering when we introduce the dynamic programming principle, the subject of the next subsection.

3.1.2 Dynamic Programming

To solve this control problem, we will make use of the dynamic programming principle (DPP). We first introduce the *value function* defined as

$$v(y,s) := \inf_{\alpha(\cdot)\in\mathcal{A}} J(\alpha; y, s). \tag{3.3}$$

This can be thought of as the minimum cost for a trajectory starting at position $y \in \mathbb{R}^n$ at time $s \in [0,T]$ and we note that the minimum cost associated with our optimal control problem (3.1) is given by $v(x,t)$. The value function satisfies the following dynamic programming principle.

Theorem 3.1.1 (Dynamic Programming)**.** *For each $h > 0$ such that $t + h \leq T$, we have that*

$$v(x,t) = \inf_{\alpha(\cdot)\in\mathcal{A}} \left\{ \int_t^{t+h} r(x(s),\alpha(s))ds + v(x(t+h), t+h) \right\}, \tag{3.4}$$

where $x(\cdot)$ solves (3.1) for the control $\alpha(\cdot)$.

For a proof, see [17]. From the dynamic programming principle, at least formally, we can derive a PDE that the value function must satisfy. Here we will present the formal derivation, as it will motivate our methods in the following chapter.

3.1.3 A formal derivation

We first re-write the dynamic programming principle using an infinitesimal change in time dt as

$$v(x,t) = \inf_{\alpha(\cdot)\in\mathcal{A}} \left\{ r(x(t),\alpha(t))dt + v(x(t+dt), t+dt) \right\}. \tag{3.5}$$

Re-arranging (3.5) and dividing by dt we get

$$\inf_{\alpha(\cdot)\in\mathcal{A}} \left\{ r(x(t),\alpha(t)) + \frac{v(x(t+dt), t+dt) - v(x,t)}{dt} \right\} = 0. \tag{3.6}$$

Now taking the limit as dt goes to zero in (3.6), it follows that

$$\inf_{\alpha(\cdot)\in\mathcal{A}} \left\{ r(x(t),\alpha(t)) + \frac{d}{dt} v(x,t) \right\} = 0. \tag{3.7}$$

By the chain rule we can write

$$\frac{d}{dt}v(x,t) = \langle Dv(x,t), \partial_t x\rangle + \partial_t v = \langle Dv(x,t), f(x(\cdot),\alpha(\cdot))\rangle + \partial_t v.$$

Substituting this into (3.7) and noting that $\partial_t v$ does not depend on α, we arrive at

$$\partial_t v(x,t) + \inf_{\alpha(\cdot)\in\mathcal{A}} \{r(x(t),\alpha(t)) + \langle Dv(x,t), f(x(\cdot),\alpha(\cdot))\rangle\} = 0. \tag{3.8}$$

By substituting $t = T$ into the definition of the value function (3.3), we can see that we have a terminal value condition of $v(x,T) = g(x)$. Putting these together, we arrive at the following terminal value problem

$$\begin{cases} \partial_t v(x,t) + \inf_{\alpha(\cdot)\in\mathcal{A}} \{r(x(t),\alpha(t)) + \langle Dv(x,t), f(x(\cdot),\alpha(\cdot))\rangle\} = 0 & \text{in } [0,T)\times\mathbb{R}^n \\ v(x,T) = g(x) & \text{on } \{T\}\times\mathbb{R}^n \end{cases} \tag{3.9}$$

which is called the Hamilton Jacobi Bellman equation for this optimal control problem. This is a particular type of Hamilton Jacobi equation, where the Hamiltonian is

$$H(Dv(x(t),t),x) = \inf_{\alpha(\cdot)\in\mathcal{A}} \{r(x(t),\alpha(t)) + \langle Dv(x,t), f(x(\cdot),\alpha(\cdot))\rangle\}.$$

We note that PDE (3.8) is all one needs in order to solve the optimal control problem, in fact, all that is needed is the Hamiltonian. To construct the optimal control value at some time s, given the initial state x at time t, all that is needed is to find the value of α that minimizes the Hamiltonian at each time.

We will end this section on the following note; the value function $v(x,t)$ does not in general solve (3.9) in the classical sense, and neither in the weak sense. However, the value function is the unique *viscosity solution* to (3.9).

Definition 3.1.1 (Viscosity Solution)**.** *Assume that v is a bounded and uniformly continuous function on $[0,T]\times\mathbb{R}^n$. We call u a viscosity solution to (3.9) if*

1. *$v = g$ on $\{T\} \times R^n$*
2. *for each $u \in C^\infty(\mathbb{R}^n \times (0,T))$ if $u - v$ has a local maximum at a point $(x_0, t_0) \in \mathbb{R}^n \times (0,T)$, then $\partial_t u(x_0,t_0) + H(Du(x_0,t_0), x_0) \geq 0$*
3. *for each $u \in C^\infty(\mathbb{R}^n \times (0,T))$ if $u - v$ has a local minimum at a point $(x_0, t_0) \in \mathbb{R}^n \times (0,T)$, then $\partial_t u(x_0,t_0) + H(Du(x_0,t_0), x_0) \leq 0$.*

It is easy to see that any classical solution is a viscosity solution, furthermore any sufficiently differentiable viscosity solution, is a classical solution. Uniqueness of viscosity solutions is also guaranteed under assumptions on the Hamiltonian, namely, that $H(p,x)$ is Lipschitz in p and satisfies for a fixed $p \in \mathbb{R}^n$ and $x, y \in \mathbb{R}^n$,

$$|H(p,x) - H(p,y)| \leq C|x-y|(1+|p|).$$

This concludes our discussion of the deterministic optimal control problem, and we now turn to the stochastic optimal control problem.

3.2 Stochastic Optimal Control

In this section we introduce the classical stochastic optimal control problem theory, so as to give an idea of how these problems have been handled in the classical case by the DPP. We will only require the framework from this section in the next chapter of this thesis and so after the problem is introduced, we will keep the remainder of this section brief and descriptive.

We first need to make clear the space on which we are working.

Definition 3.2.1 (Generalized Reference Probability Space)**.** *A generalized reference probability space, μ is the 5 tuple*

$$\gamma = (\Omega, \mathcal{F}, \mathbb{F}_s^t, \mathbb{P}, W), \tag{3.10}$$

*where Ω is the state space, $\mathcal{F}$ is the complete (with respect to $\mathbb{P}$) σ-algebra of measurable sets, $\mathbb{P}$ is the probability measure on $\mathcal{F}$ and W is the standard Wiener Process. Furthermore, $\mathbb{F}_s^t$ is a complete filtration such that $F_s^t = \cap_{r>s} \mathbb{F}_r^t$, (such a filtration is called **right continuous**). When F_s^t is the natural filtration generated by the Wiener process, γ will be called a **reference probability measure**.*

Fix some $T > 0$, then for $t \in (0, T)$ we have the state $X(t) \in \mathbb{R}^n$ which we wish to control. The state is governed by the stochastic differential equation (SDE) with an initial condition,

$$\begin{cases} dX_s = b(s, X(s), a(s))ds + \sigma(s, X(s), a(s))dW_s & s \in (0, T] \\ X(0) = x. \end{cases} \tag{3.11}$$

In (3.11) we note that $a(\cdot) : [0, T] \times \Omega \to \Lambda$ is the control process where Λ is some Polish space. The coefficients are the maps $b : [0, T] \times \mathbb{R}^n \times \Lambda \to \mathbb{R}^n$ and $\sigma : [0, T] \times \mathbb{R}^n \times \Lambda \to \mathbb{R}^{n \times m}$ which for a fixed $a \in \Lambda$ are uniformly continuous on $[0, T] \times \mathbb{R}^n$. Furthermore, to guarantee that solutions to (3.11) exist, we assume that b and σ are Lipschitz in x and satisfy the linear growth condition

$$|\mu(s, x, a)| + \|\sigma(x, s, a)\| \leq C(1 + |x|). \tag{3.12}$$

This classical existence and uniqueness result for SDEs can be found in [46]. The following definitions makes clear what it means to be a solution of 3.11, we first introduce the preliminary definition of progressively measurable.

Definition 3.2.2 (Progressively Measurable)**.** *A stochastic process $a(\cdot)$ is progressively measurable if for all $s > t$, $a(\cdot) : [t, s] \times \Omega \to \Lambda$ is $Bor([t, s]) \times \mathbb{F}_s^t$ measurable, where $Bor(S)$ is the Borel σ-algebra on the set S.*

Definition 3.2.3 (Solution to the SDE)**.** *A stochastic process $X(\cdot)$ is a solution of the (3.11) if X is a progressively measurable process and for every $s > t$ we have that*

$$X(s) = x + \int_t^s b(r, X(r), a(r))dr + \int_t^s \sigma(r, X(r), a(r))dW(r),$$

$\mathbb{P}$-almost surely.

We now introduce the cost functional which will be what is minimized in this control problem. The functional will consist of a running cost $r(s, X(s), a(s))$ and a terminal cost $g(X(T))$ which we assume to be continuous. The full cost functional is then given by

$$J(a(\cdot);t,x) = \mathbb{E}\left\{\int_t^T r(s,X(s),a(s))ds + g(X_T)\right\}. \tag{3.13}$$

Now that all the necessary notions have been introduced, we begin to outline the stochastic optimal control problem in both it's strong and weak formulations. This is strongly paralleled with the theory for strong and weak (martingale) solutions of SDE's, where the main difference is that the underlying probability space is part of the weak solution whereas it is fixed in the strong solutions.

3.2.1 Strong Formulation

In the strong formulation, we fix the Generalized Reference Probability Space γ and we define the set of **admissible controls** to be

$$\mathcal{U}_t^\gamma = \left\{a(\cdot) : [0,T]\times\Omega \to \Lambda |\ a(\cdot) \text{ is } \mathbb{F}_s^t \text{ progresively measurable}\right\}. \tag{3.14}$$

The strong formulation of the problem is to then minimize (3.13) over the set of all admissible controls, $\mathcal{U}_t^\gamma$.

3.2.2 Weak Formulation

In the weak formulation, we allow the Generalized Reference Probability Space to be chosen as part of the solution, as such, the set of all admissible controls will be given by

$$\mathcal{U}_t = \bigcup_\gamma \mathcal{U}_t^\gamma, \tag{3.15}$$

where the union is over all possible Generalized Reference Probability spaces.

Now that we have these formulations outlined, we will look at the dynamic programming principle for the stochastic optimal control problem.

3.2.3 Dynamic Programming

The weak formulation introduced allows to formulate the DPP by varying the reference probability space which means the probability space is part of the control. We introduce the value function

$$V(t,x) = \inf_{a \in \mathcal{U}_t} J(a;t,x),$$

and now state the DPP for stochastic control problems.

Theorem 3.2.1. *Let h be small enough such that $t < t+h \leq T$. Then*

$$V(t,x) = \inf_{a \in \mathcal{U}_t} \mathbb{E}\left[\int_0^h r(s, X(s), a(s))ds + V(t+h, X(t+h))\right].$$

The DPP connects the control problem to the HJB equation

$$\begin{cases} u_t + F(t,x,Du,Du^2) = 0 \\ u(T,x) = g(x) \end{cases}, \tag{3.16}$$

for some appropriate form of Hamiltonian F. This can then be used as described previously to solve for the optimal control.

Chapter 4

Mckean-Vlasov Stochastic Optimal Control

In this final chapter, we examine the progress made in the optimal control of Mckean-Vlasov SDEs, and highlight the reformulation of this problem as a deterministic optimal control problem.

4.1 Mckean-Vlasov SDEs

Mckean-Vlasov SDEs (MKVSDEs) are of the form

$$dX_t = b(t, X_t, \rho_t)dt + \sigma(t, X_t, \rho_t)dW_t. \tag{4.1}$$

In (4.1), W_t is a Brownian motion on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the filtration of Brownian motion $\mathbb{F}$ and $\rho_t = \text{law}(X_t)$. MKVSDEs are characterized by the co-efficients dependence on the underlying probability distribution. That is, b and σ are deterministic functions defined on $[0, T] \times \mathbb{R}^d \times P_2(\mathbb{R}^d)$ where $P_2(\mathbb{R}^d)$ is endowed with the W_2 metric.

SDE's of this type were originally considered by Mckean [35] in his study on the propagation of chaos and Kac [21] on his work on interacting molecules. More recently, the theory of Mean field games put forth by Lasry and Lions in [27] saw the emergence of such SDEs as the large population limit of the behavior of agents who have some mean field interaction. Indeed for this reason MKVSDEs are referred to of "Mean Field type."

The following existence and uniqueness result of strong solutions to the initial value problem

$$\begin{cases} dX_t = b(t, X_t, \rho_t)dt + \sigma(t, X_t, \rho_t)dW_t \quad t \in (0, T] \\ X(0) = X_0 \end{cases}, \tag{4.2}$$

where b and σ can be random is taken from [12].

Theorem 4.1.1. *Assume*

- *There exists some $L > 0$ such that for all $t \in [0, T]$, $\omega \in \Omega$, $x, y \in \mathbb{R}^d$ and $\mu, \nu \in P_2(\mathbb{R}^d)$,*

  $$|b(t, x, \mu) - b(t, y, \nu)| + |\sigma(t, x, \mu) - \sigma(t, y, \nu)| \leq L[|x - y| + W_2(\mu, \nu)] \tag{4.3}$$

- *For each $(x, \mu) \in \mathbb{R}^d \times P_2(\mathbb{R}^d)$ we have that*

  $$\mathbb{E}\int_0^T |b(s, x, \mu)|ds < +\infty$$

  and

  $$\mathbb{E}\int_0^T |\sigma(s, x, \mu)|ds < +\infty,$$

  and the processes $b(s, x, \mu)$ and $\sigma(s, x, \mu)$ are progressively measurable with respect to the filtration $\mathbb{F}$

- $X_0 \in L^2(\Omega; \mathbb{R}^d)$.

Under these assumptions, there exists a unique continuous solution to (4.2).

We now move on to discuss the optimal control of these SDEs.

4.2 Mckean-Vlasov Optimal Control Problem

We define the strong formulation of the stochastic optimal control problem as in section 3.2 with the following exceptions. The trajectory follows the dynamics given by

$$\begin{cases} dX_t = b(t, X_t, \rho_t, a_t)dt + \sigma(t, X_t, \rho_t, a_t)dW_t \quad t \in (0, T] \\ X(0) = X_0. \end{cases} \tag{4.4}$$

where b and σ are deterministic functions defined on $[0, T] \times \mathbb{R}^d \times P_2(\mathbb{R}) \times \Lambda$. The cost functional is

$$J(\alpha) = \mathbb{E}\left\{ \int_0^T r(t, X_t, \rho_t, \alpha_t)dt + g(X_T, \rho_T) \right\}$$

where r is the running cost and g is the end point cost.

Up until recently the study of such an optimal control problem has been left untouched in the literature. As the authors in [12] explain, this was largely due to the lack of a formalism to analyze functions of probability measures. Although this thesis is mainly concerned with the developments of such a theory from the grounds of optimal transportation, it would be dubious of us to ignore the notion of the L-derivative introduced by P.L. Lions [11]. This has become the main derivative used for functions on the space of probability measures, and for good reason. This notion of derivative relies on the lifting of functions of probability measures to functions of L^2 random variables. As such it shifts the problem from analysis on the Wasserstein space to analysis on the Hilbert space L^2, which has been routinely studied. For instance, the theory of viscosity solutions for second order Hamilton-Jacobi equations has been extensively studied on separable Hilbert spaces by Lions in [30, 31]. As mentioned in section 3.2.3, solutions to HJB equations are viscosity solutions, and so these previous results of Lions are indispensable.

Viscosity solutions to Hamilton Jacobi equations on the Wasserstein space have been studied in [20], which uses the notion of differentiability arising from optimal transportation. In their work [12], Carmona and Delarue show the equivalence between the notions of the L-derivative and differentiation originating from optimal transportation theory.

Before we detail the advances in the control problem by the dynamic programming principle, we note that the first results on optimal control of MKVSDEs in a general setting were achieved using the Pontryagin maximum principle. See [13] for the first study in a general setting and the comprehensive reference by the same authors [12]. We do not discuss these results or their progression any further as this is not the focus of this thesis.

The first results applying the DPP to this problem in a general setting were due to Pham and Wei in [38, 39]. Before their work, the dynamic programming principle was studied for specific co-efficients and cost functionals in [28]. This work depended upon the assumption that the underlying marginal distribution of the state process exists for all time. This later work relied on reformulating the problem as a deterministic control problem on the law of the state of the trajectory. In particular, this was achieved using the Kolmogorov forward equation, which was by then a standard process for optimally controlling classical SDEs [2]. This is the approach we choose to follow in the next section in a more general setting.

The more general works of Pham and Wei relied on the same principle of controlling the underlying probability distribution, however, they did not use the Kolmogorov equation. Instead they established a flow property for the marginal distribution via the push forward of measures as done in [9]. The advantage of this method is that it bypasses the need to discuss the well-posdeness of the Kolomogorov forward PDE, and without assuming restrictive regularity assumptions on the co-efficients. Furthermore, in comparison to the maximum methods used in the first papers addressing this problem in a general setting, this method does not require any convexity assumptions. We briefly mention that they successfully applied their results to two financial models, namely mean-variance portfolio selection and the inter bank systemic risk model. However, their work is restricted to closed loop feed back controls, and thus does not work for larger more general classes of controls. Closed loop feedback controls are ones which can be written as functions of the state variables, $\alpha = \alpha(x, s, \mu)$. In [4], the authors prove a version of the DPP which allowed for a broader range of controls, specifically, controls which are open loop. We note that all these worked in a Markovian framework, that is, the controls considered cannot depend on states earlier than the previous one.

In a similar vein to the control of classical SDEs in subsection 3.2.2, [14] formulated a notion of weak solution to the Mckean-Vlasov optimal control problem, providing the most recent contribution to the theory. The authors prove the DPP for this weak formulation under the most general assumptions so far obtained in the literature. They only require continuity of the co-efficients as well as a growth condition; there is no Lipschitz assumption present and no assumptions on the running cost or terminal cost. Imposing Lipschitz assumptions on the co-efficients and growth conditions on the running cost and the terminal cost, the authors proved the DPP for their strong formulations of the problem. All their

results were proven for the most general class of controls, non-Markovian controls, which none of the previous results could cover. A key feature of the analysis in [14] is that they do not rely on the notion differentiability introduced by Lions, but rather they use measurable selection arguments [15, 24, 25] giving them the ability to use less restrictive assumptions.

The following reformulation was found to be contained in the text [12]. We make two distinctions in the method; in their work, they do not use the dynamic programming principal to solve for the optimal control. Rather their derivation of the Hamiltonian was an intermediate step to use the Pontryagin maximum principle. Furthermore, they did not use the DPP in their derivation of the Hamiltonian, but rather used an adjoint variable method. As mentioned, the procedure we follow is quite standard in the literature for classical SDEs, and was carried out for specific cases in the Mckean-Vlasov setting. We note that in this work we use the existence results in [12] to guarantee the existence of the underlying probability distribution at all times, doing away with the existence assumptions made in [28].

4.3 Reformulation as a Deterministic Control Problem

The idea of this section is to re-formulate the problem into a deterministic optimal control problem. We then use the dynamic programming principle to write down the general form of the Hamiltonian derived in section 3.1.3. We assume we are dealing with a fixed complete probability space, $(\Omega, \mathcal{F}, \mathbb{F}_s^t, \mathbb{P}, W)$, where $\mathbb{F}_s^t$ is the natural filtration of the Brownian motion W. We restate the stochastic control problem here as

$$\begin{cases} dX_s = b(s, X(s), \rho_s, a(s))ds + \sigma(s, X(s), \rho_s, a(s))dW_s \quad \text{on } (0, T] \\ X(0) = X_0, \end{cases} \tag{4.5}$$

where the aim is to minimize the cost functional

$$J(\alpha(\cdot); t, x) = \mathbb{E}\left\{ \int_t^T r(s, X(s), \rho_s, \alpha(s))ds + g(X_T, \rho_T) \right\}, \tag{4.6}$$

over the set $\mathcal{A}$ of all admissible controls

$$\mathcal{A} = \left\{ a(\cdot) : [0,T] \times \Omega \to \Lambda |\ a(\cdot) \text{ is } \mathbb{F}_s^t \text{ progresively measurable} \right\}. \quad (4.7)$$

It is a well known fact that the probability distribution of the state X_t defined as $\rho_t := \mathcal{L}(X_t) = \mathbb{P} \circ X_t^{-1}$ evolves under the Kolmogorov forward equation. That is, for $t \in (0,T]$,

$$\partial_t \rho + \nabla \cdot (b(t, X(t), \rho(t), a(t))\rho_t) = \frac{1}{2}\nabla^2 (\sigma(t, X(t), \rho(t), a(t))\sigma(t, X(t), \rho(t), a(t))^T \rho_t). \quad (4.8)$$

We impose the assumptions of theorem 4.1.1 upon b and σ in order to ensure the existence and uniqueness of solutions to the SDE (4.5), which in turn will give us existence of solutions to the PDE (4.8), just take $\rho_t = \text{law}(X_t)$.

We therefore recast the problem of optimally controlling the state of the trajectory, to optimally controlling the distribution of the trajectory. This is a completely deterministic control problem. Once this is achieved, we can solve the deterministic problem to find the optimal control α^* using the dynamic programming principle. However, there are still some questions that remain: what is the initial condition of the initial value problem, and what is the cost functional which we will aim to minimize.

The first question is obvious, we simply take the initial distribution to be the probability measure $\rho_t = \mathbb{P} \circ X_0^{-1}$ where X_0 is the random variable describing the initial state of the trajectory.

Dealing with the cost functional is slightly trickier, we require to re-write (4.6) in a deterministic form. Given that we know what the distribution is at time $t \in (0,T]$, (this is given to us by (4.8)), we can calculate the expectation of the running cost and the terminal cost, even though we do not know what the state is. In light of this, we can write down the following deterministic cost functional,

$$J(\alpha(\cdot); \mu, t) = \int_t^T \tilde{r}(s, \rho(s), \alpha(s))ds + \tilde{g}(\rho(T)), \quad (4.9)$$

where $\tilde{r}(s, \rho(s), \alpha(s)) = \int r(s, x', \rho_s, \alpha(s))d\rho_s(x')$ and $\tilde{g}(\rho(T)) = \int g(x', \rho_T)d\rho_T(x')$. Here, $\rho(s)$ is the solution to

$$\begin{cases} \partial_s \rho + \nabla \cdot (b\rho_s) = \frac{1}{2}\nabla^2(\sigma\sigma^T\rho_s) \quad \text{on } (t,T] \\ \rho_t = \mu. \end{cases} \tag{4.10}$$

In light of these considerations, we can write down the deterministic optimal control problem associated with (4.5).

Given that the probability distribution of the trajectory follows the initial value problem

$$\begin{cases} \partial_s \rho + \nabla \cdot (b\rho_s) = \frac{1}{2}\nabla^2(\sigma\sigma^T\rho_t) \quad \text{on } (0,T] \\ \rho_0 = \mathbb{P} \circ X_0^{-1}, \end{cases} \tag{4.11}$$

find the optimal control $\alpha^*() \in \mathcal{A}$, such that the cost functional given in (4.9) is minimized. We can then follow the approach given in section 3.1.3 to write down a value function and find a Hamiltonian to minimize at each time step.

The value function for the cost functional in this case will be given by

$$v(\mu, t) := \inf_{\alpha(\cdot)\in\mathcal{A}} J(\alpha; \mu, t). \tag{4.12}$$

Since this is a deterministic problem, we use the dynamic programming principle as was done in section 3.1.3 to arrive at the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} 0 = \partial_t v(t,\mu) + \inf_{\alpha(\cdot)\in\mathcal{A}} \{ & r(\mu(t),\alpha(t)) + \langle \partial_\rho v(\mu(t),t), \\ & - \nabla \cdot (b\rho_t) + \frac{1}{2}\nabla^2(\sigma\sigma^T\rho_t)\rangle\}, \end{aligned} \tag{4.13}$$

where ∂_ρ is the derivative of the value function with respect to the probability measure. The Hamiltonian in this case is

$$\begin{aligned} H = \inf_{\alpha(\cdot)\in\mathcal{A}} \{ & r(\mu(t),\alpha(t)) + \langle \partial_\rho v(\mu(t),t), \\ & - \nabla \cdot (b\rho_t) + \frac{1}{2}\nabla^2(\sigma\sigma^T\rho_t)\rangle\}, \end{aligned} \tag{4.14}$$

which can now be minimized to find the optimal control α^*.

4.3.1 An explicit example

In this final section, we will provide a simple example of where the above abstract formalism becomes useful and simplifies the problem at hand. We consider we are working in the framework of section 4.3 with the exception of working in one dimension as well as that we are trying to maximize the following gain function

$$U(X_T) - \int_0^T C(\alpha)dt. \tag{4.15}$$

Here, $C(\alpha) = \int c(\alpha)\rho(x)dx$ is the cost and $U(x)$ is the end point reward. Since the state of X is governed by (4.5) then $\rho(x) = \text{law}(X)$ is governed by (4.8), which we re-arrange in the form of

$$\partial_t \rho_t + \partial_x(v_t \rho_t) = 0, \tag{4.16}$$

where $v_t := b - \frac{1}{2}\frac{1}{\rho_t}\partial_x(\sigma^2\rho_t)$ which simplifies the following computations. Taking $v(t,\mu)$ as the value function for this problem, we evaluate the inner product contained in (4.13) as follows,

$$\langle \partial_\rho v(t,\mu), \partial_t \rho \rangle_{L^2} = \int \partial_\rho v(t,\mu)\partial_t \rho$$

substituting 4.16

$$= -\int \partial_\rho v(t,\mu)\partial_x(\rho v)$$

intergrating by parts

$$\begin{aligned} &= \int \partial_x(\partial_\rho v(t,\mu))(b\rho - \frac{1}{2}\partial_x(\sigma^2\rho)) \\ &= \int \rho(\partial_x(\partial_\rho v(t,\mu)b + \frac{1}{2}\partial_{xx}(\partial_\rho v(t,\mu))\sigma^2). \end{aligned}$$

Hence, the maximization of the Hamiltonian in this instance reduces to the point wise maximization of

$$\partial_x(\partial_\rho v(t,\mu)\mu + \frac{1}{2}\partial_{xx}(\partial_\rho v(t,\mu))\sigma^2 - c(\alpha). \tag{4.17}$$

Conclusion

In this thesis we have surveyed the progress made in the theory of optimal transportation, particularly on the metric structure it motivated on the space of probability measures. The Otto calculus has seen success in a variety of applications, specifically the linear Fokker-Planck equation is realized as the gradient flow of an energy functional in the Wasserstein space. The dynamic programming principle was then studied as a tool to solving the optimal control problem in the deterministic and stochastic cases. Continuing on from this, recent advances of applying dynamic programming to the control of Mckean-Vlasov stochastic differential equations were reviewed. We carried out a reformulation of the Mckean Vlasov problem as a deterministic problem on the underlying probability measure. Using the existence and uniqueness theorem taken from [12], we ensure the existence of a probability measure at all times, which past similar approaches lacked.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[2] Mario Annunziato, Alfio Borzì, Fabio Nobile, and Raul Tempone, *On the connection between the Hamilton-Jacobi-Bellman and the Fokker-Planck control frameworks* (2014).

[3] Paul Appel, *Le problème géométrique des déblais et remblais*, Mémorial des sciences mathématiques **27** (1887).

[4] Erhan Bayraktar, Andrea Cosso, and Huyên Pham, *Randomized dynamic programming principle and Feynman-Kac representation for optimal control of McKean-Vlasov dynamics*, Transactions of the American Mathematical Society **370** (2018), no. 3, 2115–2160.

[5] Jean-David Benamou and Yann Brenier, *A numerical method for the optimal time-continuous mass transport problem and related problems*, Contemporary mathematics **226** (1999), 1–12.

[6] Patrick Billingsley, *Convergence of probability measures*, Second, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999. A Wiley-Interscience Publication. MR1700749 (2000e:60008)

[7] Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on pure and applied mathematics **44** (1991), no. 4, 375–417.

[8] Haim Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2010.

[9] Rainer Buckdahn, Juan Li, Shige Peng, Catherine Rainer, et al., *Mean-field stochastic differential equations and associated pdes*, The Annals of Probability **45** (2017), no. 2, 824–878.

[10] Luis Caffarelli, Mikhail Feldman, and Robert McCann, *Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs*, Journal of the American Mathematical Society **15** (2002), no. 1, 1–26.

[11] P Cardaliaguet, *Notes on mean field games.(2013)*, URL https://www. ceremade. dauphine. fr/~ cardaliaguet/MFG20130420. pdf (2016).

[12] René Carmona and François Delarue, *Probabilistic theory of mean field games with applications i-ii*, Springer, 2018.

[13] René Carmona, François Delarue, et al., *Forward–backward stochastic differential equations and controlled McKean–Vlasov dynamics*, The Annals of Probability **43** (2015), no. 5, 2647–2700.

[14] Mao Fabrice Djete, Dylan Possamaï, and Xiaolu Tan, *McKean-Vlasov optimal control: the dynamic programming principle*, arXiv preprint arXiv:1907.08860 (2019).

[15] Nicole El Karoui, Nguyen Du Huu, and Monique Jeanblanc-Picqué, *Compactification methods in the control of degenerate diffusions: existence of an optimal control*, Stochastics **20** (1987), no. 3, 169–219.

[16] L. C. Evans and W. Gangbo, *Differential Equations Methods for the Monge–Kantorovich Mass Transfer Problem*, Mem. Amer. Math. Soc **137** (1999), 653.

[17] Lawrence C. Evans, *Partial differential equations*, American Mathematical Society, Providence, R.I., 2010.

[18] Wilfrid Gangbo, *An elementary proof of the polar factorization of vector-valued functions*, Archive for rational mechanics and analysis **128** (1994), no. 4, 381–399.

[19] Wilfrid Gangbo and Robert J McCann, *The geometry of optimal transportation*, Acta Mathematica **177** (1996), no. 2, 113–161.

[20] Wilfrid Gangbo, Truyen Nguyen, Adrian Tudorascu, et al., *Hamilton-Jacobi equations in the Wasserstein space*, Methods and Applications of Analysis **15** (2008), no. 2, 155–184.

[21] Mark Kac, *Foundations of kinetic theory*, Proceedings of the third berkeley symposium on mathematical statistics and probability, 1956, pp. 171–197.

[22] L. V. Kantorovich, *On a problem of monge*, Uspekhi Mat. Nauk **3** (1948), no. 2, 225–226.

[23] L. V. Kantorovtich, *On the translocation of masses*, Dokl. Akad. Nauk. SSSR **37** (1942), no. 7-8, 227–229.

[24] Nicole El Karoui and Xiaolu Tan, *Capacities, measurable selection and dynamic programming part i: abstract framework*, arXiv preprint arXiv:1310.3363 (2013).

[25] ———, *Capacities, measurable selection and dynamic programming part ii: application in stochastic control problems*, arXiv preprint arXiv:1310.3364 (2013).

[26] Martin Knott and Cyril S Smith, *On the optimal mapping of distributions*, Journal of Optimization Theory and Applications **43** (1984), no. 1, 39–49.

[27] Jean-Michel Lasry and Pierre-Louis Lions, *Mean field games*, Japanese journal of mathematics **2** (2007), no. 1, 229–260.

[28] Mathieu Laurière and Olivier Pironneau, *Dynamic programming for mean-field type control*, Comptes Rendus Mathematique **352** (2014), no. 9, 707–713.

[29] Guillaume Legendre and Gabriel Turinici, *Second-order in time schemes for gradient flows in Wasserstein and geodesic metric spaces*, Comptes Rendus Mathematique **355** (2017), no. 3, 345–353.

[30] PL Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. Part I: The case of bounded stochastic evolutions*, Acta mathematica **161** (1988), no. 1, 243–278.

[31] ———, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. III. Uniqueness of viscosity solutions for general second-order equations*, Journal of Functional Analysis **86** (1989), no. 1, 1–18.

[32] R. McCann, *A convexity principle for interacting gasses*, Adv. Math. **128** (1997), 153–179.

[33] Robert J McCann, *Polar factorization of maps on riemannian manifolds*, Geometric & Functional Analysis GAFA **11** (2001), no. 3, 589–608.

[34] Robert J McCann et al., *Existence and uniqueness of monotone measure-preserving maps*, Duke Mathematical Journal **80** (1995), no. 2, 309–324.

[35] Henry P McKean, *Propagation of chaos for a class of non-linear parabolic equations*, Stochastic Differential Equations (Lecture Series in Differential Equations, Session 7, Catholic Univ., 1967) (1967), 41–57.

[36] Gaspard Monge, *Memoire sur la theorie des d´eblais et des remblais*, Histoire de l'Acad´emie Royale des Sciences de Paris, avec les M´emoires de Math´ematique et de Physique pour la mˆeme ann´ee (1781), 666–704.

[37] F. Otto, *The geometry of dissipative evolution equations*, Communications in Partial Differential Equations **26** (2001), 102–172.

[38] Huyên Pham and Xiaoli Wei, *Dynamic Programming for Optimal Control of Stochastic McKean–Vlasov Dynamics*, SIAM Journal on Control and Optimization **55** (2017), no. 2, 1069–1101.

[39] ———, *Bellman equation and viscosity solutions for mean-field stochastic control problem*, ESAIM: Control, Optimisation and Calculus of Variations **24** (2018), no. 1, 437–461.

[40] F. Otto R. Jordan D. Kindelherer, *The Variational Formulation of the Fokker-Planck Equation*, SIAM J. Math. Anala **29** (1998), 1–17.

[41] S. Rachev and L. Rüschendorf, *Mass transportation problems*, Springer, 1998.

[42] Ludger Rüschendorf and Svetlozar T Rachev, *A characterization of random variables with minimum L2-distance*, Journal of Multivariate Analysis **32** (1990), no. 1, 48–54.

[43] V. N. Sudakov, *Geometric problems in the theory of infinite-dimensional probability distributions*, Proc. Steklov Inst. Math. **2** (1979), 1–178.

[44] Andrzej Święch, *Optimal control and pde*.

[45] Cedric Villani, *Topics in optimal transport*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, 2003.

[46] Jiongmin Yong and Xun Yu Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*, Vol. 43, Springer Science & Business Media, 1999.

Appendix A

Preliminary Mathematics

In this appendix, we collect some useful results that were studied throughout the honours project and which are necessary to understand the main content of the thesis. The references consulted were [8, 17, 45].

A.1 Cauchy-Lipschitz Theory

Definition A.1.1 (Lipschitz functions)**.** *Let (X_1, d_1) and (X_2, d_2) be two metric spaces. A function $f : X_1 \to X_2$ is called Lipschitz if there exists some $K > 0$ such that for all $x, y \in X_1$,*

$$d_2(f(x), f(y)) \leq K d_1(x, y).$$

This is the central assumption in the Cauchy-Lipschitz theorem.

Theorem A.1.1 (Cauchy-Lipshcitz (1))**.** *For the initial value problem defined on $t \in [0, T]$,*

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0, \tag{A.1}$$

suppose that f is uniformly Lipschitz in x and continuous in t. Then, for some $\delta > 0$, there exists a unique solution $x(t)$ to (A.1) in the interval $[t_0 - \delta, t_0 + \delta]$.

The following theorem is an existence result for flows of vector fields.

Theorem A.1.2 (Cauchy-Lipschitz (2))**.** *Consider the initial value problem* (A.1) *and let f be uniformly Lipschitz in x and continuous in t. Furthermore, for each $t \in [0,t]$ define the map $T_t : \mathbb{R}^n \to \mathbb{R}^n$ where $T_t(x) = x(t)$ and x is the unique solution to* (A.1) *as guaranteed by theorem A.1.1. Then $(T_t)_t$ is a one parameter family of homeomorphisms. If in addition f is C^k, then the map T_t is a C^k diffeomorphism.*

A.2 Functional Analysis

In this section we recall some facts from functional analysis.

Let E be a normed vector space with norm $\|\|_E$. We call the set of all linear functionals, functions of the form $f : E \to \mathbb{R}$, the dual space, E^*. The dual norm is defined as

$$\|f\|_{E^*} = \sup_{\|X\| \leq 1, x \in E} |f(x)|. \tag{A.2}$$

Now, given a function $\phi : E \to (-\infty, \infty]$ that is not identically infinity, define $\phi* : E^* \to (-\infty, \infty]$ such that

$$\phi^*(f) = \sup_{x \in E} \{\langle f, x \rangle - \phi(x)\}, \tag{A.3}$$

where $\langle f, x \rangle = f(x)$ is the scalar product for the duality E, E^*. This is known as the Frenchel-Legendre transform of ϕ and is a convex function, so in light of this, ϕ^* is also called the convex conjugate of ϕ.

Definition A.2.1 (Lower semi-continuous)**.** *A function $\phi : E \to (-\infty, +\infty]$ is lower semi-continuous if for every sequence $x_n \to x$ in E, then*

$$\liminf_{n \to \infty} \phi(x_n) \geq \phi(x).$$

We note that the indicator function is convex if and only if it is the indicator function of a convex set. The following version of Riesz representation theorem for the dual to the space of continuous linear functionals is used in the proof of the Kantorovich duality.

Theorem A.2.1 (Riesz Representation Theorem for Measures)**.** *Let X be a compact Hausdorff space and let μ be a radon measure on X. Then there is a unique signed Borel measure ν on X such that*

$$\langle \mu, u \rangle = \int_X u d\nu,$$

for all $u \in C(X)$.

A.3 Weak formulation of Partial Differential Equations

Throughout this section X will denote some subset of $\mathbb{R}^n$.

Definition A.3.1 (Test function)**.** *A test function $\phi : X \to \mathbb{R}$ is smooth and has compact support. The set of all test functions on X is denoted as $D(X)$ and is a real vector space.*

From here on, we assume $\phi \in D(X)$.

We can define a topology on $D(X)$ in terms of convergence of sequences in this space. We say that $\phi_k \to \phi$ in $D(X)$ if there exists some $K \subset X$ compact such that $\bigcup_{k=1}^{\infty} \text{supp}(\phi_k) \subset K$ and that for every multi index α, the sequence $\partial^\alpha \phi_k \to \partial^\alpha \phi$ uniformly.

Definition A.3.2 (Distribution)**.** *A distribution F is a continuous linear functional $F : D(X) \to \mathbb{R}$ and when F acts on a test function ϕ, we write $\langle F, \phi \rangle$.*

Clearly, the space of all distributions is dual space of $D(X)$, $D(X)^*$.

A function $f : X \to \mathbb{R}$ defines the distribution F_f by the relation

$$\langle F_f, \phi \rangle = \int_X f(x)\phi(x)dx, \tag{A.4}$$

and similarly a measure μ defines F_μ by

$$\langle F_\mu, \phi \rangle = \int_X \phi(x)d\mu(x), \tag{A.5}$$

We define the derivative of a distribution to be such that

$$\langle F', \phi \rangle = -\langle F, \phi' \rangle, \tag{A.6}$$

and as such every distribution is smooth. Note that this is justified by integration by parts since all test functions vanish on the boundary of X as they have compact support. A particular example used in the thesis is the definition of $\nabla\cdot$ operator which we define to be

$$\langle \nabla \cdot F_\mu, \phi \rangle = -\langle F_\mu, \nabla \cdot \phi \rangle. \tag{A.7}$$

When $D(X)^*$ is endowed with the weak-* topology, we have the following result.

Proposition A.3.1. *A sequence of distributions T_k converges to T with respect to the weak-* topology on $D(X)^*$ if and only if*

$$\langle T_k, \phi \rangle \to \langle T, \phi \rangle,$$

for all $\phi \in D(X)$.

We can define a linear differential operator on X to be

$$P = \sum_\alpha a_\alpha(x) \partial^\alpha, \tag{A.8}$$

where $a_\alpha(x)$ are the co-efficients and α are multi indices that vary in some subset of $\mathbb{N}_0^n$. Now, a solution, $u(x)$, to the equation

$$Pu(x) = 0 \tag{A.9}$$

is called a strong solution. We can now define weak solutions to (A.9) in terms of distributions, multiplying (A.9) by $\phi \in D(X)$ and integrating by parts we get the following weak formulation

$$\int_X u(x) Q\phi(x) dx = 0, \tag{A.10}$$

and a solution $u(x)$ that satisfies (A.10) for all $\phi \in D(X)$ is called a weak solution. The operator Q is found by successive integration by parts and is known as the formal adjoint of P and is found to be

$$Q = \sum_{\alpha} (-1)^{|\alpha|} \partial^{\alpha} [a_{\alpha} \phi(x)]. \tag{A.11}$$

A.4 Convex Analysis

In this section, we state the definitions and theorems needed to understand the existence theorems of optimal transport plans in chapter 1 as well as some of the ideas used in the proofs of the theorems in chapter 2.

We first begin with the definition of small sets on a metric space $(X, \tilde{d})$. We will need to introduce the definition of Hausdorff dimension.

Definition A.4.1 (Hausdorff Dimension)**.** *Define the quantity*

$$H^d_{\delta}(S) = \inf\{\sum_{i=1}^{\infty} (diam U_i)^d : \{U_i\}_{I \in \mathbb{N}} \text{ cover } S \text{ and } diam(U_i) \leq \delta)\}.$$

Then

$$H^d(S) = \lim_{\delta \to 0} H^d_{\delta}(S)$$

is the d-dimensional Hausdorff dimension. Finally,

$$diam(S) = \inf\{d \geq 0 : H^d(S) = 0\}$$

is the Hausdorff dimension.

Once we have this definition, a small set in $\mathbb{R}^n$ is one that has Hausdorff dimension $n-1$. In what follows we recall some basic facts about convex functions. We let $U \subset \mathbb{R}^n$.

Definition A.4.2 (Convexity)**.** *A function $f : U \to \mathbb{R}$ is convex if for any $t \in (0,1)$, $x, y \in X$,*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

A convex function is locally Lipschitz continuous on the interior of its domain.

Theorem A.4.1 (Rademacher)**.** *Let $U \subset \mathbb{R}^n$. A function $f : U \to \mathbb{R}$ that is Lipschitz continuous is differentiable almost everywhere.*

So in light of this, a convex function f is automatically differentiable on the interior of it's domain and so ∇f is well defined. For points on the frontier of the domain or simply sharp points in a piece wise defined convex function, it may not be differentiable. To deal with non-differentiability of convex functions, we introduce the notion of subdifferential.

Definition A.4.3. *The subdifferential of a convex function, $f : U \to \mathbb{R}$ at some point $x_0 \in U$ and is the set of all $y \in \mathbb{R}^n$ such for all $x \in U$ we have that*

$$f(x) - f(x_0) \geq \langle y, x - x_0 \rangle.$$

We say that y belongs to the subdifferential of f at x_0 and denote it by $y \in \partial f(x_0)$.

If $\partial f(x_0)$ only contains one element, then f is differentiable at x_0 and that one element is $\nabla f(x_0)$. We follow the convention in [45] and identify the subdifferential with it's graph, $\text{Graph}(\partial f) \subset \mathbb{R}^n \times \mathbb{R}^n$.

A rich duality between convex functions and their convex conjugates (also called the Frenchel-Legendre transform) exists. We state one key result that is used throughout this thesis and that is if f is differentiable and strictly convex then

$$(\nabla f)^{-1} = \nabla f^*.$$

We will need the notion of a superlinear function, that is, a function f that satisfies

$$\lim_{|x| \to \infty} \frac{f(x)}{|x|} = +\infty.$$

In the final part of this section, we will list a series of important definitions that generalize the notion convexity and concavity.

Definition A.4.4 (λ uniformly convex)**.** *A function $\phi : U \to \mathbb{R}$ is called λ uniformly convex if the map $x \mapsto \phi(x) - c\frac{|x|^2}{2}$ is convex.*

Definition A.4.5 (Semi-convex)**.** *A function $\phi : U \to \mathbb{R}$ is called semi-convex with constant C if the map $x \mapsto \phi(x) + c\frac{|x|^2}{2}$ is convex.*

Definition A.4.6 (c-concavity)**.** *Let X, Y be non-empty sets and let $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ be some function. A function $\phi : X \to \mathbb{R} \cup \{-\infty\}$ is c-concave if there exists some $\psi : Y \to \mathbb{R} \cup \{-\infty\}$, not identically $-\infty$, such that for all $x \in X$, $\phi(x) = \inf_{y \in \mathbb{R}^n}[c(x, y) - \psi(y)]$.*