A Course in Optimal Control and Optimal Transport

Dongjun Wu

dongjun.wu@control.lth.se

August, 2023

LUND
UNIVERSITY

CONTENTS

Disclaimer

The present notes are a ongoing work in which there exist many errors and non-rigorous statements – especially about references. I recommend you to double check all the statements while reading. I believe this is also a good way of learning.

CHAPTER 1

DYNAMIC PROGRAMMING

1.1 Discrete time systems

1.1.1 Shortest path problem

To understand dynamic programming, perhaps it is best to start with the shortest path problem. The following digraph (Figure 1.1) shows some possible paths connecting the starting point F to the target T. The number on each arrow indicates the cost walking from one node to the other, and the total cost is the sum of the costs of all moves. The objective is to find the path connecting F to T which has the minimal cost.

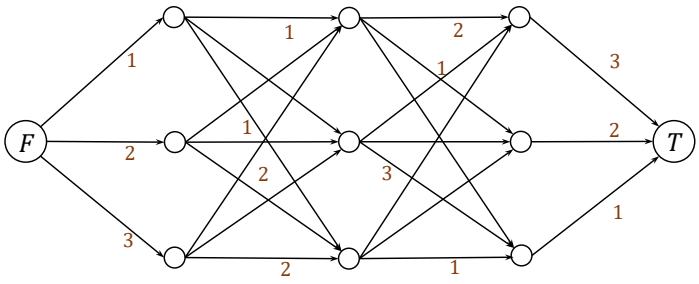

Figure 1.1: Shortest path problem.

A naive solution to this problem is via enumeration. That is, find all the paths connecting F to T, compute the cost of each path, and select the path with the minimal cost. For a problem with N layer (stage) and m states, there are m^{N-2} possible paths, and on each path, one has to do addition operation for $N-1$ times. That is, one has to do at least $(N-1)m^{N-2}$ addition operations, which grows exponentially fast as the number of stages increases. Even for small m, this is not realistic since in practice, N is usually very large.

Dynamic programming can be seen as an *algorithm* that can reduce the computational loads based on the celebrated *Bellman's principle of optimality:*

Bellman's principle of optimality

An optimal policy has the property that no matter what the previous decision have been, the remaining decisions must constitute an optimal policy with regard to the state resulting from those previous decisions.

This principle appears to be obvious and that no proof is needed, although rigorous proof is not hard to provide.

Before applying Bellman's principle of optimality to the shortest path problem, let us first highlight a basic methodology in optimal control that will be used throughout the lecture:

To derive an optimal solution, fix an optimal solution, then check the properties of this solution.

This philosophy, though naive, can sometimes provide rich information of the optimal solution which largely reduces the search space. We justify this fact by applying Bellman's principle of optimality to the shortest path problem. Let us introduce some notations. Denote $J_i(x)$ the *cost-to-go* function from state x at stage i to stage N, $\mathcal{N}(x)$ the set of neighbours of x at the next stage and $c(x,y)$ the cost going from state x (at stage i) to y (at stage $i+1$). The shortest path problem amounts to find

$$\min_{\text{paths } F \to T} J_1(F).$$

Define the *value function*

$$J_i^*(x) = \min_{\text{paths } x \to T} J_i(x)$$

which is the optimal cost going from x at stage i to T. Suppose that we have found an optimal path ℓ, then at any stage $< N$, for $x \in \ell$, according to Bellman's principle, there must hold

$$J_i^*(x) = \min_{y \in \mathcal{N}(x)} \{c(x,y) + J_{i+1}^*(y)\} \tag{1.1}$$

for $i = 1, \cdots, N-1$. The boundary condition appears at $i = N$, in which case $J_N^*(y) = 0$. In principle, one may solve the above equation backward to finally get the value $J_1^*(F)$ and the desired shortest path. Let us count the number of additions that we need to do. As before, the digraph has N stages and at each stage, there are m states. Thus to obtain $J_{N-1}^*(\cdot)$, one only needs to do m comparisons and no addition is needed. To obtain $J_{N-2}^*(\cdot)$, at most m^2 additions are needed, the same for $J_i^*(\cdot)$ when $2 \le i \le N-2$. For $J_1^*(\cdot)$, only m additions are needed. Putting together these operations, we need $(N-3)m^2 + m = O(Nm^2)$ additions. This number is much smaller than $(N-1)m^{N-2}$ when N is large. The equation (1.1), derived from Bellman's principle, is called the *Bellman equation* of this problem. Thus the shortest path problem is turned into solving a Bellman equation.

Although Bellman's equation is merely a necessary condition, it is clear that in the shortest path problem, it's also sufficient for finding the optimal path.

We underscore a crucial property of the cost function that can be easily neglected when applying Bellman's principle. That is, the fact that the total cost is a sum of the costs at each step is essential. We will come back to this point when we study continuous dynamic programming. For the moment, establishing some intuitions is enough.

1.1.2 Optimal control on finite horizon

We now dive into optimal control of discrete time systems. We will see that optimal control can be formulated as a shortest path problem, at least when the control space and state space are finite. Thus the above reasoning still holds true.

Consider the nonlinear discrete time dynamical system

$$x_{k+1} = f_k(x_k, u_k), \tag{1.2}$$

where $x_k \in X_k$ (the system state at time instant k), $u_k \in U_k$ (the input at time instant k). A control cost is a function that takes the following form

$$J = \varphi(x_N) + \sum_{k=1}^{N-1} L_k(x_k, u_k), \tag{1.3}$$

with $1 \leq N \in \mathbb{Z}$, and the initial state x_1 is assumed to fixed. Here φ and L_k are assumed to be some nonnegative functions. The control objective is to seek for a sequence of control input $\pi = (u_1, \cdots, u_{N-1})$, which is also called a *policy,* such that the cost J is minimized, while keeping the constraints $x_k \in X_k$ and $u_k \in U_k$. The cost defined as (1.3) encompasses most (if not all) optimal control costs (on finite time horizon) in the control literature. This is intimately related to the fact that the cost of a problem in real world is almost always additive in the sense we discussed in previous subsection.

Notice that the cost (1.3) is only calculated on finite time intervals, i.e., from 1 to N. We call such an optimal control problem on finite horizon. Sometimes we also consider infinite horizon optimal control where the the cost function takes the form

$$J = \sum_{k=1}^{\infty} L_k(x_k, u_k). \tag{1.4}$$

Although one may formulate the finite horizon optimal control problem as a infinite horizon one, for example, by defining $L_k = 0$ for all $k > N$ and $L_N(x_N, u_N) = \varphi(x_N)$, this may sometimes complicate the problem. As we will see later, in general, the infinite horizon problem is usually harder (at least theoretically) than the finite horizon one.

As mentioned before, when U_k and X_k are finite sets, the optimal control problem is equivalent to a shortest path problem. Thus we can immediately derive the Bellman equation. However, it is often the case that either the input space or the state space or both are continuous spaces, say for example the constraint $|u_k| \leq 1$. Although the problem is no longer a shortest path problem, we can apply Bellman's principle in almost the same manner. As before, define the cost-to-go function $J_i(x) = \sum_{k=i}^{N-1} L_k(x_k, u_k)|_{x_k=x} + \varphi(x_N)$, and the value function $J_i^*(x) = \min_{(u_i,\cdots,u_{N-1})} J_i(x)$. Then according to Bellman's principle,

$$J_i^*(x) = \min_{u_i \in U_i} \{L_i(x, u_i) + J_{i+1}^*(f_i(x, u_i))\}. \tag{1.5}$$

The above equation meets the boundary at $i = N-1$, with $J_N^*(x) = \varphi(x)$, for there is no control at the final stage. Equation (1.5) is the Bellman equation for the optimal control problem on finite horizon. Thus by solving this equation, we can, at least obtain the information (necessary condition) of the optimal policy. It is easy to notice that, this equation can be solved backward. For example, since $J_N^*(\cdot)$ is known, we deduce

$$u_{N-1}^*(x_{N-1}) = \arg\min_{u_{N-1}} \{L_{N-1}(x_{N-1}, u_{N-1}) + \varphi(f_{N-1}(x_{N-1}, u_{N-1}))\}$$

and so on. Finally, one terminates at $u_1^*(x_1) = \arg\min_{u_1}\{L_1(x_1,u_1) + J_2^*(f_1(x_1,u_1))\}$.

The function $J_1^*(x_1)$ is clearly the optimal cost and the corresponding policy $(u_1^*(x_1),\cdots,u_{N-1}^*(x_{N-1}))$ is optimal. That is, solving the Bellman equation (1.5) is necessary and sufficient for finding the optimal control.

Here we mention a difficulty in solving the Bellman equation. When no additional structures are imposed on f and L, the minimization (1.5) is often not numerically tractable. When U_i and X_i are finite with low dimension, it is not a big problem. But taking into consideration that the control law has to be digitalized at the implementation stage, the input space, as well as the state space, when continuous, need to be discretized. We may still assume that U_i and X_i are finite, but with possibly large cardinalities. For example, assume that $U_i = \prod_{k=1}^m I_k \subseteq \mathbb{R}^m$ and $X_i = \prod_{k=1}^n J_k \subseteq \mathbb{R}^n$, with I_k, J_k intervals in $\mathbb{R}$. Partition I_k and J_k into q and p intervals respectively, then there will be q^m possible inputs and p^n states at each stage. In the worst case, there will be $O(Np^nq^m)$ addition operations to do, which is intractable when p and q are large for $n,m \geq 3$. Such phenomenon is called *curse of dimensionality* noticed by Bellman in the 1960s. Today, this term is widely used in various areas to indicate the intractability of the algorithm in higher dimension.

There is, however, a special but extremely important case, that we can solve without much pain: the linear quadratic regulator (LQR) problem.

1.1.3 Example: Discrete LQR on finite horizon

Consider the constraint free linear plant

$$x_{k+1} = Ax_k + Bu_k$$

with cost function defined by

$$J = x_N^\top S_N x_N + \sum_{i=1}^{N-1}(x_i^\top Q x_i + u_i^\top R u_i)$$

with $Q \geq 0$, $S_N \geq 0$ and $R > 0$.

The optimal control problem is to find an optimal control policy such that J is minimized. Using previous notations, the Bellman equation reads

$$J_i^*(x) = \min_{u_i}\{x^\top Q x + u_i^\top R u_i + J_{i+1}^*(Ax + Bu_i)\} \tag{1.6}$$

with boundary condition $J_N^*(x) = x^\top S_N x$. We assert that $J_i^*(x)$ is of the form $x^\top S_i x$ for some $S_i \geq 0$. To see this, we calculate $J_{N-1}^*(x_{N-1})$ and the rest is justified by induction. Indeed,

$$J_{N-1}^*(x_{N-1}) = \min_{u_{N-1}}\{x_{N-1}^\top Q x_{N-1} + u_{N-1}^\top R u_{N-1} + (Ax_{N-1} + Bu_{N-1})^\top S_N (Ax_{N-1} + Bu_{N-1})\},$$

from which we see that

$$u_{N-1}^* = -(B^\top S_N B + R)^{-1} B^\top S_N A x_{N-1}$$

and it is evident that $J_{N-1}^*(x_{N-1})$ contains no first order or scalar terms. Define

$$K_{N-1} := (B^\top S_N B + R)^{-1} B^\top S_N A$$

which is called the *Kalman gain*, then $u_{N-1}^* = -K_{N-1}x_{N-1}$. Substituting u_{N-1}^* back, after direct but cumbersome calculations, we get

$$J_{N-1}^* = x_{N-1}^\top S_{N-1} x_{N-1}$$

where

$$S_{N-1} = Q + (A - BK_{N-1})^\top S_N (A - BK_{N-1}) + K_{N-1}^\top R K_{N-1}$$

or equivalently

$$S_{N-1} = Q + A^\top S_N A - A^\top S_N B (R^\top S_N R + B)^{-1} B^\top S_N A.$$

By induction, one may derive the equation for u_i^*, K_i and S_i, which we summarize in the following:

$$\begin{aligned} K_i &= (B^\top S_{i+1} B + R)^{-1} B^\top S_{i+1} A \\ u_i^* &= -K_i x_i \\ J_i^* &= x_i^\top S_i x_i \\ S_i &= Q + (A - BK_i)^\top S_{i+1} (A - BK_i) + K_i^\top R K_i \end{aligned} \tag{1.7}$$

with boundary condition S_N a known matrix. The optimal value of the problem is provided by $J_1^*(x_1) = x_1^\top S_1 x_1$. The algorithm runs as

$$S_N \to (K_{N-1}, u_{N-1}^*) \to S_{N-1} \to (K_{N-2}, u_{N-2}^*) \to \cdots \to S_2 \to (K_1, u_1^*) \to S_1$$

Although the linear plant we consider here is time-invariant, the extension to time-varying linear systems is rather straightforward: it suffices to replace A by A_i and B by B_i in the formula (1.7).

1.1.4 Infinite horizon problem

Unlike in the finite horizon case, where the time-dependence of the system is of little importance (for example, even though the system is time-invariant, the optimal policy is clearly time-dependent), the optimal control of time-invariant systems on infinite horizon is quite different from that of time-varying systems. In particular, the theory for time-invariant system is much richer than that of time-varying system. For this reason, we will focus on time-invariant system

$$x_{k+1} = f(x_k, u_k) \tag{1.8}$$

where $x_k \in X$ and $u_k \in U$ for all $k \geq 1$. The admissible control input space may be time-dependent, say $u_k \in U(x_k) \subseteq U$, a constraint. The cost function is of the form

$$J = \sum_{k=1}^{\infty} L(x_k, u_k). \tag{1.9}$$

Claim. For any stationary policy u, i.e., $u_k = u(x_k)$ for all $k \geq 1$, the cost function (1.9) under policy u has the property that

$$J_u(x) = L(x, u(x)) + J_u(f(x, u(x)))$$

In fact, $J(x) = L(x, u(x)) + \sum_{k=2}^{\infty} L(x_k, u(x_k)) = L(x, u(x)) + J_u(f(x, u(x))$, as claimed.

Recall that the cost-to-go function $J_i(x) = \sum_{k=i}^{\infty} L(x_k, u_k)|_{x_i = x}$. The value function J_i^* is the same for all i since

$$J_i^*(x) = \min_{(u_i, \cdots,} \sum_{k=i}^{\infty} L(x_k, u_k)|_{x_i = x} = \min_{(u_1, \cdots,} \sum_{k=1}^{\infty} L(x_k, u_k)|_{x_1 = x} = J_1^*(x)$$

Due to this, we may denote $J^*(x) := J_i^*(x)$, and the Bellman equation takes a very special structure:

$$J^*(x) = \min_{u \in U(x)} \{L(x, u) + J^*(f(x, u))\}. \tag{1.10}$$

The difference of (1.10) compared to the Bellman equation of finite horizon problem lies in the fact that the function J^* appears on both sides of the equation. Therefore, it seems not possible to solve equation (1.10) via backward iteration as in the finite horizon case, after all, there is no boundary condition to start with! However, one may guess that starting with $J^* = 0$ and by iteration, J^* converges to a solution. We will discuss this in more detail in next subsection. Once J^* has been found, the optimal policy is given by

$$u^*(x) = \arg\min_{u \in U(x)} \{L(x,u) + J^*(f(x,u))\}.$$

As mentioned before, Bellman equation provides necessary and sufficient condition for finite horizon optimal control problems. One may ask if this still holds for infinite horizon problem, i.e., when (1.10) is satisfied for some function $\hat{J}$, is $\hat{J}$ the optimal cost function? This is clearly untrue as one may always add a constant to the solution which produces another solution. But at least we know the following.

Proposition 1.1. *Let J^* be the optimal cost function and $\hat{J}$ a solution to the Bellman equation* (1.10)*, then $\hat{J} \geq J^*$.*

Proof. By assumption, there exists $\hat{u}(\cdot)$ satisfying $\hat{J}(x) = L(x, \hat{u}(x)) + \hat{J}(f(x, \hat{u}(x))$. Then under the policy $\hat{u}(\cdot)$, for any $x_1 \in X$, we have

$$\hat{J}(x_1) = \hat{J}(x_k) + \sum_{i=1}^{k} L(x_i, \hat{u}(x_i)),$$

which holds for all $k \geq 1$. Thus $\hat{J}(x_1) \geq \sum_{i=1}^{\infty} L(x_i, \hat{u}(x_i)) \geq J^*(x_1)$. □

On the other hand, if we know before hand that the solution to the Bellman equation is unique (at least in a certain class), then we may conclude that solving Bellman equation is sufficient to find the optimal cost function.

Contraction property

Define an operator T accordingly to

$$TJ(x) = \min_{u \in U(x)} [L(x,u) + J(f(x,u))].$$

It is not yet clear how to choose the living space for $J(\cdot)$. Let us consider a simple but illustrative case. Assume that $L(x,u)$ is uniformly bounded for all $u \in U(x)$, L and f are measurable, $U(x)$ is measurable for all x and the minimization is always achieved. Then T can be seen as a mapping on $L^\infty(X)$. We show that T is non-expansive on the Banach space $L^\infty(X)$. In fact, for any other $\tilde{J} \in L^\infty(X)$, there holds

$$\begin{aligned} TJ(x) &= \min_{u \in U(x)} [L(x,u) + \tilde{J}(f(x,u)) + (J(f(x,u)) - \tilde{J}(f(x,u)))] \\ &\leq \min_{u \in U(x)} [L(x,u) + \tilde{J}(f(x,u))] + ||J - \tilde{J}||_\infty \\ &= T\tilde{J}(x) + ||J - \tilde{J}||_\infty, \end{aligned}$$

changing the role of J and $\tilde{J}$, we immediately get

$$||TJ - T\tilde{J}||_\infty \leq ||J - \tilde{J}||_\infty,$$

as claimed. However, non-expansiveness does not necessarily imply the existence of a fixed point.

To get stronger conclusions, further assumptions on the system or cost function shall be needed. For example, this can be achieved by adding a *discount factor* $\alpha \in (0,1)$ to the cost function:

$$J(x) = \sum_{k=0}^{\infty} \alpha^k L(x_k, u_k)\Big|_{x_0 = x}.$$

In this case, it is readily checked that $J^*_{i+1}(x) = \alpha J^*_i(x)$. Hence the Bellman's equation (**??**) becomes

$$J^*(x) = \min_{u \in U(x)} [L(x,u) + \alpha J^*(f(x,u))].$$

Now define $\bar{T}J(x) = \min_{u\in U(x)}[L(x,u) + \alpha J(f(x,u))]$, similarly, we deduce that

$$||\bar{T}J_1 - \bar{T}J_2|| \leq \alpha ||J_1 - J_2||, \ \ \forall J_1, J_2 \in L^\infty(X)$$

Hence $\bar{T}$ is a Banach contraction mapping, and hence there exists a unique $J^* \in L^\infty(X)$ such that

$$\bar{T}J^* = J^*$$

For any initial function $J_0 \in L^\infty(X)$, denoting $J_n = T^n J_0$, we will get

$$||J_n - J^*|| \leq \alpha^n ||J_0 - J^*||$$

Thus J_n converges to the optimal cost exponentially as $n \to \infty$.

It has to be noted however that, the assumption of uniform boundedness of $L(x,u)$ is very strong. For example, it is often the case that $L(x,u) \to \infty$ as $|x| \to \infty$, e.g., $L(x,u) = |x|^2 + |u|^2$, obviating the assumption. Thus, in the present context, the non-expansiveness and contraction properties of the Bellman equation will not be used.

Quadratic cost function for affine nonlinear systems

There is an important class of systems, called control affine systems,

$$x_{k+1} = f(x_k) + g(x_k)u_k$$

with quadratic cost function $L(x,u) = x^\top Q x + u^\top R u$. Assume that f is C^1 and that u can be chosen freely. If the Bellman equation (1.10) admits a C^1 solution (strong assumption!), then one can "solve" for the optimal policy

$$u^*(x) = \frac{1}{2} R^{-1} g(x)^\top \frac{\partial J^*}{\partial x}(f(x) + g(x)u^*(x)). \tag{1.11}$$

Plugging $u^*(x)$ into the Bellman equation, we obtain

$$\begin{aligned} J^*(x) = {} & x^\top Q x + J^*(f(x) + g(x)u^*(x)) \\ & + \frac{1}{4}\left(\frac{\partial J^*}{\partial x}(f(x) + g(x)u^*(x))\right)^\top g(x) R^{-1} g(x)^\top \frac{\partial J^*}{\partial x}(f(x) + g(x)u^*(x)) \end{aligned} \tag{1.12}$$

which is a partial differential equation.

In general, the equations (1.11) and (1.12) cannot be solved explicitly (except in the LQR case as we will see later), thus one needs to apply numerical methods to approximate the solution. One may initialize with an arbitrary policy $u_0(\cdot)$ and a value function $J_0(\cdot)$, then iterate according to equations (1.11) and (1.12):

$$\begin{aligned}
u_1(x) &= \frac{1}{2}R^{-1}g(x)^\top \frac{\partial J_0}{\partial x}(f(x)+g(x)u_0(x)) \\
J_1(x) &= x^\top Qx + J_0(f(x)+g(x)u_1(x)) \\
&\quad + \frac{1}{4}\left(\frac{\partial J_0}{\partial x}(f(x)+g(x)u_1(x))\right)^\top g(x)R^{-1}g(x)^\top \frac{\partial J_0}{\partial x}(f(x)+g(x)u_1(x)) \\
&\vdots
\end{aligned}$$

This naive iteration scheme has no guarantee of convergence for general nonlinear affine systems. However, by slightly modifying the above iteration procedure, convergence can be guaranteed for large class of systems. The idea is, we iterate either the policy u or the value function J while the other one is calculated according to the Bellman equation. Intuitively, this has better convergence property than iterating u and J at the same time, since it may happen that both u and J are away from u^* and J^*. We study this in the next subsection.

Policy iteration and value iteration[1]

There are two basic iteration approaches for solving the Bellman equation (1.10) approximately, namely, policy iteration and value iteration.

Value iteration: start from some non-negative function $J_0 : X \to \mathbb{R}$ and iterate according to

$$J_{k+1}(x) = \min_{u\in U(x)}\{L(x,u) + J_k(f(x,u))\}. \tag{1.13}$$

The approximate optimal policy can be taken as

$$u^*_{N+1}(x) = \arg\min_{u\in U(x)}\{L(x,u) + J_N(f(x,u))\}$$

when J_N reaches a reasonable level of accuracy.

There is an important property of value iteration, called the *monotonicity* property. Being J^* the optimal cost function, if we start from $J_0 \geq J^*$, then $J_k \geq J^*$ for all $k \geq 0$. In fact,

$$\begin{aligned}
J_1(x) &= \min_{u\in U(x)}\{L(x,u) + J_0(f(x,u))\} \\
&\geq \min_{u\in U(x)}\{L(x,u) + J^*(f(x,u))\} \\
&= J^*(x).
\end{aligned}$$

Interestingly, we can get stronger result for the case $J_0 \leq J^*$. That is, the sequence $\{J_k\}$ is monotone increasing:

$$J_0 \leq J_1 \leq J_2 \leq \cdots \leq J_*$$

[1] This part is mainly taken from the paper [2].

since

$$\begin{aligned} J_1(x) &= \min_{u\in U(x)} L(x,u) \geq J_0(x) \\ J_2(x) &= \min_{u\in U(x)} \{L(x,u) + J_1(f(x,u))\} \\ &\geq \min_{u\in U(x)} \{L(x,u) + J_0(f(x,u)) \\ &= J_1(x) \\ &\vdots \end{aligned}$$

Thus, there exists a function $\tilde{J} \leq J^*$, such that $J_k \to \tilde{J}$ pointwisely, but there may exist a gap between $\tilde{J}$ and J^*. The following classical result provides a sufficient condition that $\tilde{J} = J^*$.

Proposition 1.2 (Convergence of value iteration I)**.** *If U is a metric space and the sets*

$$U_k(x,\lambda) = \{u \in U(x) : L(x,u) + J_k(f(x,u)) \leq \lambda\}$$

is compact for all $x \in X$, $\lambda \in \mathbb{R}$ and k, then the value iteration $J_k \uparrow J^*$ pointwisely for any $J_0 \geq 0$ satisfying $J_0(x) \leq \min_{u\in U(x)} L(x,u) + J_0(f(x,u))$ for all $x \in X$, e.g., $J_0 = 0$.

The proof of this proposition is a bit technical, the interesting reader is referred to [1].

We now switch to the other case: $J_0 \geq J^*$, for which we need more structures and assumptions. Assume that the set defined by

$$X_s := \{x \in X : \exists u \in U(x), \text{s.t.}\, L(x,u) = 0,\ x = f(x,u)\}$$

is non-empty. Then, if $x \in X_s$, we have $J^*(x) = 0$. We call X_s the stopping set, which is a desirable set of termination states that we try to reach or approach with minimum total cost.

For an initial state x, a policy π is said to *terminate* starting from x if the trajectory of system starting from x reaches the set X_s in finite time. Denote

$$\mathscr{J} = \{J \geq 0 : J(x) = 0,\ \forall x \in X_s\}. \tag{1.14}$$

We assume that for every $x \in X$, there is a policy which terminates and which can approximate the optimal policy as closely as possible. More precisely, we assume:

Assumption 1. The stopping set X_s is non-empty. Moreover, for every $x \in X$, with $J^*(x) < \infty$, and every $\epsilon > 0$, there exists a policy π that terminates starting from x and satisfies $J_\pi(x) \leq J^*(x) + \epsilon$.

This is a reasonable assumption which is satisfied in many important examples. Normally, the most technical part to verify is the existence of π satisfying $J_\pi(x) \leq J^*(x) + \epsilon$. Check the paper [2] for sufficient conditions that guarantee Assumption 1.

Proposition 1.3 (Uniqueness of solution of Bellman equation)**.** *Let Assumption 1 hold. The optimal cost function J^* is the unique solution of the Bellman equation* (1.10) *in the set $\mathscr{J}$.*

Note that Proposition (1.3) is NOT saying that the optimal policy terminates! It says that the optimal cost must vanish on X_s. For example, as we will see later, in the linear quadratic regulator problem, the optimal policy is a static feedback $u = -Kx$, which normally does not vanish unless one starts from $x = 0$. One should be able to derive the proof of this proposition after finishing this section.

Proposition 1.4 (Convergence of value iteration II)**.** *Let Assumption 1 hold. Then the value iteration* (1.13) J_k *converges pointwisely to* J^* *for initial* J_0 *having the following properties:*

- $J_0 : X \to \mathbb{R}_+$ is non-negative;
- $J_0(x) = 0$ for all $x \in X_s$;
- $J_0 \geq J^*$.

Proof. By monotonicity property, we know $J_k \geq J^*$ for all $k \geq 0$. Moreover, it is obvious that for $x \in X_s$, $J_k(x) = 0$ for all k.

We now take a different viewpoint of the value iteration: it can be seen as solving for a finite horizon optimal control problem with terminal cost J_0. Thus for every policy $\pi = (u_1, u_2, \cdots)$ and every initial state $x_1 \in X$, we have

$$J^*(x_1) \leq J_k(x_1) \leq J_0(x_k) + \sum_{i=1}^{k-1} L(x_i, u_i)$$

where $\{x_i\}_1^k$ is the state trajectory generated by the policy π with initial condition x_1. Note that the first inequality is due to $J_k \geq J^*$. If $J^*(x_1) = \infty$, there is nothing to prove, hence assume $J^*(x_1) < \infty$. Now, for any policy π that terminates from x_1, by definition we have $x_k \in X_s$ for k large enough, and consequently $J_0(x_k) = 0$. Thus

$$\begin{aligned} J^*(x_1) &\leq \limsup_{k\to\infty} J_k(x_1) \\ &\leq \limsup_{k\to\infty} \left\{ J_0(x_k) + \sum_{i=1}^{k-1} L(x_i, u_i) \right\} \\ &= \sum_{i=1}^{\infty} L(x_i, u_i) = J_\pi(x_1) \end{aligned}$$

Now take the infimum on π, we should get

$$J^*(x_1) \leq \limsup_{k\to\infty} J_k(x_1) \leq J^*(x_1)$$

for $J^*(x_1) < \infty$, since J_π can approximate J^* arbitrarily well. □

Let's carry on to the policy iteration scheme.

Policy iteration: start from a policy $u_1(\cdot)$, then solve

$$J_{u_k}(x) = L(x, u_k(x)) + J_{u_k}(f(x, u_k(x))) \tag{1.15}$$

for $J_{u_k}(\cdot)$. Next, iterate $u_k(\cdot)$ according to

$$u_{k+1}(x) \in \arg\min_{u\in U(x)} \{L(x, u) + J_{u_k}(f(x, u))\}. \tag{1.16}$$

We mention that $J_{u_k}(\cdot)$ is only implicitly defined by (1.15), and that J_{u_k} vanishes on X_s.

The main result of policy iteration that we are going to prove is the following.

Proposition 1.5. *Let Assumption 1 hold. A sequence* $\{J_{u_k}\}$ *generated by the policy iteration algorithm* (1.15), (1.16) *satisfies* $J_{u_k}(x) \downarrow J^*(x)$ *for every* $x \in X$.

Proof. In equation (1.16), two stationary policies, u_k and u_{k+1} are involved. Let us replace them with two arbitrary stationary policies, say μ and ν, and that ν is defined through the minimization

$$\nu \in \arg\min_{u\in U(x)} \{L(x,u) + J_\mu(f(x,u))\} \tag{1.17}$$

from which it follows that

$$J_\mu(x) = L(x,\mu(x)) + J_\mu(f(x,\mu(x))) \geq L(x,\nu(x)) + J_\mu(f(x,\nu(x)))$$

For notation ease, denote $\tilde{J}_1(x) = J_\mu(x)$ and $\tilde{J}_2(x) = L(x,\nu(x)) + J_\mu(f(x,\nu(x)))$. Continuing the above procedure inductively (viewing $f(x,\nu(x))$ as x_2), we obtain a monotone decreasing sequence

$$J_\mu(x) \geq \tilde{J}_1(x) \geq \tilde{J}_2(x) \geq \cdots \tilde{J}_i(x) \geq \cdots$$

where

$$\tilde{J}_i(x) = J_\mu(x_i) + \sum_{j=1}^{i-1} L(x_j, \nu(x_j))$$

in which the sequence $\{x_j\}$ is generated by policy ν from x. Thus

$$\begin{aligned} J_\mu(x) &\geq \tilde{J}_2(x) \\ &= \min_{u\in U(x)} \{L(x,u) + J_\mu(f(x,u))\} \text{ (see (1.17))} \\ &\geq \lim_{i\to\infty} \tilde{J}_i(x) \geq \sum_{j=1}^{\infty} L(x_j, \nu(x_j)) \\ &= J_\nu(x) \end{aligned}$$

Now substituting $\mu = u_k$, $\nu = u_{k+1}$ into the above inequality, we get

$$J_{u_k}(x) \geq \min_{u\in U(x)} \{L(x,u) + J_{u_k}(f(x,u)\} \geq J_{u_{k+1}}(x)$$

for all $x \in X$, and all $k \geq 1$. Thus $J_{u_k} \downarrow J_\infty$ for some $J_\infty \geq 0$. Taking the limit on both sides, we obtain[2]

$$J_\infty(x) = \min_{u\in U(x)} \{L(x,u) + J_\infty(f(x,u))\}.$$

That is, J_∞ is a solution to the Bellman equation (1.10). Note that $J_{u_k} \in \mathscr{J}$, hence $J_\infty \in \mathscr{J}$. Now, invoking the uniqueness of the Bellman equation (Proposition (1.3)), the conclusion follows. □

Stability issue

One of the most important problems in control theory is the problem of stabilization. Optimal control provides a way for fulfilling this purpose. For that, a cost function shall be proposed first. A widely

[2] More rigorously, the limiting procedure is divided into two parts. First, note that

$$\min_{u\in U(x)} \{L(x,u) + J_\infty(f(x,u))\} \leq \min_{u\in U(x)} \{L(x,u) + J_{u_k}(f(x,u))\} \leq J_\infty(x)$$

on the other hand,

$$L(x,u) + J_{u_k}(f(x,u)) \geq J_\infty(x)$$

for all u. Now taking the limit, we discover

$$L(x,u) + J_\infty(f(x,u)) \geq J_\infty(x), \quad \forall\, u \in U(x).$$

used function is the quadratic cost function that we have already mentioned previously, i.e., $L(x,u) = x^\top Qx + u^\top Ru$ for $Q > 0$ and $R \geq 0$.

Now if for the nonlinear system (1.8) $x_{k+1} = f(x_k, u_k)$, a solution J^* to the Bellman equation satisfies

$$c_1||x||^p \leq J^*(x) \leq c_2||x||^p, \quad \forall x \in X$$

for some positive constants c_1, c_2 and $p \geq 1$, then we assert that the stationary optimal policy $u^*(x) = \arg\min_u\{L(x,u) + J^*(x,u)\}$ is exponentially stabilizing. Indeed, under the optimal policy,

$$\begin{aligned} J^*(x_{k+1}) &= J^*(x_k) - L(x_k, u^*(x_k)) \\ &\leq (1-c)\, J^*(x_k) \end{aligned}$$

where we have used the fact that $L(x,u) \geq cJ^*(x)$ for some positive constant $c > 0$. It follows that $J^*(x_k) \leq (1-c)^k J^*(x_1) \to 0$ as $k \to \infty$. Hence $x_k \to 0$ exponentially as expected.

Infinite horizon LQR

The most well studied optimal control problem on infinite horizon is the linear quadratic regulator problem. This is mainly because of its wide range applicability and simplicity. For us, the LQR problem will serve as a concrete example to help us enhance the understandings of the materials of this section, e.g., the solution to the Bellman equation, policy and value iteration, and stability issue etc. But we also highlight that, due to the nice structure of the LQR problem, it has some interesting features that cannot be derived from the machinery that we have introduced so far. For example, we will see that *observability* will now play a role in determining the global convergence of the value iteration in for a certain class of initial functions.

Consider the linear time-invariant discrete time system

$$x_{k+1} = Ax_k + Bu_k \tag{1.18}$$

with quadratic cost

$$J = \sum_{k=1}^{\infty} x_k^T Q x_k + u_k^T R u_k \tag{1.19}$$

where $Q \geq 0$ and $R > 0$. Assume that $u \in \mathbb{R}^m$ is constraint free. In order that $J < \infty$, it is sufficient to assume that the system is stabilizable (verify!). Now the Bellman equation (1.10) reads

$$J^*(x) = \min_u\{x^\top Qx + u^\top Ru + J^*(Ax + Bu)\}. \tag{1.20}$$

The central question becomes how to solve the Bellman equation (1.20). At this stage, it is not clear how should J^* look like. Fortunately, we know that under some mild conditions, the value iteration starting from $J_1 = 0$ converge to J^*, see Proposition (1.2). Let's calculate $J_i(x)$. For $i = 1$, it is obvious that

$$J_1(x) = \min_u\{x^\top Qx + u^\top Ru\} = x^\top Qx.$$

Denote $J_1(x)$ as $J_1(x) =: x^\top P_1 x$, or $P_1 = Q$. To get $J_2(x)$, we calculate

$$J_2(x) = \min_u\{x^\top Qx + u^\top Ru + (Ax + Bu)^\top P_1 (Ax + Bu)\}.$$

Let us pause for a moment. It seems that we have done similar calculations in the finite horizon LQR problem. If wee look at the Bellman equation (1.6), we notice that the only difference there is that we solved (1.6) in backward time with a boundary condition at $t = N$. Here, the value iteration starts with an "initial" value function, like solving the Bellman equation in forward time, but it's obvious that these two are the same, only with different boundary conditions. Thus we can copy most of the calculations from there, say equation (1.7), to write $J_i(x) = x^\top P_i x$ where

$$P_{i+1} = Q + A^\top P_i A - A^\top P_i B(R + B^\top P_i B)^{-1} B^\top P_i A \tag{1.21}$$

with boundary condition $P_0 = 0$. To employ Proposition (1.2), it remains to check if the set

$$U_k(x, \lambda) = \{u \in \mathbb{R}^m : x^\top Q x + u^\top R u + x^\top P_k x \leq \lambda\}$$

is compact, but this is obvious since R is positive definite. Thus we conclude that

$$\lim_{i \to \infty} P_i = P$$

for some $P \geq 0$. In particular, the optimal cost function is a purely quadratic function. With this information in mind, now if we are to solve the Bellman equation (1.20) directly instead of by value iteration, we may substitute $J^*(x) = x^\top P x$ into (1.20), to get the discrete time *algebraic Riccati equation* (abbr. DARE):

$$P = Q + A^\top P A - A^\top P B(R + B^\top P B)^{-1} B^\top P A. \tag{1.22}$$

which can also be written as (see (1.7))

$$P = Q + (A - BK)^\top P (A - BK) + K^\top R K. \tag{1.23}$$

From the above reasoning we see that the existence of solution of the DARE is guaranteed when the system is stabilizable. The uniqueness is however a bit more tricky. The only tool available for us is Proposition 1.3, which requires Assumption 1. The reader can check that Assumption 1 is satisfied whenever $Q > 0$ (see [2]), which is a stronger requirement than needed. After all, linear systems theory tells us that the DARE admits a unique solution $P \geq 0$ whenever the pair (A, C) is detectable where C (full row rank) factors Q through $Q = C^\top C$.

One should distinguish between the uniqueness of the solution of the DARE (1.22) and the uniqueness of the solution of the Bellman equation (1.20); the former restricts implicitly on the space of purely quadratic functions.

Proposition 1.6. *Consider the LTI system* (1.18) *with cost function defined by* (1.19). *Assume that* $Q \geq 0$, $R > 0$ *and* Q *can be factored as* $Q = C^\top C$ *for some matrix* C *with full row rank. Assume further that* (A, B) *is stabilizable and* (A, C) *is detectable. Then the following properties hold:*

1. There exists $P > 0$ such that for every $P_1 \geq 0$, we have

$$\lim_{k \to \infty} P_k = P$$

where P_k is obtained by the value iteration (1.21).

2. P is the unique solution to the DARE (1.22) with the set of positive semi-positive definite matrices.

3. The optimal policy is given by the static state feedback

$$u(x_k) = -Kx_k \tag{1.24}$$

where

$$K = -(B^\top PB + R)^{-1} B^\top PA.$$

4. The closed-loop system is exponentially stable, i.e., $A - BK$ is Schur stable (the spectrum of which lies in the open unit circle of the complex plane).

Before proving this proposition, we mention that the second argument is a direct consequence of the first: suppose $\tilde{P} \geq 0$ is another solution, then by the first claim, the value iteration starting with $\tilde{P}$ will converge to P. But $\tilde{P}$ is a fixed point of the DARE, hence $P_k \equiv P$ for all $k \geq 1$, which forces $\tilde{P} = P$.

Another thing to mention is that since we are working implicitly under the space of purely quadratic functions rather than the space $\mathscr{J}$ (see (1.14)), we will adopt a new proof approach.

Proof. *Step 1*: show that P obtained by value iteration from $P_1 = 0$ is positive definite, and hence (1.24) is exponentially stabilizing; see the subsection **Stability issue**.

Notice that under the control (1.24), the closed loop system reads $x_{k+1} = (A-BK)x_k$. Then from (1.23) we see

$$x_k^\top Px_k - x_{k+1}^\top Px_{k+1} = x_k^\top (Q + K^\top RK)x_k, \quad \forall k \geq 1.$$

If P is not positive definite, then $\exists x_1 \neq 0$, such that $x_1^\top Px_1 = 0$, but this enforces that $x_{k+1}^\top Px_{k+1} = x_k^\top (Q + K^\top RK)x_k = 0$ for all $k \geq 1$ since $P \geq 0$, $Q + K^\top RK \geq 0$. Consequently,

$$Cx_k = 0, \quad Kx_k = 0, \quad \forall k \geq 1.$$

In this case, the closed loop system (along this trajectory!) is simply $x_{k+1} = Ax_k$. Now that (A, C) is observable, it follows that $x_k = 0$ for all $k \geq 1$, a contradiction. Thus $P > 0$.

Step 2: show that the value iteration from any $P_1 \geq 0$ also converges to the P above.

Denote by $P_k(P_1)$ the value iteration from P_1 at the k-th step. By monotonicity of the value iteration, we know that $P_k(P_1) \geq P_k(0)$ for all $k \geq 1$. Recall that $x_1^\top P_k(P_1)x_1$ is the minimal cost of

$$x_k^\top P_1 x_k + \sum_{i=1}^{k-1} (x_i^\top Qx_i + u_i^\top Ru_i)$$

then it must be smaller than the cost generated by the control law (1.24), i.e.,

$$\begin{aligned} x_1^\top P_k(P_1)x_1 \leq x_1^\top \Big\{ &((A-BK)^{k-1})^\top P_1 (A-BK)^{k-1} + \\ &+ \sum_{i=1}^{k-1} ((A-BK)^\top)^{i-1} (Q + K^\top RK)(A-BK)^{i-1} \Big\} x_1. \end{aligned}$$

Recall that $A - BK$ is Schur stable, it follows that $\limsup_{k\to\infty} x_1^\top P_k(P_1)x_1 \leq x_1^\top Px_1$. Combining this with the fact that $P_k(P_1) \geq P_k(0) \to P$ as $k \to \infty$, we immediately get $P_k(P_1) \to P$. □

We now turn to policy iteration for the LQR problem.

To make the problem meaningful, choose an initial policy $u_1 = -K_1 x$ such that $A - BK_1$ is Schur stable. Then according to policy iteration, in the first step we solve

$$J_{u_1}(x) = x^\top Q x + u_1^\top R u_1 + J_{u_1}(Ax + Bu_1)$$

for J_{u_1}. Arguably, we can choose J_{u_1} to be purely quadratic, say $J_{u_1}(x) = x^\top P_1 x$. Then the above procedure is equivalent to solving the linear matrix equation

$$P_1 = Q + K_1^\top R K_1 + (A - BK_1)^\top P_1 (A - BK_1). \tag{1.25}$$

In the second step, one computes u_2 according to

$$\begin{aligned} u_2 &= \arg\min_u \{x^\top Q x + u^\top R u + (Ax + Bu)^\top P_1 (Ax + Bu)\} \\ &= -(B^\top P_1 B + R)^{-1} B^\top P_1 A x \\ &=: -K_2 x. \end{aligned}$$

Repeat this procedure, one would obtain a sequence of stationary policies $\{u_k\}$.

Notice that equation (1.25) is not a Riccati equation. However, if we can show that P_k converges to some P, then obviously $K_k \to -(B^\top P B + R)^{-1} B^\top P A$ and that equation now becomes the Riccati equation (1.22). If in addition, $P \geq 0$, then P must be the same P obtained in the value iteration procedure. Indeed, we have the following result. For a proof, see [10].

Proposition 1.7. *Under the assumptions of Proposition 1.6, the matrix P_k obtained from the policy iteration converges monotonically to the matrix P obtained from the value iteration, i.e., $P_k \downarrow P$.*

Remark 1.1. The policy iteration for LQR problem has better converge rate (quadratic) than value iteration (linear), see [10].

1.1.5 Appendix: Multistage Optimization

We provide an alternative approach to optimal control of discrete times systems via multi-stage optimization. It can be seen as the mathematical foundation of Bellman principle.

A Fundamental Lemma

Consider the minimization problem

$$\inf_{(x,y)\in D} f(x,y)$$

where $D \subset X \times X$, and X is a metric space. Denote $D_x := \{y \in X : (x,y) \in D\}$, $D^y := \{x \in X : (x,y) \in D\}$.

Lemma 1.1. *The following equality holds*

$$\inf_{(x,y)\in D} f(x,y) = \inf_x \inf_{y\in D_x} f(x,y) = \inf_y \inf_{x\in D^y} f(x,y).$$

In addition, the infimums can be replaced by minimum when either of the three infimum is achieved for some $(x_*, y_*) \in D$.

Proof. Let $I_1 = \inf_{(x,y)\in D} f(x,y)$, $I_2 = \inf_x \inf_{y\in D_x} f(x,y)$. It suffices to prove $I_1 = I_2$. By definition, there exists a sequence $\{(x_k, y_k)\}_{k=1}^{\infty} \subset D$, such that $\lim_{k\to\infty} f(x_k, y_k) \searrow I_1$ [3]. Then

$$I_2 \le \inf_{y\in D_{x_k}} f(x_k, y) \le f(x_k, y_k),\ \forall k \ge 1.$$

Letting $k \to \infty$, we get $I_2 \le I_1$. For the inverse direction, note that for any $\varepsilon > 0$, one can find a pair $(x'', y'') \in D$ such that

$$\inf_{(x',y')\in D} f(x', y') \ge f(x'', y'') - \varepsilon \tag{1.26}$$

and $f(x'', y'') \ge I_1$. On the other hand, there exists an integer $K > 0$, such that for all $k \ge K$, $f(x_k, y_k) \le I_1 + \varepsilon \le f(x'', y'') + \varepsilon$. It follows from (1.26) that for any $(x, y) \in D$,

$$f(x,y) \ge \inf_{(x',y')\in D} f(x', y') \ge f(x_k, y_k) - 2\varepsilon,\ \forall k \ge K.$$

Hence

$$I_2 = \inf_x \inf_{y\in D_x} f(x,y) \ge \lim_{k\to\infty} f(x_k, y_k) - 2\varepsilon = I_1 - 2\varepsilon.$$

Since ε is arbitrary, we get $I_2 \ge I_1$, which implies that $I_1 = I_2$. □

This lemma is extremely simple but powerful as we will see next. We underscore that the interchangeability of two inf are essential. On the other hand, it is usually illegal to interchange inf and sup as in general

$$\inf_x \sup_y f(x,y) \ne \sup_y \inf_x f(x,y).$$

which makes differential games different from classical optimal control.

Multistage optimization

For most of the time, we will consider "min" in lieu of "inf", due to the consideration that the difference between the two are not essential for our discussions.

Consider the function

$$J(x,y,z) = f(x,y) + g(y,z)$$

and the minimization problem

$$J^*(x) = \min_{(y,z)\in D} f(x,y) + g(y,z).$$

Invoking Lemma 1.1, J can be rewritten as

$$\begin{aligned} J^*(x) &= \min_y \min_{z\in D_y} f(x,y) + g(y,z) \\ &= \min_y (f(x,y) + \min_{z\in D_y} g(y,z)) \end{aligned}$$

The minimization has been divided into two steps and that is why we call it a twostage minimization problem. In the first step, y is fixed, and we minimize $g(y,z)$ over D_y to get a function $\varphi(y) = \min_{z\in D_y} g(y,z)$. The second step is to minimize the function $f(x,y) + \varphi(y)$.

[3] $\lim_{k\to\infty} a_k \searrow a$ means that a_k is decreasing and converges to a.

For multistage minimization, we consider

$$\begin{aligned} J(x_0) &= \sum_{k=1}^{N} g_k(x_{k-1}, x_k) \\ J^*(x_0) &= \min_{(x_1,\cdots x_N)} J(x_0, x_1, \cdots, x_N) \end{aligned} \tag{1.27}$$

This is a multistage minimization problem. Using Lemma 1.1, we know

$$\begin{aligned} J_i(x) &= \sum_{k=i}^{N} g_k(x_{k-1}, x_k)\Big|_{x_{i-1}=x} \\ J_i^*(x) &= \min_{(x_i,\cdots,x_N)} J_i(x), \;\; 1 \le i \le N \end{aligned}$$

rewrite J as

$$\begin{aligned} J_1^*(x_0) &= \min_{x_1} \min_{(x_2,\cdots,x_N)} \sum_{k=1}^{N} g_k(x_{k-1}, x_k) \\ &= \min_{x_1} \left[g_1(x_0, x_1) + \min_{(x_2,\cdots,x_N)} \sum_{k=2}^{N} g_k(x_{k-1}, x_k) \right] \\ &= \min_{x_1} \left[g_1(x_0, x_1) + J_2^*(x_1) \right] \end{aligned}$$

Similarly,

$$\begin{aligned} J_m^*(x) &= \min_{x_m} [g_m(x, x_m) + J_{m+1}^*(x_m)], \;\; 1 \le m \le N-1 \\ J_N^*(x) &= \min_{x_N} g_N(x, x_N) \end{aligned}$$

The above algorithm is nothing but the Bellman equation that we derived earlier using Bellman's principle of optimality. Hence in the discrete time case, everything that we have done so far can be approached alternatively via multi-stage optimization. In case you feel a bit uncomfortable with Bellman's principle, which we didn't prove rigorously, then the multi-stage optimization provides you a rigorous framework that should dispel all your doubts.

1.2 Continuous time systems

1.2.1 Bellman principle and the HJB equation

Bellman principle

In this section, we begin to study dynamic programming in the setting of continuous time systems:

$$\begin{aligned} \dot{x} &= f(x, u), \\ x(t_0) &= x_0 \end{aligned} \tag{1.28}$$

where $x(t) \in X \subseteq \mathbb{R}^n$, $u(t) \in U_t \subseteq \mathbb{R}^m$ and $u(\cdot) \in \mathscr{U}$, and $\mathscr{U}$ is called the *space of admissible control input*. To avoid pathological cases, we assume that the solution to this equation exists and is unique for each $u(\cdot) \in \mathscr{U}$. When the initial condition is clear from the context or is not important to us, we simply write $x(t)$ as the solution to system. If however we want to highlight the initial condition, we may write $x(t, x_0)$ (when the initial time instant is unimportant) or $x(t; t_0, x_0)$. If, further more, we want to include the input, we can write $x(t; t_0, x_0, u)$ for $u \in \mathscr{U}$.

Recall that in the discrete time setting, the cost function for finite horizon problem is defined as $\varphi(x_N) + \sum_{k=1}^{N-1} L(x_k, u_k)$, i.e., a sum of the cost at each stage plus a boundary cost. Naturally, we can consider a similar cost in the continuous time setting by changing the sum to integration. More precisely, we will consider cost functions in *Bolza form*

$$J(u(\cdot)) = \varphi(x(T)) + \int_0^T L(x(s), u(s))\mathrm{d}s \tag{1.29}$$

where φ and L are both non-negative functions. Likewise, we can consider infinite horizon cost

$$J(u(\cdot)) = \int_0^\infty L(x(s), u(s))\mathrm{d}s$$

As before, the objective of optimal control is to seek for an admissible control $u^*(\cdot)$ such that

$$u^*(\cdot) \in \arg\min_{u\in\mathscr{U}} J(u(\cdot)). \tag{1.30}$$

Remark 1.2. The discrete cost can be seen as a special case of the Bolza form cost since the sum is simply an integration w.r.t. the counting measure.

To apply Bellman's principle, as before, we should define cost-to-go and value functions. Again, these are done by simply changing the summation to integration in the discrete time setting. In words, the *cost-to-go function* from $t = s$ with $x(s) = y$ to T is

$$J(s, y; u(\cdot)) := \varphi(x(T)) + \int_s^T L(x(t), u(t), t)dt,$$

and the *value function* is defined as the optimal value of the cost-to-go under admissible control on the interval $[s, T]$:

$$J^*(s, y) := \min_{u(\cdot)\in\mathscr{U}|_{[s,T]}} J(s, y; u(\cdot)). \tag{1.31}$$

Here, the set $\mathscr{U}|_{[s,T]}$ is the set of admissible controls that can be implemented on the interval $[s, T]$. More rigorously, $\mathscr{U}|_{[s,T]} = \{u1_{[s,T]} : u \in \mathscr{U}\}$, where $1_{[s,T]}$ stands for the characteristic function of the set $[s, T]$.

Recall that the Bellman principle says: an optimal policy has the property that no matter what the previous decision have been, the remaining decisions must constitute an optimal policy with regard to the state resulting from those previous decisions. Thus for any time instant r, there must hold

$$J^*(s, y) = \min_{u(\cdot)\in\mathscr{U}|_{[s,r]}} \left\{\int_s^r L(x(t, s, y, u), u(t))\mathrm{d}t + J^*(r, x(r; s, y, u)\right\}, \quad \forall r \in [s, T]. \tag{1.32}$$

As we have seen, in the discrete time setting, the correctness of the corresponding formula of (1.32) can be justified rigorously via multi-stage optimization. In the continuous time setting, the following lemma, which can be equally called Bellman's principle of optimality, legitimates the formula (1.32).

Lemma 1.2. *Let $J(u(\cdot))$ be a cost function, with $u(\cdot) \in \mathscr{U}|_{[t_0,t_1]}$. Assume that*

1. $J(u(\cdot))$ is separable for any time $t \in [t_0, t_1]$ in the sense that there exist functions $J_1 : U \times \mathbb{R} \to \mathbb{R}$, $J_2 : U \to \mathbb{R}$ such that

   $$J(u(\cdot)) = J_1(u_1(\cdot), J_2(u_2(\cdot)))$$

   where $u_1 = u1_{[t_0,t)}$ and $u_2 = u1_{[t,t_1]}$ for all $t \in [t_0, t_1]$, i.e., the truncations of u on the interval $[t_0, t)$ and $[t, t_1]$ respectively;

2. J_1 is nondecreasing with respect to the second argument.

Then Bellman's principle of optimality holds for the cost function $J(u(\cdot))$:

$$J^* = \min_{u(\cdot)\in\mathscr{U}|_{[t_0,t_1]}} J(u(\cdot)) = \min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J_1\left(u_1(\cdot), \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} J_2(u_2(\cdot))\right), \quad \forall\, t\in(t_0,t_1]$$

Proof. For any $u_1(\cdot), u_2(\cdot)$, we have

$$J^* \le J_1(u_1(\cdot), J_2(u_2(\cdot)))$$

hence

$$J^* \le J_1\left(u_1(\cdot), \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} J_2(u_2(\cdot))\right)$$

and

$$J^* \le \min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J_1\left(u_1(\cdot), \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} J_2(u_2(\cdot))\right).$$

On the other hand,

$$\begin{aligned}
&\min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J_1\left(u_1(\cdot), \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} J_2(u_2(\cdot))\right)\\
&\le \min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J_1\left(u_1(\cdot), J_2(u_2(\cdot))\right) \quad \text{(monotonicity)}\\
&\le \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} \min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J_1\left(u_1(\cdot), J_2(u_2(\cdot))\right)\\
&= \min_{u_2(\cdot)\in\mathscr{U}|_{[t,t_1]}} \min_{u_1(\cdot)\in\mathscr{U}|_{[t_0,t)}} J(u(\cdot))\\
&= \min_{u(\cdot)\in\mathscr{U}|_{[t_0,t_1]}} J(u(\cdot)) = J^*.
\end{aligned}$$

This completes the proof. □

To verify (1.32), let

$$\begin{aligned}
J_1(u_1, y) &= y + \int_0^t L(x(s), u(s))\,ds\\
J_2(u_2) &= \varphi(x(T)) + \int_t^T L(x(s), u(s))\,ds
\end{aligned}$$

then obviously $J(u(\cdot)) = J_1(u_1, J_2(u_2))$ and J_1 is nondecreasing with respect to the second argument. (Note that J_2 is nothing but the cost-to-go function!)

The Hamilton-Jacobi-Bellman equation

Let's recall the central result from the previous subsection: if J^* is the value function defined for the optimal control problem defined by (1.28), (1.29) and (1.30), then it satisfies the following equation:

$$J^*(s,y) = \min_{u(\cdot)\in\mathscr{U}|_{[s,r]}} \left\{\int_s^r L(x(t;s,y,u),u(t))\mathrm{d}t + J^*(r, x(r;s,y,u)\right\}, \quad \forall\, r\in[s,T]. \tag{1.33}$$

This equation looks too implicit and is hard to use in practice. The main task of this subsection is to derive the celebrated *Hamilton-Jacobi-Bellman* equation based (1.33), a more tractable form than (1.33). The key is to note that (1.33) is satisfied for all $r \ge s$ and hence one can take derivatives when J^* are assumed to be smooth.

On the one hand, for any give $u \in \mathscr{U}$, and $r > s$, we have

$$\frac{J^*(s,y) - J^*(r, x(r;s,y,u))}{r-s} - \frac{1}{r-s}\int_s^r L(x(t;s,y,u), u(t))\mathrm{d}t \le 0, \quad \forall u \in \mathscr{U}$$

Suppose that J^* is continuously differentiable, and that L, u are continuous, then the above implies

$$-\frac{\partial J^*}{\partial s}(s,y) - \frac{\partial J^*}{\partial y}(s,y) f(y,u(s)) - L(y,u(s)) \le 0, \quad \forall u(s) \in U_s$$

resulting in

$$-\frac{\partial J^*}{\partial s}(s,y) + \sup_{u \in U_s} H\left(y, u, -\frac{\partial J^*(s,y)}{\partial y}\right) \le 0. \tag{1.34}$$

where

$$H(x,u,p) = p^\top f(x,u) - L(x,u) \tag{1.35}$$

On the other hand, for any pair (r,ϵ), with $r > s$, $\epsilon > 0$, there exists a control $u_{\varepsilon,r}$ such that

$$J^*(s,y) \ge \int_s^r L(x(t,s,y,u_{\varepsilon,r}(\cdot)), u_{\varepsilon,r}(t))\mathrm{d}t + J^*(r, x(r;s,y,u_{\varepsilon,r})) - \varepsilon(r-s)$$

or

$$\begin{aligned}
-\varepsilon &\le \frac{J^*(s,y) - J^*(r, x(r,s,y,u_{\varepsilon,r})}{r-s} - \frac{1}{r-s}\int_s^r L(x(t,s,y,u_{\varepsilon,r}), u_{\varepsilon,r}(t), t)\mathrm{d}t \\
&= -\frac{1}{r-s}\int_s^r \left[\frac{\partial J^*}{\partial s}(t, x(t,s,y,u_{\varepsilon,r}) + \frac{\partial J^*}{\partial y}(t, x(t,s,y,u_{\varepsilon,r}) f(x(t,s,y,u_{\varepsilon,r}), u_{\varepsilon,r}(t))\right]\mathrm{d}t \\
&\quad - \frac{1}{r-s}\int_s^r L(x(t,s,y,u_{\varepsilon,r}), u_{\varepsilon,r}(t), t)\mathrm{d}t \\
&= \frac{1}{r-s}\int_s^r \left[-\frac{\partial J^*}{\partial s}(t, x(t,s,y,u_{\varepsilon,r}) + H\left(x(t,s,y,u_{\varepsilon,r}), u_{\varepsilon,r}(t), -\frac{\partial J^*}{\partial y}(t, x(t,s,y,u_{\varepsilon,r})\right)\right]\mathrm{d}t
\end{aligned}$$

Let $r \to s+$ while keeping ϵ fixed, we get

$$\begin{aligned}
-\varepsilon &\le -\frac{\partial J^*}{\partial s}(s,y) + H\left(y, u_{\varepsilon,r}(s), -\frac{\partial J^*}{\partial y}(s,y)\right) \\
&\le -\frac{\partial J^*}{\partial s}(s,y) + \sup_{u \in U_s} H\left(y, u, -\frac{\partial J^*}{\partial y}(s,y)\right)
\end{aligned} \tag{1.36}$$

Since ε is arbitrary, (1.36) and (1.34) together imply

$$-\frac{\partial J^*}{\partial s}(s,y) + \sup_{u \in U_s} H\left(y, u, -\frac{\partial J^*}{\partial y}(s,y)\right) = 0, \ \ \forall s \in [0,T], \ \forall y \in X$$

or equivalently

$$\frac{\partial J^*}{\partial s}(s,y) + \inf_{u \in U_s}\left\{\frac{\partial J^*}{\partial y}(s,y) f(y,u) + L(y,u)\right\} = 0, \ \ \forall s \in [0,T], \ \forall y \in X$$

This is a partial differential equation with dependent variable (s,y). By the definition of value function (see (1.31)), the PDE is accompanied with boundary condition

$$J^*(T,y) = \varphi(y), \quad \forall y \in X.$$

Summarizing, we have:

Proposition 1.8. *Suppose that $J^*(s,x)$ defined as (1.31) is continuously differentiable. Then $J^*(t,x)$ is a solution to the following Hamilton-Jacobi-Bellman (HJB) PDE on $[0,T]\times X$:*

$$-V_t(t,x)+\sup_{u\in U_t} H(x,u,-V_x(t,x))=0, \tag{1.37}$$

or its equivalent form

$$V_t(t,x)+\inf_{u\in U_t}\left\{V_x(t,x)f(x,u)+L(x,u)\right\}=0, \tag{1.38}$$

with boundary condition

$$V(t,x)=\varphi(x),$$

where H is defined in (1.37) and we have adopted the notations $\frac{\partial V}{\partial t}=V_t$ and $\frac{\partial V}{\partial x}=V_x$.

Although the optimal control problem is a "minimization", the HJB equation (1.37) may involve a maximization, see (1.37).

Suppose that $U_t=U$ for all $t\geq 0$. To obtain the optimal control law based on the solution of the HJB equation (1.37), we can follow Algorithm 1.1 (called the *verification rule*).

Algorithm 1.1 The verification rule

1. Solve the optimization problem

$$u^*(x,p)=\arg\sup_{u\in U} H(x,u,-p).$$

2. Find a continuously differentiable solution $V(t,x)$ to

$$\begin{aligned} -V_t(t,x)+&H\left(x,u^*(x,V_x(t,x)),-V_x(t,x)\right)=0,\\ V(T,x)=&\varphi(x), \end{aligned} \tag{1.39}$$

   for $(t,x)\in(0,T]\times X$.

3. Solve for the solution $x^*(t)=:x^*(t;s,x)$ to the Cauchy problem of the following ODE:

$$\begin{aligned} \dot{x}^*&=f(x^*(t),u^*(t,V_x(t,x^*(t))))\\ x^*(s)&=x \end{aligned}$$

   Then

$$u^*(t,V_x(t,x^*(t)))$$

   is an optimal control and $x^*(t;t_0,x)$ is the corresponding optimal process.

As the discrete time optimal control problem on finite horizon, solving the Bellman equation (1.37) (when the solution has some regularities) is sufficient to obtain the optimal control. For continuous problems, we have a similar result.

Proposition 1.9. *If the verification rule Algorithm 1.1 admits a C^1 solution V, then u^* obtained from the algorithm is an optimal control.*

Proof. Let V be a C^1 solution to the Bellman equation. Let $u(\cdot)$ be any admissible control and $x(\cdot)$ the corresponding trajectory. Then for initial condition x_0,

$$\begin{aligned} V(t,x(t)) - V(0,x_0) &= \int_0^t \frac{dV(s,x(s))}{ds}\mathrm{d}s \\ &= \int_0^t V_t(s,x(s)) + V_x(t,x(s))f(x(s),u(s))\mathrm{d}s \\ &\geq \int_0^t -L(x(s),u(s))\mathrm{d}s \text{ (use(1.38))} \end{aligned}$$

from which it follows that

$$V(0,x_0) \leq \varphi(x(T)) + \int_0^t L(x(s),u(s)\mathrm{d}s.$$

Since $u(\cdot)$ is arbitrary, we conclude that $V(0,x_0)$ is the optimal value function. On the other hand, it is readily checked that u^* is a control that achieves the optimal value. □

Apparently, the most challenging part of the algorithm is the second step, i.e., solving a PDE of the form $F(x,v,v_x)=0$. But even numerically solving the HJB is quite difficult, which normally incurs curse of dimensionality after discretization.

The above two main results both have a drawback: they require the value function to be continuously differentiable, which is almost never met in real applications. What's worse, the HJB equation may not have continuously differentiable solutions! This problem turns out be non-negligible and must be handled with care. We will come back to this issue later.

1.2.2 Example: Continuous LQR on finite horizon

We study the system

$$\dot{x} = A(t)x + B(t)u$$

with $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$ and cost function

$$J = x(T)^\top Q_f x(T) + \int_{t_0}^T x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t)\mathrm{d}t$$

where Q_f, $Q(t) \geq 0$ and $R(t) > 0$ for all $t \geq 0$. In addition, we assume $A(\cdot)$, $B(\cdot)$, $Q(\cdot)$ and $R(\cdot)$ are continuous. The objective is to find an optimal control u^* such that J is minimized.

The Hamiltonian function is

$$H(x,u,p,t) = p^\top(A(t)x + B(t)u) - x^\top Q(t)x - u^\top R(t)u$$

We implement the first two steps of the verification rule:

Step 1: solve the minimization $\min_u H(x,u,p,t)$, resulting in

$$u^* = \arg\max_u H(x,u,p,t) = \frac{1}{2}R(t)^{-1}B(t)^\top p$$

and

$$H(x,u^*,p,t) = p^\top A(t)x - x^\top Q(t)x + \frac{1}{4}p^\top B(t)R(t)^{-1}B(t)^\top p$$

Step 2: solve the HJB

$$-V_t - V_x A(t)x - x^\top Q(t)x + \frac{1}{4}V_x B(t)R(t)^{-1}B(t)^\top V_x^\top = 0.$$

Consider a candidate $V(t,x) = x^\top P(t)x$, in this case

$$u^* = \frac{1}{2}R(t)^{-1}B(t)^\top P(t)x$$

Substitute into the above HJB equation we get

$$-x^\top \dot{P}(t)x - x^\top [P(t)A(t) + A(t)^\top P(t)]x - x^\top Q(t)x + x^\top P(t)B(t)R(t)^{-1}B(t)^\top P(t)x = 0$$

Since this equation must be satisfied for all x, it follows that

$$-\dot{P}(t) = Q(t) + P(t)A(t) + A(t)^\top P(t) - P(t)B(t)R(t)^{-1}B(t)^\top P(t). \tag{1.40}$$

with boundary condition

$$P(T) = Q_f.$$

The first order ODE (1.40) is called *differential Riccati equation* (DRE). Thus the continuous LQR problem on finite horizon reduces to solving the DRE (1.40).

Proposition 1.10. *Suppose that $A(\cdot)$, $B(\cdot)$, $Q(\cdot)$ and $R(\cdot)$ are continuous and $Q_f \geq 0$, $Q(t) \geq 0$, $R(t) > 0$ for all $t \in \mathbb{R}$. Then the differential Riccati equation has a unique semi-positive definite solution on any interval $(t_0, T]$ for all $t_0 \in \mathbb{R}$.*

Proof. Notice that the right hand side of (1.40) is quadratic in P (thus locally Lipschitz!) and that $A(\cdot)$, $B(\cdot)$, $Q(\cdot)$ and $R(\cdot)$ are continuous, therefore local existence and uniqueness of solutions are guaranteed. This also implies that the solution to (1.40) is symmetric: if $P(t)$ is a solution, so is $P(t)^\top$, while both have the same terminal condition, thus $P(t) = P(t)^\top$.

We show next that there is no finite escape time. Suppose that the solution exists on $(t_1, T]$ for some finite $t_1 \in \mathbb{R}$. Then by construction, for any $t_2 \in (t_1, T]$, and $x(t_2) \in \mathbb{R}^n$,

$$x(t_2)^\top P(t_2)x(t_2) \leq x(T)^\top Q_f x(T) + \int_{t_2}^{T} x(t)^\top Q(t)x(t) + u(t)^\top R(t)u(t)\mathrm{d}t, \quad \forall u(\cdot).$$

(The inequality becomes equality for $u = u^*$, thus we also get $P(t_2) \geq 0$.) In particular, this is true for $u \equiv 0$, implying that one can find a constant $c > 0$ such that $P(t_2) < cI$ for all $t_2 \in (t_1, T]$ (no blow-up!). It is then routine to show that when t_2 is sufficiently close to t_1, the solution can be extended outside $(t_1, T]$. □

Remark 1.3. Note that the DRE (1.40) has a terminal condition instead of an initial condition. If one wants to solve a true ODE in forward time, one can introduce a change of variables

$$\tau = T - t,\ \tilde{P}(\tau) = P(T-\tau),\ \tilde{R}(\tau) = R(T-\tau)$$
$$\tilde{A}(\tau) = A(T-\tau),\ \tilde{B}(\tau) = B(T-\tau)$$

Then it becomes equivalent to solving

$$\dot{\tilde{P}}(\tau) = \tilde{Q}(\tau) + \tilde{P}(\tau)\tilde{A}(\tau) + \tilde{A}(\tau)^\top \tilde{P}(\tau) - \tilde{P}(\tau)\tilde{B}(\tau)\tilde{R}(\tau)^{-1}\tilde{B}(\tau)\tilde{P}(\tau),$$
$$\tilde{P}(0) = Q_f.$$

1.2.3 Method of characteristics and the Hamiltonian equation

Method of characteristics

A well-known approach to solving PDE of the form

$$F(x, v, v_x) = 0, \quad x \in \Omega \subset \mathbb{R}^n. \tag{1.41}$$

with boundary condition

$$v(x) = \tilde{v}(x), \quad x \in \partial\Omega,$$

is via the so-called *method of characteristics*. Here v is a real valued function and F is assumed to be a continuous mapping from $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ to $\mathbb{R}$. In addition, we assume Ω to be compact with smooth boundary.

The idea of the method of characteristics is to turn the first order PDE (1.41) into a set of ODEs. Given a point $y \in \partial\Omega$ and a curve $x : [0,1] \to \bar{\Omega}$, with $x(0) = y$. W examine the values of $v(x)$ along this curve, see Figure 1.2.3.

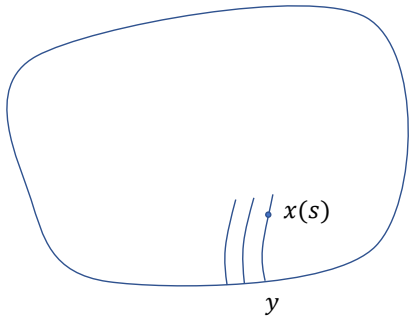

Figure 1.2: Method of characteristics.

Introduce the notation

$$(p_1, \cdots, p_n) = (v_{x_1}, \cdots, v_{x_n}).$$

For convenience, denote

$$v(s) =: v(x(s))$$
$$p(s) =: p(x(s)) = v_x(x(s)).$$

Differentiating v and p w.r.t. s, we find

$$\dot{v} = \sum_{i=1}^{n} v_{x_i} \dot{x}_i = \sum_{i=1}^{n} p_i \dot{x}_i$$
$$\dot{p}_i = \sum_{j=1}^{n} v_{x_i x_j} \dot{x}_i$$

where $\dot{x}_i$ stands for the derivative of x_i w.r.t. s. Further, differentiating (1.41) w.r.t. x_i, we get

$$\frac{\partial F}{\partial x_i} + \frac{\partial F}{\partial v} v_{x_i} + \sum_{i=1}^{n} \frac{\partial F}{\partial p_i} v_{x_i x_j} = 0.$$

Now if the curve $x(s)$ is chosen such that $\dot{x}_i = \partial F/\partial p_i$ (this time, we call x a characteristic curve), one can easily obtain the following

$$\begin{aligned}\dot{v} &= \sum_{i=1}^{n} p_i \frac{\partial F}{\partial p_i},\\ \dot{p}_i &= -\frac{\partial F}{\partial x_i} - \frac{\partial F}{\partial v} p_i, \quad i = 1, \cdots, n\end{aligned}$$

or in more compact form

$$\begin{cases} \dot{x} = F_p^\top \\ \dot{v} = p^\top \dfrac{\partial F}{\partial p} \\ \dot{p} = -F_x^\top - F_v^\top p \end{cases} \tag{1.42}$$

The above equation is a system of ordinary differential equations with boundary condition

$$x(0) = y, \quad v(0) = \bar{v}(y), \quad p(0) = v_x(y)$$

for $y \in \partial\Omega$. Thus by varying the initial condition y, we can obtain *local solutions* near $\partial\Omega$ of the PDE (1.41). In general, however, the solution cannot be extend globally to the entire region Ω. For example, when two characteristic curves meet in Ω, singularity occurs.

To solve the HJB using method of characteristics, we first need to write the equation (1.39) into the standard form $F(x, v, v_x) = 0$ for some F. For that, let $x_{n+1} = t$ and $\tilde{x} = (x, x_{n+1})$. Then (1.39) can be written as $-v_{x_{n+1}} + H(x, u^*(x, v_x), -v_x) = 0$, or $-v_{x_{n+1}} + \tilde{H}(x, v_x) = 0$ for some scalar function $\tilde{H}$, in which Dv stands for the gradient of v w.r.t. x (not $\tilde{x}$!). Let $\tilde{p} = (p_1, \cdots, p_n, p_{n+1})$, then F takes the form

$$F(\tilde{x}, v, \tilde{p}) = -p_{n+1} + \tilde{H}(x, p).$$

Hence the first line of (1.42) reads

$$\begin{aligned}\dot{x} &= F_p^\top = \tilde{H}_p^\top \\ \dot{x}_{n+1} &= \frac{\partial F}{\partial p_{n+1}} = -1\end{aligned} \tag{1.43}$$

Notice that $\partial F/\partial v = 0$, the third line of (1.42) reads

$$\begin{aligned}\dot{p} &= -\tilde{H}_x \\ \dot{p}_{n+1} &= -\frac{\partial \tilde{H}}{\partial x_{n+1}} = 0\end{aligned} \tag{1.44}$$

and the second line of $\dot{v} = p^\top \tilde{H}_p - p_{n+1}$. In the above formulas, the only relevant ones are the first lines of (1.43) and (1.44), i.e.,

$$\begin{cases} \dot{x} = \tilde{H}_p^\top \\ \dot{p} = -\tilde{H}_x^\top \end{cases} \tag{1.45}$$

This equation is the celebrated *Hamiltonian equation* which plays a fundamental role in analytic mechanics and modern physics. In next subsection, we mention some well-know properties of the Hamiltonian equation.

The Hamiltonian equation

For the moment, the tilde above H in equation (1.45) is superfluous. We consider instead

$$\begin{cases} \dot{x} = \ H_p^\top \\ \dot{p} = -H_x^\top \end{cases} \tag{1.46}$$

for $x, p \in \mathbb{R}^n$. In the sequel, we introduce some well-know properties of the Hamiltonian equation (1.46).

Energy conservation. In physics, for example in mechanics, the function H is often some form of energy of the system. If we calculate the time evolution of the energy, we discover that

$$\frac{dH}{dt} = \frac{\partial H}{\partial p}^\top \dot{p} + \frac{\partial H}{\partial x}\dot{x} = \dot{x}^\top \dot{p} - \dot{p}^\top \dot{x} = 0.$$

That is, along the trajectory of the system (1.46), the energy function keeps constant.

Volume preservation (Liouville theorem). Another remarkable property of the Hamiltonian equation is volume preservation. Consider a bounded measurable set D on the phase space $\mathbb{R}^{2n}$. Starting from $t = 0$, the set D is mapped to $\phi_t(D)$ by the flow of Hamiltonian equation at time instant t. Denote $\mathrm{vol}(\Omega)$ the volume of a measurable set Ω. Then the transport equation[4] tells us that

$$\frac{d}{dt}\mathrm{vol}(\phi_t(D)) = \int_D \mathrm{div}\left(\frac{\partial H}{\partial p}, -\frac{\partial H}{\partial x}\right) d\mathrm{vol}.$$

But

$$\mathrm{div}\left(\frac{\partial H}{\partial p}, -\frac{\partial H}{\partial x}\right) = \sum_{i=1}^{n} \frac{\partial^2 H}{\partial x_i \partial p_i} - \sum_{i=1}^{n} \frac{\partial^2 H}{\partial p_i \partial x_i} = 0$$

Thus

$$\mathrm{vol}(\phi_t(D)) = \text{constant}, \quad \forall\, t \geq 0$$

as expected.

The Liouville theorem has many interesting consequences:

- Assume that D is a bounded forward invariant set of the system (1.46). Then the system does not admit asymptotic stable point. Otherwise there exists an equilibrium point $(x_*, p_*) \in D$ and a compact set $U \subseteq D$, which contains both (x_*, p_*) and $\phi_t(D)$ for t sufficiently large. But in this case, the volume of $\phi_t(D)$ would be strictly smaller than that of D, a contradiction.

- *Poincaré recurrence theorem:* Assume that there exists a bounded forward invariant set $D \subseteq \mathbb{R}^{2n}$ of the system (1.46). Then for any open set $U \subseteq D$, and any $s > 0$, there exists at least one point $x \in U$ which returns to the set after some time $t \geq s$. To prove this, assume that $\phi_s(U) \cap U = \emptyset$, otherwise there's nothing to prove. Consider the sequence

  $$\phi_s(U), \phi_{2s}(U), \cdots, \phi_{ks}(U), \cdots$$

  Since $\phi_{js}(U)$ has the same volume for all $j \geq 1$, there must exists some integers $k > j$, such that

  $$\phi_{js}(U) \cap \phi_{ks}(U) \neq \emptyset$$

  otherwise, the above sequence generates infinite volume inside the set D, which is impossible. Thus $U \cap \phi_{(k-j)s}(U) \neq \emptyset$ since as claimed.

[4] The transport equation: consider a system $\dot{x} = f(x)$, and let $\phi_t : \mathbb{R}^n \to \mathbb{R}^n$ be its flow, then for any bounded set $D \subseteq \mathbb{R}^n$,

$$\frac{d}{dt}\mathrm{vol}(\phi_t(D)) = \int_{\phi_t(D)} \mathrm{div} f \, \mathrm{d}x.$$

1.2.4 Viscosity solution of HJB equation

Our next aim is to show that the value function defined in (1.31) is a *viscosity solution* of the HJB equation (1.37).

Define the set of super-differentials of a function $g : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ at x as

$$D^+ g(x) =: \left\{ p \in (\mathbb{R}^n)^* : \limsup_{y \to x} \frac{g(y) - g(x) - p(y-x)}{|y-x|} \leq 0 \right\}$$

and the set of sub-differentials at x as

$$D^- g(x) =: \left\{ p \in (\mathbb{R}^n)^* : \liminf_{y \to x} \frac{g(y) - g(x) - p(y-x)}{|y-x|} \geq 0 \right\}$$

As shown in the following figures.

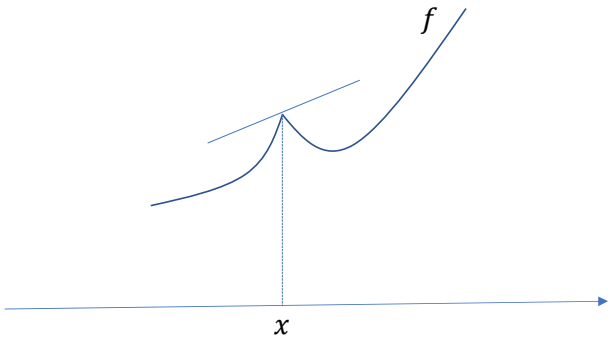

Figure 1.3: Super-differential.

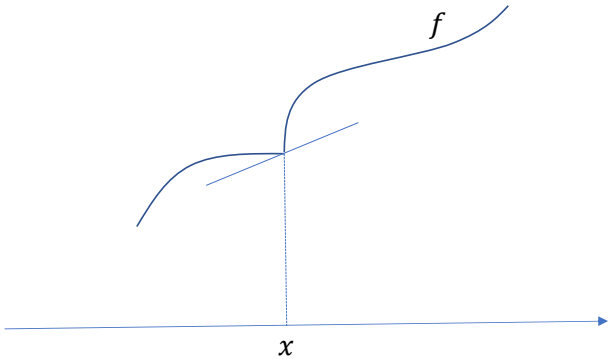

Figure 1.4: Sub-differential.

The crucial characterization of super- and sub-differentials for us is the following.

Lemma 1.3. *Let $g \in C(\Omega)$. Then*

1. $p \in D^+ g(x)$ iff there exists a function $\varphi \in C^1(\Omega)$ such that $\nabla\varphi(x) = p$ and $g - \varphi$ has a local maximum at x;

2. $p \in D^- g(x)$ iff there exists a function $\varphi \in C^1(\Omega)$ such that $\nabla\varphi(x) = p$ and $g - \varphi$ has a local minimum at x.

The proof of this lemma is easy and hence omitted. Notice that the mere regularity assumption on g is only continuity! Now we are ready to give the definition of the celebrated *viscosity solution*. One should keep in mind that there is no differentiability assumption on the solution.

Definition 1.1. Consider the first order PDE (1.41), in which F is continuous. A function $g \in C(\Omega)$ is a *viscosity sub-solution* of the PDE if

$$F(x, g(x), p) \le 0, \quad \forall x \in \Omega, \ \ p \in D^+ g(x).$$

It is a *viscosity super-solution* if

$$F(x, g(x), p) \ge 0, \quad \forall x \in \Omega, \ \ p \in D^- g(x).$$

It is a *viscosity solution* if it is both a viscosity supersolution and a viscosity subsolution.

Due to Lemma 1.3, g is a viscosity sub-solution if, for each $\varphi \in C^1(\Omega)$ such that $u - \varphi$ has a local maximum at x, there holds

$$F(x, g(x), \nabla\varphi(x)) \le 0$$

and it is a viscosity super-solution if, for each $\varphi \in C^1(\Omega)$ such that $u - \varphi$ has a local minimum at x, there holds

$$F(x, g(x), \nabla\varphi(x)) \ge 0.$$

Theorem 1.1. *Consider the system $\dot{x} = f(x,u)$ with $x \in \mathbb{R}^n$ and $u \in U \subset \mathbb{R}^m$ compact. Let $V(s,y)$ be the value function defined as (1.31). Suppose that there exists a constant $C > 0$, such that*

$$|f(x,u)|, |L(x,u)|, |\varphi(x)| < C$$
$$|f(x,u) - f(y,u)|, |\varphi(x) - \varphi(y)|, |L(x,u) - L(y,u)| < C|x - y|$$

for all $x \in \mathbb{R}^n$ and $u \in U$. Then V is the unique viscosity solution of the HJB equation

$$-V_t + \sup_{u \in U} H(x, u, -V_x) = 0, \quad (t,x) \in (0,T) \times \mathbb{R}^n$$

with boundary condition $V(T,x) = \varphi(x)$.

Proof. (We follow [6].) Let $\gamma \in C^1((0,T) \times \mathbb{R}^n)$. We need to show

1) If $V - \gamma$ attains a local maximum at $(t_0, x_0) \in (0,T) \times \mathbb{R}^n$, then

$$-\gamma_t(t_0,x_0) + \sup_{u \in U}\{-\nabla\gamma(t_0,x_0) f(x_0,u) - L(x_0,u)\} \le 0$$

or

$$\gamma_t(t_0,x_0) + \inf_{u \in U}\{\nabla\gamma(t_0,x_0) f(x_0,u) + L(x_0,u)\} \ge 0 \tag{1.47}$$

2) If $V - \gamma$ attains a local minimum at $(t_0, x_0) \in (0,T) \times \mathbb{R}^n$, then

$$-\gamma_t(t_0,x_0) + \sup_{u \in U}\{-\nabla\gamma(t_0,x_0) f(x_0,u) - L(x_0,u)\} \ge 0$$

or

$$\gamma_t(t_0,x_0)+\inf_{u\in U}\{\nabla\gamma(t_0,x_0)f(x_0,u)+L(x_0,u)\}\le 0. \tag{1.48}$$

To prove 1), assume that $V(t_0,x_0)=\gamma(t_0,x_0)$ and $V(t,x)\le\gamma(t,x)$ for all t,x. If (1.47) is not true, then there exist $\omega\in U$, $\theta>0$ such that

$$\gamma_t(t_0,x_0)+\inf_{u\in U}\{\nabla\gamma(t_0,x_0)f(x_0,u)+L(x_0,u)\}<-\theta.$$

By continuity, this inequality implies

$$\gamma_t(t,x)+\{\nabla\gamma(t,x)f(x,\omega)+L(x,\omega)\}<-\theta-L(x,\omega) \tag{1.49}$$

when

$$|t-t_0|<\delta,\ \ |x-x_0|<\delta,$$

for some $\delta>0$. Call $x(t):=x(t;t_0,x_0,\omega)$ the solution to

$$\dot{x}=f(x(t),\omega),\ \ x(t_0)=x_0.$$

We then have

$$\begin{aligned}V(t_0+\delta,x(t_0+\delta))-V(t_0,x_0)&\le\gamma(t_0+\delta,x(t_0+\delta))-\gamma(t_0,x_0)\\&=\int_{t_0}^{t_0+\delta}\frac{d}{dt}\gamma(t,x(t))dt\\&=\int_{t_0}^{t_0+\delta}\left\{\gamma_t(t,x(t))+\nabla\gamma(t,x(t))f(x(t),\omega)\right\}dt\\&\le-\int_{t_0}^{t_0+\delta}L(x(t),\omega)dt-\delta\theta.\ \ \text{(due to (1.49)).}\end{aligned}$$

On the other hand, by the definition of value function,

$$V(t_0+\delta,x(t_0+\delta))-V(t_0,x_0)\ge\int_{t_0}^{t_0+\delta}L(x(t),\omega)dt$$

which induces a contradiction. Thus $V(t,y)$ is indeed a viscosity sub-solution. Part 2) can be proved similarly.

To prove the uniqueness, one needs more effort. Interesting readers are referred to [6, Theorem 8.5.3].

□

Relation to stochastic optimal control

Let us consider two systems

$$\begin{aligned}&\mathrm{S}_1: dx(t)=f(x(t),u(t))dt\\&\mathrm{S}_2: dx(t)=f(x(t),u(t))dt+\sqrt{2\varepsilon}dB_t\end{aligned}$$

where $t\mapsto B_t$ is a standard Brownian motion. Note that S_2 is obtained by adding a stochastic term $\sqrt{2\varepsilon}dB_t$ on S_1.

Consider the cost function for the two systems

$$J_1 = \int_0^T L(x(t), u(t))\,dt + \varphi(x(T)), \;\; x(t) \text{ solves } S_1$$
$$J_2 = E\left[\int_0^T L(t, x(t), u(t))\,dt + \varphi(x(T))\right], \;\; x(t) \text{ solves } S_2$$

respectively.

The HJB for the two systems are

$$0 = V_t + \inf_u \left(\frac{\partial V(t,x)}{\partial x} f(t,x,u) + L(t,x,u) \right) \tag{1.50}$$
$$0 = W_t + \inf_u \left(\frac{\partial W(t,x)}{\partial x} f(t,x,u) + L(t,x,u) \right) + \varepsilon \frac{\partial^2 W(x,t)}{\partial x^2} \tag{1.51}$$

We observe that the stochastic HJB can be obtained from the deterministic HJB by adding the term $\varepsilon \Delta W$. It is reasonable to expect that when $\varepsilon \to 0$, W^ε (the solution to (1.51) with a given ε) converges to V in certain sense (in fact, uniformly) since the term $\varepsilon \Delta W^\varepsilon$ vanishes as $\varepsilon \to 0$. From parabolic PDE theory, (1.51) admits smooth solutions (while (1.50) doesn't!). Thus the term $\varepsilon \Delta W$ regularizes the HJB (1.50)). Since the convergence of W^ϵ is uniform, V should be continuous. One can show that this V is indeed the viscosity solution. On the other hand, the construction of the viscosity solution in this section has nothing to do with the discussion here. It is indeed a more intrinsic way of construction.

1.2.5 Infinite horizon problems

Consider the time-invariant system

$$\begin{cases} \dot{x} = f(x,u) \\ x(0) = x_0 \end{cases}$$

with cost

$$J = \int_0^\infty L(x(t), u(t))\mathrm{d}t$$

where $L \geq 0$, $u(t) \in U \subseteq \mathbb{R}^m$ for all $t \geq 0$. It is easy to notice that the value function in this case is time independent and thus can be written as $J^*(x)$. Further more, the HJB equation reads

$$\sup_{u \in U} H(x, u, -V_x) = 0$$

where $H(x,u,p) = p^\top f(x,u) - L(x,u)$, or equivalently

$$\inf_{u \in U} \{V_x f(x,u) + L(x,u)\} = 0. \tag{1.52}$$

In the LQR setting, for the system

$$\dot{x} = Ax + Bu \tag{1.53}$$

and cost

$$J = \int_0^\infty x^\top Q x + u^\top R u \mathrm{d}t \tag{1.54}$$

the HJB equation (1.52) reads

$$\inf_{u \in U} \{V_x(Ax + Bu) + x^\top Q x + u^\top R u\} = 0$$

As before, choose $V = x^\top Px$, then the above formula turns into $\inf_{u\in U}\{2x^\top P(Ax+Bu)+x^\top Qx+u^\top Ru\} = 0$. The minimum on the left hand side is achieved at

$$u^* = -R^{-1}B^\top Px$$

with minimum zero if

$$A^\top P + PA + Q - PBR^{-1}B^\top P = 0. \tag{1.55}$$

This equation in P is called *algebraic Riccati equation (ARE).*

Proposition 1.11. *Consider the LTI system* (1.53) *and cost function* (1.54) *with* $Q \geq 0$, $R > 0$. *Assume* (A, B) *is controllable,* (A, C) *is observable, where* C *is full row rank satisfying* $C^\top C = Q$. *Then the ARE has a unique symmetric solution* P *which is positive definite. Further more, the optimal control is given by a static state feedback* $u = -R^{-1}B^\top Px$ *and the optimal cost is* $x_0^\top Px_0$.

Proof. The proof of this proposition is essentially the same as the discrete time case and is thus left as an exercise. □

CHAPTER 2

MAXIMUM PRINCIPLE

In Chapter 1, we studied optimal control via dynamic programming. The are some notable features of this method.

- It is applicable to various types of problems, discrete time as well as continuous time, finite time horizon or infinite time horizon, deterministic or stochastic.
- Although the optimal control problem formulations are somewhat different, the key element used to derive the optimal controls is the same, i.e., Bellman's principle of optimality, a principle which is simple, intuitive but powerful.
- Dynamic programming provides not only necessary conditions, it also provides sufficient conditions under some mild assumptions.
- On the other hand, there are also some issues which haven't been well addressed. For example, in dynamic programming, the task is finally reduced to solving the Bellman equation (for discrete time systems), or the HJB equation (for continuous time systems). But solving these equations often runs into a generic issue: the curse of dimensionality. Even worse, for HJB equations, the existence of (classical) solutions is a subtle issue. One needs to resort to very advanced techniques from PDE theory, e.g., viscosity solution, in order to have conclusions on the existence and regularities of the solutions.

In this chapter, we are going to study a totally different approach of optimal control, which has its origin in calculus of variation. A salient feature of this approach is that it does not involve solving partial differential equations! It is hard to explain how powerful and this approach is at the current stage. We will leave the discussions to the end of this chapter.

2.1 Calculus of variation

Remember that in the beginning of this course, we mentioned one basic methodology in optimal control: fix an optimal policy and then study the property of this optimal policy. The calculus of variation adopts the same methodology, but it goes one step further. Consider an optimal control problem with input u and cost functional $J(u)$. The calculus of variation works as this:

1) fix an optimal policy u;
2) adjust slightly the optimal policy, representing by a scalar parameter ϵ: u_ϵ, with the optimal policy corresponding to $\epsilon = 0$;
3) by definition, the optimal policy should minimize the one-parameter cost functionals $J(u_\epsilon)$. Thus, if J_ϵ is differentiable w.r.t ϵ, there must hold $\left.\frac{dJ_\epsilon}{d\epsilon}\right|_{\epsilon=0} = 0$.

At the beginning, one may think that a variation using only a scalar parameter is not very useful, after all, in optimal control, the optimal policy lies in certain function space which is usually infinite dimensional. Thus it seems that one can only obtain very limited information about the optimal policy. But this conclusion is based on the fact that we use only generic variations. Later we will realize that this is not the case. In fact, by cleverly choosing some special class of variations, one may obtain very rich information of the optimal policy. It is even not rare to see that the information derived from variation is also sufficient to guarantee optimality. In optimal control, such class of variations is the "needle variations", which lie in the heart of maximum principle.

2.1.1 Motivating example: principle of least action

Assume that $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a non-negative continuously differentiable function. Here we call L the *Lagrangian*, named after the mathematician Lagrange, who laid the foundation of analytic mechanics. It is custom to write

$$L = L(q, \dot{q}) = L(q_1, \cdots, q_n, \dot{q}_1, \cdots, \dot{q}_n)$$

in which $\dot{q}_i$ is only an independent variable rather than the derivative of q. Let x, y be two points in $\mathbb{R}^n$ and define the *action* on the interval $[0, T]$ as

$$\mathscr{A}(q) = \int_0^T L(q(t), \dot{q}(t))\mathrm{d}t \tag{2.1}$$

(this time $\dot{q}$ is the time derivative of q! I have to admit that the notation is a bit misleading but it has been widely adopted) where $q : \mathbb{R} \to \mathbb{R}^n$ belongs to the set

$$\Omega = \{q \in C^2([0, T]; \mathbb{R}^n) : q(0) = x,\ q(T) = y\}.$$

The problem of least action is to find $q \in \Omega$ which minimizes the action $\mathscr{A}(q)$.

We follow the three steps in the methodology of calculus of variation.

1) Assume q is the optimal solution.

2) Choose a class of variation. For any function $y \in C^2([0, T]; \mathbb{R}^n)$ with vanishing endpoints, i.e., $y(0) = y(T) = 0$, the one-parameter family of functions $q + \epsilon y \in \Omega$, $\forall \epsilon \in \mathbb{R}$ constitute a variation of q.

3) Since L is differentiable, we can take the derivatives:

$$\begin{aligned}
&\frac{d}{d\epsilon}\bigg|_{\epsilon=0}\int_0^T L(q(t)+\epsilon y(t),\dot{q}(t)+\epsilon\dot{y}(t))\mathrm{d}t \\
&=\int_0^T \frac{\partial L}{\partial q}(q(t),\dot{q}(t))y(t)+\frac{\partial L}{\partial \dot{q}}(q(t),\dot{q}(t))\dot{y}(t)\mathrm{d}t \\
&=\int_0^T \frac{\partial L}{\partial q}(q(t),\dot{q}(t))y(t)\mathrm{d}t+\frac{\partial L}{\partial \dot{q}}(q(t),\dot{q}(t))y(t)\bigg|_0^T-\int_0^T \frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}}(q(t),\dot{q}(t))\right]y(t)\mathrm{d}t \\
&=\int_0^T \left\{\frac{\partial L}{\partial q}(q(t),\dot{q}(t))-\frac{d}{dt}\left[\frac{\partial L}{\partial \dot{q}}(q(t),\dot{q}(t))\right]\right\}y(t)\mathrm{d}t
\end{aligned}$$

where we have used integration by parts in the third line. The last line should vanish for all smooth y with compact support in $(0,T)$. It is then readily checked that the term in the brace also vanishes (*fundamental lemma)*, i.e.,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}(q(t),\dot{q}(t))\right)-\frac{\partial L}{\partial q}(q(t),\dot{q}(t))=0,\quad \forall t\in[0,T], \tag{2.2}$$

or briefly

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}-\frac{\partial L}{\partial q}=0. \tag{2.3}$$

The system of equations (2.3) really mean (2.2).

Equation (2.3) is called the Lagrangian equation. If one expand (2.3), then (2.3) is easily seen to be a second-order ordinary differential equations.

In mechanics, the Lagrangian L is defined as the difference between the total kinetic energy and potential:

$$L(q,\dot{q})=\frac{1}{2}\dot{q}^\top M(q)\dot{q}-V(q).$$

The *principle of least action* in mechanics states that

> The path taken by the system between times t_1 and t_2 and configurations q_1 and q_2 is the one for which the action is optimal.

Thus, the mechanical systems evolve according to the Lagrangian equation (2.3). If we expand (2.3), then it will look like

$$M(q)\ddot{q}+C(q,\dot{q})\dot{q}=-\nabla V(q)$$

where $C(q,\dot{q})_{ij}$ corresponds to *Coriolis and centrifulal forces.*

An important property of the Lagrangian equation for mechanical systems is energy conservation. Indeed

$$\begin{aligned}
\frac{dL}{dt}&=\frac{\partial L}{\partial q}\dot{q}+\frac{\partial L}{\partial \dot{q}}\ddot{q}=\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}+\frac{\partial L}{\partial \dot{q}}\ddot{q},\ \text{(by (2.3))} \\
&=\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\dot{q}\right)=\frac{d}{dt}(\dot{q}^\top M(q)\dot{q}) \\
&=\frac{d}{dt}(2L+2V)
\end{aligned}$$

from which it follows that

$$\frac{d}{dt}(L+2V)=0.$$

That is, the quantity $L+2V=:E$, the sum of kinetic energy and potential energy, is constant during the evolution of the system.

Consider a coordinate change $p=\frac{\partial L}{\partial \dot{q}}$, called *canonical transform* and define the *Hamiltonian*

$$H(q,p)=p^\top \dot{q}-L(q,\dot{q})$$

in which $\dot{q}$ is understood as a function of q and p. Differentiating H w.r.t. q and p, we get

$$\frac{\partial H}{\partial q}=\frac{\partial \dot{q}}{\partial q}p-\frac{\partial L}{\partial q}-\frac{\partial \dot{q}}{\partial q}\frac{\partial L}{\partial \dot{q}}=-\frac{\partial L}{\partial q}=-\frac{d}{dt}\frac{\partial L}{\partial \dot{q}}=-\dot{p}$$

and

$$\frac{\partial H}{\partial p}=\dot{q}+\frac{\partial \dot{q}}{\partial p}p-\frac{\partial \dot{q}}{\partial p}\frac{\partial L}{\partial \dot{q}}=\dot{q}.$$

Thus along the system, we obtain again the Hamiltonian equation (c.f. (1.46))

$$\begin{aligned}\dot{q}&=\frac{\partial H}{\partial p}(q,p)\\ \dot{p}&=-\frac{\partial H}{\partial q}(q,p)\end{aligned}\tag{2.4}$$

Notice that for mechanical systems, the Hamiltonian H is simply $E=\frac{1}{2}\dot{q}^\top M(q)\dot{q}+V(q)$, i.e., the total mechanical energy of the system, which we have shown to be a constant. The Hamiltonian equation (2.4) is another justification of this fact.

The canonical transform $p=\frac{\partial L}{\partial \dot{q}}$ and the definition of Hamiltonian seem a bit mysterious. It has an interesting interpretation by the so called *Legendre* transform. Given a function $f:\mathbb{R}^n\to\mathbb{R}$, the Legendre transform of f is a mapping $f\mapsto f^*$ defined by

$$f^*(x^*)=\sup_x\{x^\top x^*-f(x)\}.$$

Replace $f(x)$ by $L(q,\dot{q})$ by viewing $\dot{q}$ as the independent variable while keeping q constant, we get

$$L^*(q,p)=\sup_{\dot{q}}\{p^\top \dot{q}-L(q,\dot{q})\}.$$

The supremum in the above formula is achieved at the point such that $p=\frac{\partial L}{\partial \dot{q}}$, which is the canonical transform. Thus we see $H=L^*$. Recall that the Legendre transform is involutive when f is convex. It follows that $L=H^*$ if L is convex in $\dot{q}$, which is true for mechanical systems.

2.1.2 Euler-Lagrangian equation

In this section, we use the Euler-Lagrangian equation to solve some classical problems in calculus of variation.

Brachistochrone problem

The most well-known example in calculus of variation is probably the *problem of Brachistochrone* (problem of minimal time) which is stated as follows. Consider a bead which slides down a frictionless wire that connects two fixed points under the influence of gravity. The wire is kept in the vertical plane containing the two endpoints. The objective is find the optimal shape of the wire such that the travel time is minimal. Due to the conservation of energy

$$\frac{1}{2}mv^2 = mgy$$

thus $\dot{x} = v_x = \frac{\sqrt{gy}}{\sqrt{1+(y')^2}}$ and the travel time is

$$\int_0^a \sqrt{\frac{1+(y'(x))^2}{gy(x)}}\mathrm{d}x$$

Thus the the Lagrangian for this problem is $L(y,y') = \sqrt{\frac{1+(y')^2}{gy}}$ and the Lagrangian equation reads

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{gy(1+(y')^2}}\right) = -\frac{1}{2}\sqrt{\frac{1+(y')^2}{gy^3}}$$

or

$$2yy'' + (y')^2 + 1 = 0$$

after simplification.

Riemannian geodesic

Remember that in Euclidean spaces, the length of a piecewise smooth curve $\gamma : [0,1] \to \mathbb{R}^n$ is the integral

$$\mathscr{A}(\gamma) = \int_0^1 |\gamma'(s)|\mathrm{d}s$$

and the distance between two fixed points is defined as the minimum of $\ell(\gamma)$ when γ runs over all piecewise smooth curves. The curve that minimizes the length between the two points is called the *geodesic* (may not be unique) between the them.

To measure distance on a curved space, e.g., sphere, torus, we follow the same spirit. More precisely, the distance between two points is the minimum length of curves joining the two points. The only issue is how to define $|\gamma'(s)|$, i.e., the norm of the velocity vector of the curve. In Riemannian geometry, the norm of the velocity is defined as the square root of an inner product $|\gamma'(s)| = \sqrt{\langle \gamma'(s), \gamma'(s)\rangle}$. For example, on a Euclidean space, define a smooth positive definite function $G : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, and claim that $|\gamma'(s)| = \sqrt{\gamma'(s)^\top G(\gamma(s))\gamma'(s)}$. The function G is called a *Riemannian metric* on the space $\mathbb{R}^n$. The very case $G(x) \equiv I$ corresponds to the standard Euclidean metric. Let us derive the Euler-Lagrangian equation for the geodesic.

Fix two points $x, y \in \mathbb{R}^n$ and a smooth (for simplicity, we remove the generality of piecewise smoothness) curve $\gamma : [0,1] \to \mathbb{R}^n$ joining x and y, i.e., $\gamma(0) = x$, $\gamma(1) = y$. We are to minimize the action

$$\mathscr{A}(\gamma) = \int_0^1 L(\gamma(s), \gamma'(s))\mathrm{d}s$$

where $L(q,\dot{q}) = |\dot{q}| = \sqrt{\dot{q}^\top G(q)\dot{q}} = \sqrt{\sum_{i,j=1}^n g_{ij}(q)\dot{q}_i\dot{q}_j}$. Let $\tau(t) = \int_0^t L(\gamma(s),\gamma'(s))\mathrm{d}s$, then $\frac{d\tau}{dt} = L$. This transform will largely simplify our calculation. We calculate (using Einstein summation notation):

$$\frac{\partial L}{\partial q_k} = \frac{1}{2L}\frac{\partial g_{ij}}{\partial q_k}\frac{dq_i}{dt}\frac{dq_j}{dt} = \frac{L}{2}\frac{\partial g_{ij}}{\partial q_k}\frac{dq_i}{d\tau}\frac{dq_j}{d\tau}$$

and

$$\frac{\partial L}{\partial \dot{q}_k} = \frac{1}{L}g_{ik}\frac{dq_i}{dt} = g_{ik}\frac{dq_i}{d\tau},$$

then

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_k} = L\frac{d}{d\tau}\left(g_{ik}\frac{dq_i}{d\tau}\right) = L\left(\frac{1}{2}\frac{\partial g_{ik}}{\partial q_j}\frac{dq_j}{d\tau}\frac{dq_i}{d\tau} + \frac{1}{2}\frac{\partial g_{kj}}{\partial q_i}\frac{dq_i}{d\tau}\frac{dq_j}{d\tau} + g_{ik}\frac{d^2q_i}{d\tau^2}\right)$$

Combining those equations, we get

$$\frac{d^2q_k}{d\tau^2} + \Gamma^k_{ij}\frac{dq_i}{d\tau}\frac{dq_j}{d\tau} = 0$$

in which

$$\Gamma^k_{ij} = \frac{1}{2}g^{kr}\left(\frac{\partial g_{ir}}{\partial q_j} + \frac{\partial g_{rj}}{\partial q_i} - \frac{\partial g_{ij}}{\partial q_r}\right).$$

The coefficients Γ^k_{ij} are called the *Christoffel symbols*.

If G is constant everywhere, then $\Gamma^k_{ij} = 0$, thus geodesics are straight lines. Let us consider a non-trivial example.

Example 2.1 (Poincaré half upper plane model)**.** Consider the upper half plane $\{(x,y) \in \mathbb{R}^2 : y > 0\}$ with the Riemannian metric

$$G = \begin{bmatrix} \frac{1}{y^2} & 0 \\ 0 & \frac{1}{y^2} \end{bmatrix}.$$

There are only four non-zero Christoffel symbols, i.e., $\Gamma^1_{12} = \Gamma^1_{21} = -\frac{1}{y}$, $\Gamma^2_{11} = \frac{1}{y}$, $\Gamma^2_{22} = -\frac{1}{y}$. Thus the geodesic equation reads

$$\ddot{x} - \frac{2}{y}\dot{x}\dot{y} = 0$$
$$\ddot{y} + \frac{1}{y}(\dot{x}^2 - \dot{y}^2) = 0$$

from which one can verify that $\dot{x} = ay^2$ and $\dot{x}^2 + \dot{y}^2 = by^2$ for some constants $a, b > 0$. Then $\left(\frac{dy}{dx}\right)^2 = \left(\frac{\dot{y}}{\dot{x}}\right)^2 = \frac{by^2 - a^2y^4}{a^2y^4} = \frac{b}{a^2y^2} - 1$. Therefore $(x-c)^2 + y^2 = b/a^2$ for some c. That is, geodesics are parts of half circles.

Multi-dimension EL equation and minimal surface problem

Although we derived the Euler-Lagrangian equation under the assumption that the action is an integration over a scalar variable, it is straightforward to extend to multi-dimensional variable. In that case, the Euler-Lagrangian equation will naturally become partial differential equations.

Instead of considering the general case, we study a specific problem, i.e., the well-known minimal surface problem. Consider a surface $D \ni (x,y) \mapsto (x,y,u(x,y)) \in \mathbb{R}^3$, whose surface is calculated as

$$\mathscr{A}(u) = \int_D \sqrt{1 + u_x^2 + u_y^2}\mathrm{d}x\mathrm{d}y$$

with boundary condition $u = g$ on ∂D.

For any smooth function $\psi \in C_c^\infty(D)$, $u_\epsilon = u + \epsilon\psi$ forms a variation of u. Now

$$\mathscr{A}(u_\epsilon) = \int_D \sqrt{1 + (u_x + \epsilon\psi_x)^2 + (u_y + \epsilon\psi_y)^2)}\mathrm{d}x\mathrm{d}y$$

and

$$\frac{d}{d\epsilon}\mathscr{A}(u_\epsilon)_{\epsilon=0} = \int_D \frac{u_x\psi_x + u_y\psi_y}{\sqrt{1+u_x^2+u_y^2}}\mathrm{d}x\mathrm{d}y = 0$$

Integrating by parts and noticing that ψ has compact support in D, we get

$$\begin{aligned}\int_D \frac{u_x\psi_x + u_y\psi_y}{\sqrt{1+u_x^2+u_y^2}}\mathrm{d}x\mathrm{d}y &= -\int_D \psi\left\{\frac{d}{dx}\left(\frac{u_x}{\sqrt{1+u_x^2+u_y^2}}\right) + \frac{d}{dy}\left(\frac{u_y}{\sqrt{1+u_x^2+u_y^2}}\right)\right\}\mathrm{d}x\mathrm{d}y \\ &= -\int_D \psi \frac{u_{xx}(1+u_y^2) - 2u_xu_yu_{xy} + u_{yy}(1+u_x^2)}{(1+u_x^2+u_y^2)^{3/2}}\mathrm{d}x\mathrm{d}y = 0\end{aligned}$$

Invoking again the fundamental lemma, we arrive at the Euler-Lagrangian equation

$$u_{xx}(1+u_y^2) - 2u_xu_yu_{xy} + u_{yy}(1+u_x^2) = 0. \tag{2.5}$$

which is a second-order partial differential equation.

This equation (2.5) turns out to have many solutions.

Exercise. Derive the general Euler-Lagrangian equation in multi-dimension.

2.1.3 Other conditions

Legendre necessary condition

For a smooth function $f : U \subseteq \mathbb{R}^n \to \mathbb{R}$, where U is open, if f has a local minimum at a point x_*, then f must satisfy two necessary conditions, i.e., the first order condition

$$Df(x_*) = 0 \tag{2.6}$$

and the second order condition.

$$D^2 f(x_*) \geq 0 \tag{2.7}$$

The Euler-Lagrangian equation is a first order condition similar to (2.6). Here we introduce another necessary condition, namely, *Legendre necessary condition,* which is a reminiscent of the second order condition (2.7).

To derive this condition, let us first calculate the second-order variation along an optimal solution. Consider again the action (2.1) and suppose that $q(\cdot)$ is an optimal solution and $y \in C_c^\infty(0,T)$. Then

$$\begin{aligned}&\frac{\partial^2}{\partial\epsilon^2}\bigg|_{\epsilon=0}\int_0^T L(q(t)+\epsilon y(t), \dot{q}(t)+\epsilon\dot{y}(t))\mathrm{d}t \\ =&\int_0^T y^\top(D_{qq}L)y + \dot{y}^\top(D_{q\dot{q}}L)y + y^\top(D_{\dot{q}q}L)\dot{y} + \dot{y}^\top(D_{\dot{q}\dot{q}}L)\dot{y}\mathrm{d}t \\ =&\int_0^T \left[y^\top\left(D_{qq}L - \frac{d}{dt}(D_{q\dot{q}}L)\right)y + \dot{y}^\top(D_{\dot{q}\dot{q}}L)\dot{y}\right]\mathrm{d}t\end{aligned}$$

which should be non-negative. We assert that $D_{\dot{q}\dot{q}}L(q,\dot{q})$ must be semi-positive definite, i.e.,

$$D_{\dot{q}\dot{q}}L(q,\dot{q}) \geq 0 \tag{2.8}$$

along the optimal solution $(q,\dot{q})$. To see this, it is sufficient to note that there exist functions with small magnitude but with rather large derivatives; the converse is false, thus it may happen that $D_{qq}L - \frac{d}{dt}D_{q\dot{q}}L$ is non semi-positive definite (it is not even symmetric!).

Sufficient condition

When the optimal solution exists and is continuously differentiable, it necessarily satisfies the Euler-Lagrangian equation. On the other hand, the solutions to the Euler-Lagrangian equation may be minimizing, maximizing or neither. One good example to illustrate this is the geodesic problem on a sphere $S^2 \subseteq \mathbb{R}^n$. For any two points $x \neq -y$ on the sphere, there exist exactly two geodesics joining them, both satisfying the Euler-Lagrangian equation, but only one of them is minimizing – the one that does not contain two antipodal points. When the two points are exactly antipodal, then there are infinitely many geodesics joining them and all of them have the same length. In conclusion, a geodesic on the sphere is strictly minimizing if and only if the geodesic does not contain two antipodal points. It turns out that this is a general phenomenon and antipodal points on the sphere are a special case of a more general notion: *conjugate points*.

Proposition 2.1. *If $[a,b] \ni t \mapsto (q(t),\dot{q}(t))$ is a C^1 solution to the Euler-Lagrangian equation and $D_{\dot{q}\dot{q}}L > 0$ along the solution, then $q(\cdot)$ is a strict minimum of the action restricted to $[a,b]$ if $[a,b]$ contains no conjugate points of a.*

We are not going to give the precise definition of conjugate points nor are we going to prove the above result. After all, analyzing conjugate points is a daunting task and is out of the scope of this course. It is enough to remember the sphere example to be aware of such phenomenon.

2.1.4 Optimal control via calculus of variation

In Chapter 1, we studied optimal control using dynamic programming. In this subsection, we use a primary example to show how calculus of variation can be used to study optimal control. For that, we consider the system

$$\dot{x} = f(x,u)$$

with fixed initial condition x_0 and cost function

$$J = \varphi(x(T)) + \int_0^T L(x(t),u(t))\mathrm{d}t.$$

We impose no constraints on the input u. Fix a control $u_*(\cdot)$ and corresponding trajectory $x_*(\cdot)$. Let us construct a one-parameter family of variations of $u_*(\cdot)$. Let $v(\cdot)$ be another control input and consider $u_\epsilon = u_* + \epsilon v$. Then, we need to calculate $J(u_\epsilon)$. Denote by x_ϵ the trajectory of the system under the control u_ϵ, i.e., $\dot{x}_\epsilon = f(x_\epsilon,u_\epsilon)$. Then

$$J(u_\epsilon) = \varphi(x_\epsilon(T)) + \int_0^T L(x_\epsilon(t),u_\epsilon(t)\mathrm{d}t$$

To find the first-order necessary condition, we may differentiate $J(u_\epsilon)$ directly by brutal force, see the footnote[1]. But here we use a little trick to reduce our computational work.

Define a function $H(x,u,p) = p^\top f(x,u) - L(x,u)$ (the Hamiltonian!), and fix a C^1 curve $t \mapsto p(t)$, then

$$\begin{aligned} J(u_\epsilon) &= \varphi(x_\epsilon(T)) + \int_0^T L(x_\epsilon, u_\epsilon)\mathrm{d}t \\ &= \varphi(x_\epsilon(T)) + \int_0^T p^\top(t) f(x_\epsilon, u_\epsilon) - H(x_\epsilon, u_\epsilon, p)\mathrm{d}t \\ &= \varphi(x_\epsilon(T)) + \int_0^T p(t)^\top \mathrm{d}x_\epsilon(t) - H(x_\epsilon, u_\epsilon, p)\mathrm{d}t \end{aligned}$$

---

[1] For notational ease, write $X(t) := X(x_*(t), u_*(t))$ where X can be L, $\frac{\partial L}{\partial x}$, $\frac{\partial L}{\partial u}$ and so on. Now since

$$\left.\frac{\partial J(u_\epsilon)}{\partial \epsilon}\right|_0 = \frac{\partial \varphi(x_*(T))}{\partial x} \left.\frac{\partial x_\epsilon(T)}{\partial \epsilon}\right|_0 + \int_0^T \left(\frac{\partial L(t)}{\partial x} \left.\frac{\partial x_\epsilon(t)}{\partial \epsilon}\right|_0 + \frac{\partial L(t)}{\partial u} v(t) \right) \mathrm{d}t$$

we are led to calculate $\frac{\partial x_\epsilon(t)}{\partial \epsilon}$. Differentiate the relation $\dot{x}_\epsilon = f(x_\epsilon, u_\epsilon)$ w.r.t. ϵ, we get

$$\frac{d}{dt}\frac{\partial x_\epsilon}{\partial \epsilon} = \frac{\partial f}{\partial x}\frac{\partial x_\epsilon}{\partial \epsilon} + \frac{\partial f}{\partial u} v$$

Thus $\frac{\partial x_\epsilon(t)}{\partial \epsilon}$ satisfies the ODE: $\dot{z} = \frac{\partial f}{\partial x} z + \frac{\partial f}{\partial u} v$ with zero initial condition, from which it follows that $\frac{\partial x_\epsilon(t)}{\partial \epsilon} = \int_0^t \Phi(t,s)\frac{\partial f(s)}{\partial u} v(s)\mathrm{d}s$ in which $\Phi(t,s)$ is the state transition matrix of $\dot{z} = \frac{\partial f}{\partial x}(t) z$. Thus

$$\begin{aligned} \left.\frac{\partial J(u_\epsilon)}{\partial \epsilon}\right|_0 &= \int_0^T \frac{\partial \varphi(x_*(T))}{\partial x}\Phi(T,t)\frac{\partial f(t)}{\partial u} v(t)\mathrm{d}t \\ &\quad + \int_0^T \left[\frac{\partial L(t)}{\partial x}\int_0^t \Phi(t,s)\frac{\partial f(s)}{\partial u} v(s)\mathrm{d}s + \frac{\partial L(t)}{\partial u} v(t) \right] \mathrm{d}t \\ &= \int_0^T \left\{ \left(\frac{\partial \varphi(x_*(T))}{\partial x}\Phi(T,t) + \int_t^T \frac{\partial L(s)}{\partial x}\Phi(s,t)\mathrm{d}s \right) \frac{\partial f(t)}{\partial u} + \frac{\partial L(t)}{\partial u} \right\} v(t)\mathrm{d}t \end{aligned}$$

where we have applied Fubini's theorem to the last line:

$$\int_0^T \frac{\partial L(t)}{\partial x}\int_0^t \Phi(t,s)\frac{\partial f(s)}{\partial u} v(s)\mathrm{d}s\mathrm{d}t = \int_0^T \left[\left(\int_t^T \frac{\partial L(s)}{\partial x}\Phi(s,t)\mathrm{d}s \right) \frac{\partial f(t)}{\partial u} \right] v(t)\mathrm{d}t.$$

Invoking the fundamental lemma, we conclude that

$$\left(\frac{\partial \varphi(x_*(T))}{\partial x}\Phi(T,t) + \int_t^T \frac{\partial L(s)}{\partial x}\Phi(s,t)\mathrm{d}s \right) \frac{\partial f(t)}{\partial u} + \frac{\partial L(t)}{\partial u} = 0 \tag{2.9}$$

for all $t \geq 0$. Denote

$$-p(t) = \frac{\partial \varphi(x_*(T))}{\partial x}\Phi(T,t) + \int_t^T \frac{\partial L(s)}{\partial x}\Phi(s,t)\mathrm{d}s$$

then it is obvious that p (a row vector) satisfies the ODE

$$\dot{p} = -p\frac{\partial f}{\partial x} + \frac{\partial L}{\partial x}$$

with terminal condition $p(T) = -\frac{\partial \varphi(x_*(T))}{\partial x}$. The equation (2.9) now reads

$$p\frac{\partial f}{\partial u} - \frac{\partial L}{\partial u} = 0 \tag{2.10}$$

along the system

$$\begin{aligned} \dot{x} &= f \\ \dot{p} &= -p\frac{\partial f}{\partial x} + \frac{\partial L}{\partial x} \end{aligned}$$

This equation is obviously the Hamiltonian equation if we define $H(x,u,p) = pf(x,u) - L(x,u)$. The stationary condition (2.10) reads $\frac{\partial H}{\partial u} = 0$.

Denote $\eta_\epsilon(t) := \frac{\partial x_\epsilon(t)}{\partial \epsilon}$, then it is readily calculated

$$\begin{aligned}\frac{\partial J(u_\epsilon)}{\partial \epsilon} &= \varphi_x(x_\epsilon(T))\eta_\epsilon(T) + \int_0^T (p(t)^\top \mathrm{d}\eta_\epsilon(t) - H_x^\top \eta_\epsilon(t) - H_u^\top v)\mathrm{d}t \\ &= \varphi_x(x_\epsilon(T))\eta_\epsilon(T) + p(T)^\top \eta_\epsilon(T) - \int_0^T (\dot{p}(t)^\top + H_x^\top)\eta_\epsilon(t) + H_u^\top v \mathrm{d}t. \end{aligned} \tag{2.11}$$

Let p be such that $\dot{p} = -H_x^\top$ with terminal condition $p(T) = -\varphi_x^\top(x(T))$, the above evaluating at $\epsilon = 0$ results in

$$\left.\frac{\partial J(u_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = -\int_0^T H_u^\top v \mathrm{d}t$$

but then this implies $H_u^\top \equiv 0$ if $\frac{\partial J(u_\epsilon)}{\partial \epsilon}|_{\epsilon=0} = 0$ since v is arbitrary. To conclude, we have the following proposition:

Proposition 2.2. *Consider the system $\dot{x} = f(x,u)$, $x(0) = x_0$ fixed, with the cost $J = \varphi(x(T)) + \int_0^T L(x(t),u(t))\mathrm{d}t$, where f, φ and L are C^1 functions and $u \in \mathbb{R}^m$ is constraint free. Define $H(x,u,p) = p^\top f(x,u) - L(x,u)$. Then along the optimal process $(x_*(\cdot), u_*(\cdot))$, there hold*

1) the Hamiltonian equation

$$\begin{cases} \dot{x} = H_p^\top \\ \dot{p} = -H_x^\top \end{cases} \tag{2.12}$$

with initial and terminal conditions $x(0) = x_0$, $p(T) = -\varphi_x(x_*(T))$ and

2) the stationary condition

$$H_u = 0. \tag{2.13}$$

To derive a Legendre type second order necessary condition, we continue calculation based on (2.11):

$$\begin{aligned}\frac{\partial^2 J(u_\epsilon)}{\partial \epsilon^2} &= \eta_\epsilon(T)^\top \varphi_{xx}\eta_\epsilon(T) + [\varphi_x(x_\epsilon(T)) + p(T)^\top]\eta'_\epsilon(t) \\ &\quad - \int_0^T (\dot{p}(t)^\top + H_x^\top)\eta'_\epsilon(t)\mathrm{d}t - \int_0^T \begin{bmatrix}\eta_\epsilon(t)\\ v\end{bmatrix}^\top \begin{bmatrix}H_{xx} & H_{xu}\\ H_{ux} & H_{uu}\end{bmatrix}\begin{bmatrix}\eta_\epsilon(t)\\ v\end{bmatrix}\mathrm{d}t\end{aligned}$$

evaluating at $\epsilon = 0$, we get

$$\left.\frac{\partial^2 J(u_\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} = \eta_0(T)^\top \varphi_{xx}\eta_0(T) - \int_0^T \begin{bmatrix}\eta_0(t)\\ v\end{bmatrix}^\top \begin{bmatrix}H_{xx} & H_{xu}\\ H_{ux} & H_{uu}\end{bmatrix}\begin{bmatrix}\eta_0(t)\\ v\end{bmatrix}\mathrm{d}t$$

Using similar argument as in deriving the Legendre necessary condition, we can conclude that $H_{uu} \le 0$ in order to guarantee $\left.\frac{\partial^2 J(u_\epsilon)}{\partial \epsilon^2}\right|_{\epsilon=0} \ge 0$. (Remember $\dot{\eta}_0 = \frac{\partial f}{\partial x}\eta_0 + \frac{\partial f}{\partial u}v$, then $v \mapsto \eta_0$ can be seen as a low pass filter. If we choose v as some spiking signals, then the output η_0 will be kept relatively small. Thus the integrand in the above equation is indeed dominated by $v^\top H_{uu} v$).

Now the first and second conditions $H_u = 0$ and $H_{uu} \le 0$ together seem to imply that along the optimal solution at each time instant, u should maximize $H(x(t),u,p(t))$. This conjecture turns out to be correct; indeed it is the essential part of the celebrated maximum principle, whose proof is far from obvious. Our main objective in the next section is to prove this result.

Limitations of calculus of variation

In deriving the first and second order necessary condition. We have assumed that the variation can be arbitrary. This is a very restricted assumption. In practice, the admissible control set is often constrained, e.g., $|u| \le 1$. In this case, when u is at the boundary, say $u \equiv 1$, the variation cannot be arbitrary.

Another assumption we have imposed is the smoothness of the function f and L. It is quite often the case that L is not differentiable. Thus we cannot talk about H_x, H_{uu} etc. It turns out to prove the maximum principle, it is indispensable to develop some non-smooth techniques. What lies in the heart of these techniques is the so-called tent method.

2.2 The maximum principle

Consider the system model

$$\dot{x} = f(x,u), \tag{2.14}$$

with initial condition $x(0) = x_0$ and cost function

$$J(u) = \varphi(x(t_f)) + \int_0^{t_f} L(x(t),u(t))\mathrm{d}t \tag{2.15}$$

where φ, L are continuously continuously differentiable in x and $\varphi \ge 0$, $L \ge 0$. The terminal time instant t_f can be either free or fixed. The control input $u(t) \in U_t \subseteq \mathbb{R}^m$ for every $t \ge 0$. The objective is to find u which drives the initial state to a target set $S \subseteq \mathbb{R}^n$ with the minimum cost.

2.2.1 Statements of the maximum principle

Proposition 2.3. *Consider the system* (2.14) *with cost function* (2.15). *Let* $(x^*(\cdot), u^*(\cdot))$ *be the optimal process and define the Hamiltonian function* $H(x,u,p,p_0) = p^\top f(x,u) - p_0 L(x,u)$. *Then there exists a function* $p^* : [0,t_f] \to \mathbb{R}^n$ *and a constant* $p_0^* \le 0$, *satisfying* $(p_0^*, p^*(t)) \not\equiv (0,0)$ *such that*

1) $(x^*(\cdot), p^*(\cdot))$ *satisfies the canonical equation*

$$\dot{x} = H_p$$
$$\dot{p} = -H_x$$

with initial condition $x^*(0) = x_0$. *The second equation is called the costate equation, and* p *is the costate.*

2) The transversality condition holds:

$$p^*(t_f) + \varphi_x^\top(x(t_f)) \perp S$$

3) The maximum principle holds:

$$H(x^*(t), u^*(t), p^*(t), p_0^*) = \max_{u \in U_t} H(x^*(t), u, p^*(t), p_0^*) = \text{constant} \tag{2.16}$$

for all $t \in [0,t_f]$. *This constant is zero if* t_f *is free.*

Although the optimal control problem seeks for a *minimizing* control, equation (2.16) says that the optimal control should *maximize* the Hamiltonian function, thus the name *maximum principle*.

The following observations are in order:

1) Unlike the dynamic programming method, there is no partial differential equation to solve. Only ordinary differential equations, i.e., the Hamiltonian equation are present.

2) There is no smoothness assumption on $f(x,u)$, $L(x,u)$ as functions of u. This allows the maximum principle extremely useful for applications.

3) The first-order and second-order necessary conditions are replaced by the elegant maximum principle (2.16) which is only a finite dimensional maximization of a scalar function.

4) For the transversality condition, when $S=\mathbb{R}^n$, it reduces to $p^*(t_f)=-\varphi_x^\top(x(t_f))$. When S is a singleton, i.e., $x(t_f)$ is fixed, then there is no information about $p^*(t_f)$ (note also that in this case it suffices to consider cost $J=\int_0^{t_f}L(x,u)\mathrm{d}t$). When S is described by constraint $S=\{x:\psi(x)=0\}$ for some smooth, constant rank (on S) mapping ψ, then the transverse condition can be expressed as

$$p^*(t_f)+\varphi_x^\top(x(t_f))\perp\ker D\psi$$

or equivalently

$$p^*(t_f)+\varphi_x^\top(x(t_f))\in\mathrm{Im}(D\psi)^\top.$$

To see how to apply the maximum principle, we study some examples in next subsection.

2.2.2 Some examples

Dido's problem, t_f free, $x(t_f)\in S$

Suppose we have a string with fixed length. One end of the string is fixed at the origin, the other end point is to be placed somewhere on the x-axis. The task is to find the optimal shape which maximizes the area encircled by the string and the x-axis. See figure 2.1.

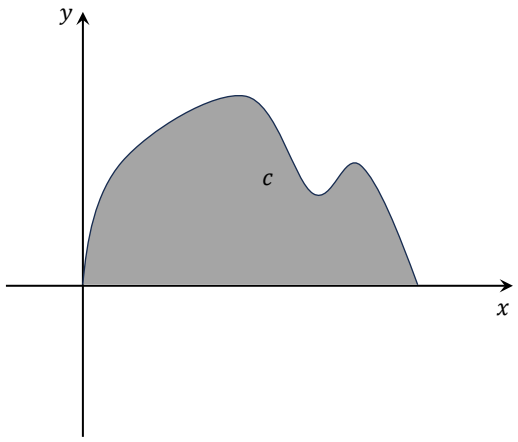

Figure 2.1: Dido's problem

Assume that the curve lies in the upper half plance (otherwise reduce to this case by reflection and translation). Let the curve be parameterized by $c:[0,t_f]\to\mathbb{R}^2$. Assume that the curve is continuously differentiable with respect to the parametrization. The the length of the curve is

$$\ell(c)=\int_c\mathrm{d}s=\int_0^{t_f}\sqrt{\dot{x}^2+\dot{y}^2}\mathrm{d}t.$$

Let D be the set enclosed by c and the x-axis. Then the area of the set D is $\Phi(c) = \int_D \mathrm{d}x \wedge dy$. Define $\alpha = \frac{1}{2}(x\mathrm{d}y - y\mathrm{d}x)$, then $\mathrm{d}x \wedge \mathrm{d}y = \mathrm{d}\alpha$. Notice that $\alpha = 0$ along the x-axis, thus by Stoke's theorem

$$\Phi(c) = \int_c \alpha.$$

Now we reformulate the problem a bit. Instead of maximize $\Phi(c)$ while fixing $\ell(c)$, we may consider minimizing $\ell(c)$ while fixing $\Phi(c)$, since the form of $\ell(c)$ seems more suitable for applying maximum principle. Introduce the dynamics

$$\dot{x} = u_1$$
$$\dot{y} = u_2$$

then the cost becomes

$$\ell(c) = \int_0^{t_f} \sqrt{u_1^2 + u_2^2}\mathrm{d}t \tag{2.17}$$

under the constraint $\Phi(c) = \text{constant}$. W.l.o.g., assume $\Phi(c) = 1$. To cope with this integration constraint, we use a small trick. Introduce another state z, satisfying $\dot{z} = \frac{1}{2}(-yu_1 + xu_2)$ with initial condition $z(0) = 0$. To see the reason behind this, we integrate the dynamics of z:

$$\begin{aligned} z(t_f) &= \int_0^{t_f} \frac{1}{2}(-y\dot{x} + x\dot{y})\mathrm{d}t \\ &= \int_0^{t_f} \frac{1}{2}(-y\mathrm{d}x + x\mathrm{d}y) \\ &= \int_c \alpha. \end{aligned}$$

Thus the problem has been shifted into the following form

$$\dot{x} = u_1$$
$$\dot{y} = u_2$$
$$\dot{z} = \frac{1}{2}(-yu_1 + xu_2)$$

with initial condition $(x(0), y(0), z(0)) = (0,0,0)$ and terminal condition $x(t_f) > 0$, $y(t_f) = 0$, $z(t_f) = 1$. This system is also known as the *Heisenberg system*. The objective is to find (u_1, u_2) that minimizes the cost (2.17).

Assume $p_0 \neq 0$. The Hamiltonian is $H = p_1u_1 + p_2u_2 + \frac{1}{2}p_3(-yu_1 + xu_2) - \sqrt{u_1^2 + u_2^2}$. The costate equation is

$$\dot{p}_1 = -\frac{1}{2}p_3u_2 = -\frac{1}{2}p_3\dot{y}$$
$$\dot{p}_2 = \frac{1}{2}p_3u_1 = \frac{1}{2}p_3\dot{x}$$
$$\dot{p}_3 = 0$$

with $p_1(t_f) = 0$ since $S = \{x(t_f) > 0, y(t_f) = 0, z(t_f) = 1\}$. Integrating the costate equation, we get: p_3 is a nonzero constant, $p_1(t) = -\frac{1}{2}p_3y(t)$ (recall that $y(t_f) = 0$), and $p_2(t) = p_2(0) + \frac{1}{2}p_3x(t)$.

To maximize H, it suffices to find $\frac{\partial H}{\partial u}$, which gives

$$p_1 - \frac{1}{2}p_3 y = \frac{u_1}{\sqrt{u_1^2+u_2^2}},$$
$$p_2 + \frac{1}{2}p_3 x = \frac{u_2}{\sqrt{u_1^2+u_2^2}}.$$

Note that this also implies

$$(x + \frac{p_2(0)}{p_3})^2 + y^2 = \frac{1}{p_3^2}.$$

Thus the optimal shape is a half circle with center $(-\frac{p_2(0)}{p_3}, 0)$ and radius $-\frac{p_2(0)}{p_3}$. Since the area of the circle is 1, it follows that $-\frac{p_2(0)}{p_3} = \sqrt{\frac{1}{\pi}}$. Thus $p_3 = \pm\sqrt{\pi}$ and $p_2(0) = \mp 1$.

Exercise 2.1. Find the explicit form of u_1 and u_2 (may not unique).

Planar elastic rod

See [7], [14].

Swtiching system

[12].

Moon lander, t_f free, $x(t_f)$ fixed

Suppose that we are to land a lunar rover on the moon. The dynamics of this model is described by

$$\ddot{y} = -g + u$$

where y is the height of the lander, $g \geq 0$ the gravitational acceleration, and u the trust, which can be up or down and is bounded $|u| \leq 1$, and $0 < g < 1$ Note that here we assume the mass of the lander is 1 (fuel loss is neglected). The initial height of the lander is $y(0) = h$ and initial velocity $\dot{y}(0) = v < 0$. In order the problem to be feasible, assume h is sufficiently large, otherwise the lander may never be able to land with zero velocity.

Find an optimal control law which minimizes the fuel consumption

$$J = \int_0^{t_f} |u| \mathrm{d}t$$

with t_f free, and which drives the system to the final state $y(t_1) = \dot{y}(t_f) = 0$.

Rewrite the system model as

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -g + u$$

with initial and terminal conditions $(x_1(0), x_2(0) = (h, v)$, $(x_1(t_f), x_2(t_f)) = (0, 0)$. The Hamiltonian is $H(x, u, p) = p_1 x_2 + p_2(-g + u) + p_0|u|$, and the costate equation

$$\dot{p}_1 = 0$$
$$\dot{p}_2 = -p_1$$

Then $p_1(t) = c_1$ and $p_2(t) = -c_1 t + c_2$ for some constants c_1 and c_2. Since there the terminal state is fixed, for the moment we don't know the terminal condition of the costate equation.

First we need to exclude abnormal extremal – left as an exercise. Henceforth let $p_0 = -1$. In this case,

$$u^*(t) = \begin{cases} -1, & p_2 < -1 \\ 0, & -1 \le p_2 < 1 \\ 1, & p_2 \ge 1 \end{cases}$$

Note that when x_1 is near zero, u must be positive, i.e., it must be in the phase $p_2 = -c_1 t + c_2 \ge 1$, for all t near t_f. This implies $c_1 < 0$ (check $c_1 = 0$ is not possible). On the other hand, since t_f is free, $H(x^*(t), u^*(t), p^*(t)) \equiv 0$. In particular, $(-c_1 t_f + c_2)(-g + 1) - 1 = 0$, from which we can solve for $t_f = \frac{1/(1-g)-c_2}{-c_1}$. Thus $c_2 < \frac{1}{1-g}$.

If $1 \le c_2 < \frac{1}{1-g}$, then $u^* \equiv 1$, then $h = \frac{v^2}{2(1-g)}$, a contradiction since h is sufficiently large.

If $c_2 < -1$, then there will be two switches: $t_1 = \frac{1+c_2}{c_1}$, $t_2 = \frac{c_2-1}{c_1}$, and

$$u^*(t) = \begin{cases} -1, & 0 \le t \le t_1 \\ 0, & t_1 < t \le t_2 \\ 1, & t_2 < t \le t_f \end{cases}.$$

Using $x_2(t_f) = 0$, we can obtain the equality

$$v + (-g - 1)t_1 - g(t_2 - t_1) + (1 - g)(t_f - t_2) = 0,$$

from which we can solve for $c_2 = \frac{c_1 v - 1}{1+g} > -1$ since $v < 0$ as assumed, a contradiction.

Thus $-1 \le c_2 < 1$, and there is only one switch at $t_s = \frac{c_2-1}{c_1}$. The corresponding optimal control is

$$u^*(t) = \begin{cases} 0, & 0 < t \le t_s \\ 1, & t_s < t \le t_f \end{cases}.$$

To find t_s, use the terminal condition $x_1(0) = x_2(0) = 0$:

$$\begin{aligned} v - g t_s + (1 - g)(t_f - t_s) &= 0 \\ v t_s - \frac{1}{2} g t_s^2 + \frac{1}{2}(1 - g)(t_f - t_s)^2 &= h \end{aligned}$$

from which we find solve for t_s, t_f and then c_1, c_2 (exercise). See Figure 2.2.

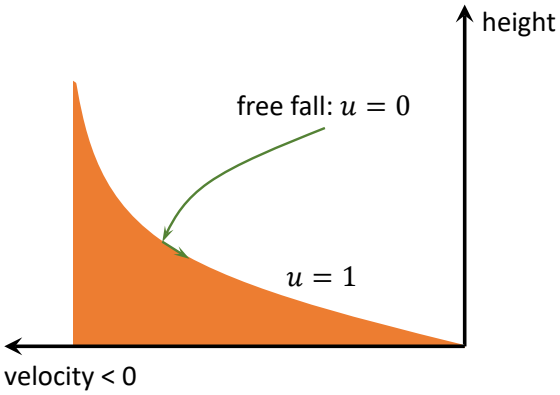

Figure 2.2: Moon lander

Insect control, t_f fixed, $x(t_f)$ free

Let $w(t)$ and $r(t)$ denote, respectively, the worker and reproductive population levels in a colony of insects, e.g. wasps. At any time t, $0 \le t \le T$ in the season the colony can devote a fraction $u(t)$ of its effort to enlarging the worker force and the remaining fraction $u(t)$ to producing reproductives. The per capita mortality rate of workers is μ and the per capita natality rate is b when full effort is put on the worker population. Assume $\mu < b$. The two populations are governed by the equations

$$\dot{w} = (bu - \mu)w$$
$$\dot{r} = c(1-u)w$$

with $(w(0), r(0) = (1,0)$, where u satisfies the constraint $0 \le u(t) \le 1$. The objective is to maximize $r(T)$ or minimize

$$J = -r(T).$$

Since $L = 0$, the Hamiltonian for this problem is $H = p_1(bu-\mu)w + p_2 c(1-u)w$. The costate equation

$$\dot{p}_1 = p_1(bu-\mu) + p_2 c(1-u)$$
$$\dot{p}_2 = 0$$

with terminal condition $p_1(T) = 0$, $p_2(T) = 1$. Thus $p_2(t) \equiv 1$ and

$$H = (p_1 b - c)wu + (c - p_1\mu)w.$$

Since $w > 0$ for all $t \ge 0$, the optimal control law is

$$u(t) = \begin{cases} 1, & p_1(t)b \ge c \\ 0, & p_1(t)b < c \end{cases}.$$

Since $p_1(T) = 0$, then near T, u should be taken as 0. Moving backward, assume t_s is the first time instance that $p_1(t_s)b = c$. Then on $[t_s, T]$,

$$\dot{p}_1 = -\mu p_1 + c$$

which results in

$$p_1(t) = e^{-\mu(t-t_s)}p_1(t_s) + \frac{c}{\mu}(1-e^{\mu(t_s-t)}) = \frac{c}{b}e^{-\mu(t-t_s)} + \frac{c}{\mu}(1-e^{\mu(t_s-t)})$$

at $t = T$,

$$0 = \frac{1}{b}e^{-\mu(T-t_s)} + \frac{1}{\mu}(1-e^{-\mu(T-t_s)})$$

from which it follows that

$$t_s = T - \frac{1}{\mu}\ln\left(1-\frac{\mu}{b}\right).$$

Continuing moving backward, the costate equation becomes

$$\dot{p}_1 = p_1(b-\mu)$$

with terminal condition $p_1(t_s) = \frac{c}{b} > 0$. Thus p_1 increases as t decreases. Hence

$$u^*(t) = \begin{cases} 1, & 0 \le t < t_s \\ 0, & t_s \le t \le T \end{cases}.$$

2.2.3 Time optimal control

Time optimal control is an important problem in engineering, which seeks for the optimal control that renders the system from current state to the target in minimal time under given constraints. The cost function for time optimal control is

$$J = t_f = \int_0^{t_f} 1 \mathrm{d}t.$$

Thus this is an optimal control problem with free t_f and $x(t_f) \in S$. In this subsection, we focus on the case when S is a singleton. The general case is essentially similar.

It is clear that the problem described above is closely related to the problem of stabilization. Loosely speaking, the system is stabilizable (to the target) if and only if $\min J < \infty$.

In this subsection, we focus on affine control systems:

$$\dot{x} = f(x) + g(x)u$$

where $u \in \mathbb{R}^m$. The constraint for u is $|u_i| \le 1$ for all $i = 1, \cdots m$. The Hamiltonian for the system is

$$H = p^\top(f(x) + g(x)u) + p_0$$

and the costate equation is

$$\dot{p} = -(f_x^\top + \sum_{i=1}^m u_i g_{ix}^\top)p.$$

Recall that $|u_i| \le 1$, then the optimal control should has the following form:

$$u_i^*(t) = \begin{cases} 1, & p^\top(t)g_i(x^*(t)) > 0 \\ -1, & p^\top(t)g_i(x^*(t)) < 0 \\ ? & p^\top(t)g_i(x^*(t)) = 0 \end{cases}.$$

Thus typically, the optimal control switches between 1 and -1, except at those time instants such that $p^\top(t)g_i(x^*(t)) = 0$. Such control is named *bang-bang control* (a control whose components are either 1

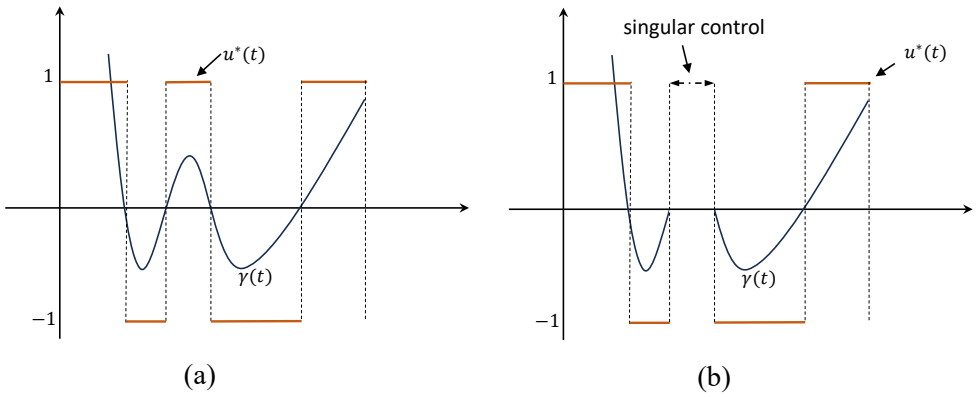

Figure 2.3: Normal control and singular control

or -1). If the function $\gamma(t) = p^\top(t) g_i(x^*(t))$ has only finite zeros, we say that the system is *normal*. If γ is zero on some interval $[t_1, t_2]$, then the optimal control on $[t_1, t_2]$ is called *singular*, and the corresponding trajectory $x^*|_{[t_1,t_2]}$ is called a *singular arc*.

Example 2.2 (Double integrator, normal system)**.** Consider a double integrator

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = u$$

with unknown initial condition (ξ, η). The objective is to drive the initial condition to the origin $x(0) = (0,0)$ in minimal time under the constraint $|u| \le 1$. The Hamiltonian is $H = p_1 x_2 + p_2 u + p_0$. The costate equation reads

$$\dot{p}_1 = 0$$
$$\dot{p}_2 = -p_1$$

Thus $p_1 = c_1$, $p_2 = -c_1 t + c_2$ for some constants c_1 and c_2. Thus the input u switches at most once at t_s, when $-c_1 t_s + c_2 = 0$. Singular control exists only if $c_1 = c_2 = 0$, which is not possible. Thus the optimal control with switch should have the following form

$$u^*(t) = \begin{cases} -1, & -c_1 t + c_2 < 0 \\ 1, & -c_1 t + c_2 > 0 \end{cases}$$

If $c_1 < 0$, then $c_2 < 0$ and $u^*|_{[0,t_s]} = -1$ and $u^*|_{(t_s,t_f)} = 1$. Under this control, we can calculate

$$x_2(t) = t - t_f$$
$$x_1(t) = \frac{1}{2}(t - t_f)^2$$

for $t \in (t_s, t_f]$ and

$$x_2(t) = \eta - t$$
$$x_1(t) = \xi + \eta t - \frac{1}{2} t^2$$

for $t \in [0, t_s]$. By continuity of the state trajectory, we know

$$\eta - t_s = t_s - t_f$$
$$\xi + \eta t_s - \frac{1}{2} t_s^2 = \frac{1}{2}(t_s - t_f)^2$$

from which we can solve for $t_s = \eta + \sqrt{\xi + \frac{1}{2}\eta^2}$, $t_f = \eta + 2\sqrt{\xi + \frac{1}{2}\eta^2}$ provided that $\xi + \frac{1}{2}\eta^2 > 0$, $\eta + \sqrt{\xi + \frac{1}{2}\eta^2} > 0$, or

$$\xi + \frac{1}{2}\eta^2 > 0, \text{ if } \eta > 0$$
$$\xi - \frac{1}{2}\eta^2 > 0, \text{ if } \eta \leq 0$$

and

$$c_1 = \frac{-1}{\sqrt{\xi + \frac{1}{2}\eta^2}}, \quad c_2 = -\left(1 + \frac{\eta}{\sqrt{\xi + \frac{1}{2}\eta^2}}\right)$$

(here $p_0 \neq 0$ otherwise $c_1 = c_2 = 0$). The situation for $c_1 > 0$ can be discussed in the same fashion.

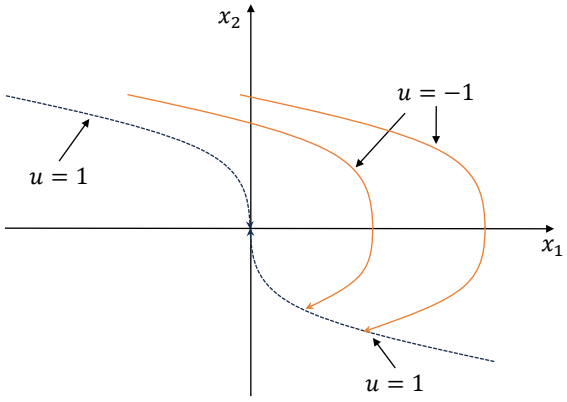

Figure 2.4: Minimal time double integrator

Exercise 2.2. Finish the case for $c_1 > 0$.

Example 2.3 (Singular control)**.** Consider the time optimal control of the system

$$\dot{x}_1 = x_2^2 - 1$$
$$\dot{x}_2 = u$$

with initial condition $x = (1,0)$ and target $x(t_f) = (0,0)$, $u \in [-1,1]$. We argue that the unique optimal control is $u^*(t) \equiv 0$. Indeed, when $u^* \equiv 0$, x_1 decreases to 0 in time 1, while keeping $x_2(t) \equiv 0$. If u is not zero, then the decay of x_1 should be slower and the time to go to 0 is longer. Thus we have a singular control $u^* \equiv 0$ on $[0,1]$.

Singular optimal controls are generally hard to compute. It is natural to ask when can we exclude the existence of singular controls beforehand. To that end, let's reexamine the singularity condition

$$\gamma_i(t) = p^\top(t) g_i(x^*(t)) = 0, \quad \forall t \in [t_1, t_2]. \tag{2.18}$$

For simplicity, consider single input system, i.e., $m = 1$ and $g_i = g$. Differentiate $\gamma_i(t)$ again,

$$\begin{aligned}
\dot{\gamma}_i &= \dot{p}^\top g + p^\top \dot{g} \\
&= -p^\top (f_x + u g_x) g + p^\top g_x (f + ug) \\
&= p^\top (g_x f - f_x g) \\
&= p^\top [f, g]
\end{aligned}$$

Denote $\mathrm{ad}^i(f)(g) := [\mathrm{ad}^{i-1}(f), g]$ and $\mathrm{ad}^1(f)(g) = [f, g]$, then it is easily seen that

$$\frac{d^k \gamma_i}{dt^k} = p^\top \mathrm{ad}^k(f) g.$$

Now, the singularity condition for single input system would imply

$$p \perp \mathrm{span}\{\mathrm{ad}^1(f)g, \mathrm{ad}^2(f)g, \cdots, \}, \quad \forall t \in [t_1, t_2].$$

Thus a sufficient condition which guarantees no existence of singular control is that the

$$\mathrm{rank}\{\mathrm{span}\{\mathrm{ad}^1(f)g, \mathrm{ad}^2(f)g, \cdots, \}\} = n, \quad \forall x \tag{2.19}$$

since this implies $p = 0$, a contradiction.

For single input linear system, the rank condition (2.19) reads

$$\mathrm{rank}\{b, Ab, \cdots\} = n$$

which is the controllability condition.

Although for single input LTI systems, controllability is sufficient to exclude singular controls, this is not true for multi-input LTI systems as we will see in next subsection.

2.2.4 LQR with constraints

In Chapter 1, we studied LQR under several circumstances, all of which didn't consider input constraints. In this subsection, we study optimal LQR controller under input constraints.

Stabilization via time optimal control

Consider the LTI system

$$\dot{x} = Ax + Bu \tag{2.20}$$

with initial condition x_0. The objective is to find an optimal control u which drives x_0 to the origin under the constraint $|u_i| \le 1$ for all i in minimum time. The system is assumed to be controllable.

Proposition 2.4. *Let $B = [b_1, \cdots, b_m]$ be full rank and assume that the system has no abnormal extremals. Then the time optimal control problem of the system* (2.20) *admits no singular arcs if and only if*

$$\operatorname{rank}\{b_i, Ab_i, \cdots, A^{n-1}b_i\} = n, \quad \forall i = 1, \cdots, m. \tag{2.21}$$

Proof. The Hamiltonian is $H = p^\top(Ax + Bu) - 1$, which is zero along the optimal solution. From previous subsection, we know that singularity appears when

$$\gamma_i(t) = p^\top(t)b_i = 0, \quad \forall t \in [t_1, t_2]$$

for some interval $[t_1, t_2]$. Notice that the costate equation for the LTI system is

$$\dot{p} = -A^\top p$$

thus $\frac{d^k\gamma_i}{dt^k} = (-1)^k A^k b_i$. Thus singular control exists if and only if

$$p(t) \perp \operatorname{span}\{b_i, \cdots, A^{n-1}b_i\}$$

which is equivalent to saying that either $p(t) \equiv 0$ or the rank condition (2.21) holds. But $p(t) \equiv 0$ can never happen since $p^\top(Ax + Bu) = 0$ along the optimal process and p satisfies a linear system. □

We say that the linear system (2.20) is *normal* if it satisfies the rank condition (2.21). Note that this requirement is stronger than controllability.

Note that the above results does not imply that non-normal system has no bang-bang optimal controller. In fact, we have the following theoretical result:

Proposition 2.5. *Consider the system* (2.20) *with control $|u_i| \le 1$, $\forall i$. If $x_0 \in \mathbb{R}^n$ is reachable from the origin, and $T > 0$ a real number, then there exists a bang-bang control that steers x_0 to 0 at time T.*

Example 2.4 (Harmonic oscillator)**.** Consider a harmonic oscillator $\ddot{x} + x = u$ whose control is constrained in the interval $[-1, 1]$. It is desired to find a control u which drives the system to the origin in minimal time. Write the system in standard form

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -x_1 + u.$$

The Hamiltonian is $H = p_1x_2 + p_2(-x_1 + u) + p_0$ and the costate equation is

$$\dot{p}_1 = p_2$$
$$\dot{p}_2 = -p_1$$

which is again a harmonic oscillator. Thus $p_2 = r\cos(t + \alpha_0)$ for some constants $r > 0$ and $\alpha_0 \in (-\pi, \pi)$ and the control input switches exactly once for every π elapsed time. Since u is piece-wise constant, along the optimal process, $\frac{d}{dt}[(x_1 - u)^2 + x_2^2] = 2(x_1 - u)\dot{x}_1 + 2x_2\dot{x}_2 = 0$, thus $(x_1 - u)^2 + x_2^2$ is also piece-wise constant. These constitute arcs on a circle, whose radius are determined by the initial condition. We can draw these circles on the plane as in Figure 2.5.

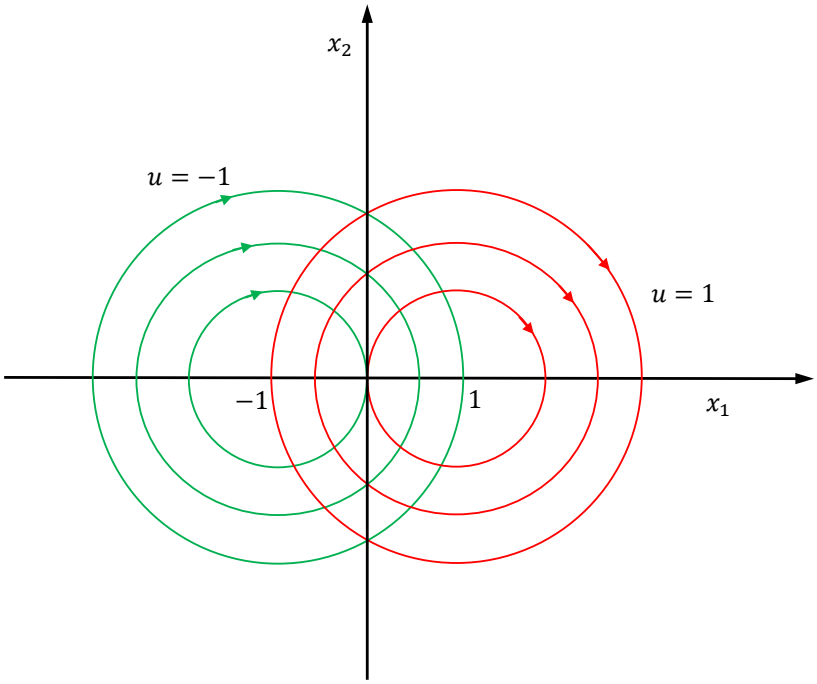

Figure 2.5: The phase plot for $u = \pm 1$.

Write $x_1 = u + \cos\theta$ and $x_2 = \sin\theta$ and substitute in to the system dynamics, we get $\dot{\theta} = -1$. Hence the system trajectories travel clockwise with velocity 1. But to due switches, the state cannot be kept on the same circle for angle more than π rad.

Let us trace back from $t = t_f$. At the final stage, in order to reach the origin, only two arcs are possible, see Figure 2.6.

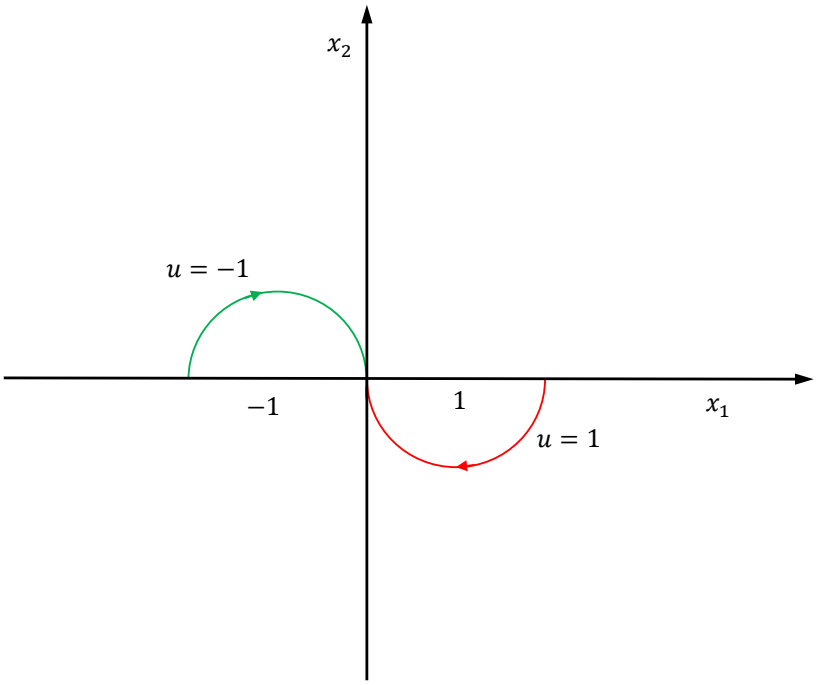

Figure 2.6: The phase plot at the final stage.

To find the previous arc, choose a point A as in Figure 2.7, then draw a line passing through A and $(-1,0)$. The intersection of this dashed line with the circle determined by A and center $(-1,0)$ is denoted A', which lies on the circle $(x_1+3)^2+x_2^2=C^2$. Thus for all initial states on the arch between A' and A, they should flow along the arch and then reach point A and goes to zero following the final stage arc. Continuing this procedure, we can find the optimal trajectory for all for arbitrary initial condition.

Singular control

Singular optimal control of linear system (2.20) considers minimizing the cost of the following form

$$J = x(t_f)^\top Q_f x(t_f) + \int_0^{t_f} x(t)^\top Q x(t) \mathrm{d}t \tag{2.22}$$

where Q and Q_f are symmetric non-negative definite matrices. The control is constrained by $|u_i| \le 1$, for all $i = 1, \cdots m$.

Note that the cost function (2.22) is different from the standard one in LQR control, where the integrand in the cost is of the form $x^\top Qx + u^\top Ru$ with Q semi-positive definite and R positive definite. In other words, in the standard LQR problem, the control u is penalized through the term $u^\top Ru$ whereas in singular control, the control u is penalized by direct constraint $|u_i| \le 1$. Note also that here Q is not required to be semi-positive definite.

For this problem, the Hamiltonian is

$$H = p^\top (Ax + Bu) - x^\top Qx$$

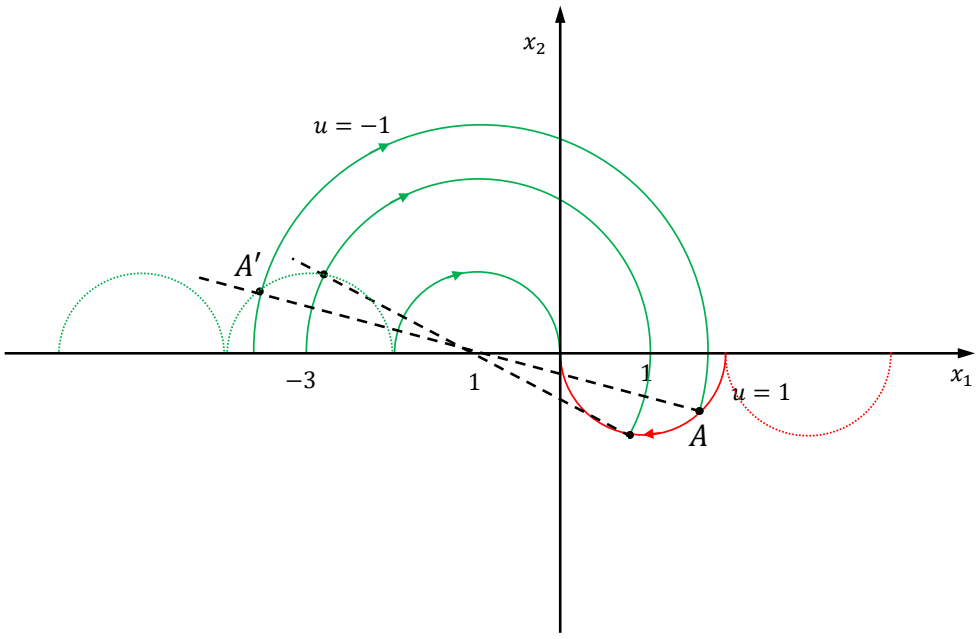

Figure 2.7: The phase plot of the last two stages.

and the costate equation is

$$\dot{p} = -A^\top p + Qx$$

Similar to time optimal control, the optimal control should satisfy

$$u_i^*(t) = \begin{cases} -1, & p^\top(t) b_i < 0 \\ 1, & p^\top(t) b_i > 0 \\ ?, & p^\top(t) b_i = 0 \end{cases}.$$

thus the optimal policy may be singular. However, different from time optimal control, for singular control, it is generally more difficult to exclude the existence of singular controls.

Example 2.5. Consider the system

$$\dot{x} = u$$

with initial condition $x(0) = 1$ and cost function

$$J = \frac{1}{2}\int_0^2 x(t)^2 \mathrm{d}t$$

Find an optimal control u which drives the system to $x(2) = 0$ under the constraint $|u| \le 1$.

The Hamiltonian is $H = pu + \frac{1}{2}p_0 x^2$. The costate equation is $\dot{p} = p_0 x$. If $p_0 = 0$, then p must a nonzero constant. Thus u is either 1 or -1 on the interval $[0,2]$, but in either case the control cannot bring $x(0)$ to the origin. Thus assume $p_0 = -1$. Applying maximum principle yields

$$u^*(t) = \begin{cases} 1, & p(t) > 0 \\ -1, & p(t) < 0 \\ ?, & p(t) = 0 \end{cases}$$

Thus when at the time instant when $p(t) = 0$, the maximum principle provides no information about the optimal control $u^*(t)$.

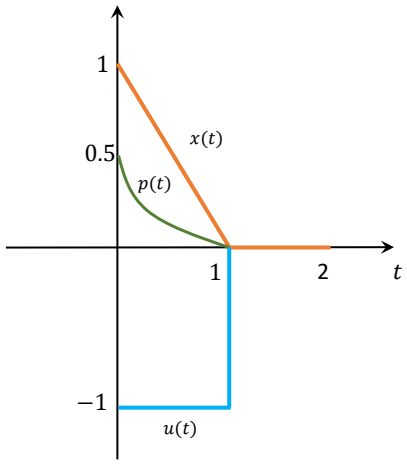

Figure 2.8: Singular control

When $p(t) < 0$, $\dot{x} = -1$, then

$$\dot{x} = -1,$$
$$\dot{p} = -x.$$

Suppose that the first switch happens at t_s. There are two possible cases.

Case 1: $p(t) < 0$ on $[0, t_s]$ with $p(t_s) = 0$. Then

$$x(t) = 1 - t$$
$$p(t) = p(0) - t + \frac{1}{2}t^2$$

for $t \in [0, t_s]$. Suppose that $[t_s, t_r]$ is a singular arch, i.e., $p(t) \equiv 0$ for $t \in [t_s, t_f]$. Thus $x = -\dot{p} \equiv 0$ and $u^* = 0$ on $[t_s, t_f]$. In particular, $1 - t_s = 0$ and $p(0) - t_s + \frac{1}{2}t_s^2 = 0$, which yields $t_s = 1$, $p(0) = \frac{1}{2}$, and $x(1) = 0$.

Case 2: $p(t) > 0$ on $[0, t_s]$ with $p(t_s) = 0$. One can they verify that $t_s = -1$, a contradiction.

To conclude, the first switch happens at $t_s = 1$, and $p(t) < 0$ on $[0, 1]$ while $p(1) = 0$. Obviously, for the rest of the time $t \in (1, 2]$, no control should be added, i.e., $u^*(t) = 0$ for $t \in (1, 2]$. Thus $x^*|_{(1,2]}$ is a singular arc with singular control $u^*|_{(1,2)} \equiv 0$. See Figure 2.8.

For singular control problem, it may happen that the bang-bang control law switches infinitely many times and that the law fails to be piece-wise constant. Such phenomenon is called *Fuller's phenomenon*.

Example 2.6 (Fuller's problem)**.** Consider the double integrator

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = u$$

with constraint $|u| \le 1$ and cost function

$$J = \int_0^{t_f} x_1^2(t)\mathrm{d}t$$

where t_f is free. The objective is to drive the system to the origin with minimal cost. The Hamiltonian is $H = p_1 x_2 + p_2 u + p_0 x_1^2$ and the costate equation is

$$\dot{p}_1 = -2p_0 x_1$$
$$\dot{p}_2 = -p_1$$

One can check easily that there is no singular arcs. Therefore, the control is bang-bang. It remains to compute the switching condition. A. T. Fuller showed that 1) the switching curve for this problem is $x_1 + \gamma x_2|x_2| = 0$ for some constant $\gamma > 0$; 2) there are infinitely many switches for along this curve; 3) the time intervals between two consecutive switches decrease geometrically. For details, see [9].

2.2.5 State constraints

State constraints appear quite naturally in many practical applications, such as in obstacle avoidance problems and medical/biological systems (many biological states are required to be positive).

There are two main classes of state constraints that arise frequently applications:

A. pure state constraints: $s(x(t)) \leq 0$, $\forall\, t \geq 0$ for some function $s : \mathbb{R}^n \to \mathbb{R}^p$.

B. mixed state-control constraints: $s(x(t), u(t)) \leq 0$, $\forall\, t \geq 0$ for some function $s : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$.

We state the maximum principle under smooth pure state constraints.

Theorem 2.1 (MP under pure state constraints)**.** *Consider the system $\dot{x} = f(x,u)$, where $u(t) \in U$ and the state is constrained according to $s(x(t)) \leq 0$ for some smooth function $s : \mathbb{R}^n \to \mathbb{R}^p$. If u^* is an optimal control that minimize the cost function $J = \varphi(x(t_f)) + \int_0^{t_f} L(x,u)\mathrm{d}t$, then*

1) There exists a costate function $p^*(\cdot)$ and a function $\lambda : [0, t_f] \to \mathbb{R}^p$ such that

$$\dot{p} = -H_x^\top - s_x^\top \lambda$$

where $H = p^\top f(x,u) - p_0 L(x,u)$, $p_0 \in \{0,-1\}$ and $(p_0, p(t)) \not\equiv 0$.

2) The maximum principle holds: $H(x^*(t), u^*(t), p^*(t)) = \max_{u\in U} H(x^*(t), u, p^*(t))$.

Optimal control problems with state constraints are generally quite difficult to solve. In most cases, one should not expect to derive analytic solutions and should resort to numerical methods instead.

The following academic example shows how to use the maximum principle under constraints.

Example 2.7. Consider the system $\dot{x} = x^2 - u$ with initial condition $x(0) = 1$ and cost function

$$J = \int_0^2 x^2 + u^2 \mathrm{d}t.$$

The control has not constraint. The objective is to find an optimal control u such that $x(2) = 1$ while keeping $x(t) \geq a$ with minimal cost. Here a is some real constant.

The state constraint can be equally described by $s(x) := a - x \leq 0$. The Hamiltonian is $H = p(x^2 - u) + p_0(x^2 + u^2)$ and the costate equation is

$$\dot{p} = -2(p + p_0)x + \lambda.$$

If $p_0 = 0$, then $|u| = +\infty$, which is impossible; thus $p_0 = -1$ and $u^*(t) = -\frac{1}{2}p(t)$. When the system stays on the boundary $a = x$, then $\dot{s}$ must also vanishes, in which case $x^2 - u = 0$, or $u = x^2 = a^2$. It follows that $p(t) = -2a^2$ on the boundary. Hence $\lambda = -2a(2a^2 + 1)$. Now the costate equation can be rewritten as

$$\dot{p} = -2(p-1)x - 2a(2a^2 + 1)$$

To determine the initial condition of p, substitute the relation $p = 2\dot{x} - 2x^2$ into the costate equation which yields a second order ODE

$$\ddot{x} = (2x^2 + 1)x - a(2a^2 + 1)$$

with boundary condition $x(0) = x(2) = 1$. Thus the problem reduces to solving a boundary value problem, which can be done via numerical methods.

2.2.6 Infinite horizon problem

2.2.7 Appendix: reachability and controllability

(This subsection is largely taken from [5].) Reachability and controllability are closely to Lie algebra of vector fields, which we recall briefly. Let $\Omega \subseteq \mathbb{R}^n$ be an open set, $\mathscr{F}(\Omega)$ the space of smooth real value functions on Ω and $\mathscr{X}(\Omega)$ the space of smooth vector fields on Ω. For two vector fields $f, g \in \mathscr{X}(\Omega)$, the *Lie bracket* $[f, g]$ is defined as $(Dg)f - (Df)g$, where Df represents the Jacobian of f. Some immediate observations of the Lie bracket: 1) $[f, f] = 0$; 2) $[f, g] = -[g, f]$; 3) $[\lambda f + \mu g, h] = \lambda[f, h] + \mu[g, h]$.

Note that both $\mathscr{F}(\Omega)$ and $\mathscr{X}(\Omega)$ are (infinite dimensional) real vector spaces. The following definition plays a fundamental role in nonlinear controllability theory.

Definition 2.1. A *Lie algebra* of vector fields on Ω is a linear subspace $\mathscr{A} \subseteq \mathscr{X}(\Omega)$ which is closed under Lie bracket operation, i.e., $[f, g] \in \mathscr{A}$ if $f, g \in \mathscr{A}$. For any set $S \subseteq \mathscr{X}(\Omega)$, the *Lie algebra generated by* S is the smallest Lie algebra containing S, denoted by $\mathrm{Lie}(S)$. We say that S is *Lie bracket generating at point* x if the dimension of $\mathrm{Lie}_x(S) = \{f(x) \mid f \in \mathrm{Lie}(S)\}$ is n.

The dimension of $\mathrm{Lie}(S)$ is in general infinite, but $\mathrm{Lie}_x(S)$ is a linear subspace of the tangent space at x, which is finite dimensional.

It is routine to verify that $\mathrm{Lie}(S)$ can be constructed through the following procedure. Let $\mathscr{A}_1 = \mathrm{span}_{\mathbb{R}} S$, then construct $\mathscr{A}_k$ recursively via

$$\mathscr{A}_{k+1} = \{[f, g] : f \in \mathscr{A}_k, g \in S\}, \quad k = 1, \cdots.$$

Then

$$\mathrm{Lie}(S) = \bigcup_{k \geq 1} \mathscr{A}_k.$$

For example, if $S = \{f_1, \cdots, f_m\}$, then S is spanned by all brackets of the form

$$[[\cdots [f_i, f_j], f_k], \cdots, f_\ell]$$

for f_is in S.

Consider a forward complete system defined by vector field f:

$$\dot{x} = f(x)$$

We use a more suggestive notation to denote the flow of the system: $e^{tf}(x) := \phi(t, x)$. It is then immediate to note

- $e^{(t+s)f} = e^{tf} \circ e^{ts}$, $\forall t, s \in \mathbb{R}$;

- e^{tf} is a diffeomorphism whose inverse is e^{-tf}, $\forall\, t \in \mathbb{R}$;
- $\frac{d}{dt} e^{tf}(x) = f(e^{tf}x)$, $\forall\, t \in \mathbb{R}$.

Definition 2.2 (Reachability)**.** Consider the system $\dot{x} = f(x,u)$, $u \in U$. Define

1) the *reachable set* from x_0 *at* time $t \geq 0$ is

$$\mathscr{A}(t, x_0) := \{\phi(t, x_0; u) : \exists u : [0,t] \to U\};$$

2) the *reachable set* from x_0 *within* time $t \geq 0$ is

$$\mathscr{A}(\leq t, x_0) := \bigcup_{\tau \in [0,t]} \mathscr{A}(\tau, x_0);$$

3) the *reachable set* from x_0 is

$$\mathscr{A}(x_0) = \bigcup_{t \geq 0} \mathscr{A}(t, x_0).$$

We say that the system is completely controllable if $\mathscr{A}(x_0) = \mathbb{R}^n$ for all $x_0 \in \mathbb{R}^n$.

The following lemma shows how Lie bracket is related to reachability.

Lemma 2.1. *For any two vector fields f and g,*

$$e^{t[f,g]}x = e^{-\sqrt{t}g} e^{-\sqrt{t}f} e^{\sqrt{t}g} e^{\sqrt{t}g} x + o(t)$$

for $|t|$ sufficiently small.

This lemma can be proved easily by Taylor expansion and is left as an exercise. The lemma can be understand through the driftless control system

$$\dot{x} = u_1 f(x) + u_2 g(x)$$

Then the lemma suggests that it is possible, by switching the input u_1 and u_2, to reach points that is reachable by the system $\dot{x} = [f,g]$.

From now on, we will focus our attention on affine control systems

$$\dot{x} = f_0(x) + \sum_{i=1}^{m} u_i f_i(x), \tag{2.23}$$

where f_i are smooth vector fields and $u = (u_1, \cdots, u_m) : \mathbb{R}_{\geq 0} \to U \subseteq \mathbb{R}^m$. We assume that U contains an open neighbourhood of the origin. Define

$$\Sigma := \mathrm{Lie}\{f_0 + \sum_{i=1}^{m} u_i f_i : u \in U\}$$

then it is easy to show that

$$\Sigma = \mathrm{Lie}\{f_0, f_1, \cdots f_m\}.$$

We call Σ the Lie algebra associated with the system (2.23).

Exercise 2.3. Consider the single input LTI system $\dot{x} = Ax + bu$, $x \in \mathbb{R}^n$. Show that the Lie algebra associated with the system is

$$\mathrm{Lie}\{Ax, b\} = \mathrm{span}\{Ax, b, Ab, \cdots, A^{n-1}b\}.$$

Proposition 2.6 (Krener). *If the Lie algebra associated with the system* (2.23) *is Lie bracket generating at* x_0*, then for every* $t > 0$*,* x_0 *belongs to the closure of the interior of* $\mathscr{A}(\leq t, x_0)$.

Krener Theorem says that $\mathscr{A}(\leq t, x_0)$ contains an open set O having x_0 in its closure.

Definition 2.3. A family of vector fields S is said to be *symmetric* if $f \in S$ implies $-f \in S$.

For example, for driftless system $\dot{x} = \sum u_i f_i$, if both u and $-u$ are admissible controls, then the system is symmetric.

The following is the fundamental theorem regarding nonlinear controllability.

Theorem 2.2 (Chow-Rashevskii). *For the system* (2.23)*, if* $\{f_0, f_1, \cdots, f_m\}$ *is Lie bracket generating and symmetric, then the system is completely controllable, i.e., for every* $x_0 \in \mathbb{R}^n$*,* $\mathscr{A}(x_0) = \mathbb{R}^n$.

Example 2.8 (Dubins car). Consider a model for a two wheel cart on the plane

$$\begin{aligned} \dot{x} &= u_1 \cos\theta \\ \dot{y} &= u_1 \sin\theta \\ \dot{\theta} &= u_2 \end{aligned}$$

where (x, y) represents the position of the cart and θ the heading angle. There are two controls, u_1 the driving speed, and u_2 the turning rate. Suppose that the cart can be driven either back and forward and the turning rate can be either negative or positive. Thus the system is symmetric. Let $f_1 = [\cos\theta, \sin\theta, 0]$ and $f_2 = [0, 0, 1]$. Then $[f_1, f_2] = -[\sin\theta, -\cos\theta, 0]$. It follows that $\text{rank}\{f_1, f_2, [f_1, f_2]\} = 3$ for all (x, y, θ). Thus the system is completely controllable.

Exercise 2.4 (Nelson's car). Consider a front-wheel drive car shown in Figure 2.9. The control input are: 1) the front wheel turning rate; 2) the driving speed. Derive the motion dynamics of this model and check its controllability.

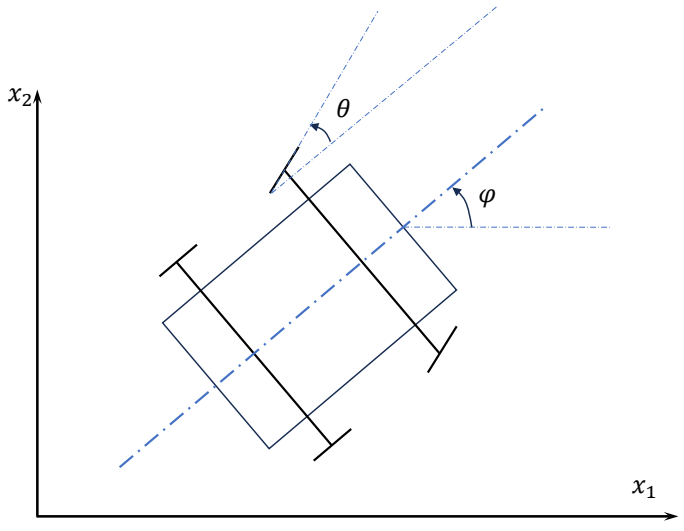

Figure 2.9: Nelson's car

2.3 Proof of the maximum principle

In this section, we prove the maximum principle following Boltyanskii [4].

2.3.1 Nonlinear optimization

Motivating example

The methodology that we are going to use to prove the maximum principle can be illustrated through static nonlinear optimization problem:

$$\begin{aligned} &\min\ g_0(x) \\ &\text{subject to } g_i(x) \le 0, \quad i = 1, \cdots, m \end{aligned} \tag{LM}$$

in which $\{g_i\}_{i=0}^m \in C^1(\mathbb{R}^n;\mathbb{R})$. Assume $\operatorname{rank}(Dg(x)) = m$ on the set $S = \{x : g_i(x) = 0,\ i = 1, \cdots, m\}$, where $g = [g_1, \cdots, g_m]^\top$. Suppose that the problem is feasible, i.e., there exists an admissible x_* which minimizes $g_0(x)$.

To solve this optimization problem, it is standard practice to use the so called *Lagrangian multiplier* method. Other than that, one may use calculus of variation that we have introduced previously to derive first order necessary conditions.

Exercise 2.5. Derive the first order necessary condition of the (LM) problem using calculus of variation.

Here, we adopt a completely new approach, which bears the name *method of tent* introduced by Boltyanskii and his colleagues when proving the maximum principle.

Define the following sets:

$$\Omega_i = \{x \in \mathbb{R}^n : g_i(x) \le 0\}, \quad i = 1, \cdots, m$$

and for $x_1 \in \mathbb{R}^n$, let

$$\Omega_0 = \{x : g_0(x) < g_0(x_1)\} \cup \{x_1\}.$$

Take the intersection of all these sets

$$\Sigma := \Omega_0 \cap \Omega_1 \cap \cdots \cap \Omega_m$$

We claim that x_1 is a minimizer *if and only if* $\Sigma = \{x_1\}$. To see this, suppose x_1 is a minimizer, then $g_i(x_1) = 0$ for $i \ge 1$ and $g_0(x_1) \le g_0(x)$ for all $x \in S$. Thus $x_1 \in \Sigma$. If there is another point $x_2 \in \Sigma$, then x_2 is feasible and $g_0(x_2) < g_0(x_1)$, a contradiction, thus if x_1 is a minimizer, there must hold $\Sigma = \{x_1\}$. Conversely, suppose that $\Sigma = \{x_1\}$, if x_1 is not a minimizer, then either x_1 is not feasible or there exists $x_2 \ne x_1$, both feasible such that $g_0(x_2) < g_0(x_1)$. For the first case, $x_1 \notin \Omega_1 \cap \cdots \cap \Omega_m$, thus $x_1 \notin \Sigma$, a contradiction. For the second case, $\{x_1, x_2\} \subseteq \Sigma$, a contradiction.

As an example, let $m = 1$ and Figure 2.12 is a sketch of the sets Ω_0 and Ω_1. In this figure, Ω_1 and Ω_0 intersects on the curve γ. In order that $\Omega_0 \cap \Omega_1 = \{x_1\}$, then the two sets must separate in the sense that they intersect only at point x_1. To go one step further, let us recall the definition of a tangent cone.

Given a set $\Omega \subseteq \mathbb{R}^n$ (may be non-convex), the *tangent cone* at $x \in \Omega$ is defined as

$$T_x\Omega := \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{array}{c} \exists \{x_i\}_{i=1}^\infty \subseteq \Omega, \exists \{t_i\}_{i=1}^\top \subseteq \mathbb{R}_{>0}, \text{s.t.} \\ t_i \downarrow 0,\ x_i \to x, \text{ and } (x_i - x)/t_i \to v \end{array} \right\}$$

see Figure 2.10.

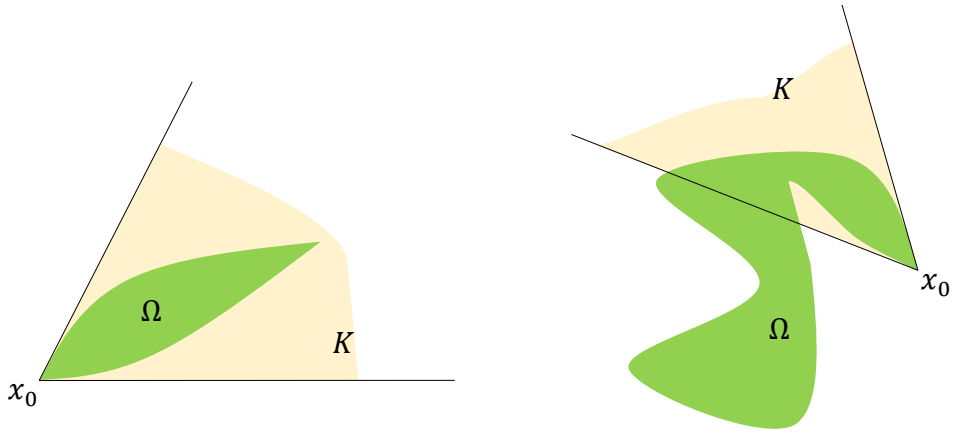

Figure 2.10: Tangent of convex and non-convex sets Ω.

A convex cone $K \subseteq T_x\Omega$ with apex x is called a *tent*. Note that although a tangent cone may be non-convex, a tent is required to be convex. In Figure 2.11, K_0 represents the tangent cones while K_1 some tents.

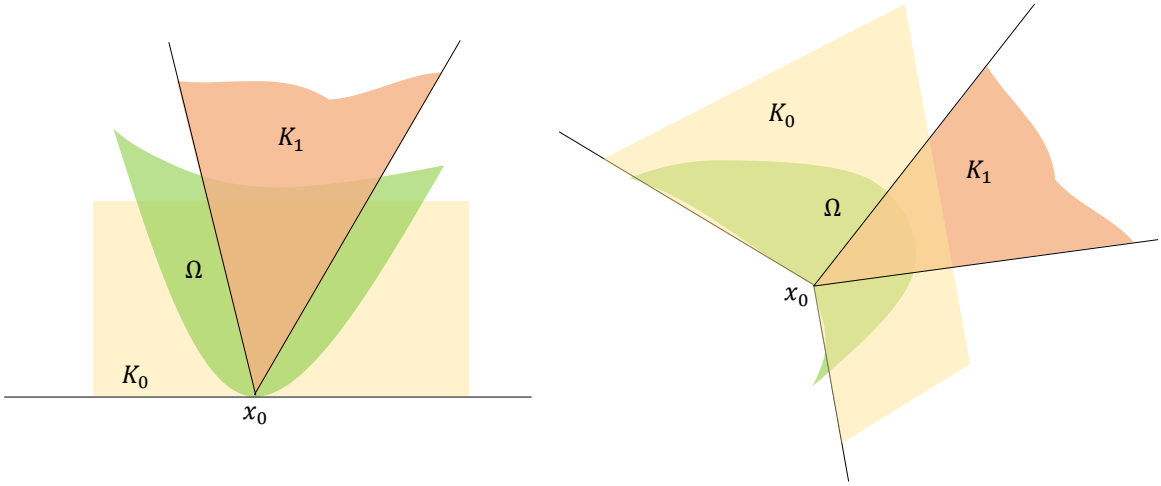

Figure 2.11: Tents.

Intuitively, to be able to "separate" Ω_0 and Ω_1, the tangent cone of the two sets should be separable in the sense that they intersect only at the apex. Or equivalently, there is a hyperplane passing through x_1 which separates $T_{x_1}\Omega_0$ and $T_{x_1}\Omega_1$, see Figure 2.13.

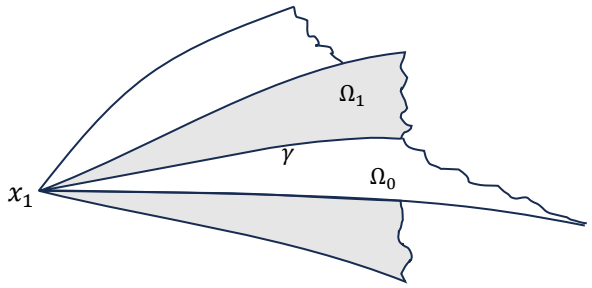

Figure 2.12: The sets Ω_0 and Ω_1.

In Figure 2.13, let us choose two arbitrary nonzero vectors a_0 and a_1 perpendicular to the separating hyperplane such that $a_0 + a_1 = 0$, and it is easy to see that such vectors always exist. Furthermore, we see that

$$a_i^\top (x - x_1) \geq 0, \quad \forall x \in K_i,\ i = 0, 1. \tag{2.24}$$

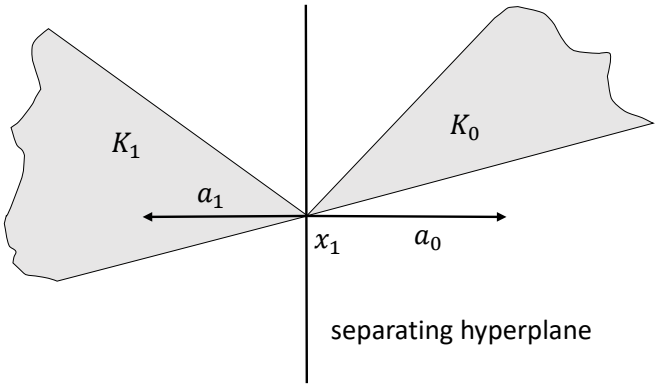

Figure 2.13: Separating 2-dim tents.

Thus if we can find out K_0 and K_1, we can obtain a necessary condition based on the relation (2.24). For problem (LM), this is easy since g_0 and g_1 are smooth:

$$K_i = \{x : \nabla g_i(x_1)(x - x_1) \leq 0\}, \quad i = 0, 1$$

That is, K_i are half spaces, see Figure 2.14.

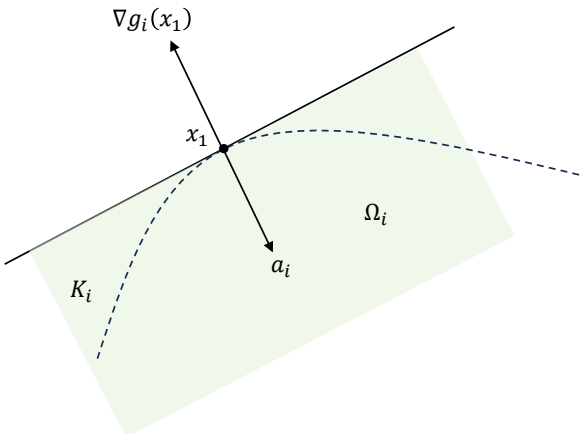

Figure 2.14: The tents are half spaces.

Therefore, a_i must be of the form

$$a_i = \lambda_i \nabla g_i(x_1)$$

for $\lambda_i \leq 0$. Since λ_i cannot be zero at the same time, $\lambda_i < 0$ for $i = 0, 1$. Thus the relation (2.24) becomes

$$\nabla g_0(x_1) + \lambda \nabla g_1(x_1) = 0$$

for some $\lambda > 0$. This is a special case of the famous KKT (Karush-Kuhn-Tucker) condition which we will be able to prove once we have generalize the above reasoning.

The separability of tents

We generalize our previous discussions to arbitrary finite many tents.

Definition 2.4 (Separability)**.** Let $K_0, \cdots, K_p$ be some closed, convex cones with a common apex x in $\mathbb{R}^n$. They are said to be *separable* if there exists a hyper plane Γ through x that separates one of the cones from the intersection of the others.

Lemma 2.2. *Let $K_0, \cdots, K_p$ be some closed, convex cones with a common apex x in $\mathbb{R}^n$. Then they are separable if and only if there exist dual vectors a_i, $i = 0, 1, \cdots, p$ fulfilling*[2]

$$a_i^\top (y - x) \leq 0, \quad \forall y \in K_i$$

and at least one of which is not zero and such that

$$a_0 + \cdots + a_p = 0.$$

Lemma 2.3. *Let $\Omega_0, \cdots, \Omega_p$ be sets in $\mathbb{R}^n$ satisfying*

$$\Omega_0 \cap \cdots \cap \Omega_p = \{x\},$$

and $K_0, \cdots, K_p$ be tents of these sets at x. If all the tents are convex and that at least one of the tents is distinct from its affine hull. Then $K_0, \cdots, K_p$ is separable.

[2]Note that we can also use $a_i^\top (y - x) \geq 0$ by reversing the sign of a_i, see (2.24).

The proofs of the above two results are quite technical and are hence omitted. Interested readers are referred to [4].

We now have all the ingredients to derive the KKT condition. First, recall the problem formulation:

Problem. Let f, g_i, h_j be continuously differentiable real functions. Derive the necessary condition for the following minimization problem:

$$\begin{aligned} &\min\ f(x) \\ &\text{subject to } g_i(x) \le 0, \quad i = 1, \cdots, p \\ &\qquad\qquad h_j(x) = 0, \quad j = 1, \cdots, q \end{aligned} \tag{2.25}$$

To solve this problem, let x_* be a minimizer and define

$$\begin{aligned} \Omega_i &= \{x : g_i(x) \le 0\}, \quad i = 1, \cdots, p \\ \Xi_j &= \{x : h_j(x) = 0\}, \quad j = 1, \cdots, q \\ \Theta &= \{x : f(x) \le f(x_*)\} \cup \{x_*\} \end{aligned}$$

then

$$\Sigma := \bigcap_i \Omega_i \bigcap_j \Xi_j \bigcap \Theta = \{x_*\}.$$

The tents of the defined sets are

$$\begin{aligned} K^{\Omega_i} &= \{x : \nabla g_i(x_*)(x - x_*) \le 0\} \\ K^{\Xi^j} &= \{x : \nabla h_j(x_*)(x - x_*) = 0\} \\ K^{\Theta} &= \{x : \nabla f(x_*)(x - x_*) \le 0\} \end{aligned}$$

By Lemma 2.3, there exists non-negative vectors ω_i, ξ_i, θ satisfying

$$\begin{aligned} \omega_i^\top (x - x_*) \le 0, \quad \forall x \in K^{\Omega_i} \\ \xi_j^\top (x - x_*) \le 0, \quad \forall x \in K^{\Xi_j} \\ \theta^\top (x - x_*) \le 0, \quad \forall x \in K^{\Theta} \end{aligned}$$

and

$$\sum_i \omega_i + \sum_j \xi_j + \theta = 0 \tag{2.26}$$

Since K^{Ω_i} and K^{Θ} are half spaces, it follows that

$$\omega_i = \mu_i \nabla g_i(x_*), \quad \xi_j = \nu_j \nabla h_j(x_*), \quad \theta = \theta_0 \nabla f(x_*)$$

in which $\mu_i \ge 0$, $\theta_0 \ge 0$ and the signs of ν_j are undetermined. Plugging into (2.26), we get the KKT condition:

$$\theta_0 \nabla f(x_*) + \sum_i \mu_i \nabla g_i(x_*) + \sum_j \nu_j \nabla h_j(x_*) = 0.$$

2.3.2 Proof of the maximum principle

Problem statement

We start by introducing the optimal control problem under fixed terminal time. First, let us recall our optimal control problem. We focus on time-invariant control systems:

$$\dot{x} = f(x,u), \tag{2.27}$$

where $x(t) \in \mathbb{R}^n,\ u(t) \in U \subset \mathbb{R}^m$ for all $t \in [0, t_f]$, the initial condition $x(0) = x_0$ is assumed to be fixed. The cost function is

$$J(u(\cdot)) = \varphi(x(t_f)) + \int_0^{t_f} L(x(s), u(s))ds,$$

where $\varphi(\cdot)$, $f(\cdot,u)$, $L(\cdot,u)$ are continuously differentiable for all u. The optimal control problem amounts to finding a process $u_*(t),\ x_*(t)$, $0 \le t \le t_f$, with a (measurable) controller $u_*(t)$ such that $x_*(t_f) \in M$ for some manifold M, and $J(u_*(\cdot))$ attains a minimum. We say that the problem is in 1) *Mayer form* if $L = 0$; 2) *Lagrange form* if $\varphi = 0$; 3) *Bolza form* if neither L nor φ is zero.

We claim that the preceding three types of optimal control problems can all be reduced to Mayer form. In fact, let

$$x_{n+1}(t) = \int_0^t L(x(s), u(s))ds$$

the system becomes

$$\begin{cases} \dot{x} = f(x,u) \\ \dot{x}_{n+1} = L(x,u) \end{cases} \tag{2.28}$$

\with initial condition $(x_0, 0)$, and the cost function becomes

$$J = \varphi(x(t_f)) + x_{n+1}(t_f). \tag{2.29}$$

This is an optimal control problem of the Mayer form of a time-invariant system. Due to this reason, it suffices to study the optimal control problem with cost function:

$$J = \varphi(x(t_f)).$$

Introduce the following notations which will be used in the proof:

$$x_1 := x_*(t_f)$$
$$\Omega_0 = \{x_1\} \cup \{x : \varphi(x) < \varphi(x_1)\}$$
$$\Omega_1 : \text{ reachability set from } x_0$$
$$\Omega_2 = M\text{: the terminal manifold}$$

Let $u_*(t)$, $x_*(t)$, $0 \le t \le t_f$ be an optimal process. Then it is easily seen that

$$\Omega_0 \cap \Omega_1 \cap \Omega_2 = \{x_1\}. \tag{2.30}$$

The reader should immediately realize that such type of condition implies separability of tents of the three sets, this is the content of Lemma 2.3. Denote K_i the tent of Ω_i at x_1. It thus remains to find the tents K_i. The tents K_0 and K_2 can be easily computed:

$$K_0 = \left\{x \in \mathbb{R}^n : \nabla\varphi(x_1)(x - x_1) \le 0\right\}$$
$$K_2 = T_{x_1}\Omega_2$$

(note that Ω_2 is a fixed manifold).

Therefore, our problem boils down to calculating the tent of Ω_1 at x_1: K_1. By definition, a tent is only a convex subcone of the tangent cone of Ω_1 at x_0, we should however, try to find a tent as big as possible, since the bigger the tent, the more necessary information it conveys. This is the main non-trivial step in proving the maximum principle (if we already know Lemma 2.2, 2.3) and was first achieved by Boltyanskii and his colleagues using the so called needle variation.

Needle variation

Suppose at the moment that the optimal control $u_* : [0, t_f] \to U$ is continuous. Fix $\tau \in (0, t_f]$ and consider the following *needle shaped variation* of u_* for small $\varepsilon > 0$:

$$u_\varepsilon(t) = \begin{cases} w, & t \in (\tau - \varepsilon, \tau] \\ u_*(t), & \text{otherwise} \end{cases}$$

where $w \in U$ is some constant, see Figure 2.15.

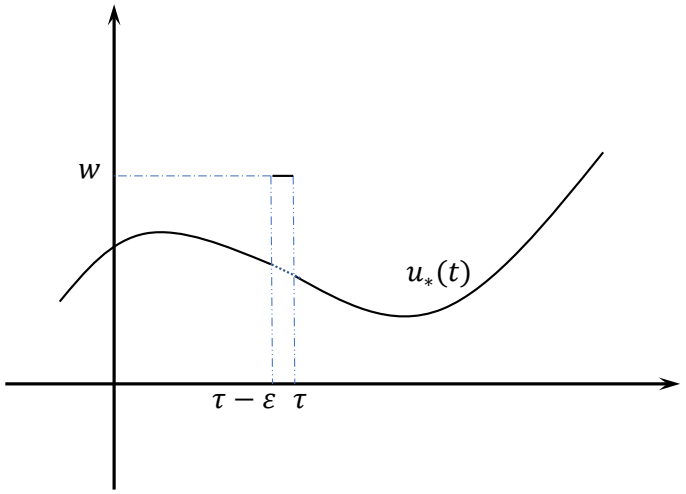

Figure 2.15: Needle variation.

Denote $t \mapsto x_\varepsilon(t)$ the solution to $\dot{x} = f(x, u_\varepsilon)$. Obviously, $u_\epsilon(\cdot)$ is admissible, thus $x_\epsilon(t_f)$ belongs to the reachable set at t_f, i.e., $x_\epsilon(t_f) \in \Omega_1$ for all ϵ chosen above. Thus by definition, $\frac{\partial x_\epsilon(t_f)}{\partial \varepsilon}\Big|_{\varepsilon=0+}$ must belong to the tangent cone of Ω_1. Denote

$$v(t) = \frac{\partial x_\varepsilon(t)}{\partial \varepsilon}\Big|_{\varepsilon=0+}, \quad t \in [\tau, t_f]$$

then it remains to find $v(t_f)$. We call $v(t_f)$ a *deviation vector*. To find the deviation vector, first we need to characterize $x_\epsilon(t)$. Denote $v_\epsilon(t) = \frac{\partial x_\epsilon(t)}{\partial \varepsilon}$, since $u_\epsilon(t) = u_*(t)$ for $t \geq \tau$, it follows that

$$\begin{aligned} \frac{d v_\epsilon(t)}{d t} &= \frac{\partial}{\partial \epsilon} f(x_\epsilon(t), u_*(t)) = \frac{\partial f}{\partial x}(x_\epsilon(t), u_*(t)) \frac{\partial x_\epsilon(t)}{\partial t} \\ &= \frac{\partial f}{\partial x}(x_\epsilon(t), u_*(t)) v_\epsilon(t), \quad \forall\, t \in (\tau, t_f] \end{aligned}$$

Evaluating at $\epsilon = 0+$, we get $\dot{v}(t) = \frac{\partial f}{\partial x}(x_*(t), u_*(t))v(t)$. That is, $v(t)$ satisfies a linear ODE. It still remains to find the initial condition $v(\tau)$. Note that

$$\begin{aligned} x_\varepsilon(\tau) &= x_*(\tau-\epsilon) + \int_{\tau-\epsilon}^{\tau} f(x_\epsilon(s), w)\mathrm{d}s, \\ &= x_*(\tau-\epsilon) + \int_{\tau-\epsilon}^{\tau} f(x_*(s), u_*(s))\mathrm{d}s + \int_{\tau-\epsilon}^{\tau} [f(x_\epsilon(s), w) - f(x_*(s), u_*(s))]\mathrm{d}s \\ &= x_*(\tau) + \int_{\tau-\epsilon}^{\tau} [f(x_\epsilon(s), w) - f(x_*(s), u_*(s))]\mathrm{d}s \end{aligned}$$

thus

$$\begin{aligned} v(\tau) &= \lim_{\varepsilon\to 0+} \frac{x_\varepsilon(\tau) - x_*(\tau)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0+} \frac{1}{\varepsilon}\left[\int_{\tau-\varepsilon}^{\tau} f(x_\varepsilon(t), w)\,dt - f(x_*(t), u_*(t))\mathrm{d}t\right] \\ &= f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)). \end{aligned} \tag{2.31}$$

To summarize, $v(\cdot)$ is the solution to the following Cauchy problem

$$\begin{cases} \dot{v} = \dfrac{\partial f}{\partial x}(x_*(t), u_*(t))\, v, \quad \forall\, t \in [\tau, t_f] \\ v(\tau) = f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)). \end{cases}$$

To construct more deviation vectors, let $v_1(t_f), \cdots, v_r(t_f)$ be some different deviation vectors obtained as above corresponding to some distinct time instants $\tau_1 < \cdots < \tau_r$ and constant inputs $w_1, \cdots, w_r$. Consider the combined needle variation

$$u_{\varepsilon,k}(t) = \begin{cases} w_i, & t \in (\tau_i - k_i\varepsilon, \tau_i] \text{ for some } i \in \{1, \cdots, r\} \\ u_*(t), & \text{otherwise} \end{cases}$$

where k_i are non-negative constants satisfying $\sum_{i=1}^r k_i = 1$. One can show that

$$\sum_{i=1}^{r} k_i v_i(t_f) = \left.\frac{\partial x(t_f, u_{\varepsilon,k})}{\partial \varepsilon}\right|_{\varepsilon=0+}$$

which implies that $\sum_{i=1}^r k_i v_i(t_f)$ are again in $T_{x_1}\Omega_1$. Still call these vectors deviation vectors and define K_1 to be the set of all deviation vectors, i.e.,

$$K_1 = \left\{ \sum_{i=1}^{r} k_i v_i(t_f) \;\middle|\; \begin{array}{l} \exists r \in \mathbb{Z}_+,\, \tau_i \in [0, t_f),\, w_i \in U,\, k_i \geq 0,\, \sum_{i=1}^r k_i = 1, \\ v_i(t_f) \text{ the deviation vector obtained from needle} \\ \text{variation at} \tau_i \text{ with spike } w_i \end{array} \right\}$$

Then K_1 is a tent of Ω_1 at x_1.

Final step: the costate equation and the maximum principle

Condition (2.30) implies that K_0, K_1, K_2 are separable. Invoking Lemma 2.2 and Lemma 2.3, we deduce that there exist three vectors a_i, at least one of which is nonzero, satisfying

$$a_i^\top v \leq 0, \quad v \in K_i,\, i = 0, 1, 2 \tag{2.32}$$

and

$$a_0 + a_1 + a_2 = 0. \tag{2.33}$$

In particular, $a_1^\top v(t_f) \le 0$ for any deviation vector $v(t_f)$. Now we introduce a small trick: if we are able to construct some function $p : [0, t_f] \to \mathbb{R}^n$ such that $p(t)^\top v(t) \equiv$ constant with $p(t_f) = a_1$, then we obtain immediately $p(t)^\top v(t) = a_1^\top v(t_f) \le 0$ for all $t \in [0, t_f]$. In particular, if v is the deviation vector obtained by needle variation at time τ with spike w, then $v(\tau) = f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)$. Thus at $t = \tau$, $p(\tau)^\top [f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau))] \le 0$ or

$$p(\tau)^\top f(x_*(\tau), u_*(\tau)) \ge p(\tau)^\top f(x_*(\tau), w)) \tag{2.34}$$

For convenience, define

$$H(x, u, p) := p^\top f(x, u)$$

which is the Hamiltonian associated with the system . Now that the spike can be any $w \in U$ and $t \in [0, t_f)$, it follows from (2.34) that

$$H(x_*(t), u_*(t), p(t)) = \max_{u \in U} H(x_*(t), u, p(t)) = \text{constant}, \quad \forall\, t \in [0, t_f). \tag{2.35}$$

This is the maximum principle that we have been looking for! Except two things: the interval $[0, t_f)$ doesn't include the endpoint t_f and the function p hasn't been determined yet. The first issue can be fixed if everything is continuous in the above formula, which is indeed true as long as we have shown p is, since f, x_* and u_* are continuous as assumed. For the second issue, let us recall the following simple fact:

Lemma 2.4. *Consider two linear ODE*

$$\dot{x} = A(t)x$$
$$\dot{p} = -A(t)^\top p$$

where $x, p \in \mathbb{R}^n$. Then $p(t)^\top x(t) = p(t')^\top x(t')$ for any $t, t' \in \mathbb{R}$.

With this lemma, we can now construct p to be the solution of the following ODE

$$\begin{aligned} \dot{p} &= -\left[\frac{\partial f}{\partial x}(x_*(t), u_*(t))\right]^\top p \\ &= -H_x^\top (x_*, u_*, p) \end{aligned} \tag{2.36}$$

with terminal state $p(t_f) = a_1$ (note that this is exactly the costate equation).

Recall that

$$K_0 = \left\{x \in \mathbb{R}^n : \nabla\varphi(x_1)(x - x_1) \le 0\right\}$$
$$K_2 = T_{x_1}\Omega_2$$

For a_0, since K_0 is a half space, $a_0^\top v \le 0$ for $v \in K_0$ implies $a_0 = \lambda\nabla\varphi(x_1)^\top$ for some constant $\lambda \ge 0$. For a_2, since K_2 is a sub-manifold, $a_2 \perp K_2$. It follows from (2.33) that (recall $a_1 = p(t_f)$):

$$\lambda\nabla\varphi(x_*(t_f))^\top + p(t_f) \perp \Omega_2 \tag{2.37}$$

for some constant $\lambda \ge 0$.

Up to now, we have prove the maximum principle for the Mayer problem under the assumption that u_* is continuous.

For u not continuous, only the condition (2.35) needs to be modified by noticing that the limits in (2.31) exist for almost all $t \in [0, t_f]$. Summarizing, we have proved the following.

Theorem 2.3. *Suppose that the Mayer form optimal control problem admits an admissible measurable optimal law $u_*(\cdot)$ with corresponding trajectory $x_*(\cdot)$. Then there is a solution to the costate equation (2.36), such that the triple $(x_*(t), u_*(t), p(t))$ satisfies the maximum principle (2.35) for almost all t (all t on the interval on which $u_*(\cdot)$ is continuous) and the transversality condition (2.37).*

Some variants

We have so far considered the optimal control problem under the condition that t_f is fixed. It can be easily extended to the case of free terminal time: it is obvious that all the necessary conditions of Theorem 2.3 still need to be hold. The mere difference is that now one can also make the variation of the terminal time. For example, consider a needle variation at τ, let $v(t_f)$ be the corresponding deviation vector. Fix some $\mu > 0$, since $x_\epsilon(t_f + \epsilon\mu) \in \Omega_1$, $\left.\frac{\partial x_\epsilon(t_f+\epsilon\mu)}{\partial \epsilon}\right|_{0+}$ must also lie in the tangent cone of Ω_1, but

$$\left.\frac{\partial x_\varepsilon(t_f+\varepsilon\mu)}{\partial \varepsilon}\right|_{\varepsilon=0+} = \left.\frac{\partial x_\varepsilon(t_f)}{\partial \varepsilon}\right|_{\varepsilon=0+} + \left.\frac{\partial x_*(t_f+\epsilon\mu)}{\partial \epsilon}\right|_{\varepsilon=0+} = v(t_f) + \mu f(x_*(t_f), u_*(t_f))$$

Thus we can construct another tent of Ω_1 at x_1 as

$$K_1' = \{v(t_f) + \mu f(x_*(t_f), u_*(t_f)) : v(t_f) \in K_1,\ \mu \in \mathbb{R}\}.$$

It follows that one can obtain a finer condition than (2.35):

$$H(x_*(t), u_*(t), p(t)) = \max_{u\in U} H(x_*(t), u, p(t)) = 0, \quad \forall\, t \in [0, t_f).$$

Indeed, take $v(t_f) = 0$ (no needle variation), then $a_1^\top(\mu f(x_*(t_f), u_*(t_f))) \le 0$ for any $\mu \in \mathbb{R}$ implies that $a_1^\top f(x_*(t_f), u_*(t_f)) = 0$.

Let us use Theorem 2.3 to derive the maximum principle for Bolza form. Recall that the system model and cost function of the Bolza problem can be equally written as (2.28) and (2.29). Suppose that the terminal manifold $\Omega_2 = M$, then for the augmented system (2.28), the terminal manifold is $\tilde{\Omega}_2 = \Omega_2 \times \mathbb{R}$. The Hamiltonian becomes

$$H(x, u, p, p_0) = p^\top f(x, u) + p_0 L(x, u)$$

and the costate equation still reads $\dot{p} = -H_x$, and $\dot{p}_0 = 0$ since H doesn't depend on x_{n+1}. Thus p_0 is a constant. The transversal condition reads

$$\lambda \begin{bmatrix} \nabla\varphi(x_*(t_f))^\top \\ 1 \end{bmatrix} + \begin{bmatrix} p(t_f) \\ p_0 \end{bmatrix} \perp T_{x_1}\Omega_2 \times \mathbb{R}$$

for some $\lambda \ge 0$, from which it follows that $p_0 = -\lambda \le 0$ and $p(t_f) + \lambda\nabla\varphi(x_*(t_f))^\top \perp \Omega_2$. When p_0 is nonzero, one can take $p_0 = -1$ by modifying λ. Thus we are done with the general Bolza form problem.

2.4 Some advanced topics

2.4.1 Maximum principle on manifolds

The Poisson bracket and symplectic geometry

We include a short introduction to symplectic geometry. The main reference of this subsection is [13].

Definition 2.5 (Poisson bracket)**.** Let M be a manifold and $\mathscr{F}(M)$ the set of smooth real-valued functions on M (an $\mathbb{R}$-algebra under piecewise product and sum). A *Poisson bracket* is a binary operation

$$\{,\} : \mathscr{F}(M) \times \mathscr{F}(M) \to \mathscr{F}(M)$$

which satisfies the following properties for all $f, g, h \in \mathscr{F}(M)$:

1. bilinearity: $\{f, g\}$ is $\mathbb{R}$-bilinear in f and g;
2. anticommutativity: $\{f, g\} = -\{g, f\}$;
3. Jacobi's identity: $\{\{f, g\}, h\} + \{\{h, f\}, g\} + \{\{g, h\}, f\} = 0$;
4. Leibnitz' rule $\{fg, h\} = f\{g, h\} + g\{f, h\}$.

The manifold M is said to be a *Poisson manifold* if it is equipped with a Poisson bracket.

Definition 2.6. Let $(M_1, \{,\}_1)$ and $(M_2, \{,\})$ be two Poisson manifolds. A mapping $\varphi : M_1 \to M_2$ is called *Poisson* if for all $f, h \in \mathscr{F}(M_2)$, we have

$$\{f, h\}_2 \circ \varphi = \{f \circ \varphi, h \circ \varphi\}_1.$$

Definition 2.7 (Symplectic manifold)**.** Let M be a manifold and Ω is a 2-form ($(0,2)$-tensor). The pair (M, Ω) is called a *symplectic manifold* if Ω satisfies

1. $d\Omega = 0$ (i.e., Ω is closed) and
2. Ω is nondegenerate in the sense that $\Omega(v, w) = 0$ for all w implies that v is a zero tangent vector.

Definition 2.8 (Hamiltonian vector field)**.** Let (M, Ω) be a symplectic manifold and let $f \in \mathscr{F}(M)$. Let X_f be the unique vector field on M satisfying

$$\Omega_z(X_f(z), v) = df(z)(v), \quad \text{for all } v \in T_z M.$$

We call X_f the *Hamiltonian vector field* of f. *Hamilton's equations* are the differential equations on M given by

$$\dot{z} = X_f(z).$$

Remark 2.1. The existence and uniqueness of X_f is guaranteed by the non-degeneracy of Ω.

If (M, Ω) is a symplectic manifold. Then one can define a Poisson bracket as

$$\{f, g\} = \Omega(X_f, X_g)$$

which we call the Poisson bracket associated with the symplectic manifold (M, Ω). Therefore, every symplectic manifold is also Poisson. The converse is not true. However, Hamiltonian vector fields can still be defined on Poisson manifold.

Definition 2.9. Let $(M, \{,\})$ be a Poisson manifold and let $f \in \mathscr{F}(M)$. Define X_f be the unique vector field on M satisfying

$$dk(X_f) = \{k, f\} \text{ for all } k \in \mathscr{F}(M)$$

we call X_f the *Hamiltonian vector field* of f.

This definition coincides with Definition 2.8 when the Poisson manifold is symplectic. Now given $H \in \mathscr{F}(M)$, the Hamilton's equation is $\dot{z} = X_H(z)$. By the chain rule, for any $f \in \mathscr{F}(M)$, we have $\frac{df(z(t))}{dt} = df(X_H(z)) = \{f, H\}$. Thus the Hamilton's equation can be written in the following three (equivalent) ways:

1. $\dot{z} = X_H(z)$;
2. $\dot{f} = df(X_H(z))$ for all $f \in \mathscr{F}(M)$;
3. $\dot{f} = \{f, H\}$ for all $f \in \mathscr{F}(M)$.

The following is a basic fact about Hamiltonian system.

Proposition 2.7. *Let $\phi_t : M \to M$ be the flow of the Hamilton's equation $\dot{z} = X_H(z)$. Then*

1. ϕ_t is a Poisson map;
2. $H \circ \phi_t = H$ (conservation of energy).

The cotangent bundle

We now come to one of the most important constructions of symplectic manifold, namely, the cotangent bundle.

Consider an n dimensional manifold Q and its cotangent bundle T^*Q. Let (q_i) be the coordinates on Q and (q^i, p_j) the induced coordinate on T^*Q. More precisely, for any $\omega \in T^*Q$, $p_j(\omega) = \omega\left(\frac{\partial}{\partial q^j}\right)$. Next, define a 2-form ω on T^*Q by

$$\omega = \sum_{i=1}^{n} dq^i \wedge dp_i \tag{2.38}$$

One can check that ω is well-defined (coordinate-free). As an alternative, we consider the 1-form

$$\Theta = \sum_{i=1}^{n} p_i dq^i$$

and $\omega = -d\Theta$. Thus, it suffices to show that Θ is coordinate-free. (The notation $p_i dq^i$ is a little ambiguous since it may also be understood as a dual vector in T^*Q instead of in T^*T^*Q! We adopt this notation anyway since it is standard. The function p_i in front of dq^i should remind the reader that it is a dual vector in T^*T^*Q.) To show that Θ is well-defined, let $(\tilde{q}^i, \tilde{p}_j)$ be another coordinate, where $p_i = \tilde{p}_j \frac{\partial \tilde{q}_j}{\partial q^i}$. Since $dq^i = \sum_{j=1}^{n} \frac{\partial q^i}{\partial \tilde{q}^j} d\tilde{q}^j$, we have

$$\Theta = \sum_{i=1}^{n} p_i dq^i = \sum_{i,j,r=1}^{n} \tilde{p}_j \frac{\partial \tilde{q}^j}{\partial q^i} \frac{\partial q^i}{\partial \tilde{q}^r} d\tilde{q}^r = \sum_{j,r=1}^{n} \delta^r_j \tilde{p}_j d\tilde{q}^r = \sum_{i=1}^{n} \tilde{p}_i d\tilde{q}^i$$

The 1-form Θ is the *tautological form* or *Liouville 1-form* and the 2-form $\omega = -d\Theta$ is the *canonical symplectic form*. To summarize:

Proposition 2.8. *Let Q be a smooth manifold. Then (T^*Q, ω) is a symplectic manifold, where ω is defined as (2.38).*

Let's calculate in coordinates. Let $H \in \mathscr{F}(T^*Q)$, and $X_H = X^i \frac{\partial}{\partial q^i} + X^{n+i} \frac{\partial}{\partial p_i}$ (we use Einstein summation natation: the repeated index is summed). By definition, for any $v = v^i \frac{\partial}{\partial q^i} + v^{n+i} \frac{\partial}{\partial p_i}$, there holds

$$dq^i \wedge dp_i (X_H, v) = dH(v),$$

or

$$X^i v^{n+i} - X^{n+i} v^i = \frac{\partial H}{\partial q^i} v^i + \frac{\partial H}{\partial p^i} v^{n+i},$$

from which it follows that

$$X^i = \frac{\partial H}{\partial p^i}, \quad X^{n+i} = -\frac{\partial H}{\partial q^i}$$

Hence

$$X_H(q,p) = \sum_{i=1}^{n} \left(\frac{\partial H}{\partial p^i} \frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i} \frac{\partial}{\partial p_i} \right).$$

And the Hamilton's equation reads

$$\dot{q}^i = \frac{\partial H}{\partial p^i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q^i}.$$

Further more, for $H_1, H_2 \in \mathscr{F}(T^*Q)$, the Poisson bracket reads

$$\{H_1, H_2\} = dH_1(X_{H_2}) = \sum_{i=1}^{n} \left(\frac{\partial H_1}{\partial q^i} \frac{\partial H_2}{\partial p^i} - \frac{\partial H_1}{\partial p^i} \frac{\partial H_2}{\partial q_i} \right).$$

In the context of canonical symplectic manifold T^*Q, definition 2.8 defines a map $f \mapsto X_f$ from $\mathscr{F}(T^*Q)$ to $\mathscr{X}(T^*Q)$, where $\mathscr{X}(T^*Q)$ stands for the set of smooth vector fields on T^*Q. We define a map from $\mathscr{X}(Q)$ to $\mathscr{F}(T^*Q)$.

Definition 2.10 (Momentum function)**.** Given a smooth vector field X on Q, i.e., $X \in \mathscr{X}(Q)$, define the *momentum function of* X as as the unique function $P_X \in \mathscr{F}(T^*Q)$ satisfying

$$P_X(q,p) = p(X_q)$$

for all $p \in T_q^*Q$.

In coordinates, the momentum function reads

$$P_X(p_i dq_i) = p_i X^i$$

where $X = X^i(q) \frac{\partial}{\partial q^i}$.

The momentum function has the following important property: let X, Y be two smooth vector fields, then

$$\{P_X, P_Y\} = -P_{[X,Y]} \tag{2.39}$$

this property is called the *anti-homomorphism* (from the Lie bracket to the Poisson bracket) of the momentum function.

Exercise 2.6. Verify the formula (2.39).

With these preparations, we are ready to state the maximum principle on manifolds.Consider the optimal control problem in Mayer form. The maximum principle can be stated as follows (we omit the transversal condition as they are the same as before).

Theorem 2.4. *Equip T^*M with the canonical symplectic structure. Let $u_* : [t_0, t_f] \to U$ be the optimal control and $x_* : [t_0, t_f] \to M$ the optimal process of the Mayer problem. Define a Hamiltonian $H : T^*M \times U \to \mathbb{R}$ by*

$$H(x, \lambda, u) = \langle \lambda, f(x, u) \rangle .$$

Then there exists a curve $\Lambda : [t_0, t_f] \to T^*M$, with $\Lambda(t) = (x_*(t), \lambda(t))$ for $t \in [t_0, t_f]$ such that Λ is the solution to the Hamilton's equation

$$\dot{\Lambda} = X_H(\Lambda)$$

Moreover, along Λ the Hamiltonian H satisfies the maximum principle

$$H(x_*(t), \lambda(t), u_*(t)) = \max_{u \in U} H(x_*(t), \lambda(t), u).$$

2.4.2 Nonholonomic systems and sub-Riemannian geometry

There is a large class of control systems which can be written as

$$\dot{x} = \sum_{i=1}^{m} u_i f_i(x) \tag{2.40}$$

where $x \in \mathbb{R}^n$, $m < n$, f_i some C^1 vector fields and $u_i(t) \in U \subseteq \mathbb{R}$, $\forall\, t \geq 0$ the inputs. We assume that the input space U is symmetric in the sense that $u_i \in U$ implies $-u_i \in U$. We call (2.40) a *kinematic control system* or a *control system without drift*. The term kinematic control system originates from mechanical systems, which is in contrast with dynamics control system, where the system is of second order and the input (force) is imposed on the acceleration. Thus we can think of (2.40) as controlling directly the velocity of a mechanical system. We have already seen example of kinematic control system in Section 2.2.2 – Dido's problem and Section 2.2.7 – Dubins car.

For system (2.40), we are interested in those that are completely controllable. According to Chow-Rashevskii's Theorem (Theorem 2.2), this occurs when the set of vector fields $\{f_1, \cdots, f_m\}$ is Lie bracket generating, or the linear span of the set of vector fields of the form $[[\cdots [f_i, f_j], f_k], \cdots, f_\ell]$ has rank n, for more details, see Section 2.2.7. When the system (2.40) satisfies the Lie bracket generating property, we call it a *kinematic holonomic system*, or in short *holonomic system* (we are not going to cover mechanical holonomic systems, interested readers are referred to [3]). One can verify that both Dido's system (the one with state (x, y, z)) and Dubins car are holonomic systems and hence are completely controllable.

The optimal control problem regarding nonholomic system (2.40) that we are going to study is to minimize the following cost:

$$J_0(u) = \int_0^{t_f} \sqrt{\sum_{i=1}^{m} u_i^2(t)}\, \mathrm{d}t$$

for $x(0)$ and $x(t_f)$ fixed.[3] This optimal control problem is the content of the so called *sub-Riemannian geometry*. To gain some insight, let $\gamma : [0, t_f] \to \mathbb{R}^n$ be a state trajectory of the system (2.40) under some control input $u(t)$ – such a curve is called a *horizontal curve* – then we define the length of the curve γ as

$$\ell(\gamma) = \int_0^{t_f} ||\gamma'(t)||\, \mathrm{d}t$$

[3] For this problem, the Euler-Lagrangian equation is degenerate as L. Thus the usual calculus of variation does not tell us useful information.

where the "norm" $||\gamma'(t)||$ is defined as $\sqrt{\sum_{i=1}^m u_i^2(t)}$. The curve with the minimal length is called a *geodesic*. Thus, the optimal control problem is equivalent to finding the *geodesic* between two given points. Clearly, when $m = n$, sub-Riemannian geometry becomes Riemannian geometry.

We remark that J_0 is invariant under time reparametrization in the following sense: if $t = t(\tau)$ such that $t(0) = 0$ and $dt/d\tau > 0$, then

$$J_0' = \int_0^{\tau_f} \sqrt{\sum_{i=1}^m \tilde{u}_i^2(\tau)} \mathrm{d}\tau = J_0$$

where $t(\tau_f) = t_f$ and $\tilde{u}_i$ is the the new input for the system $\frac{dx}{d\tau} = \sum_{i=1}^m \tilde{u}_i(\tau) f_i(x(t(\tau))$. To see this, calculate directly: $\frac{dx}{dt} = \frac{dx}{d\tau}\frac{d\tau}{dt} = \sum_{i=1}^m u_i(t(\tau)) f_i(x(t(\tau))$, from which we deduce that $\tilde{u}_i(\tau) = u_i(t(\tau))\frac{dt}{d\tau}$. On the other hand, by change of variable, $J_0 = \int_0^{\tau_f} \sqrt{\sum_{i=1}^m u_i^2(t(\tau))}\frac{dt}{d\tau}\mathrm{d}\tau = \int_0^{\tau_f} \sqrt{\sum_{i=1}^m [u_i(t(\tau))\frac{dt}{d\tau}]^2}\mathrm{d}\tau = J_0'$. In particular, choose $\tau(t) = \int_0^t \sqrt{\sum_{i=1}^m u_i^2(s)}\mathrm{d}s$, then $\sum_{i=1}^m \tilde{u}_i^2(\tau) = 1$. Thus we may conclude that for any input u, one can find another input $\tilde{u}$ such that $\sum_{i=1}^m \tilde{u}_i^2(t)$ is constant and $\tilde{u}$ generates the same cost as u.

For this reason, we claim that:

Claim. Let t_f be fixed, then u minimizes J_0 *if and only if* u minimizes

$$J = \frac{1}{2}\int_0^{t_f} \sum_{i=1}^m u_i^2(t)\mathrm{d}t.$$

To see this, by Cauchy-Schwarz inequality, we have $J_0^2(u) \le 2t_f J(u)$ with equality holds if and only $\sum_{i=1}^m u_i^2(t)$ is constant for all $t \in [0, t_f]$.[4] Suppose now that u minimizes J, and let $\hat{u}$ be any other admissible controller, by previous discussion, there exists $\tilde{u}$ satisfying $J_0(u) = J_0(\tilde{u})$ and $\sum_{i=1}^m \tilde{u}_i^2(t) =$ constant. Hence $2t_f J(u) \le 2t_f J(\tilde{u}) = J_0^2(\tilde{u}) = J_0^2(\hat{u})$. Thus $J_0(u) \le J_0(\hat{u})$, as desired. Conversely, suppose that u minimizes J_0, then we can find $\tilde{u}$ such that $\sum_{i=1}^m \tilde{u}_i^2(t) =$ constant, then for any other $\hat{u}$,

$$J(u) = \frac{1}{2t_f}J_0^2(u) = \frac{1}{2t_f}J_0^2(\tilde{u}) \le J(\hat{u}),$$

as desired. As we will see, it is analytically more convenient to work with J rather than J_0, which we adopt hereafter.

Let us now apply the maximum principle to the optimal control problem of nonholomic systems. The Hamiltonian is (we assume there is no abnormal extremal):

$$H(x, u, p) = \sum_{i=1}^m u_i(p^\top f_i) - \frac{1}{2}u^\top u$$

where we have denoted $u = (u_1, \cdots, u_m)^\top$. For convenience, denote $H_i(x, p) = p^\top f_i(x)$. Using the momentum function introduced in previous subsection, $H_i(x, p)$ can also be written as $P_{f_i}(p)$. Then

$$u_i^* = (\arg\max_u H(x, u, p))_i = H_i$$

and along the optimal trajectory,

$$H(x^*(t), u^*(t), p^*(t)) = \frac{1}{2}\sum_{i=1}^m H_i^2(x^*(t), u^*(t)) = \text{constant}.$$

[4] Cauchy-Schwarz inequality: for $f, g \in L^2$, we have

$$\int fg \le \sqrt{\int f^2}\cdot\sqrt{\int g^2}$$

with equality holds if and only if $f = \lambda g$ for some constant λ.

By the way, this again justifies that minimizing J is equivalent to minimizing J_0 since along the optimal solution $\sum_{i=1}^m u_i^{*2} = \sum_{i=1}^m H_i^2$ is a constant.

Unlike in the usual case, we do not write the costate equation, instead, we propose another more useful type of equations which are equivalent to the costate equation. Since by Theorem 2.4, (x, p) satisfies the Hamiltonian equation, thus H_i satisfies the Poisson equation:

$$\dot{H}_i = \{H_i, H\}, \quad \forall i = 1, \cdots m.$$

To see how these equation can be used, we revisit Dido's problem.

Revisit of Dido's problem

Remember that the equation for Dido's problem reads (the Heisenberg system)

$$\frac{d}{dt}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = u_1 f_1 + u_2 f_2$$

where $f_1 = [1, 0, -\frac{1}{2}y]$, $f_2 = [0, 1, \frac{1}{2}x]$. Let us consider the equivalent cost

$$J = \frac{1}{2}\int_0^{t_f} u_1^2 + u_2^2 \mathrm{d}t.$$

If we use the costate equation, what we get is

$$\begin{aligned} \dot{p}_1 &= -\frac{1}{2}p_3(p_2 + \frac{1}{2}p_3 x) \\ \dot{p}_2 &= \frac{1}{2}p_3(p_1 - \frac{1}{2}p_3 y) \\ \dot{p}_3 &= 0 \end{aligned}$$

However, the Poisson equations read

$$\begin{cases} \dot{H}_1 = \{H_1, H_2\}H_2 \\ \dot{H}_2 = -\{H_1, H_2\}H_1 \end{cases} \tag{2.41}$$

which is much simpler. Denote $H_3 = \{H_1, H_2\}$ and $f_3 = [f_1, f_2]$, then

$$\begin{aligned} \dot{H}_3 = \{H_3, H\} &= \{\{H_1, H_2\}, H\} \\ &= \{-P_{f_3}, \frac{1}{2}(P_{f_1})^2 + \frac{1}{2}(P_{f_2})^2\} \\ &= P_{f_1}P_{[f_3,f_1]} + P_{f_2}P_{[f_3,f_2]} = 0 \end{aligned}$$

since $[f_3, f_1] = [f_3, f_2] = 0$, where $P_\cdot$ is the momentum function, see Definition 2.10. Thus the Poisson equation can be rewritten as

$$\begin{cases} \dot{H}_1 = H_3 H_2 \\ \dot{H}_2 = -H_3 H_1 \\ \dot{H}_3 = 0 \end{cases} \tag{2.42}$$

which look more promising than the costate equation. Moreover, the system dynamics along the optimal trajectory has the form

$$\begin{cases} \dot{x} = H_1 \\ \dot{y} = H_2 \\ \dot{z} = \dfrac{1}{2}(-yH_1 + xH_2) \end{cases} \tag{2.43}$$

Claim. Let us now try to solve the equations (2.41, 2.42). The first observation is that H_3 is constant. Define $w = H_1 + iH_2$, where $i^2 = -1$. Then we find $\dot{w} = -H_2H_3 + iH_1H_3 = iH_3(H_1 + iH_2) = iH_3w$, thus $w(t) = e^{iH_3t}w(0)$. On the other hand $\frac{d}{dt}(x + iy) = w$, thus (remember $x(0) = y(0) = 0$),

$$x(t) + iy(t) = w(0)\frac{e^{iH_3t} - 1}{iH_3} = -i\frac{w(0)}{H_3} - i\frac{w(0)e^{iH_3t}}{H_3}$$

write $-iw(0)/H_3 = r_0e^{i\theta_0}$ for $r_0 \geq 0$, then the above equation can be rewritten as

$$(x(t) + c_1) + i(y(t) + c_2) = r_0e^{i(H_3t+\theta_0)}$$

for some real constants c_1 and c_2. Thus $(x(t), y(t))$ lies on a circle.

Dubins car

Dubins car model has the same form as Dido's problem for with

$$f_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \\ 0 \end{bmatrix}, \quad f_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The only thing we need to modify in equation (2.42) is the third line. First, we calculate $f_3 = [\sin\theta, -\cos\theta, 0]$, it follows that $[f_3, f_1] = 0$ and $[f_3, f_2] = -f_1$. Thus

$$\dot{H}_3 = P_{f_1}P_{[f_3,f_1]} + P_{f_2}P_{[f_3,f_2]} = -H_2H_1.$$

Combining together, the Poisson equation reads

$$\begin{aligned} \dot{H}_1 &= H_3H_2 \\ \dot{H}_2 &= -H_3H_1 \\ \dot{H}_3 &= -H_1H_2 \end{aligned}$$

this equation however, is much harder to solve than that of the Dido's problem.

Note carefully that although $H_1^2 + H_2^2$ is a constant along the optimal trajectory, $H_1^2 + H_2^2 + H_3^3$ needn't be .

2.5 Appendix: Maximum Principle of Discrete Time Systems

2.5.1 Fixed control region

Historically, the dynamic principle was first developped for continuous time systems. This however, doesn't mean that MP for discrete time system is harder. We will see now for fixed control region, the problem is in fact easy.

Consider the discrete time system

$$x_{k+1} = f_k(x_k, u_k), \tag{2.44}$$

with input sequence

$$(u_1, \cdots, u_N)$$

and resulting process

$$(x_1, \cdots, x_N, x_{N+1}).$$

Assume

$$u_t \in U_t \subset \mathbb{R}^m, t = 1, \cdots, N \tag{2.45}$$

in which U_t can be state dependent. But in this subsection, we assume that $U_t \equiv U$ is fixed. The state is under constraint

$$x_t \in M_t \subset \mathbb{R}^n, \ \ t = 1, \cdots, N+1. \tag{2.46}$$

Problem 1. The optimal control problem of the discrete time system (2.44) with cost function

$$J(u) = \varphi(x_{N+1}) + \sum_{k=1}^{N} L_k(x_k, u_k)$$

consists in finding a policy

$$u^* = (u_1^*, u_2^*, \cdots, u_N^*)$$

with $u_i^* \in U$ and N fixed, such that $x_t^* \in M_t$ for $t = 1, \cdots, N+1$, and $J(u^*)$ attains a minimum. We say that the problem is in

- *Mayer form* if $L = 0$,
- *Lagrange form* if $\varphi = 0$,
- *Bolza form* if neither L nor φ is zero.

Like in the continuous time case, we consider only the Mayer form as the other two forms are equivalent to it.

For discrete time system, the sets Ω_0, Ω_2 and the tents K_0, K_2 are the same as before. The only difference is the reacheability region Ω_1 and its tent K_1. To calculate K_1, define a variation similar to the needle variation of continuous signal at the instant $i \in \{1, \cdots, N\}$:

$$u_k^{(i,\varepsilon)} = \begin{cases} u_k^* + \varepsilon u, & k = i, \\ u_k^*, & \text{otherwise} \end{cases}$$

where $u \in U$. Let $x_{k+1}^{(i,\varepsilon)}$ be the solution to

$$\begin{aligned} x_{k+1}^{(i,\varepsilon)} &= f_k(x_k^{(i,\varepsilon)}, u_k^{(i,\varepsilon)}), \ \ k = \{i, \cdots, N\}. \\ x_i^{(i,\varepsilon)} &= x_i^*. \end{aligned}$$

Then

$$\left.\frac{\partial x_k^{(i,\varepsilon)}}{\partial \varepsilon}\right|_{\varepsilon=0}, \ \ k \in \{i+1, \cdots, N\}$$

is the solution to the discrete time variational system

$$v_{k+1} = \frac{\partial f_k}{\partial x}(x_k^*, u_k^*)v_k, \ \ k \in \{i, \cdots, N\}$$
$$v_{i+1} = \frac{\partial f_i}{\partial u_i}(x_k^*, u_k^*)u$$

Similar to the continuous time case, we call v_{N+1} a *deviation vector* under the variation. Then the convex cone $K_1^{''}$ generated by the deviation vectors is a tent of the reacheability region. By Theorem 2.3, there exist three covectors $a_0 \in K_0^*$, $a_1 \in \left(K_1''\right)^*$, $a_2 \in K_2^*$, not all zero, such that $a_0 + a_1 + a_2 = 0$ and $a_0 = \lambda \operatorname{grad} \varphi(x_{N+1})$ with $\lambda \geq 0$. To characterize K_1'', introduce the *adjoint system*

$$p_k = p_{k+1}\frac{\partial f_k}{\partial x_k}(x_k^*, u_k^*), \ \ k \in \{i+1, \cdots, N\}$$
$$p_{N+1} = a_1$$

The following lemma is obvious, which is the discrete time version of Lemma 2.4:

Lemma 2.5. *Consider the system*

$$x_{k+1} = A_k x_k$$
$$p_k = p_{k+1} A_k$$

where $x_k \in \mathbb{R}^n$, $p_k \in (\mathbb{R}^n)^*$. Then $p_k x_k = p_{k'} x_{k'}$ for all integers k, k'.

Then

$$0 \leq a_1 v_{N+1} = p_{N+1} v_{N+1} = p_{i+1} v_{i+1} = p_{i+1}\frac{\partial f_i}{\partial u_i}(x_i^*, u_i^*)u, \ \ \forall u \in U.$$

which implies

$$p_{k+1}\frac{\partial f_k}{\partial u_k}(x_k^*, u_k^*) = 0, \ \forall k \in \{1, \cdots, N\},$$

since U is open. Since x_1 is not fixed, we can take $\pm v_1 \in T_{x_1^*} M_0$. Then $0 \leq a_1 v_{N+1} = p_1 v_1 \leq 0$, or $p_1 v_1 = 0$, which is equivalent to

$$p_1 \perp M_0$$

The transversal condition is

$$\lambda \operatorname{grad} \varphi(x_{N+1}) + p_{N+1} \perp M_1$$

with $\lambda \leq 0$. If the terminal state is free, i.e., $M_1 = \mathbb{R}^n$, then $p_{N+1} = -\lambda \operatorname{grad}(x_{N+1})$.

2.5.2 Variable control region

In this subsection we consider the Mayer problem with $J(u) = J(x_{N+1})$.

Let

$$\Phi_t(x(t)) = \bigcup_{u \in U_t} \{f_t(x(t), u)\} \subset \mathbb{R}^m, \ t = 1, \cdots, N.$$

Assume that the sets $\Phi_t(x)$ are compact, convex and continuously dependent on $x \in \mathbb{R}^n$ for every $t = 1, \cdots, N$. We say a trajectory

$$(x(1), \cdots, x(N))$$

admits a local section if for every $t = 1, \cdots, N$, there is a smooth function $\sigma_t : U \subset \mathbb{R}^m \to \mathbb{R}^m$, where U is a neighbourhood of $x(t)$, such that

$$\sigma_t(x) \in \Phi_t(x),\ \forall x \in U \text{ and } \sigma_t(x(t)) = x(t+1) = f_t(x(t), u(t))$$

Introduce the following notations for each $\theta \in \{1, \cdots, N\}$ in which $n = (N+1)m$:

$$\begin{aligned}
z &= (x_1, \cdots, x_{N+1}) \in \mathbb{R}^n, \text{ with } x_\theta \in \mathbb{R}^m,\ \theta \in \{1, \cdots, N+1\} \\
\Xi_\theta &= \{z \in \mathbb{R}^n : x_{\theta+1} \in \Phi_\theta(x_\theta)\},\ \theta \in \{1, \cdots, N\} \\
\Omega_\theta^* &= \{z \in \mathbb{R}^n : x_\theta \in M_\theta\},\ \theta \in \{1, \cdots, N+1\} \\
P_\theta &: \text{ a tent of } M_\theta \text{ at } x_\theta,\ \theta \in \{1, \cdots, N+1\} \\
P_\theta^* &= \{z \in \mathbb{R}^n : x_\theta \in P_\theta\} \text{ (then } P_\theta^* \text{ is a tent of } \Omega_\theta^*),\ \theta \in \{1, \cdots, N+1\}
\end{aligned}$$

Assume that

$$\bar{z} = (\bar{x}_1, \cdots, \bar{x}_{N+1})$$

is the optimal process.

With these notations, the problem of finding an optimal trajectory for the system (2.44) reduces to the problem of minimizing $J(x(N+1))$ on the set

$$\Sigma = \left(\bigcap_{k=1}^{N} \Xi_k\right) \cap \left(\bigcap_{k=1}^{N+1} \Omega_k^*\right).$$

Since the tents of Ω_θ^* are known as P_θ^* for $\theta = 1, \cdots, N+1$, it remains to calculate the tents of Ξ_θ. Define

$$\begin{aligned}
Q_\theta = \left\{ \bar{z} + \delta z : \bar{x}_{\theta+1} + \delta x_{\theta+1} - \frac{\partial \sigma_\theta(\bar{x}_\theta)}{\partial x} \delta x_\theta \in L_\theta \right\}, \\
\theta = 1, \cdots, N
\end{aligned}$$

where

$$L_\theta : \text{ supporting cone of } \Phi_\theta(\bar{x}_\theta) \text{ at } \bar{x}_{\theta+1},\ \theta \in \{1, \cdots, N\}$$

We claim that Q_θ is a tent of Ξ_θ. Assume this fact, we would deduce the following.

There is a number $\psi_0 \le 0$ and dual vectors

$$\begin{aligned}
a_t &\in D(P_t^*) \subset \mathbb{R}^{m(N+1)},\ \ t = 1, \cdots, N+1 \\
b_t &\in D(Q_t) \subset \mathbb{R}^{mN},\ \ t = 1, \cdots, N
\end{aligned}$$

such that

$$\psi_0 \text{grad}_z J(\bar{x}_{N+1}) + \sum_{t=1}^{N+1} a_t + \sum_{t=1}^{N} b_t = 0 \tag{2.47}$$

If we write

$$\begin{aligned}
a_t &= (a_t^1, \cdots, a_t^{N+1}) \\
b_t &= (b_t^1, \cdots, b_t^N)
\end{aligned}$$

then $a_t^i \neq 0$ only when $i = t$, and $b_t^i \neq 0$ only when $i = t,\ t+1$. By assumption $\langle b_t, \delta z\rangle \geq 0$ for any $\delta z \in Q_t$. Take a δz satisfying $\delta x_{t+1} = \frac{\partial \sigma_t(\bar{x}_t)}{\partial x}\delta x_t$ and $\delta x_i = 0$ for $i \neq t$, then

$$b_t^t \delta x_t + b_t^{t+1}\frac{\partial \sigma_t(\bar{x}_t)}{\partial x}\delta x_t \geq 0,\ \ \delta x_t \in \mathbb{R}^m$$

Let $\varphi_t = b_t^t$ and $\psi_t = b_t^{t+1}$, the above implies

$$\varphi_t + \psi_t \frac{\partial \sigma_t(\bar{x}_t)}{\partial x} = 0.$$

Hence the condition (2.47) can be written as

$$\begin{aligned}
\psi_{t-1} &= -\lambda_t + \psi_t \frac{\partial \sigma_t(\bar{x}_t)}{\partial x},\ \ t = 1, \cdots, N \\
\psi_0 &= 0 \\
\psi_N &= -\lambda_{N+1} - \psi_0 \mathrm{grad}_x J(\bar{x}_{N+1})
\end{aligned}$$

since

$$\begin{aligned}
&\mathrm{grad}_z J(\bar{x}_{N+1}) = (0, \cdots, 0, \mathrm{grad}_x J(\bar{x}_{N+1})) \\
&\sum_{t=1}^{N+1} a_t = (\lambda_1, \cdots, \lambda_{N+1}) \\
&\sum_{t=1}^{N} b_t = \left(\varphi_1, \psi_1 + \varphi_2, \cdots, \psi_{N-1} + \varphi_N, \psi_N\right)
\end{aligned}$$

where we have denoted $\lambda_t = a_t^t$.

Further, choose δz is such a way that $x_{t+1} = \bar{x}_{t+1} + \delta x_{t+1} \in L_t$ and $\delta x_i = 0$ for $i \neq t+1$. Therefore

$$0 \leq \psi_t \delta x_{t+1}$$

In other words, the function $\eta_t(v) = \psi_t v$ achieves minimum at the point $\bar{x}_{t+1}$. Since $\Phi_t(\bar{x}_t)$ is contained in L_t, it follows that

$$\psi_t \bar{x}_{t+1} = \min_{x \in \Phi_t(\bar{x}_t)} \psi_t x = \min_{u \in U_t} \psi_t f_t(\bar{x}_t, u),\ \ t = 1, \cdots, N$$

Thus we are left to show that Q_θ is a tent of Ξ_θ.

Choose $z \in Q_\theta$ arbitrary ($x_{\theta+1}$ is not necessarily in $\Phi_\theta(x_\theta)$). Define

$$\varphi_\theta(z) \text{ the projection of } x_{\theta+1} \text{ to } \Phi_\theta(x_\theta),$$

and

$$\Psi_\theta(z) = (x_0, \cdots, x_\theta, \varphi_\theta(z), x_{\theta+2}, \cdots, x_N) \in \mathbb{R}^n$$

Then since $\varphi_\theta(z) \in \Phi_\theta(x_\theta)$, $\Psi_\theta(z) \in \Xi_\theta$ for any $z \in \mathbb{R}^n$. It remains to show

$$\Psi_\theta(z) = z + o(z - \bar{z})$$

or

$$\varphi_\theta(z) = x_{\theta+1} + o(z - \bar{z}).$$

Consider the point

$$s_\theta(x_\theta) = \sigma(x_\theta) + \delta x_{\theta+1} - \frac{\partial \sigma_\theta(\bar{x}_\theta)}{\partial x}\delta x_\theta.$$

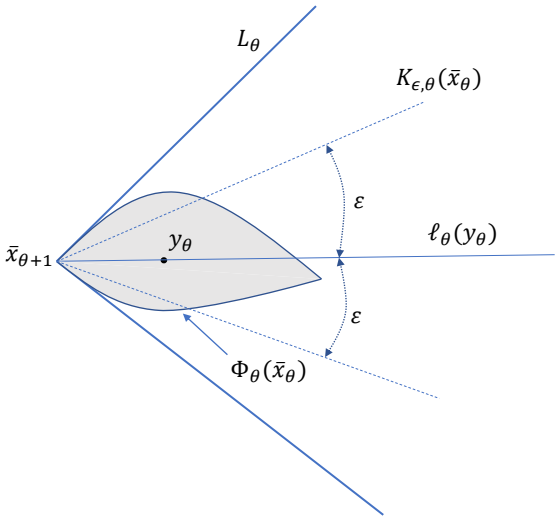

Figure 2.16: Illustration I of the proof.

By definition of Q_θ, $s_\theta(x_\theta) \in L_\theta$, the supporting cone of $\Phi_\theta(\bar{x}_\theta)$ at $\bar{x}_{\theta+1}$. Since z is close to $\bar{z}$, we have

$$\begin{aligned} s_\theta(\delta x_\theta) &= \sigma(\bar{x}_\theta + \delta x_\theta) + \delta x_{\theta+1} - \frac{\partial \sigma_\theta(\bar{x}_\theta)}{\partial x}\delta x_\theta \\ &= \sigma(\bar{x}_\theta) + \delta x_{\theta+1} + o(\delta x_\theta) \\ &= \bar{x}_{\theta+1} + \delta x_{\theta+1} + o(\delta x_\theta) \\ &= x_{\theta+1} + o(\delta x_\theta). \end{aligned}$$

Suppose now that $s_\theta \in \Phi_\theta(x_\theta)$, then the conclusion would follow as $|\varphi_\theta(z) - x_{\theta+1}| \le |s_\theta - x_{\theta+1}|$ since $\varphi_\theta(z)$ is the projection of $x_{\theta+1}$ to $\Phi_\theta(x_\theta)$. Unfortunately, s_θ may not be in $\Phi_\theta(x_\theta)$, therefore, it should be replaced by some other point s'_θ. For this, we notice that $s_\theta(\bar{x}_\theta)$ is in L_θ. We draw a ray emanating from $\bar{x}_{\theta+1}$ passing through this point $(s_\theta(\bar{x}_\theta))$ (see Figure 2.16), and then the rays emanating from $\bar{x}_{\theta+1}$ with angle $\varepsilon > 0$ (sufficiently small) form a cone, which we denote as $K_{\varepsilon,\theta}(\bar{x}_\theta)$ and that

$$\mathrm{Int} K_{\varepsilon,\theta}(\bar{x}_\theta) \cap \Phi_\theta(\bar{x}_\theta) \neq \varnothing.$$

In the similar way, one can define a cone at $\sigma_\theta(x_\theta)$ with the direction of $s_\theta(x_\theta)$ as axis and angle radius ε, see Figure 2.17. By continuity and compactness, one can then show that

$$\mathrm{Int} K_{\varepsilon,\theta}(x_\theta) \cap \Phi_\theta(x_\theta) \neq \varnothing.$$

Now we project $s_\theta(x_\theta)$ to an interior point of $\Phi_\theta(x_\theta)$, say s'_θ, and

$$|s_\theta(x_\theta) - s'_\theta| \le \left| \delta x_{\theta+1} - \frac{\partial \sigma_\theta(\bar{x}_\theta)}{\partial x}\delta x_\theta \right| \sin(\varepsilon)$$

The conclusion follows by noticing that

$$|\varphi_\theta(z) - x_{\theta+1}| \le |s'_\theta - x_{\theta+1}| \le |s'_\theta - s_\theta(x_\theta)| + |s'_\theta(x_\theta) - x_{\theta+1}| = O(\varepsilon) + o(\delta x_\theta)$$

and $\varepsilon > 0$ is arbitrary.

To summarize, we have proven the following theorem.

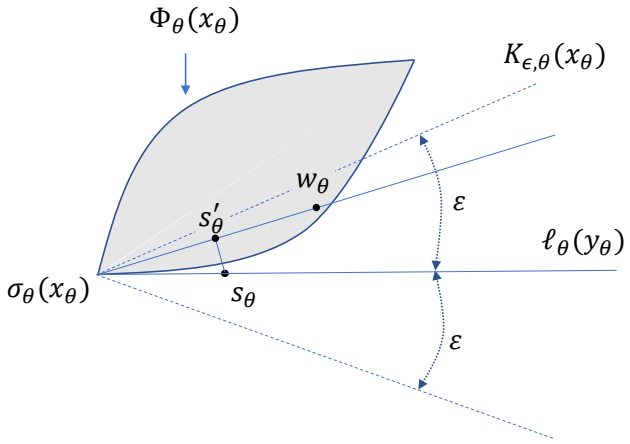

Figure 2.17: Illustration II of the proof.

Theorem 2.5. *Consider the control system (2.44), with state constraint (2.46) and input constraint (2.45). Assume that $\Phi_t(x)$ are compact, convex and continuously dependent on $x \in \mathbb{R}^m$ for all $t = 1, \cdots, N$. Assume P_t is a tent of M_t for each $t = 1, \cdots, N$. Let $(x_1, \cdots, x_{N+1})$ be an optimal process under control sequence $(u_1, \cdots, u_N)$ which minimizes the cost $J(u) = J(x_{N+1})$. Then there is a number $\lambda_0 \geq 0$ and vectors $\psi_t \in (\mathbb{R}^m)^*$, $\lambda_t \in D(P_t)$ such that*

$$\psi_{t-1} = -\lambda_t + \psi_t \frac{\partial f_t(x_t, x_t)}{\partial x}, \quad t = 1, \cdots, N$$
$$\psi_0 = 0$$
$$\psi_N = -\lambda_{N+1} + \lambda_0 \text{grad}_x J(x_{N+1})$$

and

$$H(t, x_t, u_t) = \min_{u \in U_t} H(t, x_t, u).$$

where $H(t, x, u) = \psi_t f_t(x, u)$.

Remark 2.2. It is immediately to see that when the initial state x_1 is fixed, $x_2, \cdots, x_{N+1}$ are not constraint and $U_t = U$ is an open set, then the above condition reduces to

$$\psi_{t-1} = \psi_t \frac{\partial f_t(x_t, u_t)}{\partial x}, \quad t = 2, \cdots, N$$
$$\psi_N = \lambda_0 \text{grad}_x J(x_{N+1})$$
$$\psi_t \frac{\partial f_t(x_t, u_t)}{\partial u_t} = 0, \quad t = 1, \cdots, N$$

2.5.3 Discussions

Bolza form

Now we return to the general form of optimal control, i.e, the Bolza form

$$J(u) = \varphi(x_{N+1}) + \sum_{k=1}^{N} L_k(x_k, u_k).$$

To transform it into the Mayer form, let $y_{k+1} = y_k + \tilde{L}_k(x_k, u_k)$ where

$$\tilde{L}_k(x_k, u_k) = \begin{cases} L_k(x_k, u_k), & k = 1, \cdots, N-1 \\ L_N(x_N, u_N) + \varphi(f_N(x_N, u_N)) & k = N \end{cases}$$

Therefore we obtain an augmented system in $\mathbb{R}^{m+1}$:

$$\begin{aligned} x_{k+1} &= f_k(x_k, u_k) \\ y_{k+1} &= y_k + \tilde{L}_k(x_k, u_k) \end{aligned}$$

with $y_1 = 0$. Let

$$z_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix} \in \mathbb{R}^{m+1}$$

The cost function becomes

$$J(u) = \tilde{\varphi}(z_{N+1}) = y_{N+1}$$

Suppose that x_1 is fixed, invoking Theorem 2.5, there is a number $\lambda_0 \geq 0$, vectors $\psi_t = (\alpha_t, \beta_t) \in (\mathbb{R}^m)^* \times (\mathbb{R}^1)^*$, $\lambda_t \in D(P_t)$ such that

$$\begin{aligned} &(\alpha_{t-1}, \beta_{t-1}) = (\alpha_t, \beta_t) \begin{bmatrix} \frac{\partial f_t(x_t, u_t)}{\partial x} & 0 \\ \frac{\partial \tilde{L}_t(x_t, u_t)}{\partial x} & 1 \end{bmatrix}, \quad t = 2, \cdots, N \\ &(\alpha_N, \beta_N) = \lambda_0(0, 1) \\ &(\alpha_t, \beta_t) \begin{bmatrix} \frac{\partial f_t}{\partial u} \\ \frac{\partial L_t}{\partial u} \end{bmatrix} = 0. \end{aligned}$$

from which we see

$$\begin{aligned} \alpha_{N-1} &= \lambda_0 \left(\text{grad}_x \varphi(x_{N+1}) \frac{\partial f_N(x_N, u_N)}{\partial x} + \frac{\partial L_N(x_N, u_N)}{\partial x} \right) \\ &= \lambda_0 \text{grad}_x \varphi(x_{N+1}) \frac{\partial f_N(x_N, u_N)}{\partial x} + \lambda_0 \frac{\partial L_N(x_N, u_N)}{\partial x} \end{aligned}$$

Thus redefining $\alpha_N =: \lambda_0 \,\text{grad}\, \varphi(x_{N+1})$ we obtain

$$\begin{aligned} &\alpha_{t-1} = \alpha_t \frac{\partial f_t(x_t, u_t)}{\partial x} + \lambda_0 \frac{\partial L_t(x_t, u_t)}{\partial x} = \frac{\partial H_t}{\partial x}, \quad t = 2, \cdots, N \\ &\alpha_N = \lambda_0 \text{grad}_x \varphi(x_{N+1}), \\ &\alpha_t \frac{\partial f_t}{\partial u} + \lambda_0 \frac{\partial L_t}{\partial u} = \frac{\partial H_t}{\partial u} = 0, \quad t = 1, \cdots, N \end{aligned} \tag{2.48}$$

The Hamiltonian function is $H_k(x, y), = \alpha_k f_k(x, u) + \lambda_0(y + L_k(x, u))$. But since y is independent of u, one can also define $H_k(x, y, u) = \alpha_k f_k(x, u) + \lambda L_k(x, u)$ and the maximum condition becomes

$$H_k(x_k, y_k, u_k) = \min_{u \in U} H_k(x_k, y_k, u).$$

Connections to DP

To illustrate the connections to dynamic programming, we show that dynamic programming algorithm (1.5) and discrete time maximum principle (2.48) give the same optimal control law. We consider only the Mayer case as it is equivalent to Bolza form.

(DP ⇒ MP): Assume that

$$u_t(x) = \arg\min_u [J^*_{t+1}(f_t(x,u))]$$

then we have

$$J^*_t(x) = J^*_{t+1}(f_t(x, u_t(x)))$$

from which it follows

$$\begin{aligned}\frac{\partial J^*_t}{\partial x} &= \frac{\partial J^*_{t+1}}{\partial x}\left(\frac{\partial f_t}{\partial x} + \frac{\partial f_t}{\partial u_t}\frac{\partial u_t}{\partial x}\right)\\ &= \frac{\partial J^*_{t+1}}{\partial x}\frac{\partial f_t}{\partial x} + \frac{\partial J^*_{t+1}}{\partial x}\frac{\partial f_t}{\partial u_t}\frac{\partial u_t}{\partial x}\\ &= \frac{\partial J^*_{t+1}}{\partial x}\frac{\partial f_t}{\partial x}\end{aligned}$$

since $\frac{\partial J^*_{t+1}}{\partial u} = 0$. Letting $\lambda_0 = 1$, $\alpha_t = \frac{\partial J^*_{t+1}}{\partial x}$, we deduce

$$\begin{aligned}\alpha_{t-1} &= \alpha_t \frac{\partial f_t}{\partial x}\\ 0 &= \alpha_t \frac{\partial f_t}{\partial u_t}\end{aligned}$$

Let $H_t(x,u) = \frac{\partial J^*_{t+1}(x_{t+1})}{\partial x} f_t(x,u)$. Since $J^*_{t+1}(f_t(x_t,u_t)) \leq J^*_{t+1}(f_t(x_t,u))$ or $J^*_{t+1}(v_t) \leq J^*_{t+1}(v)$, $\forall v \in V_t = \cup_{u\in U_t}\{f_t(x_t,u)\}$ from which it follows

$$\frac{\partial J^*_{t+1}(x_{t+1})}{\partial x}(v_t - v) \leq 0, \ \forall v \in V_t$$

or $\frac{\partial J^*_{t+1}(x_{t+1})}{\partial x} v_t \leq \frac{\partial J^*_{t+1}(x_{t+1})}{\partial x} v$ (we have used the fact that V_t is convex). Hence $H_t(x_t,u_t) = \min_{u\in U_t} H_t(x_t,u)$.

(MP ⇒ DP) It is sufficient to notice that $\frac{\partial J^*_{t+1}(x_{t+1})}{\partial x}(v_t - v) \leq 0, \ \forall v \in V_t$ implies $J^*_{t+1}(v_t) \leq J^*_{t+1}(v)$, $\forall v \in V_t$.

Remark 2.3. Notice that in the discrete time maximum principle, we need the assumption of convexity, which is not the case for dynamic programming! Consider for example (a common case), when the input set U_t is only a finite set, then V_t won't be convex and the discrete time maximum principle does not say anything!

CHAPTER 3

OPTIMAL FILTERING AND STOCHASTIC OPTIMAL CONTROL

3.1 Stochastic calculus: a modern construction of stochastic integral

3.1.1 Motivations

Consider the system $\dot{x} = f(t, x(t))$ with a noise $v : \mathbb{R}_+ \to \mathbb{R}^n$

$$\dot{x} = f(t,x) + v, \quad x \in \mathbb{R}^n, \ t \geq 0$$

We sample the system under sample time Δt, and let $x_k = x(k\Delta t)$ for $k \in \mathbb{N}$.

Then

$$\begin{aligned} x_{k+1} &= x_k + \int_{k\Delta t}^{(k+1)\Delta t} f(s, x(s)) + v(s)ds \\ &= x_k + \int_{k\Delta t}^{(k+1)\Delta t} f(s, x(s))ds + \int_{k\Delta t}^{(k+1)\Delta t} v(s)ds \end{aligned}$$

For discrete time system, it is custumary to model a system with noise as

$$x_{k+1} = f(k\Delta t, x_k) + n_k \tag{3.1}$$

where n_k is a "white noise" in the sense that $n_k \sim N(0, \sigma^2)$ and that $n_1, \cdots, n_k, \cdots$ are independent. If the above is a sample system of the continuous time system, then the variance of the Gaussian variable n_k should be made to depend on the sample time since if

$$\int_{k\Delta t}^{(k+1)\Delta t} v(s)ds \sim N(0, \sigma^2)$$

then

$$\begin{aligned} \int_{k\Delta t}^{(k+2)\Delta t} v(s)ds &= \int_{k\Delta t}^{(k+1)\Delta t} v(s)ds + \int_{(k+1)\Delta t}^{(k+2)\Delta t} v(s)ds \\ &= n_k + n_{k+1} \sim N(0, 2\sigma^2) \end{aligned}$$

which can be viewed as n_j under sample time $2\Delta t$ for some j. Thus the variance of n_k should be proportional to the square root of the sample time. Hence we may assume $n_k \sim N(0, c\Delta t)$ for some $c > 0$. Now that

$$N(0, c\Delta t) \sim \int_{k\Delta t}^{(k+1)\Delta t} v(s)ds$$

it is reasonable to come up with a function $w : \mathbb{R}_+ \to \mathbb{R}^n$ with $\frac{dw(s)}{ds} = v(s)$ and that

$$\begin{aligned}\int_{k\Delta t}^{(k+1)\Delta t} v(s)ds &= \int_{k\Delta t}^{(k+1)\Delta t} \frac{dw(s)}{ds} ds = w((k+1)\Delta t) - w(k\Delta t) \\ &:= w_{k+1} - w_k \sim N(0, c\Delta t).\end{aligned}$$

Thus a reasonable noise model could be written as

$$\dot{x} = f(t, x(t)) + \frac{dw(t)}{dt}$$

where w should have the following property: $w(t_m) - w(t_{m-1})$, $w(t_{m-1}) - w(t_{m-2})$, $\cdots$ are independent Gaussian variables and that $w(t) - w(s) \sim N(0, c(t-s))$. By doing this, we are in fact constructing a stochastic process: namely, a Brownian motion. It is called a standard Brownian motion when $c = 1$. The above equation is usually written in the following form

$$dx(t) = f(t, x(t))dt + dw(t). \tag{3.2}$$

Suppose now that in (3.1), the variance of n_k is time dependent, namely $n_k \sim N(0, \sigma^2(k\Delta t)\Delta t)$ for some real function σ. Hence

$$\int_{k\Delta t}^{(k+1)\Delta t} v(t, x(t))dt \sim N(0, \sigma(k\Delta t, x(k\Delta t)))\Delta t)$$

which implies

$$\int_{k\Delta t}^{(k+1)\Delta t} v(t, x(t))dt = \sigma(k\Delta t, x(k\Delta t))[w((k+1)\Delta t) - w(k\Delta t)], \tag{3.3}$$

where w is the standard Brownian motion. Therefore it is suggestive to write

$$\int_{k\Delta t}^{(k+1)\Delta t} v(t, x(t))dt =: \int_{k\Delta t}^{(k+1)\Delta t} \sigma(t, x(t))dw(t)$$

when Δt is small. We underscore that the integral on the right hand side is not a Stieltjes integral as the Brownian motion does not have finite variation. Instead, the integral should be exactly understood as the right hand side of (3.3). The above discussions motivate to write down the following equation as an extension of (3.2) with a diffusion coefficient σ:

$$dx(t) = f(t, x(t))dt + \sigma(t, x(t))dw(t). \tag{3.4}$$

We call the equation (3.4) a stochastic differential equation (SDE). The solution to this SDE is written as

$$x(t) = x(s) + \int_s^t f(r, x(r))dr + \int_s^t \sigma(r, x(r))dw(r)$$

and the integral of the last term on the right hand side is understood as (3.3) when $|t - s|$ is small. Now since

$$\int_s^t \sigma(r, x(r))dw(r) = \sum_k \sigma(k\Delta t, x(k\Delta t))(w((k+1)\Delta t) - w(k\Delta t)) \tag{3.5}$$

the integral on the left hand side for arbitrary $s < t$ should be defined as the limit (in certain sense) of the right hand side when $\Delta t \to 0$.

Remark 3.1. We call the integral defined above *Itô integral* of σ. It is important to keep in mind that the Itô integral should always be evaluated at the left end points of the partitioned intervals, as in (3.3). One would obtain a totally different integral if one evaluate at the right end points, which is a clear difference between Rieman-Stietjes integral.

The rigorous constructions of Brownian motion and Itô integral are quite technical and out of the scope of this note. We refer the reader to the excellent text [11]. In the next part, we review some important notions from stochastic calculus, especially the Itô formula and the notion of Markov process.

3.1.2 Martingale

Throughout this subsection, we consider a probability space $(\Omega, \mathscr{F}, P)$ with Ω the sample space, $\mathscr{F}$ the sigmal algebra and P the probability measure.

Definition 3.1. A *filtration* on $(\Omega, \mathscr{F}, P)$ is a collection $(\mathscr{F}_t)_{0\le t\le\infty}$ indexed by $[0, +\infty]$ of sub-σ-algebras of $\mathscr{F}$, such that for every $0 \le s \le t$

$$\mathscr{F}_0 \subset \mathscr{F}_s \subset \mathscr{F}_t \subset \mathscr{F}_\infty \subset \mathscr{F}$$

Definition 3.2. A stochastic process $(X_t)_{t\ge0}$ with values in a measurable space $(E, \mathscr{E})$ ($\mathscr{E}$ is the σ-algebra on E) is called *adapted* (to $(\mathscr{F}_t)_{0\le t\le\infty}$) if for every $t \ge 0$, X_t is $\mathscr{F}_t$-measurable. This process is *progressive* if, for every $t \ge 0$, the mapping

$$(\omega, s) \mapsto X_s(\omega)$$

defined on $\Omega \times [0, t]$ is measurable w.r.t. the σ-algebra $\mathscr{F}_t \otimes \mathscr{B}([0, t])$. ($\mathscr{B}([0, t])$ is the Borel algebra on $[0, t]$)

Another important notion is stopping time.

Definition 3.3. A r.v. $T : \Omega \to [0, \infty]$ is a *stopping time* of the filtration $(\mathscr{F})_t$ if $\{T \le t\} \in \mathscr{F}_t$ for every $t \ge 0$. The σ-algebra of the past before T is then defined by

$$\mathscr{F}_T = \{A \in \mathscr{F}_\infty : \forall\, t \ge 0,\ A \cap \{T \le t\} \in \mathscr{F}_t\}.$$

As usual, for a r.v. X, we say that $X \in L^p$ if $E|X|^p < \infty$. Given a process $(X_t)_{t\ge0}$ adapted to $(\mathscr{F}_t)_{t\ge0}$, we adopt the notation $E_s[X_t]$ to mean $E[X_t|\mathscr{F}_s]$. Next we introduce one of the most important notions in stochastic calculus: martingale.

Definition 3.4. An adapted real-valued process $(X_t)_{t\ge0}$ such that $X_t \in L^1$ for every $t \ge 0$ is called

1. a *martingale* if, for every $0 \le s < t$, $E_s[X_t] = X_s$; (implies $EX_t = EX_s$])
2. a *supermartingale* if, for every $0 \le s < t$, $E_s[X_t] \le X_s$; (implies $EX_t \le EX_s$)
3. a *submartingale* if, for every $0 \le s < t$, $E_s[X_t] \ge X_s$ (implies $EX_t \ge EX_s$)

Definition 3.5. A real-valued process $B = (B_t)_{t\ge0}$ is a *Brownian motion* started from 0 if

1. $B_0 = 0$ almost surely (a.s.);

2. for every $0 \le s < t$, the r.v. $B_t - B_s$ is independent of $\sigma(B_r,\ r \le s)$ and distributed according to $N(0, t-s)$;

3. all sample paths $(t \mapsto B_t(\omega))$ of B are continuous;

if additionally B is adapted to $(\mathscr{F}_t)_{t\ge 0}$, we say that B is an $(\mathscr{F}_t)$-*Browninan motion*. Similarly, a process $B = (B_t)_{t\ge 0}$ with values in $\mathbb{R}^d$ is a d-*dimensional* (F_t)-*Brownian motion* if its components are independent Brownian motion and B is adapted to $(\mathscr{F}_t)$ and has independent increments with respect to $(\mathscr{F}_t)$.

Obviously, a Brownian motion is a martingale. But one can construct many more martingales using Brownian motion, among which the most important one is the stochastic integral that we will construct later. For the moment, one can easily verify that both $B_t^2 - t$ and $e^{\theta B_t - \frac{\theta^2}{2}t}$ for any $\theta \in \mathbb{R}$ are martingales.

Given a stochastic process (X_t), there is an obvious way of constructing a filtration such that the process is adapted: $\mathscr{F}_t = \sigma(X_s;\ s \le t)$. Hence, when not specified, one may always assume that a process is adapted to the filtration constructed such. Due to this reason, in the rest of this note, a process is always assume to be adpated.

Proposition 3.1. *Consider a real process $(X_t)_{t\ge 0}$ and a convex function $f : \mathbb{R} \to \mathbb{R}_+$ such that $E[f(X_t)] < \infty$ for every $t \ge 0$.*

1. If (X_t) is a martingale, then $(f(X_t))$ is a submartingale;

2. If (X_t) is a submartingale, and if f is nondecreasing, then $f(X_t)$ is a submartingale.

3.1.3 Stochastic integration

As we know from integration theory, to define abstract integration, one starts with some kind of simple functions. Then since the integration is a linear operator, there is a unique extension of this operator to the closure (under certain topology) of simple functions. The stochastic integration is also defined in this way. But what kind of "simple functions" should we start with? More generally, the stochastic integration should be defined for what kind of functions?

To get some intuiation, we go back to the formula

$$x(t) = x(s) + \int_s^t f(r, x(r))\,dr + \int_s^t \sigma(r, x(r))\,dw(r).$$

The last term on the right hand side suggests that the stochastic integration should preserve certain properties of stochastic process. For example, take $\sigma(r,x) = x$, $f = 0$, and $x(0) = 0$, then $x(t) = \int_0^t x(r)\,dw(r)$. Then if $x(t)$ is a martingale, $\int_s^t x(r)\,dw(r)$ should also be a martingale.

The goal of this subsection is to define stochastic integration for a rather general class of functions – semimartingales.

Consider an "elementary process"

$$H_s(\omega) = \sum_{i=0}^{p-1} H_i(\omega) 1_{(t_i, t_{i+1}]}(s) \tag{3.6}$$

where $0 = t_0 < t_1 < \cdots < t_p$ and for each $i \in \{0, \cdots, p-1\}$, H_i is bounded and $\mathscr{F}_{t_i}$-measurable. Obviously, H is a progressive process (Definition 3.2). Then invoking the formula (3.5), the integration of H w.r.t. a

process $M = (M_t)_{t\geq 0}$ should be defined as

$$\left(\int H dM\right)_t =: \sum_{i=0}^{p-1} H_i(M_{t_{i+1}\wedge t} - M_{t_i\wedge t}). \tag{3.7}$$

Easy calculations show that $\int H dM$ so defined is a martingale (since $H_i(M_{t_{i+1}\wedge t} - M_{t_i\wedge t})$ is for each i).

It remains to extend the "elementary processes" to some closed set under certain norm. Some preparations are needed.

Definition 3.6. An adapted continuous process $A = (A_t)_{t\geq 0}$ is called a *finite variation process* if all its sample paths are finite variation functions [1] on $\mathbb{R}_+$. If in addition the sample paths are nondecreasing functions, the process A is called an increasing process.

Given a process $M = (M_t)_{t\geq 0}$ and a stopping time T, define the stopped process at T as

$$M_t^T = M_{t\wedge T}$$

more precisely, letting $X =: M^T$, then $X_t(\omega) = M_{t\wedge T(\omega)}(\omega)$.

Definition 3.7. A continuous adpated process $M = (M_t)_{t\geq 0}$ with $M_0 = 0$ a.s. is called a *continuous local martingale* if there exists a nondecreasing sequence $(T_n)_{n\geq 0}$ of stopping times such that $T_n \uparrow \infty$ and for every n, the stopped process (M^{T_n}) is a uniformly integrable martingale. When $M_0 \neq 0$, M is called a continuous local martingale if $M - M_0$ is such. In both cases, we say that the sequence of stopping times (T_n) reduces M.

Definition 3.8. A process $X = (X_t)_{t\geq 0}$ is a *continuous semimartingale* if it can be written in the form

$$X_t = M_t + A_t$$

where M is a continuous local martingale and A a finite variation process.

The next lemma indicates that the decomposition above is unique up to indistinguishability.

Lemma 3.1. *Let M be a CLM. Assume that M is also a FVP with $M_0 = 0$. Then $M_t = 0$ for every $t \geq 0$ a.s.*

Now we go back to define a norm for the "elementary processes", a crucial task toward to definition of stochastic integration.

Theorem 3.1. *Let $M = (M_t)_{\geq 0}$ be a continuous local martingale. There exists an increasing process denoted by $(\langle M, M\rangle_t)_{t\geq 0}$, which is unique up to indistinguishability, such that $M_t^2 - \langle M, M\rangle_t$ is a continuous local martingale. Furthermore, for every fixed $t > 0$, if $0 = t_0^n < t_1^n < \cdots < t_{p_n}^n = t$ is an increasing sequence of subdivisions of $[0, t]$ with mesh tending to 0, we have*

$$\langle M, M\rangle_t = \lim_{n\to\infty} \sum_{i=1}^{p_n} (M_{t_i^n} - M_{t_{i-1}^n})^2 \tag{3.8}$$

in probability. The process $\langle M, M\rangle$ is called the quadratic variation of M.

We can make the following observations.

[1] We say that a right continuous function $a : [0, T] \to \mathbb{R}$ with $a(0) = 0$ has finite variation if there exists a signed measure μ on $[0, T]$ such that $a(t) = \mu([0, t])$ for every $t \in [0, T]$.

- It can be easily checked that for a standard Brownian motion B, we have $\langle B,B\rangle_t = t$.
- The quadratic variation of a process does not depend on the initial value M_0 by (3.8). In fact, if $M_t = M_0 + N_t$, then $\langle M,M\rangle = \langle N,N\rangle$.
- In the formula (3.8), if M is a finite variation process, then

$$\begin{aligned}\sum_{i=1}^{p_n}(M_{t_i^n}-M_{t_{i-1}^n})^2 &\le \left(\sup_{1\le i\le p_n}|M_{t_i^n}-M_{t_{i-1}^n}|\right)\sum_{i=1}^{p_n}|M_{t_i^n}-M_{t_{i-1}^n}|\\ &\le \left(\sup_{1\le i\le p_n}|M_{t_i^n}-M_{t_{i-1}^n}|\right)\left(\int_0^t|dM_s|\right)\to 0\end{aligned}$$

  in probability as $n\to\infty$. Hence we can define quadratic variation for finite variation process. But can we define it for semimartingales?

  That is, if $X = M + A$, with M a local continuous martingale and A a finite variation process. Then to define $\langle X,X\rangle = \langle M+A, M+A\rangle$, we shall define $\langle M,A\rangle$ (the impose linearlity on the bracket is "natural") i.e., the "bracket" between a local martingale and a finite variation process. But this can be simply defined as

$$\langle M,A\rangle = \lim_{n\to\infty}\sum_{i=1}^{p_n}(M_{t_i^n}-M_{t_{i-1}^n})(A_{t_i^n}-A_{t_{i-1}^n}).$$

  But

$$\left|\sum_{i=1}^{p_n}(M_{t_i^n}-M_{t_{i-1}^n})(A_{t_i^n}-A_{t_{i-1}^n})\right| \le \left(\int_0^t|dA_s|\right)\sup_{1\le i\le p_n}|M_{t_i^n}-M_{t_{i-1}^n}|\to 0$$

  in probability as $n\to\infty$.

  To go one step further, this motivates us to define the bracket between two local continuous martingale as

$$\langle M,N\rangle = \lim_{n\to\infty}\sum_{i=1}^{p_n}(M_{t_i^n}-M_{t_{i-1}^n})(N_{t_i^n}-N_{t_{i-1}^n})$$

  with mesh tending to 0. The above discussions show that the finite variation parts of M and N do not contribute to the bracket, i.e., if $M = X + A$, $N = X' + A'$, with X, X' CLM and A, A' FVP. Then $\langle M,N\rangle = \langle X,X'\rangle$.

Theorem 3.2. *Given two CLMs M,N. Then*

1. $\langle M,N\rangle$ is the unique (up to indistinguishability) FVP such that $M_tN_t - \langle M,N\rangle_t$ is a CLM.
2. The mapping $(M,N)\mapsto\langle M,N\rangle$ is bilinear and symmetric.
3. For every stopping time T, $\langle M^T,N^T\rangle_t = \langle M^T,N\rangle_t = \langle M,N\rangle_{t\wedge T}$.
4. If M and N are two continuous martingales bounded in L^2, $M_tN_t - \langle M,N\rangle_t$ is a uniformly integrable martingale. Consequently, $\langle M,N\rangle_\infty$ is well defined as the almost sure limit of $\langle M,N\rangle_t$ as $t\to\infty$ is integrable, and satisfies

$$E[M_\infty N_\infty] = E[M_0N_0] + E[\langle M,N\rangle_\infty].$$

Consider the space of all CLM bounded in L^2 with 0 as initial distribution, which we denote as $\mathbb{H}$. Define an inner product on $\mathbb{H}$ as

$$(M,N)_{\mathbb{H}} = E[M_\infty N_\infty] = E[\langle M,N\rangle_\infty]$$

then one can show that $\mathbb{H}$ is a Hilbert space under this inner product. Now fix a CLM M, define an inner product on the space of progressive processs as

$$(H,K)_{L^2(M)} = E\left[\int_0^\infty H_s K_s d\langle M,M\rangle_s\right] \tag{3.9}$$

where

$$L^2(M) = \left\{H \text{ is progressive and } (H,H)_{L^2(M)} < \infty\right\}.$$

As usual, $L^2(M)$ is a Hilbert space. Note that in (3.9), because $t \mapsto \langle M,M\rangle_t$ is a continuous increasing function, the integral inside the expectation is Stieltjes integral and hence well-defined. Thus, we have constructed two Hilbert spaces, namely, $L^2(M)$ and $\mathbb{H}$. Recall that the RHS of (3.7) is a martingale. Further more, it is bounded in $\mathbb{H}$, more precisely

$$\begin{aligned}
\left(\int_0^\cdot H dM, \int_0^\cdot H dM\right)_{\mathbb{H}} &= \left(\sum_{i=0}^{p-1} H_i(M_{t_{i+1}\wedge\cdot} - M_{t_i\wedge\cdot}), \sum_{i=0}^{p-1} H_i(M_{t_{i+1}\wedge\cdot} - M_{t_i\wedge\cdot})\right)_{\mathbb{H}} \\
&= E\left[\left\langle \sum_{i=0}^{p-1} H_i(M_{t_{i+1}} - M_{t_i}), \sum_{i=0}^{p-1} H_i(M_{t_{i+1}} - M_{t_i})\right\rangle\right] \\
&= E\left[\sum_{i=0}^{p-1} H_i^2(\langle M,M\rangle_{t_{i+1}} - \langle M,M\rangle_{t_i})\right] \\
&= \int_0^\infty H_s^2 d\langle M,M\rangle_s \\
&= (H,H)_{L^2(M)}
\end{aligned}$$

Thus the linear mapping

$$H \mapsto \int_0^\cdot H dM$$

is an isometry (hence continuous) from the set of elementary processes $\subset L^2(M)$ into $\mathbb{H}$. Then one can extend the integral to $L^2(M)$ in a unique way if elementary processes are dense in $L^2(M)$, which is indeed the case. Thus for any $H \in L^2(M)$, the integral $\int H dM$ is defined as the limit of $\int H_n dM$ where H is the limit of elementary processes in $L^2(M)$. (Note that since $\mathbb{H}$ is Hilbert (complete), the limit is still a martingale!) For convenience, $\int_0^\cdot H dM$ is also written as $H \cdot M$.

The following are some properties of the stochastic integral:

- Let $H \in L^2(M)$, $M, N \in \mathbb{H}$. Then

  $$\langle H\cdot M, N\rangle = H \cdot \langle M,N\rangle$$

  and $H \cdot M$ is the unique element in $\mathbb{H}$ such that the above holds for all $N \in \mathbb{H}$. From this formula, we can deduce that

$$\begin{aligned}
\langle H\cdot M, H\cdot M\rangle &= H\cdot\langle M, H\cdot M\rangle \\
&= H\cdot\left(\int_0^\cdot H_s d\langle M,M\rangle_s\right) \\
&= H^2\cdot\langle M,M\rangle
\end{aligned} \tag{3.10}$$

where the equality (3.10) is justified first for elementary processes and then one extend it to $L^2(M)$. Written out explicitly ,the above relation reads

$$\left\langle \int_0^\cdot H dM, \int_0^\cdot H dM \right\rangle_t = \int_0^t H_s^2 d\langle M, M\rangle_s .$$

More generally, for $K \in L^2(N)$, we have

$$\langle H \cdot M, K \cdot N\rangle = HK \cdot \langle M, N\rangle .$$

- Let $M, N \in \mathbb{H}$, and $H \in L^2(M)$, $K \in L^2(N)$. Then since $H \cdot M$ and $K \cdot N$ are martingales in $\mathbb{H}$, we have for every $t \in [0,\infty]$,

$$\begin{aligned} E\left[\int_0^t H_s dM_s\right] &= 0, \\ E_s\left[\int_0^t H_r dM_r\right] &= \int_0^s H_r dM_r, \ \ \forall 0 \le s \le t \\ E_s\left[\int_s^t H_r dM_r\right] &= 0 \end{aligned}$$

- More over

$$E[(H \cdot M)_t (K \cdot N)_t] = E[((HK) \cdot \langle M, N\rangle)_t]$$

or

$$E\left[\left(\int_0^t H_s dM_s\right)\left(\int_0^t K_s dN_s\right)\right] = E\left[\int_0^t H_s K_s d\langle M, N\rangle_s\right].$$

In particular

$$E\left[\left(\int_0^t H_s dM_s\right)^2\right] = E\left[\int_0^t H_s^2 d\langle M, M\rangle_s\right]$$

In the above, we have defined stochastic integral for martingales bounded in L^2, i.e. $\mathbb{H}$. Now we generalize the stochastic integral to CLMs.

Given a CLM M, define

$$\begin{aligned} L^2_{\text{loc}}(M) &= \left\{ H : \int_0^t H_s^2 d\langle M, M\rangle_s < \infty, \ \ \forall t \ge 0 \right\} \ \text{a.s.} \\ L^2(M) &= \left\{ H : \int_0^\infty H_s^2 d\langle M, M\rangle_s < \infty \right\} \end{aligned}$$

(Since $\langle M, M\rangle$ is FVP, both spaces are well defined). We point out that $L^2(M)$ is still a Hilbert space.

Theorem 3.3. *Let M be a CLM. For every $H \in L^2_{loc}(M)$, there exists a unique CLM with initial value 0, which is denoted by $H \cdot M$, such that, for every CLM N,*

$$\langle H \cdot M, N\rangle = H \cdot \langle M, N\rangle .$$

If $H \in L^2_{loc}(M)$ and K is a progressive process, we have $K \in L^2_{loc}(H \cdot M)$ if and only if $HK \in L^2_{loc}(M)$ and then

$$H \cdot (K \cdot M) = HK \cdot M.$$

We write

$$(H \cdot M)_t = \int_0^t H_s dM_s$$

and call it the stochastic integral of H w.r.t. M.

Now that a semimartingale X can be decomposted as the sum of a CLM and a FVP, namely, $X = M+V$. Then for any locally bounded progressive process H, one can define

$$H \cdot X := H \cdot M + \int H_s dV_s$$

where $\int H_s dV_s$ is the usual Stieltjes integral.

As before, this integral has the following properties:

1. Let X be a continuous semimartingale, and K, H two locally bounded progressive processes. Then $KH \cdot X = K \cdot (H \cdot X)$.

2. Let H be a locally bounded progressive process. If X is a CLM or FVP, the same holds for $H \cdot X$.

3.1.4 Itô's formula

Itô's formula will be our most useful tool in this text. Even if one does not know the rigorous construction of stochastic integral, Itô's formula will be sufficient for the study of stochastic optimal control.

Theorem 3.4. *Let $X^1, \cdots, X^p$ be p continuous semimartingales, and let F be a twice continuously differentiable real function on $\mathbb{R}^p$. Then for every $t \geq 0$,*

$$\begin{aligned} F(X_t^1, \cdots, X_t^p) = F(X_0^1, \cdots, X_0^p) \\ + \sum_{i=1}^{p} \int_0^t \frac{\partial F}{\partial x^i}(X_s^1, \cdots, X_s^p) dX_s^i \\ + \frac{1}{2} \sum_{i,j=1}^{p} \int_0^t \frac{\partial^2 F}{\partial x^i \partial x^j}(X_s^1, \cdots, X_s^p) d\left\langle X^i, X^j \right\rangle_s . \end{aligned}$$

We mention a few consequences of Itô's formula.

1. $F(X_t^1, \cdots, X_t^p)$ is a semimartingale. This is what we had expected in the beginning of the last subsubsection!

2. Let $F(x, y) = xy$. Then we see that

$$\begin{aligned} X_t Y_t &= X_0 Y_0 + \int_0^t X_s dY_s + \int_0^t Y_s dX_s + \int_0^t d\langle X, Y \rangle_s \\ &= X_0 Y_0 + \int_0^t X_s dY_s + \int_0^t Y_s dX_s + \langle X, Y \rangle_t \end{aligned}$$

   This formula can be viewed as the *formula of integration by parts*. In particular, if $Y = X$,

$$X_t^2 = X_0^2 + 2\int_0^t X_s dX_s + \langle X, X \rangle_t .$$

   We know that when X is a CLM, then $\langle X, X \rangle$ is the unique FVP such that $X^2 - \langle X, X \rangle$ is a CLM. The above formula tells us that

$$\langle X, X \rangle_t = X_t^2 - X_0^2 - 2\int_0^t X_s dX_s.$$

3. Let $X_t^1 = t$, $X_t^2 = B_t$ (standard Brownian motion), and $F \in C^2(\mathbb{R}_+ \times \mathbb{R})$. Then

$$F(t, B_t) = F(0, B_0) + \int_0^t \frac{\partial F}{\partial x}(s, B_s) dB_s + \int_0^t \left(\frac{\partial F}{\partial t} + \frac{1}{2} \frac{\partial^2 F}{\partial x^2} \right)(s, B_s) ds.$$

3.1.5 Theory of Markov process

Let $(E,\mathscr{E})$ be a measurable space. A *Markovian transition kernel* from E into E is a mapping $Q : E \times \mathscr{E} \to [0,1]$ satisfying the following properties:

1. For every $x \in E$, the mapping $\mathscr{E} \ni A \mapsto Q(x,A)$ is a probability measure on $(E,\mathscr{E})$.
2. For every $A \in \mathscr{E}$, the mapping $E \ni x \mapsto Q(x,A)$ is $\mathscr{E}$-measurable.

Given a transition kernel Q, if $f : E \to \mathbb{R}$ is bounded measurable, we define the function $Qf : E \to \mathbb{R}$ by

$$Qf(x) = \int_E Q(x,dy) f(y) \tag{3.11}$$

which is still bounded measurable.

Definition 3.9. A collection $(Q_{s,t})_{0\le s\le t}$ of transition kernels on E is called a transition semigroup if the following properties hold.

1. For every $x \in E$ and $t \in \mathbb{R}$, $Q_{t,t}(x,dy) = \delta_x(dy)$.
2. For all $0 \le s \le r \le t$ and $A \in \mathscr{E}$,

$$Q_{s,t}(x,A) = \int_E Q_{s,r}(x,dy) Q_{r,t}(y,A) \tag{3.12}$$

   (Chapman-Kolmogorov identity).

3. For every $A \in \mathscr{E}$, the function $(s,t,x) \mapsto Q_{s,t}(x,A)$ is measurable w.r.t. the σ-algebra $\mathscr{B}(\mathbb{R}_+)\times\mathscr{B}(\mathbb{R}_+)\times\mathscr{E}$.

When $Q_{s,t} = Q_{s+r,t+r}$ for all $r \in \mathbb{R}$, we say that the transition semigroup is time independent and we simply write $Q_{t-s} := Q_{s,t}$. Now given $f \in B(E)$, $0 \le s \le r \le t$, by Chapman-Kolmogorov identity, we have

$$\begin{aligned} Q_{s,r}Q_{r,t}f(x) &= \int_E Q_{s,r}(x,dy) Q_{r,t}f(y) \\ &= \int_E Q_{s,r}(x,dy) \int_E Q_{r,t}(y,dw) f(w) \\ &= \int_E f(w) \int_E Q_{s,r}(x,dy) Q_{r,t}(y,dw) \\ &= \int_E f(w) Q_{s,t}(x,dw) \\ &= Q_{s,t}f(x). \end{aligned}$$

Hence we get the identity

$$Q_{s,r}Q_{r,t} = Q_{s,t}, \quad \forall 0 \le s \le r \le t$$

which is equivalent to the Chapman-Kolmogorov identity when $Q_{s,t}$ is understood as operators from $B(E)$ to $B(E)$. Since $A \mapsto Q_{s,t}(x,A)$ is a probability measure, it is easily seen from (3.11) that $Q_{s,t} : B(E) \to B(E)$ is non-expansive (i.e., $||Q_{s,t}|| \le 1$) when $B(E)$ is equipped with norm $||f|| = \sup\{|f(x)| : x \in E\}$.

Now we are ready to define Markov process.

Definition 3.10. A process $(X_t)_{t\geq 0}$ with values in E is called a *Markov process* with transition semigroup $(Q_{s,t})_{0\leq s\leq t}$ if

$$E_s[f(X_t)] = Q_{s,t}f(X_s), \quad \forall s,t \geq 0 \tag{3.13}$$

for each $f \in B(E)$.

When the transition semigroup is time independent, the Markov property (3.13) becomes

$$E_s[f(X_{s+t})] = Q_t f(X_s), \quad \forall s,t \geq 0.$$

Now take $f = 1_A$, with $A \in \mathscr{E}$. Then (3.13) implies

$$P(X_{t+s} \in A|\mathscr{F}_s) = Q_{s,s+t}1_A(X_s) = Q_{s,s+t}(X_s, A)$$

from which we deduce that

$$P(X_t \in A|X_{t_1}, \cdots, X_{t_m}) = P(X_t \in A|X_{t_m})$$

whenever $t_1 \leq \cdots \leq t_m \leq t$. In other words, the conditional distribution of X_{s+t} knowing the past $(X_r,\ 0 \leq r \leq s)$ before time s depends only on the present state X_s. In particular, when $X_s = x$, we get

$$Q_{s,t}(x, A) = P(X_t \in A|X_s = x)$$

Let $C_0(E)$ be the set of continuous real functions on E that vanish at infinity. It is common knowledge that $C_0(E)$ is a Banach space for the norm $||f|| = \sup\{|f(x)| : x \in E\}$.

Definition 3.11. Let $(Q_{s,t})$ be a transition semigroup on E. We say that it is a *Feller semigroup* if

1. $\forall f \in C_0(E),\ Q_{s,t}f \in C_0(E)$ for all $0 \leq s \leq t$.
2. $\forall f \in C_0(E),\ ||Q_{s,s+h}f - f|| \to 0$ as $h \to 0$.

Define the operators $A(t)$ by

$$A(t)f = \lim_{h\to 0+} \frac{Q_{t,t+h}f - f}{h}$$

where the limit is taken in $C_0(E)$ and the domain of $A(t)$ is such that the above limit exists, i.e.,

$$D(A(t)) = \left\{ f \in C_0(E) : \frac{Q_{t,t+h}f - f}{h} \text{ converges in } C_0(E) \text{ when } h \to 0+ \right\}.$$

3.1.6 Stochastic differential equation

Let d and m be positive integers, and let σ and b be locally bounded measurable functions defined on $\mathbb{R}_+ \times \mathbb{R}^d$ and taking values in $\mathbb{R}^{d\times m}$ and in $\mathbb{R}^d$ respectively. We write $\sigma = (\sigma_{ij})_{1\leq i\leq d, 1\leq j\leq m}$ and $b = (b_i)_{1\leq i\leq d}$.

A *solution of the stochastic differential equation*

$$\begin{aligned} dX_t &= b(t, X_t)dt + \sigma(t, X_t)dB_t \\ X_0 &\text{ is } \mathscr{F}_0\text{-measurable} \end{aligned} \tag{3.14}$$

consists of

1. a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,\infty]}, P)$ (where the filtration is always assumed to be complete);

2. an m-dimensional $(\mathscr{F}_t)$-Brownian motion $B = (B^1, \cdots, B^m)$ started from 0;

an $(\mathscr{F}_t)$-adapted process $X = (X^1, \cdots, X^d)$ with values in $\mathbb{R}^d$, with continuous sample paths, such that

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dB_s.$$

The solution is called a *strong solution* if $(\mathscr{F}_t)_{t\in[0,\infty]}$ is specified *a priori*. Otherwise it is called a *weak solution*, i.e., the filtration is part of the solution. In this note, we are mainly interested in strong solution. If for any two strong solutions X, Y we have

$$P(X(t) = Y(t),\ 0 \le t < \infty) = 1,$$

we say that the solution is unique.

Theorem 3.5. *If there exists a constant $K > 0$ such that for every $t \ge 0$, $x, y \in \mathbb{R}^d$,*

$$|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \le K|x - y|$$

then (3.14) has a unique strong solution.

We show that the solution of SDE is a Markov process (Definition 3.10).

To that end, define

$$Q_{s,t}(x, A) := P(X(t; s, x) \in A)$$

we show that

$$Q_{s,t}f(X_s) = E_s[f(X_t)] \tag{3.15}$$

(note that this would imply $Q_{s,t}f(x) = E[f(X(t; s, x))]$).

In fact,

$$\begin{aligned}
Q_{s,t}1_A(X_s) &= Q_{s,t}(X_s, A) \\
&= P(X(t; s, x) \in A)|_{x=X_s} \\
&\quad (X_s \text{ is a r.v. so must be put outside } P(\cdot)!) \\
&= E_s 1_A(X(t; s, X_s)) \\
&\quad (X_s \text{ is } \mathscr{F}_s\text{-measurable}) \\
&= E_s 1_A(X_t). \\
&\quad (\text{uniqueness of the solution enforces that } X(t; s, X_s) = X_t \text{ for } t \ge s)
\end{aligned}$$

A standard argument using monotone class lemma will finalize the proof of the formula (3.15). It remains to show that $Q_{s,t}$ is a transition group, i.e.,

$$Q_{s,r}Q_{r,t} = Q_{s,t}. \tag{3.16}$$

But

$$\begin{aligned}
Q_{s,t}f(x) &= E[f(X(t;s,x)) \\
&\quad \text{(see the remark after (3.15))} \\
&= E[E_r[f(X(t;s,x))]] \\
&= E[Q_{r,t}f(X_r)] \\
&\quad \text{(again, by (3.15))} \\
&= \int Q_{r,t}f(y)P(X_r \in dy) \\
&= \int Q_{r,t}f(y)Q_{s,r}(x,dy) \\
&\quad \text{(since } P(X_r \in A) = P(X(r;s,x) \in A) = Q_{s,r}(x,A)) \\
&= \int Q_{s,r}(x,dy)Q_{r,t}f(y)
\end{aligned}$$

which indeed verifies (3.16). For time dependent function, evidently we should define $Q_{s,t}f(s,t) := E[f(t,X(t;s,x))]$.

In the literature, it is common to denote

$$P(s,x;t,A) := Q_{s,t}(x,A) = P(X_t \in A | X_s = x)$$

and the property (3.16) can now be expressed as

$$P(s,x;t,A) = \int P(s,x;r,dy)P(r,y;t,A).$$

Our next task is to find the generator of the Markov process (transition group) $(Q_{s,t})_{0\le s\le t}$.

Since

$$X_t^x = x + \int_0^t b(r,X_r^x)dr + \int_0^t \sigma(r,X_r^x)dB_r.$$

we find the quadratic variation (when X is of one-dimension)

$$\begin{aligned}
\left\langle X^x, X^x \right\rangle_t &= \left\langle \int_0^\cdot \sigma(r,X_r^x)dB_r, \int_0^\cdot \sigma(r,X_r^x)dB_r \right\rangle_t \\
&= \int_0^t \sigma(r,X_r^x)^2 dr.
\end{aligned}$$

More generally, we have $d\langle X^x, X^x\rangle_t = \sigma(t,X_t^x)\sigma^T(t,X_t^x)dt =: (a_{ij})dt$.

Now given a function $\varphi \in C^{1,2}$ (C^1 in w.r.t. to the first variable and C^2 w.r.t the second), by Itô's formula

$$\begin{aligned}
\varphi(t,X_t) &= \varphi(s,X_s) + \int_s^t \frac{\partial\varphi}{\partial t}(r,X_r)dr + \int_s^t \frac{\partial\varphi}{\partial x}(r,X_r)dX_r + \frac{1}{2}\sum_{i,j}\int_s^t \frac{\partial^2\varphi}{\partial x^2}(r,X_r)a_{ij}(r,X_r)dr \\
&= X_s + \int_s^t \frac{\partial\varphi}{\partial t}(r,X_r)dr + \int_s^t \frac{\partial\varphi}{\partial x}(r,X_r)b(r,X_r)dr + \int_s^t \frac{\partial\varphi}{\partial x}(r,X_r)\sigma(r,X_r)dB_r \\
&+ \frac{1}{2}\int_s^t \mathrm{tr}\left(\frac{\partial^2\varphi}{\partial x^2}(r,X_r)\sigma(r,X_r)\sigma^T(r,X_r)\right)dr
\end{aligned}$$

Then

$$E_s[\varphi(t,X_t) - \varphi(s,X_s)] = E_s\left[\int_s^t \frac{\partial\varphi}{\partial t}(r,X_r) + A(r)\left(\varphi(r,X_r)\right)dr\right]$$

in which

$$A(r)(\varphi(r,x)) := \frac{1}{2}\mathrm{tr}\left(\sigma\sigma^T\Delta\varphi\right) + (\nabla\varphi)b(t,x) \tag{3.17}$$
$$= \frac{1}{2}\sum a_{ij}(t,x)\frac{\partial^2\varphi}{\partial x_i\partial x_j}(t,x) + \sum b_i(t,x)\frac{\partial\varphi}{\partial x_i}.$$

Therefore, by fixing $X_s = x$, we obtain

$$\begin{aligned}\frac{Q_{s,s+h}\varphi(s,x)-\varphi(s,x)}{h} &= \frac{E[\varphi(s+h,X(s+h;s,x))]-\varphi(s,x)}{h}\\ &= \frac{1}{h}E\left[\int_s^{s+h}\frac{\partial\varphi}{\partial t}(r,X_r)+A(r)\left(\varphi(r,X_r)\right)dr\right]\\ &\to \varphi_s(s,x)+A(s)\left(\varphi(s,x)\right) \text{ as } h\to 0+\end{aligned}$$

Hence the generator of $Q_{s,t}$ is $\varphi_s + A(s)\varphi$ where A is defined as (3.17). When considering only time independent functions φ, then $A(s)$ alone is the generator since $\varphi_s = 0$ for all $s \geq 0$.

3.1.7 Girsanov theorem

$\mathscr{M}^c_{\mathrm{loc}}$: continuous local martingale

$\mathscr{M}^{2,c}_{\mathrm{loc}}$: continuous local martingale s.t. $\sup_{0\leq s\leq t} E|X_s|^2 < \infty$, $\forall\, t\in\mathbb{R}_+$

For $X \in \mathscr{M}^{2,c}_{\mathrm{loc}}$ with $X_0 = 0$, define

$$\mathscr{E}(X)_t =: \exp\left(X_t - \frac{1}{2}\langle X,X\rangle_t\right) \tag{3.18}$$

where $\langle X,X\rangle$ is the quadratic variation.

Lemma 3.2. $\mathscr{E}(X) \in \mathscr{M}^c_{loc}$.

Proof. By Ito formula,

$$\begin{aligned}\mathscr{E}(X)_t &= 1 + \int_0^t \mathscr{E}(X)_s(dX_s - \frac{1}{2}d\langle X,X\rangle_s)\\ &\quad + \frac{1}{2}\int_0^t \mathscr{E}(X)_s d\left\langle X_s - \frac{1}{2}\langle X,X\rangle_s, X_s - \frac{1}{2}\langle X,X\rangle_s\right\rangle\end{aligned}$$

but

$$\left\langle X_s - \frac{1}{2}\langle X,X\rangle_s, X_s - \frac{1}{2}\langle X,X\rangle_s\right\rangle = \langle X,X\rangle_s$$

therefore

$$\mathscr{E}(X)_t = 1 + \int_0^t \mathscr{E}(X)_s dX_s$$

which is a local continuous martingale. □

Theorem 3.6. *Let $X \in \mathscr{M}^c_{loc}$ with $X_0 = 0$. Consider the following properties:*

1. $E[\exp\frac{1}{2}\langle X,X\rangle_\infty] < \infty$ (Novikov's condition);
2. X is a uniformly integrable martingale, and $E[\exp\frac{1}{2}L_\infty] < \infty$ (Kazamaki's condition);

3. $\mathscr{E}(X)$ is a uniformly integrable martingale.

Remark 3.2. When we consider local martingales on finte interval, say on $[0,T]$, the conditions in the above theorem changes accordingly, e.g., the Novikov condition becomes $E[\exp\frac{1}{2}\langle X,X\rangle_T]<\infty$.

Theorem 3.7. *Let $(X_t)_{t\in[0,T]}$ be a continous local martingale, and assume that $\mathscr{E}(X)_t$ is a martingale on $[0,T]$. Define a process*

$$D_t =: \mathscr{E}(X)_t = E[\mathscr{E}(X)_T|\mathscr{F}_t]$$

then (D_t) is a uniformly integrable martingale. Further, define a probability measure Q by

$$\frac{dQ}{dP} = D_T$$

Then for any martingale Y on $[0,T]$, the process $\tilde{Y}_t = Y_t - \langle X,Y\rangle_t$ is a martingale under Q on $[0,T]$.

Example 3.1. Let $Y = W$ be a Brownian motion, and

$$X_t = \int_0^t \beta_s dW_s$$

then

$$\tilde{W}_t = W_t - \int_0^t \beta_s ds$$

is a martingale under $dQ = z_T dP$ where

$$z_T = \mathscr{E}\left(\int_0^\cdot \beta_s dW_s\right)_T = \exp\left(\int_0^T \beta_s dW_s - \frac{1}{2}\int_0^T |\beta_s|^2 ds\right).$$

Clearly, $\mathscr{E}(X)$ is a martingale if $E\exp\frac{1}{2}\langle X,X\rangle_T = E\exp\frac{1}{2}\int_0^T|\beta_s|^2ds<\infty$. In fact, we can say more: $(\tilde{W}_t)_{[0,T]}$ is a Brownian motion. In particular, $\tilde{W}_t$ is independent of $\mathscr{F}_0$.

Let $\mathscr{G}\subset\mathscr{F}$, and $P<<Q$ such that $dP = MdQ$, then

$$E^P[X|\mathscr{G}] = \frac{E^Q[X\frac{dP}{dQ}|\mathscr{G}]}{E^Q[\frac{dP}{dQ}|\mathscr{G}]} \tag{3.19}$$

This is called the abstract *Bayes formula*.

3.2 Stochastic optimal control

3.2.1 Stochastic principle of optimality

The formulation of stochastic optimal control problem is somewhat the same as the deterministic case. Given a filtered probability space $(\Omega,\mathscr{F},(\mathscr{F}_t)_{t\geq0},P)$ on which an m-dimensional standard Brownian motion B is defined. Consider the following controlled SDE:

$$\begin{aligned} dx(t) &= b(t,x(t),u(t))dt + \sigma(t,x(t),u(t))dB_t \\ x(0) &= x_0\in\mathbb{R}^n \end{aligned} \tag{3.20}$$

where $b:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}^n$, $\sigma:[0,T]\times\mathbb{R}^n\times U\to\mathbb{R}^{n\times m}$, with U being a given separable metric space and $u:[0,T]\times\Omega\to U$ is called the control. Define the *feasible control* set as

$$\mathscr{U}[0,T] = \{u:[0,T]\times\Omega\to U \mid u(\cdot) \text{ is } (\mathscr{F}_t)\text{-adapted}\}.$$

The cost functional for stochastic optimal control is defined as

$$J(u(\cdot)) = E\left\{h(x(T)) + \int_0^T L(t, x(t), u(t))\,dt\right\} \tag{3.21}$$

and call

$$\mathscr{U}_{\text{ad}}[0,T] = \left\{u \in \mathscr{U}[0,T] : \text{ the solution of (3.20) is unique and } J(u(\cdot)) < \infty\right\}$$

the s-admissible control set. It is also natural to consider state feedback control and call

$$\mathscr{U}_{\text{f}}[0,T] = \left\{u \in \mathscr{U}_{\text{ad}}[0,T] : \ u(t) = \phi(t, X_t) \text{ for some continuous function } \phi\right\}$$

the f-admissible control set.

As in the deterministic case, we derive the principle of optimality, i.e., the stochastic version of (**??**). The stochastic optimal control problem is find $\bar{u}(\cdot) \in \mathscr{U}_{\text{f}}[0,T]$ (if exists) such that

$$J(\bar{u}(\cdot)) = \inf_{u(\cdot)\in\mathscr{U}_{\text{f}}[0,T]} J(u(\cdot))$$

Assumption 1. (A1) U is a Polish space (separable Banach space).

(A2) The maps b, σ, h, L are uniformly continuous, and there exists a constant $K > 0$, such that for $\varphi(t,x,u) = b(t,x,u), \sigma(t,x,u), h(x), L(t,x,u)$,

$$|\varphi(t,x,u) - \varphi(t,y,u)| \le K|x-y|, \ \ \forall\, t \in [0,T], \ x, y \in \mathbb{R}^n, \ u \in U$$
$$|\varphi(t,0,u)| \le K, \ \ \forall (t,u) \in [0,T] \times U$$

Let

$$J(s,y;u(\cdot)) = E\left\{h(x(T)) + \int_s^T L(t, x(t), u(t))\,dt\right\}$$

and define the value function as

$$V(s,y) = \inf_{u(\cdot)\in\mathscr{U}_{\text{f}}[s,T]} J(s,y;u(\cdot)), \ \ \forall (s,y) \in [0,T) \times \mathbb{R}^n,$$
$$V(T,y) = h(y), \ \ \forall\, y \in \mathbb{R}^n.$$

We have the following proposition.

Proposition 3.2. *Let (A1)-(A2) hold. Then for any $(s,y) \in [0,T) \times \mathbb{R}^n$ and $s \le \hat{s} \le T$*

$$V(s,y) = \inf_{u(\cdot)\in\mathscr{U}_f[s,T]} E\left\{\int_s^{\hat{s}} L(t, X(t;s,y,u(\cdot)), u(t))\,dt + V(\hat{s}, X(t;s,y,u(\cdot)))\right\}. \tag{3.22}$$

*Formula (3.22) enjoys the same structure as (**??**), but since the cost function (3.21) does not admit the splitting in Theorem (1.2), it is not a immediate consequence of that theorem.*

Proof. Let $\mathscr{F}_{\hat{s}}^{s} = \sigma\{B_r : s \le r \le \hat{s}\}$. Denote right hand side of (3.22) by $\bar{V}(s,y)$. For any $\varepsilon \ge 0$, there exists $u(\cdot) \in \mathscr{U}_{\mathrm{f}}[s,T]$ such that

$$\begin{aligned}
V(s,y) + \varepsilon &> J(s,y;u(\cdot)) \\
&= E\left\{\int_s^T L(t, X(t;s,y,u(\cdot)), u(t))\,dt + h(X(T;s,y,u(\cdot))\right\} \\
&= E\left\{\int_s^{\hat{s}} L(t, X(t;s,y,u(\cdot)), u(t))\,dt\right\} \\
&\quad + EE_{\mathscr{F}_{\hat{s}}^{s}}\left[\int_{\hat{s}}^T L(t, X(t;s,y,u(\cdot)), u(t))\,dt + h(X(T;s,y,u(\cdot))\right] \\
&= E\left\{\int_s^{\hat{s}} L(t, X(t;s,y,u(\cdot)), u(t))\,dt\right\} \\
&\quad + EE_{\mathscr{F}_{\hat{s}}^{s}}\left[\int_{\hat{s}}^T L(t, X(t;\hat{s},X_{\hat{s}},u(\cdot)), u(t))\,dt + h(X(T;\hat{s},X_{\hat{s}},u(\cdot))\right] \\
&\quad \text{(uniqueness of solution)} \\
&= E\left\{\int_s^{\hat{s}} L(t, X(t;s,y,u(\cdot)), u(t))\,dt + J(\hat{s}, X(\hat{s};s,y,u(\cdot));u(\cdot))\right\} \\
&\ge E\left\{\int_s^{\hat{s}} L(t, X(t;s,y,u(\cdot)), u(t))\,dt + V(\hat{s}, X(\hat{s};s,y,u(\cdot)))\right\} \\
&\ge \bar{V}(s,y).
\end{aligned}$$

To prove the converse, we need a technical result regarding the regularity of J and V: Given a constant $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that whenever $|x-y| < \delta$,

$$|J(\hat{s},y;u(\cdot)) - J(\hat{s},x;u(\cdot))| + |V(\hat{s},y) - V(\hat{s},x)| \le \varepsilon, \ \ \forall u(\cdot) \in \mathscr{U}_{\mathrm{f}}[\hat{s},T].$$

Next, choose a partition of $\mathbb{R}^n$ with $\mathbb{R}^n = \cup_j D_j$, $D_i \cap D_j = \varnothing$ if $i \ne j$ and $\mathrm{diam}(D_j) < \delta$. Then there exist $(u_j)_{j\ge 1} \in \mathscr{U}_{\mathrm{f}}[\hat{s},T]$ such that

$$J(\hat{s},x_j;u_j(\cdot)) \le V(\hat{s},x_j) + \varepsilon, \ \ \forall x_j \in D_j.$$

Hence for any $x \in D_j$, we have

$$J(\hat{s},x,u_j(\cdot)) \le J(\hat{s},x_j,u_j(\cdot)) + \varepsilon \le V(\hat{s},x_j) + 2\varepsilon \le V(\hat{s},x) + 3\varepsilon.$$

Now for any $u(\cdot) \in \mathscr{U}_{\mathrm{f}}[s,T]$, define

$$\tilde{u}(t) = \begin{cases} u(t), & t \in [s,\hat{s}) \\ u_j(t), & t \in [\hat{s},T] \text{ and } x(t) \in D_j \end{cases}$$

Then

$$\begin{aligned}
V(s,y) &\le J(s,y;\tilde{u}(\cdot)) \\
&= E\left\{\int_s^T L(t,X(t;s,y,\tilde{u}(\cdot)),u(t))\,dt + h(X(T;s,y,\tilde{u}(\cdot))\right\} \\
&= E\left\{\int_s^{\hat{s}} L(t,X(t;s,y,u(\cdot)),u(t))\,dt\,x\right\} \\
&\quad + EE_{\mathscr{F}_{\hat{s}}^s}\left[\int_{\hat{s}}^T L(t,X(t;s,y,\tilde{u}(\cdot)),u(t))\,dt + h(X(T;s,y,\tilde{u}(\cdot))\right] \\
&= E\left\{\int_s^{\hat{s}} L(t,X(t;s,y,u(\cdot)),u(t))\,dt + J(\hat{s},X(\hat{s};s,y,u(\cdot));\tilde{u}(\cdot))\right\} \\
&\le E\left\{\int_s^{\hat{s}} L(t,X(t;s,y,u(\cdot)),u(t))\,dt + V(\hat{s},X(\hat{s};s,y,u(\cdot))) + 3\varepsilon\right\}.
\end{aligned}$$

□

Again, as in the deterministic case, based on the above proposition, one can easily prove the following theorem.

Theorem 3.8. *Suppose that (A1)-(A2) hold and the value function $V \in C^{1,2}([0,T]\times\mathbb{R}^n)$. Then V is a solution of the following PDE (stochastic HJB equation):*

$$\begin{aligned}
&-V_t + \sup_{u\in U} G(t,x,u,-V_x,-V_{xx}) = 0 \\
&V(x,T) = h(x), \quad x\in\mathbb{R}^n
\end{aligned} \tag{3.23}$$

where

$$G(t,x,u,p,P) = \frac{1}{2}\operatorname{tr}\left(P\sigma(t,x,u)\sigma(t,x,u)^T\right) + \left\langle p, b(t,x,u)\right\rangle - L(t,x,u).$$

Invoking the infinitesimal generator $A(\cdot)$ defined as (3.17), the stochastic HJB equation can also be written as

$$\begin{aligned}
0 &= V_t + \inf_{u\in U}[A^u(t)V + L(t,x,u)], \\
V(x,T) &= h(x), \quad x\in\mathbb{R}^n.
\end{aligned} \tag{3.24}$$

where

$$A^u(t) := \frac{1}{2}\sum a_{ij}(t,x,u)\frac{\partial^2}{\partial x_i\partial x_j}(t,x) + \sum b_i(t,x,u)\frac{\partial}{\partial x_i}.$$

Notice that when $\sigma = 0$, (3.23) reduces exactly to the deterministic HJB (c.f. (1.37)). Thus the stochastic principle of optimality is a generalization of the deterministic one.

3.2.2 Full state LQG control

Consider now the linear controlled stochastic system

$$dx(t) = [A(t)x(t) + B(t)u(t)]\,dt + \sigma(t)\,dB_t \tag{3.25}$$

on the interval $[0,T]$ with $A(\cdot)\in L^\infty([0,T];\mathbb{R}^{n\times n})$, $B(\cdot)\in L^\infty([0,T];\mathbb{R}^{n\times m})$, $u(\cdot)\in\mathscr{U}_{\mathrm{f}}[0,T]$ and $\sigma\in L^\infty([0,T];\mathbb{R}^{n\times d})$, $B.$ is a d-dimensional Brownian motion. The cost function of interest for this system is

$$J(s,x,u) = E\left\{x(T)^T Dx(T) + \int_s^T [x(t)^T M(t)x(t) + u(t)^T R(t)u(t)]\,dt\right\},$$

in which $x(s) = x$, $M(t) \geq aI_{n\times n}$, $R(t) \geq bI_{m\times m}$ and $D > cI_{n\times n}$ for some constants $a, b, c \in \mathbb{R}_{>0}$.

In order to solve the stochastic HJB (3.23) or (3.24), it is natural to propose the following candidate

$$V(t,x) = x^T K(t)x + q(t)$$

for some functions $K : [0,T] \to \mathbb{R}^{n\times n}$ (symmetric), and $q : [0,T] \to \mathbb{R}$.

Now

$$\begin{aligned}
&A^u(t)V(t,x) + L(t,x,u)\\
=&A^u(t)[x^T K(t)x + q(t)] + x^T M(t)x + u^T R(t)u\\
=&2x^T K(t)[A(t)x + B(t)u] + \mathrm{tr}(\sigma(t)\sigma(t)^T K(t)) + x^T M(t)x + u^T R(t)u
\end{aligned}$$

which is a quadratic function of u. By the fact that $R(t) \geq bI_{m\times m}$ we know $\inf[A^u(t)V(t,x) + L(t,x,u)]$ is achieved at

$$\{u : \frac{\partial}{\partial u}[A^u(t)V(t,x) + L(t,x,u)] = 0\}$$

or

$$2R(t)u_* + 2B(t)^T K(t)x = 0$$

which results in a static feedback control law

$$u_*(t,x) = -R(t)^{-1}B(t)^T K(t)x.$$

Substituting u_* into the stochastic HJB, we get

$$0 = x^T[\dot{K} + KA + A^T K - KBR^{-1}B^T K + M]x + \dot{q}(t) + \mathrm{tr}(\sigma\sigma^T K).$$

Hence a sufficient condition for the optimal law is

$$\begin{aligned}
\dot{K}(t) &= -K(t)A(t) - A(t)^T K(t) + K(t)B(t)R^{-1}(t)B(t)^T K(t) - M(t)\\
K(T) &= D\\
\dot{q}(t) &= \mathrm{tr}(\sigma(t)\sigma(t)^T K(t))\\
q(T) &= 0
\end{aligned}$$

and that the resulting solution $K(t)$ being symmetric positive definite.

3.2.3 Revisit of viscosity solution of HJB

Let us consider two systems

$$\begin{aligned}
\mathrm{S}_1 &: dx(t) = f(t,x(t),u(t))dt\\
\mathrm{S}_2 &: dx(t) = f(t,x(t),u(t))dt + \sqrt{2\varepsilon}dB_t
\end{aligned}$$

i.e., S_2 is obtained by adding a stochastic term $\sqrt{2\varepsilon}dB_t$ on S_1.

Consider the cost function for the two systems

$$\begin{aligned}
J_1(s,y,u(\cdot)) &= \int_s^T L(t,x(t),u(t))dt + \varphi(x(T)),\ \ x(t) \text{ solves } \mathrm{S}_1\\
J_2(s,y,u(\cdot)) &= E\left[\int_s^T L(t,x(t),u(t))dt + \varphi(x(T))\right],\ \ x(t) \text{ solves } \mathrm{S}_2
\end{aligned}$$

respectively.

The HJB for the two systems are

$$0 = V_t + \inf_u \left(\frac{\partial V(t,x)}{\partial x} f(t,x,u) + L(t,x,u) \right) \tag{3.26}$$

$$0 = W_t + \inf_u \left(\frac{\partial W(t,x)}{\partial x} f(t,x,u) + L(t,x,u) \right) + \varepsilon \frac{\partial^2 W(x,t)}{\partial x^2} \tag{3.27}$$

We observe that the stochastic HJB can be obtained from the deterministic HJB by adding the term $\varepsilon \Delta W$. It is reasonable to expect that when $\varepsilon \to 0$, W^ε (the solution to (3.27) with a given ε) converges to V in certain sense (in fact, uniformly) since the term $\varepsilon \Delta W^\varepsilon$ vanishes as $\varepsilon \to 0$. From parabolic PDE theory, (3.27) admits smooth solutions (while (3.26) doesn't! Thus the term $\varepsilon \Delta W$ regularizes the HJB (3.26)). Since the convergence of W^ε is uniform, V should be continuous. One can show that this V is indeed the viscosity solution that we have introduced in Section 1.2.4. On the other hand, the construction of the viscosity solution in Section 1.2.4 has nothing to do with the discussion here. It is indeed a more intrinsic way of construction.

3.3 Theory of optimal filtering

3.3.1 Kallianpur-Striebel formula

Give a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]}, P)$ and

$$\begin{aligned} &\text{System: } dX_t = b(t,X_t)dt + \sigma(t,X_t)dW_t \\ &\text{Observable: } dY_t = h(t,X_t)dt + dB_t \end{aligned} \tag{3.28}$$

Assume $(B_t)_{t\in[0,T]}$ and $(W_t)_{t\in[0,T]}$ are independent d and p dimensional Brownian motions adapted to $(\mathscr{F}_t)$, $X_0 \in \mathscr{F}_0$ and $Y_0 = 0$ a.s.

The mappings

$$\begin{aligned} b &: [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^d \\ \sigma &: [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^{m\times d} \\ h &: [0,\infty) \times \mathbb{R}^d \to \mathbb{R}^p \end{aligned}$$

are assumed to be measurable. Without further assumption, we assume that the equation for (X_t, Y_t) has a unique (strong) solution.

Denote

$$\mathscr{F}_t^Y = \sigma\{Y_s : 0 \le s \le t\}$$

The goal of the filtering problem is to compute the optimal estimates $\pi_t(f) := E[f(X_t)|\mathscr{F}_t^Y]$ when $f(X_t) \in L^1$.

The idea is to construct a probability measure Q, such that X and Y are independent under Q. Then by the Bayes formula (3.19), we would have

$$\begin{aligned} \pi_t(f)(\omega) &= \frac{E^Q[f(X_t)\frac{dP}{dQ}|\mathscr{F}_t^Y]}{E^Q[\frac{dP}{dQ}|\mathscr{F}_t^Y]} \\ &= \frac{\tilde{E}^Q[f(X_t(\tilde{\omega}))\frac{dP}{dQ}(X(\tilde{\omega}),Y(\omega))]}{\tilde{E}^Q[\frac{dP}{dQ}(X(\tilde{\omega}),Y(\omega))]} \end{aligned}$$

where $X(\tilde{\omega}), Y(\omega) \in C[0,T]$. Our main tool to construct Q is the Girsanov theorem (see Theorem 3.7 in the Appendix). Define

$$\begin{aligned}\Lambda_t &= \mathscr{E}\left(-\int_0^t h(s,X_s)dB_s\right)_t \\ &= \exp\left(-\int_0^t h(s,X_s)dB_s - \frac{1}{2}\int_0^t |h(s,X_s)|^2 ds\right) \\ &= \exp\left(-\int_0^t h(s,X_s)dY_s + \frac{1}{2}\int_0^t |h(s,X_s)|^2 ds\right)\end{aligned}$$

see (3.18). Then since $Y_t = W_t - \left(-\int_0^t h(s,X_s)dB_s\right)$, it follows from Girsanov theorem (see Example 3.1) that (Y_t) is a Brownian motion under Q defined by $dQ = \Lambda_T dP$ whenever

$$E\left[\exp\left(\frac{1}{2}\int_0^T |h(s,X_s)|^2 ds\right)\right] < \infty$$

Next, we show that X and Y are indeed independent under Q. We have to prove

$$E^Q[\Phi(X)\Psi(Y)] = E^Q[\Phi(X)]E^Q[\Psi(Y)]$$

for any bounded measurable functions Φ and Ψ on $C[0,T]$. The following relations are trivial:

$$\begin{aligned}E^Q[\Phi(X)\Psi(Y)] &= E^P[\Lambda_T(X,Y)\Phi(X)\Psi(Y)] \\ &= E^P[E^P[\Lambda_T(X,Y)\Phi(X)\Psi(Y)|X]] \\ &= E^P[\Phi(X)E^P[\Lambda_T(X,Y)\Psi(Y)|X]]\end{aligned}$$

To continure, observe that

$$\begin{aligned}E^P[\Lambda_T(X,Y)\Psi(Y)|X](\omega) &= \tilde{E}^P[\Lambda_T(X(\omega), Y^{X(\omega)}(\tilde{\omega}))\Psi(Y^{X(\omega)}(\tilde{\omega}))] \\ &= \tilde{E}^{\tilde{Q}}[\Psi(Y^{X(\omega)}(\tilde{\omega}))] \\ &= \int_{C[0,T]} \Psi(y)\mu^{\bar{W}}(y)\end{aligned}$$

where

$$Y_t^{X(\omega)}(\tilde{\omega}) = \int_0^t h(s,X_s(\omega))ds + B_t(\tilde{\omega}),$$

$\mu^{\bar{W}}$ is the measure on $C[0,T]$ induced by a Brownian motion $\bar{W}$ and that $Y_t^{X(\omega)}(\tilde{\omega})$ is a Brownian motion under $d\tilde{Q} = \Lambda_T(X(\omega), Y^{X(\omega)}(\tilde{\omega}))dP$ by Girsanov theorem. Also, we see that $E^P[\Lambda_T(X,Y)\Psi(Y)|X](\omega)$ does not dependent on ω and hence is deterministic! Thus we obtain

$$E^Q[\Phi(X)\Psi(Y)] = E^P[\Phi(X)]\int_{C[0,T]} \Psi(y)\mu^W(y).$$

Choose $\Psi \equiv 1$, we get $E^Q[\Phi(X)] = E^P[\Phi(X)]$. Choose $\Phi \equiv 1$, we get $E^Q[\Psi(Y)] = \int_{C[0,T]} \Psi(y)\mu^W(y)$, which shows that Y is independent of X.

Therefore

$$\begin{aligned}\pi_t(f)(\omega) &= \frac{E^Q\left[f(X_t)\Lambda_T^{-1}|\mathscr{F}_t^Y\right]}{E^Q[\Lambda_T^{-1}|\mathscr{F}_t^Y]} \\ &= \frac{E^Q\left[f(X_t)\Lambda_t^{-1}|\mathscr{F}_t^Y\right]}{E^Q[\Lambda_t^{-1}|\mathscr{F}_t^Y]}(\omega) \\ &= \frac{\tilde{E}^Q\left[f(X_t(\tilde{\omega})\Lambda_t^{-1}(X(\tilde{\omega}),Y(\omega))\right]}{\tilde{E}^Q[\Lambda_t^{-1}(X(\tilde{\omega}),Y(\omega))]} \\ &= \frac{\int_{C[0,T]} f(\iota_t(x))\Lambda_t^{-1}(x,Y(\omega))\mu^X(dx)}{\int_{C[0,T]}\Lambda_t^{-1}(x,Y(\omega))\mu^X(dx)}\end{aligned}$$

due to the independence of X and Y. $\iota_t(x) = x_t$. The second equality follows from the following fact:

$$\begin{aligned}E^Q[Z\Lambda_t^{-1}] &= E^P[Z\Lambda_T\Lambda_t^{-1}] \\ &= E^P[Z\Lambda_t^{-1}E^P[\Lambda_T|\mathscr{F}_t]] \\ &= E^P[Z] \\ &= E^Q[Z\Lambda_T^{-1}] \\ &= E^Q[ZE^Q[\Lambda_T^{-1}|\mathscr{F}_t]]\end{aligned}$$

hence $E^Q[\Lambda_T^{-1}|\mathscr{F}_t] = \Lambda_t^{-1}$, i.e., Λ_t is an $\mathscr{F}_t$ martingale under Q.

Kallianpur-Striebel formula

$$E[f(X_t)|\mathscr{F}_t^Y](\omega) = \frac{\tilde{E}^Q\left[f(X_t(\tilde{\omega})\Lambda_t^{-1}(X(\tilde{\omega}),Y(\omega))\right]}{\tilde{E}^Q[\Lambda_t^{-1}(X(\tilde{\omega}),Y(\omega))]}$$

where

$$\Lambda_t^{-1} = \exp\left(\int_0^t h(s,X_s)dY_s - \frac{1}{2}\int_0^t |h(s,X_s)|^2 ds\right)$$

As a biproduct, we also see

$$\begin{aligned}\Lambda_t^{-1} &= \exp\left(\int_0^t h(s,X_s)dY_s - \frac{1}{2}\int_0^t |h(s,X_s)|^2 ds\right) \\ &= \mathscr{E}\left(\int_0^\cdot h(s,X_s)dY_s\right)_t\end{aligned}$$

Hence Λ_t^{-1} is an $\mathscr{F}_t$ martingale under P on $[0,T]$ i.e, $E^P[\Lambda_T^{-1}|\mathscr{F}_t] = \Lambda_t^{-1}$.

3.3.2 Zakai and FKK equation

Keep the notations as in the previous section and introduce a new one:

$$\sigma_t(f) = E^Q\left[f(X_t)\Lambda_t^{-1}|\mathscr{F}_t^Y\right]$$

then $\pi_t(f) = \frac{\sigma_t(f)}{\sigma_t(1)}$. We derive an equation for $\sigma_t(f)$. For convenience, put $z_t = \Lambda_t^{-1}$, then $dz_t = z_t h_t^T dY_t$ where we write for convenience $h_s = h_s(s,X_s)$.

then by Ito's formula

$$\begin{aligned}
df(X_t)z_t &= z_t\nabla f(X_t)^T dX_t + f(X_t)dz_t + \\
&\quad + \frac{1}{2}\sum_{i,j=1}^{d} z_t\partial_{ij}f(X_t)d\left\langle X^i, X^j\right\rangle_t + \frac{1}{2}\sum_{i=1}^{d}\partial_i f(X_t)d\left\langle X_t^i, z_t\right\rangle \\
&= z_t\nabla f(X_t)^T[b(t,X_t)dt + \sigma(t,X_t)dW_t] + f(X_t)z_t h_t^T dY_t + \frac{1}{2}\sum_{i,j,k=1}^{d}\sigma^{ik}\sigma^{jk}\partial_{ij}f(X_t)dt \\
&= z_t\left[\nabla f(X_t)^T b(t,X_t) + \frac{1}{2}\sum_{i,j,k=1}^{d}\sigma^{ik}\sigma^{jk}\partial_{ij}f(X_t)\right]dt \\
&\quad + z_t\nabla f(X_t)^T\sigma(t,X_t)dW_t + f(X_t)z_t h_t^T dY_t \\
&= z_t Lf(X_t) + z_t\nabla f(X_t)^T\sigma(t,X_t)dW_t + f(X_t)z_t h_t^T dY_t
\end{aligned}$$

or

$$\begin{aligned}
f(X_t)z_t &= f(X_0) + \int_0^t \Lambda_s^{-1}Lf(X_s)ds \\
&\quad + \int_0^t \Lambda_s^{-1}\nabla f(X_s)^T\sigma(s,X_s)dW_s + \int_0^t f(X_s)z_s h_s^T(s,X_s)dY_s.
\end{aligned} \tag{3.29}$$

where we have used:

$$\begin{aligned}
Lf &= \frac{1}{2}\sum_{i,j,k=1}^{d}\sigma^{ik}\sigma^{jk}\partial_{ij}^2 f + \sum_{i=1}^{d} b_i\partial_i f \\
\left\langle X^i, X^j\right\rangle_t &= \left\langle \int b^i dt + \int\sum_k \sigma^{ik}dW_t^k, \int b^j dt + \int\sum_k \sigma^{jk}dW_t^k\right\rangle \\
&= \left\langle \int\sum_k \sigma^{ik}dW_t^k, \int\sum_k \sigma^{jk}dW_t^k\right\rangle = \sum_k\int\sigma^{ik}\sigma^{jk}dt \\
\left\langle X_t^i, z_t\right\rangle &= 0
\end{aligned}$$

Take the conditional expectation on (3.29), we obtain

$$\begin{aligned}
&E^Q[f(X_t)\Lambda_t^{-1}|\mathscr{F}_t^Y] \\
=&E^Q[f(X_0)] + \int_0^t E^Q[\Lambda_s^{-1}Lf(X_s)|\mathscr{F}_s^Y]ds + \int_0^t E^Q[\Lambda_s^{-1}f(X_s)h^T(s,X_s)|\mathscr{F}_s^Y]dY_s
\end{aligned}$$

or

$$\sigma_t(f) = \sigma_0(f) + \int_0^t \sigma_s(Lf)ds + \int_0^t \sigma_s(h_s f)^T dY_s \tag{3.30}$$

where $(h_s f)(x) = f(x)h(s,x)$, $\sigma_0(f) = E^P[f(X_0)]$. In differential form, it also reads

$$d\sigma_t(f) = \sigma_t(Lf)dt + \sigma_t(h_t f)dY_t \tag{3.31}$$

which is an SDE. Equation (3.30) is called the Zakai equation.

Zakai equation

$$\sigma_t(f) = \sigma_0(f) + \int_0^t \sigma_s(Lf)ds + \int_0^t \sigma_s(h_s f)dY_s$$

Differential form

$$d\sigma_t(f) = \sigma_t(Lf)dt + \sigma_t(h_t f)dY_t$$

where

$$Lf = \frac{1}{2}\sum_{i,j,k=1}^{d} \sigma^{ik}\sigma^{jk}\partial^2_{ij} f + \sum_{i=1}^{d} b_i \partial_i f$$

$$(h_s f)(x) = h^T(s,x) f(x)$$

With the Zakai equation, we can now derive a equation for $\pi_t(f)$ using Ito's formula:

$$\begin{aligned} d\pi_t(f) &= d\left(\frac{\sigma_t(f)}{\sigma_t(1)}\right) \\ &= \frac{d\sigma_t(f)}{\sigma_t(1)} - \frac{\sigma_t(f)d\sigma_t(1)}{\sigma_t(1)^2} + \frac{\sigma_t(f)|\sigma_t(h)|^2}{\sigma_t(1)^3}dt - \frac{\sigma_t(h)^T\sigma_t(h_t f)}{\sigma_t^2(1)}dt \\ &= \pi_t(Lf) + [\pi_t(h_t f) - \pi_t(f)\pi_t(h)]^T [dY_t - \pi_t(h)dt] \end{aligned}$$

or

$$\pi_t(f) = \pi_0(f) + \int_0^t \pi_s(L_s f)ds + \int_0^t [\pi_s(h_s f) - \pi_s(f)\pi_s(h)]^T d\bar{B}_s \tag{3.32}$$

where

$$\bar{B}_t = Y_t - \int_0^t \pi_s(h)ds \tag{3.33}$$

or

$$d\bar{B}_t = dY_t - \pi_s(h)dt$$

and we have used:

$$\begin{aligned} &L1 = 0 \\ &h_t 1(x) = h(t,x)^T \Rightarrow \sigma_t(h_t 1) = \sigma_t(h^T) \\ &d\sigma_t(1) = \sigma_t(h^T)dY_t \\ &d\left(\frac{x_t}{y_t}\right) = \frac{dx_t}{y_t} - \frac{x_t dy_t}{y_t^2} - \frac{d\langle x,y\rangle_t}{y_t^2} + \frac{x_t d\langle y,y\rangle_t}{y_t^3} \\ &d\langle \sigma_t(f), \sigma_t(1)\rangle_t = |\sigma_t(h_t f)|^2 d\langle Y,Y\rangle_t = |\sigma_t(h_t f)|^2 dt \end{aligned}$$

The process $\bar{B}_t$ is so important that it has a name: the *innovation process* of the filter.

The formula (3.32) is called te *FKK equation.*

FKK equation

$$\pi_t(f) = \pi_0(f) + \int_0^t \pi_s(L_s f)ds + \int_0^t [\pi_s(h_s f) - \pi_s(f)\pi_s(h)]^T d\bar{B}_s$$

where

$$\bar{B}_t = Y_t - \int_0^t \pi_s(h)ds$$

Proposition 3.3. *The innovation process $(\bar{B}_t)_{t\in[0,T]}$ is a Brownian motion adpated to $(\mathscr{F}_t^Y)_{t\in[0,T]}$.*

Proof. Clearly $\bar{B}_0 = 0$ a.s, and

$$\bar{B}_t = \int_0^t [h_s - \pi_s(h)]\,ds + B_t$$

where we have written $h_s = h(s, X_s)$ for convenience. Recall that $\pi_t(h) = E[h_t|\mathscr{F}_t^Y]$. It suffices to show

$$E\left[e^{i\alpha^T(\bar{B}_t-\bar{B}_s)}|\mathscr{F}_t^Y\right] = e^{-|\alpha|^2(t-s)/2}.$$

For this, we apply Ito's formula to $\eta_t = \exp(i\alpha^T\bar{B}_t)$:

$$\begin{aligned} e^{i\alpha^T\bar{B}_t} = e^{i\alpha^T\bar{B}_s} &+ i\int_s^t e^{i\alpha^T\bar{B}_u}\alpha^T dB_u \\ &+ i\int_s^t e^{i\alpha^T\bar{B}_u}\alpha^T(h_u-\pi_u(h))du - \frac{1}{2}|\alpha|^2\int_s^t e^{i\alpha^T\bar{B}_u}du \end{aligned}$$

An immediate observation is that $E\left[\int_s^t e^{i\alpha^T\bar{B}_u}\alpha^T dB_u|\mathscr{F}_s^Y\right] = 0$ since $\int_s^t e^{i\alpha^T\bar{B}_u}\alpha^T dB_u$ is an $\mathscr{F}_t \supset \mathscr{F}_t^Y$ martingale. Further, for $u \geq s$,

$$\begin{aligned} E\left[e^{i\alpha^T\bar{B}_u}\pi_u(h)|\mathscr{F}_s^Y\right] &= E\left[e^{i\alpha^T\bar{B}_u}E[h_u|\mathscr{F}_u^Y]|\mathscr{F}_s^Y\right] \\ &= E\left[e^{i\alpha^T\bar{B}_u}h_u|\mathscr{F}_s^Y\right] \end{aligned}$$

thus $E\left[\int_s^t e^{i\alpha^T\bar{B}_u}\alpha^T(h_u-\pi_u(h))du|\mathscr{F}_s^Y\right] = \alpha^T\int_s^t E\left[e^{i\alpha^T\bar{B}_u}h_u - \pi_u(h)|\mathscr{F}_s^Y\right]du = 0$. Combining these two, we arrive at

$$E\left[e^{i\alpha^T\bar{B}_t}h_t|\mathscr{F}_s^Y\right] = e^{i\alpha^T\bar{B}_s} - \frac{1}{2}|\alpha|^2\int_s^t E\left[e^{i\alpha^T\bar{B}_u}h_u|\mathscr{F}_s^Y\right]du$$

and the proof is completed. □

Suppose that there is a density $p_t(x)$ such that

$$p_t(x) = \frac{dP(X_t \preceq x|\mathscr{F}_t^Y)}{dx}$$

then

$$\pi_t(f) = E[f(X_t)|\mathscr{F}_t^Y] = \int_{\mathbb{R}^d} f(x)p_t(x)\mathrm{d}x$$

Substitute this into (3.32) and suppose that $f \in C_c^2(\mathbb{R}^d)$, then

$$\begin{aligned} \int_{\mathbb{R}^d} f(x)p_t(x)dx &= \int_{\mathbb{R}^d} f(x)p_0(x)dx + \int_0^t\int_{\mathbb{R}^d}(L_sf)(x)p_s(x)dxds \\ &\quad + \int_0^t\left[\int_{\mathbb{R}^d}(h_sf)(x)p_s(x)dx - \left(\int_{\mathbb{R}^d} f(x)p_s(x)dx\right)\pi_s(h)^T\right]d\bar{B}_s \\ &= \int_{\mathbb{R}^d} f(x)p_0(x)dx + \int_{\mathbb{R}^d} f(x)\left(\int_0^t L_s^*p_s(x)ds\right)dx \\ &\quad + \int_{\mathbb{R}^d} f(x)\left(\int_0^t h_s^Tp_s(x)d\bar{B}_s\right)dx - \int_{\mathbb{R}^d} f(x)\left(\int_0^t p_s(x)\pi_s(h)^Td\bar{B}_s\right)dx \\ &= \int_{\mathbb{R}^d} f(x)\left[p_0(x) + \int_0^t L_s^*p_s(x)ds + \int_0^t p_s(x)[h_s-\pi_s(h)]^Td\bar{B}_s\right]dx \end{aligned}$$

Thus

$$dp_t(x) = L_t^*p_t(x)dt + p_t(x)[h(t,x)-\pi_t(h)]^Td\bar{B}_t. \tag{3.34}$$

Since we have assumed that $f \in C_c^2(\mathbb{R}^d)$, the above equation should be understood in the week sense.

If however, the density for the unnormalized quantity $\sigma_t(f)$ is quested, i.e., search for $q_t(x)$ such that

$$\sigma_t(f) = \int_{\mathbb{R}^d} f(x) q_t(x) dx, \quad \forall f \in C_c^2(\mathbb{R}^d)$$

Note that if this is the case, then

$$\int_{\mathbb{R}^d} f(x) p_t(x) \mathrm{d}x = \int_{\mathbb{R}^d} f(x) \frac{q_t(x)}{\int_{\mathbb{R}^d} q_t(x) dx} dx$$

hence

$$p_t(x) = \frac{q_t(x)}{\int_{\mathbb{R}^d} q_t(x) dx}$$

which implies

$$\pi_t(f) = \int_{\mathbb{R}^d} f(x) p_t(x) \mathrm{d}x = \frac{\int_{\mathbb{R}^d} f(x) q_t(x) \mathrm{d}x}{\int_{\mathbb{R}^d} q_t(x) dx}$$

Thus, if the equation for $q_t(x)$ is simpler than (3.34), we can calculate $q_t(x)$ first and then use the last formula to calculate $p_t(x)$. Using (3.31), we easily find

$$dq_t(x) = L_t^* q_t(x) dt + h(t,x)^T dY_t \tag{3.35}$$

of which the initial distribution $q_0(x)$ is determined by the distribution of X_0. This equation is called the *Zakai-PDE*.

Equations for condtional density

Define

$$\pi_t(f) = \int_{\mathbb{R}^d} f(x) p_t(x) \mathrm{d}x$$
$$\sigma_t(f) = \int_{\mathbb{R}^d} f(x) q_t(x) \mathrm{d}x$$

then

$$\text{normalized: } dp_t(x) = L_t^* p_t(x) dt + p_t(x)[h(t,x) - \pi_t(h)]^T d\bar{B}_t$$
$$\text{unnormalized: } dq_t(x) = L_t^* q_t(x) dt + h(t,x)^T dY_t$$

3.3.3 Kalman-Bucy filter

Zero input

In this section, we consider filtering problem of the linear model (zero input):

$$\begin{aligned} dX_t &= [A(t)X_t + D(t)u_t]dt + C(t)dW_t \\ dY_t &= H(t)X_t dt + dB_t \end{aligned} \tag{3.36}$$

where $A(t), C(t), H(t)$ are deterministic real matrices of dimensions $n \times n$, $n \times m$, $l \times n$ respectively. W_t and B_t are Brownian motions adpated to filtration $\mathscr{F}_t$ (w.l.o.g, one can take $\mathscr{F}_t = \mathscr{F}_t^W \vee \mathscr{F}_t^B$). $Y_0 = 0$ a.e. and X_0, Y_0 are independent. u_t is the control input adpated to $\mathscr{F}_t^Y$.

It is customary in the linear case to assume $X_0 \sim N(\hat{X}_0, \hat{P}_0)$, i.e., the initial distribution of X_t is Gaussian and that the equation for (X_t, Y_t) has a unique strong solution adapted to $\mathscr{F}_t$.

In this subsection, we restrict ourselves to the zero input case, i.e., $u_t \equiv 0$ for all $t \geq 0$.

Define $\hat{X}_t := E[X_t|\mathscr{F}_t^Y]$ (notice that this is consistent with the notation $\hat{X}_0$ introduced earlier). Since the system (3.36) forms a Gaussian system, we are also interested in the covariance matrix of the conditional mean: $\hat{P}_t := E[(X_t - \hat{X}_t)(X_t - \hat{X}_t)^T|\mathscr{F}_t^Y]$. Due to Proposition 3.4 below, $\hat{P}_t$ is a deterministic function, i.e., $\hat{P}_t := E[(X_t - \hat{X}_t)(X_t - \hat{X}_t)^T]$.

To apply FKK equation, let $f(x) = x^i$, then

$$Lf(x) = \sum_k A_{ik} x_k, \;\; h_t f(x) = x^i H(t) x, \;\; h(t,x) = H(t)x$$

hence

$$\hat{X}_t^i = E[X_t^i|\mathscr{F}_t^Y] = E[X_0^i] + \int_0^t \sum_k A_{ik} \hat{X}_s^k ds + \int_0^t \left[E[X_s^i H(s) X_s|\mathscr{F}_s^Y] - \hat{X}_s^i H(s) \hat{X}_s \right]^T d\bar{B}_s.$$

Align all $\hat{X}_t^i$ as a column vector, we get

$$\begin{aligned}
\hat{X}_t &= E[X_0] + \int_0^t A\hat{X}_s ds + \int_0^t \left[E[H(s) X_s X_s^T|\mathscr{F}_s^Y] - H(s)\hat{X}_s \hat{X}_s \right]^T d\bar{B}_s \\
&= E[X_0] + \int_0^t A\hat{X}_s ds + \int_0^t \left[E[X_s X_s^T|\mathscr{F}_s^Y] - \hat{X}_s \hat{X}_s \right] H(s)^T d\bar{B}_s \\
&= E[X_0] + \int_0^t A\hat{X}_s ds + \int_0^t E[(X_s - \hat{X}_s)(X_s - \hat{X}_s)^T|\mathscr{F}_s^Y] H(s)^T d\bar{B}_s \\
&= E[X_0] + \int_0^t A\hat{X}_s ds + \int_0^t \hat{P}_s H(s)^T d\bar{B}_s
\end{aligned}$$

or equivalently

$$d\hat{X}_t = A(t)\hat{X}_t dt + \hat{P}_t H(t)^T d\bar{B}_t \tag{3.37}$$

$$d\bar{B}_t = dY_t - H(t)\hat{X}_t dt \tag{3.38}$$

with $\hat{X}_0 = E[X_0]$.

To derive the equation for $\hat{P}_t$, first notice that

$$\begin{aligned}
\hat{P}_t &= E[(X_t - \hat{X}_t)(X_t - \hat{X}_t)^T] \\
&= EX_t X_t^T - E[\hat{X}_t \hat{X}_t^T],
\end{aligned}$$

and then we apply Ito's formula to $\hat{P}_t$:

$$\begin{aligned}
dX_t^i X_t^j &= X_t^i dX_t^j + X_t^j dX_t^i + d\left\langle X^i, X^i \right\rangle_t \\
&= X_t^i dX_t^j + X_t^j dX_t^i + C_i C_j^T dt \\
&= X_t^i (A^j X_t dt + C^j dW_t) + X_t^j (A^i X_t dt + C^i dW_t) + C_i C_j^T dt \\
d\hat{X}_t^i \hat{X}_t^j &= \hat{X}_t^i d\hat{X}_t^j + \hat{X}_t^j d\hat{X}_t^i + d\left\langle \hat{X}^i, \hat{X}^j \right\rangle_t \\
&= \hat{X}_t^i d\hat{X}_t^j + \hat{X}_t^j d\hat{X}_t^i + \hat{P}_t^i H(t)^T H(t) (\hat{P}_t^j)^T dt \\
&= \hat{X}_t^i (A^j \hat{X}_t dt + \hat{P}_t^j H^T d\bar{B}_t) + \hat{X}_t^j (A^i \hat{X}_t + \hat{P}_t^i H^T d\bar{B}_t) + \hat{P}_t^i H(t)^T H(t) (\hat{P}_t^j)^T dt
\end{aligned}$$

hence

$$dE[X_t^i X_t^j] = [E(A^j X_t^i X_t + A^i X_t^j X_t) + C_i C_j^T]dt$$
$$dE[\hat{X}_t^i \hat{X}_t^j] = [E(A^j \hat{X}_t^i \hat{X}_t + A^i \hat{X}_t^j \hat{X}_t) + \hat{P}_t^i H(t)^T H(t)(\hat{P}_t^j)^T]dt$$

and

$$\frac{d\hat{P}_t^{ij}}{dt} = A^j E[X_t^i X_t - \hat{X}_t^i \hat{X}_t] + A^i E[X_t^j X_t - \hat{X}_t^j \hat{X}_t] + C_i C_j^T - \hat{P}_t^i H(t)^T H(t)(\hat{P}_t^j)^T$$

or equivalently

$$\frac{d\hat{P}_t}{dt} = A(t)\hat{P}_t + \hat{P}_t(t)A(t)^T + C(t)C(t)^T - \hat{P}_t H(t)^T H(t)\hat{P}_t \tag{3.39}$$

with $\hat{P}_0 = P_0$, i.e., the covariance matrix of X_0.

Kalman-Bucy filter

System:

$$dX_t = A(t)X_t dt + C(t)dW_t$$
$$dY_t = H(t)X_t dt + dB_t$$

Filter:

$$d\hat{X}_t = A(t)\hat{X}_t dt + \hat{P}_t H(t)^T d\bar{B}_t$$
$$d\bar{B}_t = dY_t - H(t)\hat{X}_t dt$$

where

$$\frac{d\hat{P}_t}{dt} = A(t)\hat{P}_t + \hat{P}_t(t)A(t)^T + C(t)C(t)^T - \hat{P}_t H(t)^T H(t)\hat{P}_t$$

Proposition 3.4. *The process $X_t - \hat{X}_t$ is independent of $\mathscr{F}_t^Y$, i.e., $E[f(X_s - \hat{X}_s)|\mathscr{F}_s^Y] = E[f(X_s - \hat{X}_s)]$ a.s. for any bounded measurable f.*

Proof. Step 1: we show that the conditional distribution $X_t|\mathscr{F}_t^Y$ is Gaussian. Fix t and let $Y_n^k = X_{kt/2^n}$, $n \geq 1$. Define

$$\mathscr{E}_n = \sigma\left\{Y_n^k : k = 1, \cdots, 2^n\right\}$$

Then $\mathscr{E}_n$ is a filtration ($\mathscr{E}_\infty = \mathscr{F}_t^Y$). Since (X, Y) is a Gaussian process on $[0, t]$, the joint distribution of $\{(X_{kt/2^n}, Y_{kt/2^n})\}_{k=1}^{2^n}$ is a Gaussian vector and thus the conditional expectation $X_t|\mathscr{E}_n$ is Gaussian with mean $E[X_t|\mathscr{E}_n] =: \hat{X}_t^n$ and covariance $\hat{P}_t^n$. Let $\pi_t^n(A) := P\{X_t \in A|\mathscr{E}_n\}$, then

$$\phi_n(\lambda) = E[\exp(i\lambda^T X_t)|\mathscr{E}_n] = \int_{\mathbb{R}^d} \exp(i\lambda^T x)\pi_t^n(dx)$$
$$= \exp\left(\lambda^T \hat{X}_t^n - \frac{1}{2}\lambda^T \hat{P}_t^n \lambda\right).$$

Since $\{\phi_n(\lambda)\}_{n=1}^{\infty}$ and $\{\hat{X}_t^n\}_{n=1}^{\infty}$ are both uniformly integrable martingales adapted to $\mathscr{E}_n$, $\phi_\infty(\lambda)$ and $\hat{X}_t^\infty$ exist and are in L^1. Thus $\hat{P}_t^\infty$ also exists. To sum up, in a.s. sense,

$$\begin{aligned} E[\exp(i\lambda^T X_t)|\mathscr{F}_t^Y] &= E[\exp(i\lambda^T X_t)|\mathscr{E}_\infty] \\ &= \lim_{n\to\infty} \phi_n(\lambda) \\ &= \exp\left(\lambda^T \hat{X}_t - \frac{1}{2}\lambda^T \hat{P}_t \lambda\right) \end{aligned}$$

here we have omitted the superscript "∞". Thus $X_t|\mathscr{F}_t^Y \sim N(\hat{X}_t, \hat{P}_t)$.Step 2: $X_t - \hat{X}_t$ is independent of $\mathscr{F}_t^Y$. It is known from elementary probability theory that when (X, Y) are jointly Gaussian, then $X - E[X|Y]$ is independent of Y. From Step 1, we know that $X_t - \hat{X}_t^n$ is independent of $\mathscr{E}_n$ for all n. For any bounded measurable function f and $A \in \mathscr{F}_t^Y$, let $A_n = E[1_A|\mathscr{E}_n]$, we have

$$E[f(X_t - \hat{X}_t^n)1_{A_n}] = E[f(X_t - \hat{X}_t^n)]P(A_n)$$

but

$$\begin{aligned} \lim_{n\to\infty} E[f(X_t - \hat{X}_t^n)1_{A_n}] &= E[f(X_t - \hat{X}_t)1_A] \\ \lim_{n\to\infty} E[f(X_t - \hat{X}_t^n)]P(A_n) &= E[f(X_t - \hat{X}_t)]P(A) \end{aligned}$$

thus $E[f(X_t - \hat{X}_t)1_A] = E[f(X_t - \hat{X}_t)]P(A)$. The conclusion now follows. □

To find the conditional density, we use the Zakai-PDE, which reads

$$dq_t(x) = q_t(x)x^T H(t)^T dY_t + \mathrm{tr}\left[\frac{1}{2}C(t)C(t)^T \mathrm{Hess}(q_t(x)) - \nabla(q_t(x)A(t)x)\right]dt$$

which has a solution of the form

$$q_t(x) = \mathrm{const} \times \exp\left(-\frac{1}{2}(x - \hat{X}_t)^T \hat{P}_t^{-1}(x - \hat{X}_t)\right).$$

Non-zero input

The non-zero input case is also important, which is not evident right now but will be clear in the next section.

We use a superscript "^{u}" to indicate the signal under control input u. A first observation is that

$$\begin{aligned} X_t^u &= \int_0^t A(s)X_s \mathrm{d}s + \int_0^t D(s)u_s \mathrm{d}s + \int_0^t C(s)dW_s \\ \hat{X}_t^u &= E\left[\int_0^t A(s)X_s \mathrm{d}s|\mathscr{F}_t^Y\right] + \int_0^t D(s)u_s \mathrm{d}s + E\left[\int_0^t C(s)dW_s|\mathscr{F}_t^Y\right] \end{aligned}$$

and then

$$X_t^u - \hat{X}_t^u = X_t^0 - \hat{X}_t^0.$$

There are two implifications from the above formula: first, the covariance matrix $\hat{P}_t^u$ does not depend on u, i.e., $\hat{P}_t^u = \hat{P}_t$; second, the differential of the above formula results in

$$\begin{aligned} d\hat{X}_t^u &= d\hat{X}_t^0 + dX_t^u - dX_t^0 \\ &= d\hat{X}_t^0 + D(t)u_t dt. \end{aligned}$$

Hence, the only thing we need to do to obtain the Kalman-Bucy filter with control is to add a term $D(t)u_t\mathrm{d}t$ in (3.37) and keep the innovation process and covariance matrix unchanged.

Kalman-Bucy filter with input

System:

$$dX_t = [A(t)X_t + D(t)u_t]dt + C(t)dW_t$$
$$dY_t = H(t)X_t dt + dB_t$$

Filter:

$$d\hat{X}_t = [A(t)\hat{X}_t + D(t)u_t]dt + \hat{P}_t H(t)^T d\bar{B}_t$$
$$d\bar{B}_t = dY_t - H(t)\hat{X}_t dt$$

where

$$\frac{d\hat{P}_t}{dt} = A(t)\hat{P}_t + \hat{P}_t(t)A(t)^T + C(t)C(t)^T - \hat{P}_t H(t)^T H(t)\hat{P}_t$$

3.3.4 Numerical method

Particle filter

Monte Carlo method

3.4 Partial State LQG and Separation Principle

This section is devoted to linear quadratic Gaussian control of the linear system

$$\begin{aligned} &\text{system: } dX_t = (A(t)X_t + B(t)u_t)dt + C(t)dW_t \\ &\text{observable: } dY_t = H(t)X_t dt + dB_t \end{aligned} \tag{3.40}$$

under the optimal cost

$$J[u] = E\left[\int_0^T (X_t^u)^T Q(t)X_t^u + u_t^T R(t)u_t \mathrm{d}t + X_T^u Q_f X_T^u\right] \tag{3.41}$$

where $Q(t)$, $R(t)$ and Q_f are all semi-positive definite. The term X_t^u represents the solution of the system under control u_t, which is required to depend only on the information $\{Y_s\}_{s\in[0,t]}$. In other words, u_t is $\mathscr{F}_t^Y$ measurable.

The first and the third terms in the cost functional are somewhat annoying since they are not observable. However, we can perform an easy manipulation to transform $J[u]$ into a more tractable form. This is achieved by applying the tower property of conditional expection:

$$J[u] = E\left[\int_0^T E[(X_t^u)^T Q(t)X_t^u|\mathscr{F}_t^Y] + E[u_t^T R(t)u_t]\mathrm{d}t + E[X_T^u Q_f X_T^u|\mathscr{F}_t^Y]\right]$$

Now the terms in the cost function are all observable! Instead of viewing this as a "magic", we would rather say that this somewhat natural. If we admit this fact, then the famous "separation principle" will

come as a natural consequence. To see this, let $\hat{X}_t^u = E[X_t^u|\mathscr{F}_t^Y]$, then

$$\begin{aligned}
&E[(X_t^u)^T Q(t) X_t^u|\mathscr{F}_t^Y] \\
&= E[(X_t^u - \hat{X}_t^u + \hat{X}_t^u)^T Q(t)(X_t^u - \hat{X}_t^u + \hat{X}_t^u)|\mathscr{F}_t^Y] \\
&= E[(X_t^u - \hat{X}_t^u)^T Q(t)(X_t^u - \hat{X}_t^u)|\mathscr{F}_t^Y] \\
&\quad + 2E[(X_t^u - \hat{X}_t^u)^T Q(t)\hat{X}_t^u|\mathscr{F}_t^Y] + E[(\hat{X}_t^u)^T Q(t)\hat{X}_t^u|\mathscr{F}_t^Y] \\
&= \mathrm{tr}(E[Q(t)(X_t^u - \hat{X}_t^u)(X_t^u - \hat{X}_t^u)^T|\mathscr{F}_t^Y]) + E[(\hat{X}_t^u)^T Q(t)\hat{X}_t^u] \\
&= \mathrm{tr}(E[Q(t)\hat{P}_t^u]) + E[(\hat{X}_t^u)^T Q(t)\hat{X}_t^u]
\end{aligned}$$

As mentioned in the previous subsection, the term $\mathrm{tr}(E[Q(t)\hat{P}_t^u])$ is deterministic and does not depend on u. Hence it does not affect the optimal value of $J[u]$. To make this precise, rewrite $J[u]$ as

$$J[u] = E\left[\int_0^T (\hat{X}_t^u)^T Q(t)\hat{X}_t^u + u_t^T R(t) u_t \mathrm{d}t + \hat{X}_T^u Q_f \hat{X}_T^u\right] + \mathrm{tr}(E[Q(t)\hat{P}_t] + E[Q_f \hat{P}_T])$$

and we can claim

$$\arg\min_u J[u] = \arg\min_u \bar{J}[u]$$

where $\bar{J}(u)$ is

$$\bar{J}(u) = E\left[\int_0^T (\hat{X}_t^u)^T Q(t)\hat{X}_t^u + u_t^T R(t) u_t \mathrm{d}t + \hat{X}_T^u Q_f \hat{X}_T^u\right].$$

Now the optimal control problem has been transformed into a "full-state observable" one. Thus invoking the results for full-state LQG, we can immediately state the following theorem.

Theorem 3.9. *Let $\hat{P}_t$ be the solution of the Riccati equation*

$$\begin{aligned}
\frac{d\hat{P}_t}{dt} &= A(t)\hat{P}_t + \hat{P}_t A^T(t) - \hat{P}_t H(t)^T H(t)\hat{P}_t + C(t)C(t)^T \\
\hat{P}_0 &= \mathrm{Cov}(X_0)
\end{aligned}$$

and K_t be the solution of the time-reversed Riccati equation

$$\begin{aligned}
\frac{dK_t}{dt} &= -A(t)^T K_t - K_t A(t) + K_t D(t) R^{-1}(t) D(t)^T K_t - R(t) \\
K_T &= Q_f
\end{aligned}$$

Then the partial state LQG has a solution

$$u_t^* = -R^{-1}(t)D(t)^T K_t \hat{X}_t$$

where $\hat{X}_t$ satisfies

$$\begin{aligned}
d\hat{X}_t &= (A(t) - D(t)R^{-1}(t)D(t)^T K_t)\hat{X}_t dt + \hat{P}_t H(t)^T d\bar{B}_t \\
d\bar{B}_t &= dY_t - H(t)\hat{X}_t dt.
\end{aligned}$$

This theorem has the spirit of "separation" since as we know K_t is the optimal gain for full-state LQG and $\hat{X}_t$ is the output of the optimal filter. Thus the theorem suggests that we can divide the design of the partial-state LQG into two parts. The first part is filtering, i.e., to obtain $\hat{X}_t$ and the filter gain $\hat{P}_t$, the second part amounts to the design of a full-state LQG based on the filter state $\hat{X}_t$. These two parts can be designed separately.

CHAPTER 4

OPTIMAL TRANSPORT

4.1 Monge and Kantorovich problem

4.1.1 The Kantorovich problem

Consider an extremely simplified model for the power grid in an isolated region which consists n power plants and m transformer stations located at different places. Label the power plants and transformer stations as p_i and t_j, $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, m\}$. Assume that the amount of electricity that the plant p_i can generate each day is a fixed value a_i, and the electricity that the transformer station t_j should receive each day is fixed at b_j. Assume additionally that there is no loss during the electricity transfer and that all the electricity will be sent to the transformer stations, in other words,

$$\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_j. \tag{4.1}$$

The cost of sending unit electricity from plant p_i to transformer station t_j is $c(p_i, t_j)$, where c is a non-negative real function.

Now the state grid corporation needs the decide a power transfer plan with the minimum cost. That is, how much electricity should power plant p_i send to transformer station t_j? Let us denote the amount sent from p_i to t_j as π_{ij}. Then the total cost is

$$J(\pi) = \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} c(p_i, t_j) \tag{4.2}$$

in which π is the compact notation for the decision variables (π_{ij}). See Figure 4.1.

Let us check the constraint on π. On the sender side, each plant p_i should send out all the power (i.e., a_i) it generates, which means that

$$\sum_{j=1}^{m} \pi_{ij} = a_i \tag{4.3}$$

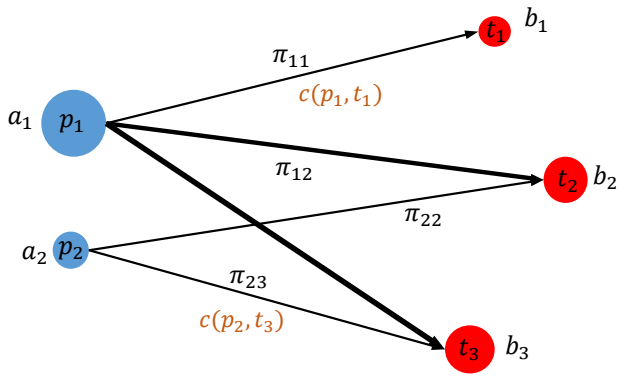

Figure 4.1: Kantorovich problem.

and on the receiver side, the amount of electricity that transformer station t_j needs is b_j, which implies that

$$\sum_{i=1}^{n} \pi_{ij} = b_j. \tag{4.4}$$

Note that constraint (4.1) is now satisfied automatically. Putting together equations (4.2-4.4), we arrive at the following optimization problem

$$\begin{aligned} \min_{\pi} J(\pi) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} c(p_i,t_j) \\ \text{subject to:} \quad & \sum_{j=1}^{m} \pi_{ij} = a_i \\ & \sum_{i=1}^{n} \pi_{ij} = b_j \\ & \pi_{ij} \geq 0 \end{aligned} \tag{SP}$$

in which $\{a_i\}_{i=1}^m$, $\{b_j\}_{j=1}^n$ and $\{c(p_i,t_j)\}_{i=1,\cdots,n;\, j=1,\cdots,m}$ are known coefficients. Let us denote this problem as (SP). Obviously, the SP problem is a linear programming problem. Define

$$\Pi(a,b) = \{P \in \mathbb{R}^{n\times m} : P_{ij} \geq 0,\ \sum_{j=1}^{m} P_{ij} = a_i,\ \sum_{i=1}^{n} P_{ij} = b_j\} \tag{4.5}$$

and for $P \in \Pi(a,b)$, denote

$$\langle P, C\rangle := \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} C_{ij}.$$

With these notations, the SP problem can be conveniently written as

$$\min_{P\in\Pi(a,b)} \langle P, C\rangle. \tag{4.6}$$

As we have mentioned, the SP problem is a linear programming. Thus it can be solved by all linear programming algorithms, e.g., network simplex method. Typically, the computational complexity is of order $O(d^3 \log d)$ where $d = m + n$.

4.1.2 The Monge problem

We now put another constraint on the transshipment problem. Suppose that the electricity generated by each one of the power plant is to be sent to only one power transformer station. For example, this may happen when building transmission lines to multiple transformer stations is impossible or too expensive. For each power plant p_i, denote $T(p_i) \in \{t_1, \cdots, t_m\}$ as its target. Then the total cost can now be written as

$$J(T) = \sum_{i=1}^{n} c(p_i, T(p_i)) a_i,$$

and the constraints (4.3,4.4) are replaced by

$$\sum_{i: T(p_i)=t_j} a_i = b_j$$

accordingly. The objective now is to seek for a map T which minimizes $J(T)$ under the above constraint:

$$\begin{aligned} &\min_T J(T) = \sum_{i=1}^{n} c(p_i, T(p_i)) a_i \\ &\text{subject to:} \quad \sum_{i: T(p_i)=t_j} a_i = b_j \end{aligned} \tag{4.7}$$

See Figure 4.2.

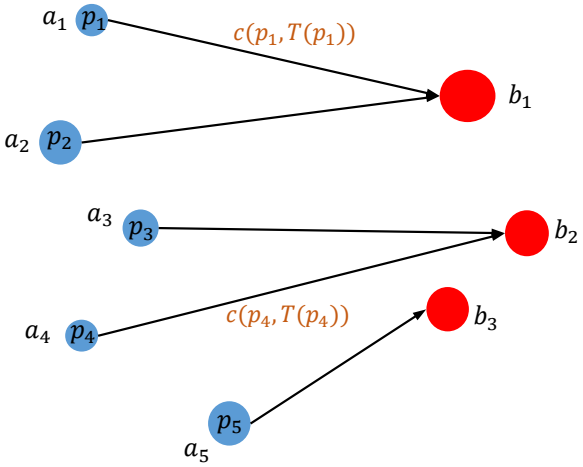

Figure 4.2: Kantorovich problem.

We call this problem a Monge problem. Unlike the SP problem that we discussed earlier which is a linear programming, the MP problem is nonlinear: the cost function $c(\cdot,\cdot)$ and $T(\cdot)$ itself may both be nonlinear. Hence the Monge problem is much more difficult and less well-behaved. But this problem is still important in applications.

It is interesting to note that the optimal value of the MP problem is always bigger than that of the SP problem, since

$$P_{ij} = \begin{cases} a_i, & \text{if } T(p_i) = t_j \\ 0, & \text{else} \end{cases}$$

is always an admissible plan for admissible T. A natural question to ask is: when are the two optimal values equal? This question is very important because it tells us when can we recast the MP problem, which is ill-behaved as an SP problem, which is a linear programming. This question is however, not obvious at all. Later, we will work this out in a systematic manner under a more general framework of optimal transportation.

4.1.3 The dual of Kantorovich problem

Associate with every linear programming problem, there is a dual problem. We derive this dual from scratch since the methodology will be used later to derive more general optimal transport dual problems.

To streamline the derivations, we introduce some useful notations that will be frequently used in the sequel. Let $C \in \mathbb{R}^{n\times m}$ be the matrix whose component at the i-th row and j-th column is $c(p_i, t_j)$, and a, b two column vectors whose i-th and j-th element is a_i and b_j respectively.

Introduce the indicator function of a set A:

$$I_A(x) = \begin{cases} 0, & \text{if } x \in A \\ +\infty, & \text{if } x \notin A \end{cases}.$$

Then the problem (4.6) is equivalent to

$$\min_{P\in\mathbb{R}^{n\times n}_{\geq 0}} \left\{\langle P, C\rangle + I_{\Pi(a,b)}(P)\right\}$$

For $a \in \mathbb{R}^n_{\geq 0}$ and $b \in \mathbb{R}^m_{\geq 0}$, define $a \oplus b$ as the matrix whose i-th row and j-th column element is $a_i + b_j$.

$$\begin{aligned} I_{\Pi(a,b)}(P) &= \sup_{f\in\mathbb{R}^n, g\in\mathbb{R}^m} \sum_{i=1}^n (a - \sum_{j=1}^m P_{ij}) f_i + \sum_{j=1}^m (b_j - \sum_{i=1}^n P_{ij}) g_j \\ &= \sup_{f\in\mathbb{R}^n, g\in\mathbb{R}^m} a^\top f + b^\top g - \left\langle P, f \oplus g\right\rangle \end{aligned}$$

thus

$$\begin{aligned} \min_{P\in\mathbb{R}^{n\times n}_{\geq 0}} \left\{\langle P, C\rangle + I_{\Pi(a,b)}(P)\right\} &= \min_{P\in\mathbb{R}^{n\times m}_{\geq 0}} \sup_{f\in\mathbb{R}^n, g\in\mathbb{R}^m} a^\top f + b^\top g - \left\langle P, f\oplus g - C\right\rangle \\ &= \sup_{f\in\mathbb{R}^n, g\in\mathbb{R}^m} \min_{P\in\mathbb{R}^{n\times m}_{\geq 0}} a^\top f + b^\top g - \left\langle P, f\oplus g - C\right\rangle \\ &= \sup_{f\oplus g\leq C} \left\langle f, a\right\rangle + \left\langle g, b\right\rangle \end{aligned}$$

where in the second equality, we swapped the minimization and maximization which is legitimate due to *minimax theorem* of linear programming problem.

4.1.4 From "discrete" to "continuous" optimal transport

Suppose now that we are going to move a pile of sand from X to Y to construct certain structures, see Figure 4.3. The sand on the left of the figure can be described by a density function $f : X \to \mathbb{R}$ and the sand on the right is described by some density function $g : Y \to \mathbb{R}$. Suppose that the unit cost of moving the sand from the interval $(x, x+dx)$ to interval $(y, y+dy)$ is $c(x, y)dxdy$, and that there is $\Pi(dx, dy)$ amount of sand moving from $(x, x + dx)$ to $(y, y + dy)$. When dx and dy are sufficiently small, we may assume

the existence of some function $\pi(x,y)$, satisfying $\Pi(dx,dy)=\pi(x,y)dxdy$. Due to mass preservation, we must have

$$\int_X \Pi(dx,dy)=g(y)dy, \quad \int_Y \Pi(dx,dy)=f(x)dx$$

or

$$\int_X \pi(x,y)\mathrm{d}x=g(y), \quad \int_Y \pi(x,y)\mathrm{d}y=f(x).$$

The total cost is

$$\int_X\int_Y c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y=\int_{X\times Y} c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y.$$

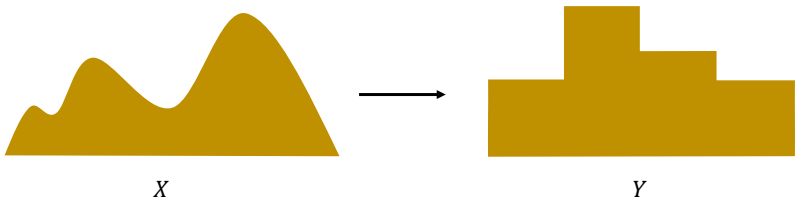

Figure 4.3: Moving a continuous distribution.

Thus the problem is formulated as

$$\begin{aligned}&\min_\pi J(\pi)=\int_{X\times Y} c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y\\ &\text{subject to:}\quad \int_X \pi(x,y)\mathrm{d}x=g(y), \quad \int_Y \pi(x,y)\mathrm{d}y=f(x)\\ &\qquad\qquad \pi(x,y)\geq 0\end{aligned} \tag{4.8}$$

This is the Kantorovich version of the optimal transport problem. We derive next the corresponding Monge problem. Suppose that the sand in the interval $(x,x+dx)$ are all sent to $(T(x),T(x)+dT(x))$ for some continuously differentiable function $T:X\to Y$. Then the mass preservation constraint is now

$$\int_{T(x)\in(y,y+\mathrm{d}y)} f(x)\mathrm{d}x=g(y)\mathrm{d}y$$

By change of variable formula (holds when T is a diffeomorphism, for the shape drawn on the right of Figure 4.3, such T clearly does not exist! We neglect this issue though), the left is

$$\int_{z\in(y,y+\mathrm{d}y)} f(T^{-1}(z))|\det DT^{-1}(z)|\mathrm{d}z=\frac{f(T^{-1}(y))}{|\det DT(T^{-1}(y))|}\mathrm{d}y$$

Hence

$$|\det DT(x)|=\frac{f(x)}{g(T(x))}, \quad \forall x\in X. \tag{4.9}$$

This equation is called the *Monge-Ampère equation.*

The total cost is

$$J(T)=\int_X c(x,T(x))f(x)\mathrm{d}x$$

To summarize, the Monge problem is the following optimization problem

$$\begin{aligned} &\min_T J(T) = \int_X c(x, T(x)) f(x) \mathrm{d}x \\ &\text{subject to:} \quad |\det DT(x)| = \frac{f(x)}{g(T(x))} \end{aligned} \tag{4.10}$$

Note that this problem is highly nonlinear and is extremely hard to solve. Indeed, the Monge-Ampère equation is a nonlinear PDE which is difficult to solve even numerically.

Exercise 4.1 (Transport maps between Gaussian distributions)**.** Consider two Gaussian distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ on $\mathbb{R}^n$, where μ_1, μ_2 are the mean vectors and Σ_1, Σ_2 the covariant matrices. Find an admissible map of the form $T(x) = \mu_2 + A(x - \mu_1)$ for some constant matrix A.

4.1.5 A quick review of measure and integration theory

Measure

To efficiently describe the optimal transport problem, it is inevitable to use some measure theory.

Given a set X, a *measure* on X is some extended real value *function* which measures the sizes of the subsets in X, e.g., the number of points in the set, the length of a curve, the area of a surface, the volume of a polyhedron, etc. Obviously, a measure, say μ, should have the following properties: 1) $\mu(\emptyset) = 0$; 2) for any finite collection of disjoint subsets $A_1, \cdots A_m$ in X, the finite additive property should hold $\sum_{i=1}^m \mu(A_i) = \mu\left(\bigcup_{i=1}^m A_i\right)$. It turns out that the finite additive property 2) is too weak to work with; just think of the case that we need to take limits when doing improper Riemann integration. Therefore, 2) is asked to be replaced by a stronger requirement, the so-called *countably additive* property: 2') for any countable[1] collection of disjoint sets $A_1, \cdots A_i, \cdots$, there hold $\sum_{i=1}^m \mu(A_i) = \mu\left(\bigcup_{i=1}^\infty A_i\right)$.

It seems that we are done with the definition of a measure, i.e., an extended real value function on 2^X (the set of all subsets of X) satisfying properties 1) and 2'). Unfortunately, such a function in general does not exist. The reason is that the set 2^X is too big which contains some "bad sets" that hinders us from defining a meaningful function having properties 1) and 2'). To cope with this, the strategy is to restrict the definition of a measure on a smaller class of sets. Let us check what kind of sets should be included in this class. First, the empty set $\emptyset$ should be in this class. Second, if $A_1, \cdots, A_i, \cdots$ are in this class, 2') is meaningful only if $\bigcup_{i=1}^\infty A_i$ is also in this class. In other words, the class should be closed under countable union operation. Apart from these, we also require that 3) the class is closed under complement; in words, if A is in the class, so is A^c. In particular, $X = \emptyset^c$ is in the class, and 2'), 3) together imply that countable intersection operation is also closed. The reason to include this is that not only we need to do addition (union of sets) in the class, but also we need be able to do subtraction. A class with properties 1), 2') and 3) is called a σ-algebra:

Definition 4.1 (σ-algebra and measurable space)**.** Given a set X, a *σ-algebra* $\mathscr{A}$ on X is a collection of subsets of X satisfying the following properties:

1) $\emptyset \in \mathscr{A}$;

2) If $A_1, \cdots A_i, \cdots \in \mathscr{A}$, then $\bigcup_{i=1}^\infty A_i \in \mathscr{A}$.

3) If $A \in \mathscr{A}$, then $A^c \in \mathscr{A}$.

We call $(X, \mathscr{A})$ a *measurable space*.

[1] A set A is said to be *countable* if there exists a bijective mapping betwee A and the set of natural numbers.

The first example of a σ-algebra is 2^X, i.e., the collection of all subsets of X. As we have mentioned before, this σ-algebra is often too big to work with. However, when X is a discrete set, either finite or countable, this σ-sigma algebra will be in effect for in this course.

It turns out on the one hand, a σ-algebra is big enough to contain the sets that we are interested in, and on the other, it is small enough for us to define a measure (i.e., such measure exists). However, the justification of the latter fact is not as obvious. Interested readers are referred to [8].

An important class of measurable spaces is the Borel measurable space.

Definition 4.2 (Borel measurable space)**.** Let X be a topological space. The Borel σ-algebra on X, denoted $\mathscr{B}(X)$, is the smallest σ-algebra containing all the open sets of X. A measure on $\mathscr{B}(X)$ is called a *Borel measure*.

Typical sets in Borel σ-algebra include: 1) all the open sets; 2) G_δ sets: countable intersection of open sets; 3) F_σ sets: countable union of closed sets, and so on.

Definition 4.3 (Measure)**.** Given a measurable space $(X,\mathscr{A})$, a *measure* $\mu:\mathscr{A}\to[0,\infty]$ is a function satisfying

1) $\mu(\emptyset)=0$;

2) for countable collection of disjoint measurable sets $\{A_i\}$, $\mu\left(\bigcup_{i=1}^{\infty}A_i\right)=\sum_{i=1}^{\infty}\mu(A_i)$.

We call $(X,\mathscr{A},\mu)$ a *measure space*.

A measure is sometimes written as $d\mu$.

Example 4.1 (Dirac measure)**.** Given measurable space $(X,\mathscr{A})$, we can define for every $x\in X$ a measure $\delta_x:\mathscr{A}\to\{0,1\}$ by

$$\delta_x(A)=1_A(x)=\begin{cases}1, & \text{if } x\in A\\ 0, & \text{else}\end{cases}.$$

Example 4.2 (Counting measure)**.** Given a set X and 2^X its σ-algebra, for $S\subseteq 2^X$, define the *counting measure* $\#S$ as the cardinality of the set S. If the cardinality of S is infinite, set $\#S=\infty$. It is plain to verify that this is indeed a measure.

Example 4.3 (Probability measure)**.** When the measure satisfies $\mu(X)=1$, then μ is called a *probability measure*. The set of probability measures on X is denoted as $\mathscr{P}(X)$.

Sets with zero measure play an important role in measure theory, we call such sets *null sets*. On a measure space X, we say that a property is satisfied for almost every $x\in X$ (abbreviated as a.e.) if the property holds for all $x\in X/N$ for some null set N, i.e., $\mu(N)=0$. If for every null set N, every subset of N is measurable, we say that the measure μ is *measure complete*. Given a measure space $(X,\mathscr{A},\mu)$, one can extend the σ-algebra $\mathscr{A}$ to make X a measure complete space $(\bar{X},\bar{\mathscr{A}},\mu)$. We call $\bar{X}$ the *measure completion* of X.

Example 4.4 (Lebesgue measure)**.** Let $X=\mathbb{R}^n$ be equipped with the normal topology. The completion of the Borel measure[2] $\mathscr{L}^n$ on $\mathscr{B}(\mathbb{R}^n)$ with the property that $\mathscr{L}^n(S)=\mathrm{vol}(S)$ for every cubic set $S\subseteq\mathbb{R}^n$ is called the Lebesgue measure. Notice that cubic sets generate the topology of $\mathbb{R}^n$, the Lebesgue measure is uniquely defined.

[2] A Borel measure needn't be complete.

Integration

A function $f:(X,\mathscr{A})\to(Y,\mathscr{B})$ is said to be *measurable* if $S\in\mathscr{B}\Rightarrow f^{-1}(S)\in\mathscr{A}$. We define integration of Borel measurable functions from $(X,\mathscr{B}(X),\mu)$ to $(\mathbb{R},\mathscr{B}(\mathbb{R}),\mathscr{L}^1)$ where $\mathscr{L}^1$ is the Lebesgue measure on the real line. The following is a characterization of such functions.

Proposition 4.1. *A function $f:(X,\mathscr{B}(X),\mu)\to(\mathbb{R},\mathscr{B}(\mathbb{R}),m)$ is measurable if and only if any of the following is measurable for all $a\in\mathbb{R}$: 1) $\{x\in X:f(x)\le a\}$; 2) $\{x\in X:f(x)<a\}$; 3) $\{x\in X:f(x)\ge a\}$; 4) $\{x\in X:f(x)>a\}$.*

Unless otherwise specified, the term "Borel function" means "real-valued Borel measurable function" henceforth.

Proposition 4.2. *If $\{f_j\}$ is a sequence of Borel functions, then the following functions are also measurable:*

$$\begin{aligned} g_1(x)&=\sup_j f_j(x), & g_2(x)&=\inf_j f_j(x)\\ g_3(x)&=\limsup_{j\to\infty} f_j(x), & g_4(x)&=\liminf_{j\to\infty} f_j(x)\end{aligned}.$$

To define the integration, one starts with non-negative simple functions of the form

$$\phi(x)=\sum_{i=1}^m a_i 1_{A_i}(x)$$

where $\{A_i\}$ are some measurable sets and a_i some non-negative coefficients. Define the integration of ϕ as

$$\int \phi \mathrm{d}\mu=\sum_{i=1}^m a_i\mu(A_i).$$

This number is set to zero if $a_i>0$ and $\mu(A_i)=\infty$ for some i.

Then one argue that any Borel function $f\ge 0$ is a limit of an increasing sequence of non-negative simple functions:

$$f=\lim_{i\to\infty} f_i$$

and therefore the integration of f can be defined as the limit

$$\int f\mathrm{d}\mu=\lim_{i\to\infty}\int f_i\mathrm{d}\mu.$$

The following are some equivalent notations for integration $\int f\mathrm{d}\mu$ when there is no danger of ambiguities:

$$\int f,\quad \int_X f\mathrm{d}\mu,\quad \int_X f(x)\mathrm{d}\mu(x),\quad \int_X f(x)\mu(\mathrm{d}x).$$

Example 4.5. Let $\mathbb{N}=\{1,\cdots,n,\cdots\}$ be the set of natural numbers equipped with the counting measure. Let $\{a_i\}_{i=1}^\infty$ be a non-negative sequence, which can be viewed as a measurable function $i\mapsto a_i$. It is readily checked that the integration of this function is simply

$$\int a_\cdot=\sum_{i=1}^\infty a_i$$

Example 4.6 (Absolutely continuous measures)**.** Given a (base) measure μ on $(X,\mathscr{B}(X))$, and a measurable function $f:X\to\mathbb{R}$, the following formula

$$\nu(A)=\int_A f(x)\mathrm{d}\mu(x):=\int_X f(x)1_A(x)\mathrm{d}\mu(x),\quad \forall A\in\mathscr{B}(X)$$

defines a new measure on $(X, \mathscr{B}(X))$. We shall denote this measure as $f(x)\mathrm{d}\mu(x)$ and say that it is *absolutely continuous* w.r.t. the measure μ. For measurable function $g : X \to \mathbb{R}$, it can be easily verified that the integration of g is

$$\int_X g(x)\mathrm{d}\nu(x) = \int_X g(x)f(x)\mathrm{d}\mu(x),$$

i.e., one simply replaces $\mathrm{d}\nu(x)$ by $f(x)\mathrm{d}\mu(x)$, this justifies our notation. The proof strategy is to first to prove for simple functions and then use simple functions to approximate general measurable functions.

Example 4.7. Let $X = [0,1]$ be equipped with the Lebesgue measure $\mathscr{L}^1|_{[0,1]}$. We calculate the integration of the function $f(x) = \sqrt{x}$ on X. For $n \geq 1$, define sets

$$E_k = f^{-1}([\frac{k}{2^n}, \frac{k+1}{2^n})), \quad 0 \leq k \leq 2^n - 1$$

and functions

$$f_n(x) = \sum_{k=0}^{2^n-1} \frac{k}{2^n} \cdot 1_{E_k}(x)$$

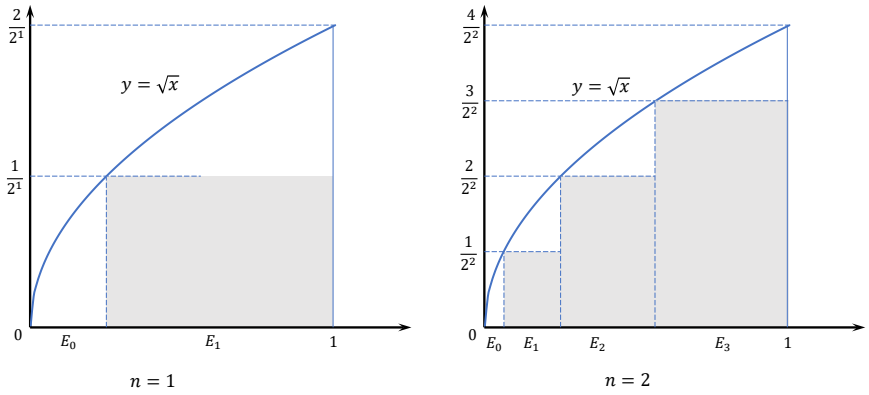

Figure 4.4: Construction of f_n. On the left, $n = 1$; on the right, $n = 2$.

On can verify that $f_n \uparrow f$ as $n \to \infty$ for all $x \in [0,1)$. Next, it is readily calculated that $\mathscr{L}^1(E_k) = \frac{2k+1}{2^{2n}}$ and

$$\int f_n = \sum_{k=1}^{2^n-1} \frac{(2k+1)k}{2^{3n}} = \frac{2}{2^{3n}} \cdot \frac{(2^n-1)2^n(2^{n+1}-1)}{6} + \frac{1}{2^{3n}} \cdot \frac{(2^n-1)2^n}{2} = \frac{2}{3} + O(\frac{1}{2^n})$$

Thus by definition

$$\int f = \lim_{n\to\infty} \int f_n = \frac{2}{3}$$

which coincides with the Riemann integral.

To define integration of a function f with possibly negative values, first decompose f as $f = f^+ - f^-$, where $f^+ \geq 0$, $f^- \geq 0$ (this decomposition may not be unique). If one of $\int f^+$ and $\int f^-$ is finite (otherwise we will run into the pathological case $\infty - \infty$), we define

$$\int f\mathrm{d}\mu = \int f^+\mathrm{d}\mu - \int f^-\mathrm{d}\mu.$$

We say that f is *integrable* as long as both $\int f^+$ and $\int f^-$ are finite, or equivalently $\int |f| \mathrm{d}\mu < \infty$. Denote the set of integrable functions on X as $L^1(X,\mu)$ or simply $L^1(X)$ or even L^1.

The integration defined above is called Lebesgue integration, which (with Lebesgue measure $\mathscr{L}^n$) coincides with the Riemannian integration when restricted piece-wise continuous functions on compact sets in $\mathbb{R}^n$. On the other hand, it is defined for much larger class of functions. Indeed, it is not an easy task to construct a function which is not measurable; almost all functions in real life are measurable and can be integrated. What's more, technically, the Lebesgue integration is much more flexible and more convenient to use. In particular, while it is often a subtle issue to exchange limit and integration in Riemannian integration (one often needs certain uniform convergence), the requirement to exchange limit and integration is much less strict. One of the most useful criteria is the following:

Theorem 4.1 (Dominated convergence theorem)**.** *Let (X,μ) be a measure space and $\{f_n\}$ a sequence of integrable functions such that $f_n(x) \to f(x)$ pointwisely as $n \to \infty$ for a.e. $x \in X$. If there exists a non-negative $g \in L^1(\mu)$ such that $|f_n| \le g$ a.e. for all n. Then $f \in L^1(\mu)$ and*

$$\int f \mathrm{d}\mu = \lim_{n\to\infty} \int f_n(x) \mathrm{d}\mu(x)$$

Another two useful results for non-negative functions are the monotone convergence theorem and Fatou lemma:

Theorem 4.2 (Monotone convergence theorem)**.** *If $\{f_n\}$ is a non-negative sequence such that $f_j \le f_{j+1}$ for all j and $f = \lim_{n\to\infty} f_n$, then*

$$\int f \mathrm{d}\mu = \lim_{n\to\infty} \int f_n \mathrm{d}\mu.$$

Lemma 4.1 (Fatou's lemma)**.** *If $\{f_n\}$ is a non-negative measurable sequence, then*

$$\int \liminf_{n\to\infty} f_n(x) \mathrm{d}\mu(x) \le \liminf_{n\to\infty} \int f_n \mathrm{d}\mu$$

4.1.6 General formulation of optimal transport

We are now ready to introduce the general formulation of the optimal transport problem. There are two approaches we may adopt. Either by abstracting the reasoning of the continuous version of optimal transport in Section **??** or introducing directly abstract optimal transport problem using measure theoretical terms. We here adopt the second approach to help the readers familiarize a bit the measure theory (the first approach is rather easy and the reader should also do it).

Pushforward of measures

Throughout this course, X, Y will be denoted as the source and target spaces of the optimal transportation respectively. They are assumed to be a complete metric spaces[3]. We equip X and Y with non-negative complete Borel measures, say μ and ν respectively.

Given a measurable function $f : (X, \mathscr{B}(X), \mu) \to (Z, \mathscr{B}(Z))$ (Z hasn't been assigned a measure yet), we can define a measure on Z by

$$f_\# \mu(B) := \mu(f^{-1}(B)), \quad \forall B \in \mathscr{B}(Z).$$

[3]A metric space X is said to be complete if for any sequence $\{x_n\}$, $d(x_n, x_m) \to 0$ as $n, m \to \infty$ (such sequence is called a Cauchy sequence) implies the existence of a point $x \in X$ such that $d(x_n, x) \to 0$ as $n \to \infty$.

One can check that $f_\#\mu$ is a well-defined measure, called the pushforward measure of μ by f. If $g : Z \to W$ is another measurable function, then it is easy to see that

$$(f \circ g)_\#\mu = f_\#(g_\#\mu).$$

The following formula will be used frequently:

Proposition 4.3 (Change of measure formula)**.** *For any measurable function $f : X \to Y$ and any measurable function $\phi : Y \to [0,\infty]$, one has*

$$\int_Y \phi \mathrm{d} f_\#\mu = \int_X (\phi \circ f)\mathrm{d}\mu.$$

The proof strategy is first to check the formula for simple functions, and then approximate Borel functions by simple functions.

Proof. If $\phi = 1_A$ for some measurable $A \subseteq Y$, then on the left

$$\int_Y 1_A(y)\mathrm{d} f_\#\mu(y) = f_\#\mu(A) = \mu(f^{-1}(A))$$

and on the right,

$$\int_X 1_A(f(x))\mathrm{d}\mu(x) = \int_X 1_{f^{-1}(A)}(x)\mathrm{d}\mu(x) = \mu(f^{-1}(A))$$

Next, it is obvious to see that this also holds for all non-negative simple functions. Now for non-negative Borel function ϕ, choose a sequence of simple functions such that $\phi_n \uparrow \phi$, then by monotone convergence theorem, the formula also holds. The proof is finalized by decomposing general ϕ into positive and negative parts. □

The abstract Monge problem

Monge problem

Let X, Y be two complete metric space, $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$ two probability measures and $c(x,y) : X \times Y \to [0,\infty]$ a Borel cost function, representing the cost of shipping a unit mass from x to y. The Monge problem is

$$\inf_T \left\{ \int_X c(x, T(x))\mathrm{d}\mu(x) : T : X \to Y \text{ Borel}, T_\#\mu = \nu \right\}. \tag{M}$$

A map T satisfying the constraint $T_\#\mu = \nu$ is called a *transport map.*

We henceforth denote this problem as (M). The critical part in (M) is the constraint $T_\#\mu = \nu$. By definition, this is equivalent to saying $\nu(A) = \mu(T^{-1}(A))$ for all measurable $A \subseteq Y$. Thinking $\nu(A)$ as the mass of set A in the target set Y, then the constraint $\nu(A) = \mu(T^{-1}(A))$ says that if we trace back the sources of the elements in A (i.e., the preimage of A under T), then they have the same mass as in the target. This coincides with the underlying assumption of optimal transport. To see that, we revisit the optimal transport problems that we defined previously.

Discrete case: If μ and ν are discrete probability measures on $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ respectively:

$$\mu = \sum_{i=1}^n a_i \delta_{x_i}, \quad \nu = \sum_{j=1}^m b_j \delta_{y_j}$$

in which $a_i, b_j \geq 0$ and $\sum a_i = \sum b_j = 1$ (the requirement that they sum to 1 is not essential; it can be replaced by any other fixed numbers). Then the cost can be calculated:

$$J(T) = \int_X c(x, T(x)) \mathrm{d}\mu(x) = \sum_{i=1}^n c(x_i, T(x_i)) a_i \tag{4.11}$$

For a mapping $T : X \to Y$, the pushforward of T is

$$T_\# \mu(A) = \mu(T^{-1}(A)) = \sum_{i=1}^n a_i \delta_{x_i}(T^{-1}(A)) = \sum_{i=1}^n a_i \delta_{T(x_i)}(A)$$

hence the requirement $T_\# \mu = \nu$ forces the following to hold

$$\sum_{i=1}^n a_i \delta_{T(x_i)} = \sum_{j=1}^m b_j \delta_{y_j}$$

which happens if and only if

$$\sum_{i:\, T(x_i) = y_j} a_i = b_j. \tag{4.12}$$

The equations (4.11, 4.12) are exactly those in (4.7).

Continuous case: This time, let us consider the absolutely continuous measures $\mathrm{d}\mu(x) = f(x)\mathrm{d}x$, $\mathrm{d}\nu(y) = g(y)\mathrm{d}y$ form some Lebesgue measurable functions f and g, where $\mathrm{d}x$ is the Lebesgue measure on $\mathbb{R}^n$. Then

$$J(T) = \int_X c(x, T(x)) \mathrm{d}\mu(x) = \int_X c(x, T(x)) f(x) \mathrm{d}x. \tag{4.13}$$

When T is a diffeomorphism, the constraint $T_\# \mu = \nu$ imposes the following

$$\begin{aligned} T_\# \mu(A) = \mu(T^{-1}(A)) &= \int_{T^{-1}(A)} f(x) \mathrm{d}x \\ = \nu(A) = \int_A g(y) \mathrm{d}y &= \int_{T^{-1}(A)} g(T(x)) |\det DT(x)| \mathrm{d}x \end{aligned}$$

hence

$$|\det DT(x)| = \frac{f(x)}{g(T(x))}, \quad \forall x \in X. \tag{4.14}$$

Equations (4.13, 4.14) are exactly (4.10) introduced in Section **??**.

One important question in Monge problem is when the set $\{T : T_\# \mu = \nu\}$ is non-empty. A further question is, how to construct a transport map from the given data μ and ν. It is easy to construct measures μ and ν such that T does not exist:

Example 4.8 (Nonexistence of transport map)**.** Let $X = \{0\}$, $Y = \{0, 1\}$ and $\mu = \delta_0$ and $\nu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Obviously there is no map $T : X \to Y$ satisfying $T_\# \mu = \nu$, because the point 0 can only be mapped to either 0 or 1, but not both. More generally, if the cardinality of the support of ν is larger than that of μ, transport map does not exist. In particular, when μ is a discrete measure, i.e., its support is a discrete set, and ν a continuous measure, there is not admissible transport map.

Example 4.9 (Transport map on the real line)**.** Consider two probability measures on the real line, μ, $\nu \in \mathscr{P}(\mathbb{R})$. The compatibility condition $T_\# \mu = \nu$ says $\mu(T^{-1}(A)) = \nu(A)$ for all Borel sets A. Since intervals $(-\infty, a]$ generates the Borel measure, it is sufficient to require

$$\mu\{T^{-1}(-\infty, a]\} = \nu\{(-\infty, a]\}, \quad \forall a \in \mathbb{R}.$$

The right hand side is simply the *cumulative distribution function*, which we denote as $F_\nu(a) := \nu\{(-\infty, a]\}$. On the left, if T is a strictly increasing map, then $T^{-1}(-\infty, a] = (-\infty, T^{-1}(a)]$. And it follows that the left hand side is $F_\mu(T^{-1}(a))$. Equating the two terms, we get

$$F_\mu(T^{-1}(a)) = F_\nu(a)$$

thus T can be taken as $T(x) = F_\nu^{-1} \circ F_\mu(x)$. In general, this map needn't be strictly increasing and F_ν needn't be invertible, e.g., when ν is supported only on finite intervals. A better definition for T is

$$T(x) := \inf\{y \in \mathrm{supp}\nu : F_\nu(y) \geq F_\mu(x)\}. \tag{4.15}$$

Indeed, if $\mathrm{supp}F_\nu = \mathbb{R}$, then F_ν is invertible and $T(x) = F_\nu^{-1} \circ F_\mu(x)$. But definition (4.15) makes sense even if F_ν is supported only on subset of $\mathbb{R}$. It remains to verify that T is an admissible transport map, i.e., $F_\nu(y) = \mu(\{x : T(x) \leq y\})$ for all $y \in \mathbb{R}$. Obviously, T is nondecreasing, and hence there exists $a \in \mathbb{R}$, such that $\{x : T(x) \leq y\}$ contains $(-\infty, a)$ and is contained in $(-\infty, a]$. Since μ is atomless, in either case they have the same measure. Thus it suffices to prove $F_\nu(y) = F_\mu(a)$. On the one hand, since $(-\infty, a) \subseteq \{x : T(x) \leq y\}$, then for any $a' < a$ and $\epsilon > 0$, we have by definition of T, $F_\mu(a') \leq F_\nu(T(a') + \epsilon) \leq F_\nu(y + \epsilon)$. Letting $\epsilon \to 0$ and $a' \to a$, by continuity of F_μ (since μ is atomless) and right continuity of F_ν, we get $F_\mu(a) \leq F_\nu(y)$. On the other hand, for any $a' > a$, we have $T(a') > y$ because $\{x : T(x) \leq y\} \subseteq (-\infty, a]$, which is equivalent to $\{x : T(x) > y\} \supseteq (a, \infty)$. Hence $F_\nu(y) \leq F_\mu(a')$, and $F_\nu(y) \leq F_\mu(a)$ by continuity of F_μ, see Figure 4.5.

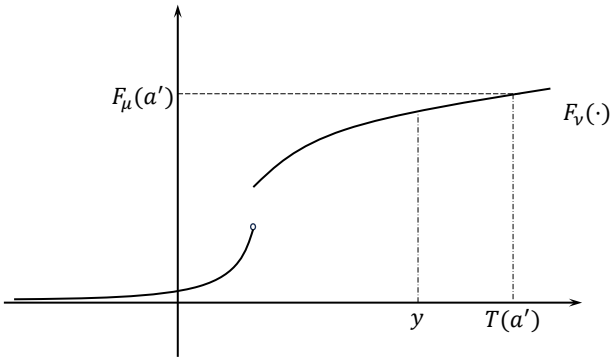

Figure 4.5: Illustration of the proof.

The following result says this also holds in higher dimension.

Proposition 4.4. *Given two probability measures μ and ν, if μ is atomless, then there always exists a Borel map T such that $T_\#\mu = \nu$.*

The abstract Kantorovich problem

Let us recall the Kantorovich problem in continuous case (see Section **??**, problem (4.8)):

$$\begin{aligned} \min_p J(\pi) &= \int_{X\times Y} c(x,y)p(x,y)\mathrm{d}x\mathrm{d}y \\ \text{subject to:} \quad & \int_X p(x,y)\mathrm{d}x = g(y), \quad \int_Y p(x,y)\mathrm{d}y = f(x) \\ & p(x,y) \geq 0 \end{aligned}$$

Here we changed a bit the notation (replace π with p) to avoid confusion. Suppose that $f(x)\mathrm{d}x$ and $g(y)\mathrm{d}y$ are two probability measures. The equations on the second line above motivate us to define a measure $\pi \in \mathscr{P}(X \times Y)$ by

$$\pi(S) = \int_S p(x,y)\mathrm{d}x\mathrm{d}y$$

for $S \subseteq X \times Y$. π is easily seen to be a probability measure. We claim that with this notation, the constraint on the second line can be recast as

$$\pi(A \times Y) = \mu(A), \quad \pi(X \times B) = \nu(B) \tag{4.16}$$

for all measurable sets $A \subseteq X$, $B \subseteq Y$. To see this, first notice

$$\pi(A \times Y) = \int_{A\times Y} p(x,y)\mathrm{d}x\mathrm{d}y = \int_A \int_Y p(x,y)\mathrm{d}y\mathrm{d}x$$
$$\mu(A) = \int_A f(x)\mathrm{d}x$$

then equate the rightmost terms of the above two lines to get

$$\int_X 1_A(x)\left(\int_Y p(x,y)\mathrm{d}y - f(x)\right)\mathrm{d}x = 0.$$

One then argue that

$$\int_X \phi(x)\left(\int_Y p(x,y)\mathrm{d}y - f(x)\right)\mathrm{d}x = 0$$

for all Borel functions ϕ and conclude that $\int_Y p(x,y)\mathrm{d}y = f(x)$, as expected. The reverse direction is straightforward.

We remark that the equations (4.16) can be written more concisely as

$$(p_X)_\# \pi = \mu, \quad (p_Y)_\# \pi = \nu$$

where $p_X : X \times Y \to X$ and $p_Y : X \times Y \to Y$ are the projection maps. Call the following set

$$\Gamma(\mu,\nu) = \{\pi \in \mathscr{P}(X \times Y) : (p_X)_\# \pi = \mu, \quad (p_Y)_\# \pi = \nu\} \tag{4.17}$$

the set of *transport plans* between μ and ν. We claim that $\Gamma(\mu,\nu)$ is never empty. Define a measure $\mu \otimes \nu$ as follows

$$\mu \otimes \nu(A \times B) = \mu(A)\nu(B)$$

which is uniquely determined since $A \times B$ generates the Borel σ-algebra of $X \times Y$. By this definition, it is immediate that $\mu \otimes \nu$ is a transport plan.

Example 4.10. Let $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_i \delta_{y_j}$ two discrete probability measures on $X = \{x_1, \cdots, x_n\}$ and $Y = \{y_1, \cdots, y_m\}$ respectively. Let $\pi \in \Gamma(\mu,\nu)$. Then π is a probability measure on $X \times Y$, or

$$\pi = \sum_{i=1}^n \sum_{j=1}^m P_{ij} \delta_{(x_i,y_j)}.$$

for some non-negative numbers $\{P_{ij}\}$ satisfying $\sum_{i,j} P_{ij} = 1$. For any $x_i \in X$, by definition of a transport plan, we have

$$a_i = \mu(\{x_i\}) = \pi(\{x_i\} \times Y) = \sum_{j=1}^m P_{ij}$$
$$b_j = \nu(\{y_j\}) = \pi(X \times \{y_j\}) = \sum_{i=1}^n P_{ij}$$

which coincides with 4.5. Thus $\Gamma(\mu,\nu) = \Pi(a,b)$, as expected.

The Kantorovich problem is formulated as follows:

Kantorovich problem

Given two probability measures $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$, the Kantorovich problem is to seek a probability measure $\pi \in \mathscr{P}(X \times Y)$ to the following minimization problem:

$$\inf_{\pi} \left\{ \int_{X \times Y} c(x,y) \mathrm{d}\pi(x,y) : \pi \in \Gamma(\mu,\nu) \right\}. \tag{K}$$

We henceforth denote this problem as (K). The Kantorovich has the following important properties:

1) The set $\Gamma(\mu,\nu)$ is never empty since $\mu \otimes \nu$ is always in $\Gamma(\mu,\nu)$. Thus the problem is always well-defined.

2) It is a convex optimization problem over a convex set. Indeed, for $\pi_1, \pi_2 \in \Gamma(\mu,\nu)$ and $\lambda \in [0,1]$, it is easily seen that $\pi = \lambda\pi_1 + (1-\lambda)\pi_2)$ is a transport plan, e.g.,$\pi(A \times Y) = \lambda\pi_1(A \times Y) + (1-\lambda)\pi_2(A \times Y) = \mu(A)$. On the other hand, the mapping $\pi \to \int c \mathrm{d}\pi$ is affine since $\int c \mathrm{d}(\lambda\pi_1 + (1-\lambda)\pi_1) = \lambda \int c \mathrm{d}\pi_1 + (1-\lambda) \int c \mathrm{d}\pi_2$. (Note that we cannot talk about linearality since π is restricted to be probability measures).

Problem (M) versus problem (K)

We mentioned earlier that the Monge problem can be viewed as adding an additional constraint on the Kantorovich problem, or in other words, Kantorovich problem is a relaxation of the Monge problem. This still holds for abstract optimal transportation problems. In fact, we can associate every transport map $T : X \to Y$ with a transport plan π by [4]

$$\pi = (\mathrm{id} \times T)_{\#}\mu$$

where id is the identity mapping on X. To see that π is a transport plan, notice that $\pi(A \times Y) = \mu((\mathrm{id} \times T)^{-1}(A \times Y)) = \mu\{x \in A; T(x) \in Y\} = \mu(A)$. The other equation is similar. Thus we can conclude

$$\inf_{\pi} \text{(K)} \leq \inf_{T} \text{(M)}. \tag{4.18}$$

We point out that even though μ has a density, e.g., $d\mu(x) = f(x)dx$, the map $\pi = (\mathrm{id} \times T)_{\#}\mu$ needn't do. In fact, π is concentrated on the *graph* of T:

$$\mathrm{Gr}(T) := \{(x, T(x)) \in X \times Y\},$$

see Figure 4.6.

[4] Given two mappings $T_1 : X \to Y_1$, $T_2 : X \to Y_2$, the mapping $T_1 \times T_2 : X \to Y_1 \times Y_2$ is defined as $T_1 \times T_2(x) = (T_1(x), T_2(x))$.

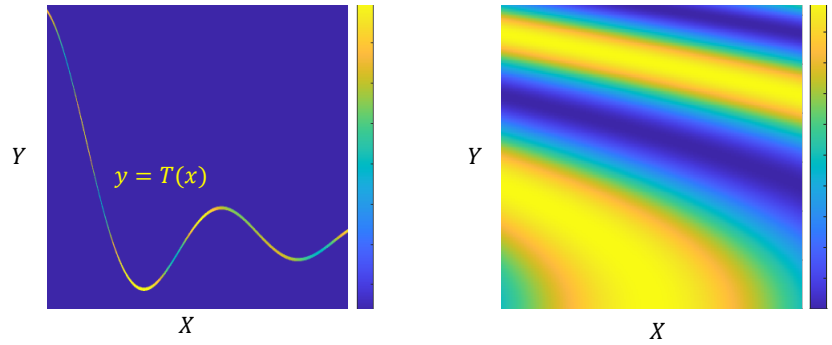

Figure 4.6: Monge problem versus Kantorovich problem. On the left, the mass is concentrated on the graph of T. On the right, we draw the continuous density function of a probability measure π. The brighter the value, the higher the density value at that point.

It now seems that the (K) problem is so much more general than (M) that the one would normally not expect the reverse direction to hold. But quite surprisingly, the reverse direction holds under very mild conditions: if 1) μ does not assign positive measure to singletons, i.e., $\mu\{x\} = 0$ for any $x \in X$ (we call such μ *atomless* or μ has no atom) and 2) the cost function $c(x, y) : X \times Y \to \mathbb{R}$ is continuous, then the inequality (4.18) becomes equality:

Theorem 4.3 (Pratelli)**.** *If μ is atomless and $c : X \times Y \to \mathbb{R}$ is continuous, then*

$$\min_{\pi} \ (K) = \inf_{T} \ (M).$$

The proof of the theorem is quite technical which relies essentially on Proposition 4.4: based on that proposition, for any $\pi \in \Gamma(\mu, \nu)$ – when μ is atomless – we can find a sequence of transport maps $\{T_n\}$ such that $(\mathrm{id} \times T_n)_{\#}\mu$ converges to π in certain sense. The continuity of c then will allow us to take the limit

$$\lim_{n\to\infty} \int_X c(x, T_n(x)) \mathrm{d}\mu(x) = \lim_{n\to\infty} \int_{X\times Y} c(x, y) \mathrm{d}(\mathrm{id} \times T_n)_{\#}\mu(x, y) = \int_{X\times Y} c(x, y) \mathrm{d}\pi(x, y).$$

Observe from above that given a sequence of measures $\{\mu_n\}$, the convergence we need is the following: there exists a measure μ such that for any continuous function ϕ, $\int \phi \mathrm{d}\mu_n$ converges to $\int \phi \mathrm{d}\mu$. In fact, a weaker requirement is sufficient, i.e., the weak convergence of measures:

Definition 4.4 (Weak convergence)**.** A sequence of measures $\{\mu_n\}$ on X is said to *converge weakly to μ* and is denoted $\mu_n \rightharpoonup \mu$, if

$$\int_X \phi \mathrm{d}\mu_n \to \int_X \phi \mathrm{d}\mu$$

for all bounded continuous functions $\phi \in C_b(X)$.

4.2 Structures of the minimizer

4.2.1 Existence of optimal transport plan

For both Kantorovich and Monge problem, the first question needs to be addressed is the existence of minimizers. We study the Kantorovich problem first.

Recall that in the discrete measure setting, the Kantorovich problem is a linear programming on a convex compact set $\Pi(a,b)$. Thus a minimizer is guaranteed to exist. In the general setting, we will see that the set $\Gamma(\mu,\nu)$ is still compact – under the weak topology introduced earlier. Recall that:

1) a set X is compact if for any sequence $\{x_n\}$ in X, there is a convergent subsequence $\{x_{n_k}\}$ whose limit lies in X.

2) the weak topology on $\Gamma(\mu,\nu)$ is defined by: $\mu_n \rightharpoonup \mu$ iff $\int \phi \mathrm{d}\mu_n \to \int \phi \mathrm{d}\mu$ for all bounded continuous ϕ.

Now that $\Gamma(\mu,\nu)$ is compact, if the functional $\pi \mapsto \int c\mathrm{d}\pi$ is continuous, we can conclude that there exists at least one minimizer in $\Gamma(\mu,\nu)$. However, sometimes continuity is a strong requirement, and a weaker condition is enough, i.e., *lower semi-continuity.*

Definition 4.5 (Lower semi-continuity)**.** On a metric space X, a function $f: X \to \mathbb{R}\cup\{+\infty\}$ is said to be *lower semi-continuous* (l.s.c. for short) if for every sequence $x_n \to x$, we have

$$f(x) \le \liminf_{n\to\infty} f(x_n).$$

By definition, a continuous function is l.s.c. Increasing left continuous functions and decreasing right continuous functions on the real line are also l.s.c.

Theorem 4.4 (Weierstrass extreme point theorem)**.** *If f is l.s.c. on a compact metric space X, then f achieves minimum on X, i.e., there exists $x_* \in X$ such that*

$$f(x_*) = \min_X f(x).$$

Next, we show that $\pi \mapsto \int c\mathrm{d}\pi$ is l.s.c. when c is on $X\times Y$:

Proposition 4.5 (L.s.c. of $\pi \to \int c\mathrm{d}\pi$)**.** *If $c: X\times Y \to [0,\infty]$ is l.s.c., the mapping $\pi \mapsto \int c\mathrm{d}\pi$ is also l.s.c. in $\mathscr{P}(X\times Y)$ w.r.t. the weak topology.*

Proof. We need to show that, for a sequence $\{\pi_i\} \in \mathscr{P}(X\times Y)$ converging weakly to π, there holds

$$\int c\mathrm{d}\pi \le \liminf_{i\to\infty} \int c\mathrm{d}\pi_i.$$

If c is continuous, then we can prove this rather easily. Define a sequence $c_k(x,y) = c(x,y)\wedge k \le c(x,y)$,[5] which is bounded continuous on $X\times Y$ and $c_k \uparrow c$ as $k\to\infty$ pointwisely. By definition of weak convergence,

$$\lim_{i\to\infty} \int c_k \mathrm{d}\pi_i = \int c_k \mathrm{d}\pi.$$

Apply monotone convergence theorem

$$\int c\mathrm{d}\pi = \lim_{k\to\infty}\int c_k\mathrm{d}\pi = \lim_{k\to\infty}\lim_{i\to\infty}\int c_k\mathrm{d}\pi_i = \lim_{k\to\infty}\liminf_{i\to\infty}\int c_k\mathrm{d}\pi_i \le \liminf_{i\to\infty}\int c\mathrm{d}\pi_i$$

as desired.

For l.s.c., the proof strategy is the same: approximate c by some continuous bounded functions and then tend to limit. The following clever construction produces a Lipschitz continuous function from a l.s.c. function:

$$c_k(x,y) := \inf_{x'\in X,\, y'\in Y} \{c(x',y') \wedge k + k d_X(x,x') + k d_Y(y,y')\} \tag{4.19}$$

[5] We denote $x\wedge y := \min\{x,y\}$, and $x\vee y = \max\{x,y\}$.

where d_X and d_Y are the metric on X and Y respectively. We assert that $c_k \uparrow c$. Indeed, $0 \le c_k \le c_{k+1} \le c \wedge k \le c$, it suffices to prove $c(x,y) \le \sup_k c_k(x,y)$ since this implies (by monotonicity of c_k)

$$\lim_k c_k \le c \le \sup_k c_k = \lim_k c_k.$$

Fix x, y, by definition of c_k, for any $k \ge 1$, there exists x_k, y_k such that

$$c(x_k, y_k) \wedge k + k d_X(x, x_k) + k d_Y(y, y_k) \le c_k(x,y) + \frac{1}{k}.$$

Let $k \to \infty$, we discover that $d_X(x, x_k) \to 0$, $d_Y(y, y_k) \to 0$. Thus by definition of l.s.c., (w.l.o.g., assume $\sup_k c(x_k, y_k)$ is finite):

$$c(x,y) \le \liminf_{k\to\infty} c(x_k, y_k) = \liminf_{k\to\infty} c(x_k, y_k) \wedge k \le \sup_k c_k(x,y).$$

The Lipschitz continuity of $c_k(x,y)$ is left as an exercise (see below). □

Exercise 4.2. If $\{f_\alpha\}_{\alpha\in A}$ is a family of Lipschitz continuous functions on X with a common Lipschitz constant – we call the family *equi-Lipschitz* – then $f(x) := \inf_\alpha f_\alpha(x)$ is also Lipschitz continuous. In particular, if f_α has the same Lipschitz constant, say L, then the Lipschitz constant of f is also L.

The compactness of the set $\Gamma(\mu,\nu)$ is much more technical. We state the following theorem without proof. Before that we need the notion of *separable spaces*. A metric space X is said to be separable if it has a countable dense set. For example, a Hilbert space admitting countable basis is separable. $L^p(X,\mu)$ is also separable for $p \in [1,\infty)$ if X is, e.g., $L^p(\mathbb{R}^n, \mathscr{L}^n)$. A separable complete metric space is called a *Polish space*.

Theorem 4.5 (Compactness of $\Gamma(\mu,\nu)$)**.** *Let X, Y be Polish spaces, and $\mu \in \mathscr{P}(X)$, $\nu \in \mathscr{P}(Y)$. Then $\Gamma(\mu,\nu)$ is compact w.r.t. the weak topology.*

Finally, we can conclude with the help of Proposition 4.5 and Theorem 4.5 the following result:

Theorem 4.6. *Let X, Y be Polish spaces and $c : X \times Y \to [0,\infty]$ l.s.c., then the Problem (K) has a minimizer.*

Although the existence of optimal transport plan for the Kantorovich problem is guaranteed in most reasonable cases and is rather easy to analyze, it is not the case for the Monge problem. We will only be able to prove the existence of optimal transport maps for much narrower class of problems and even in those cases, the proof is non-trivial and requires deeper understandings of the minimizer of the Kantorovich problem.

4.2.2 Duality theory I: $X \times Y$ compact

Remember that in Section 4.1.3, we derived the dual formula for discrete Kantorovich problem $\min \langle P, C\rangle$, which is $\sup_{f\oplus g\le C} \langle f, a\rangle + \langle g, b\rangle$, and that strong duality holds

$$\min_{P\in\Pi(a,b)} \langle P, C\rangle = \sup_{f\oplus g\le C} \langle f, a\rangle + \langle g, b\rangle. \tag{4.20}$$

With a bit insight, one may write the abstraction of the above formula in the general setting

$$\min_{\pi\in\Gamma(\mu,\nu)} \int c \mathrm{d}\pi = \sup_{\phi,\psi} \left\{ \int_X \phi(x)\mathrm{d}\mu(x) + \int_Y \psi(y)\mathrm{d}\nu(y) : \phi(x) + \psi(y) \le c(x,y) \right\} \tag{4.21}$$

This formula is called *Kantorovich-Rubinstein duality*. Since both sides of (4.21) are linear programming problems (infinite dimensional though), the duality has a high chance to be true. Recall that the non-trivial step in proving (4.20) involves an exchange of "inf" and "sup". This is also the case for (4.21). Indeed,

$$\min_{\pi\in\Gamma(\mu,\nu)}\int c\mathrm{d}\pi = \min_{\pi}\int c\mathrm{d}\pi + I_{\Gamma(\mu,\nu)}(\pi)$$

but

$$\begin{aligned} I_{\Gamma(\mu,\nu)}(\pi) &= \sup_{\phi}\int \phi(x)\mathrm{d}\left[\mu-(p_X)_{\#}\pi\right](x) + \sup_{\psi}\int \psi(y)\mathrm{d}\left[\nu-(p_Y)_{\#}\pi\right](y) \\ &= \sup_{\phi}\left[\int \phi(x)\mathrm{d}\mu(x) - \int \phi(x)\mathrm{d}\pi(x,y)\right] + \sup_{\psi}\left[\int \psi(y)\mathrm{d}\nu(y) - \int \psi(y)\mathrm{d}\pi(x,y)\right] \end{aligned}$$

Thus

$$\begin{aligned} \min_{\pi\in\Gamma(\mu,\nu)}\int c\mathrm{d}\pi &= \min_{\pi}\sup_{\phi,\psi}\int (c-\phi-\psi)\mathrm{d}\pi + \int \phi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \\ &\geq \sup_{\phi,\psi}\min_{\pi}\int (c-\phi-\psi)\mathrm{d}\pi + \int \phi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \,(\text{weak duality}) \\ &= \sup_{\phi(x)+\psi(y)\leq c(x,y)}\int \phi(x)\mathrm{d}\mu(x) + \int \psi(y)\mathrm{d}\nu(y). \end{aligned}$$

If equality is met for the inequality on the second line, i.e., when we can swap the min and sup, we will get the formula (4.21).

The device to prove the duality relation is the the following Fenchel-Rockafellar duality theorem:

Theorem 4.7 (Fenchel-Rockafellar duality)**.** *Let E be a normed vector space, E^* its topological dual (the space of bounded linear functionals on E), and Θ, Ξ two convex functions on E with values in $\mathbb{R}\cup\{+\infty\}$. Let Θ^* and Ξ^* be the Legendre-Fenchel transform of Θ, Ξ respectively. If $\exists z_0$, s.t.*

$$\Theta(z_0) < +\infty, \quad \Xi(z_0) < +\infty$$

and

$$\Theta(z) \text{ is continuous at } z_0$$

Then there holds

$$\inf_{z\in E}\{\Theta(z)+\Xi(z)\} = \max_{z^*\in E^*}\{-\Theta^*(-z^*)-\Xi^*(z^*)\}$$

To gain some insight on how to prove the Kantorovich-Rubinstein duality, we use Theorem 4.7 to justify the minimax property of the discrete duality relation (4.20). Rewrite the right hand side of (4.20):

$$\begin{aligned} \sup_{f\oplus g\leq C}\langle f,a\rangle + \langle g,b\rangle &= \sup_{f,g}\left(\langle f,a\rangle + \langle g,b\rangle + \begin{cases} 0, & \text{if } f\oplus g\leq C \\ -\infty, & \text{else} \end{cases}\right) \\ &= \sup_{P^*}\left(\begin{cases} \langle f,a\rangle + \langle g,b\rangle, & \text{if } P^* = f\oplus g \\ -\infty, & \text{else} \end{cases} + \begin{cases} 0, & \text{if } P^*\leq C \\ -\infty, & \text{else} \end{cases}\right) \end{aligned}$$

Denote

$$-\Theta^*(-P^*) = \begin{cases} \langle f,a\rangle + \langle g,b\rangle, & \text{if } P^* = f\oplus g \\ -\infty, & \text{else} \end{cases}, \qquad -\Xi^*(P^*) = \begin{cases} 0, & \text{if } P^*\leq C \\ -\infty, & \text{else} \end{cases}$$

then one can readily check that Θ^* and Ξ^* are the Legendre transform of Θ and Ξ defined as follows (calculate Θ^{**} and Ξ^{**} first and then argue $\Theta = \Theta^{**}$ and $\Xi^{**} = \Xi$):

$$\Theta(P) := \begin{cases} 0, & \text{if } \sum_{j=1}^m P_{ij} = a_i, \quad \sum_{i=1}^n P_{ij} = b_j \\ +\infty, & \text{else.} \end{cases}, \quad \Xi(P) = \begin{cases} \langle P, C\rangle, & \text{if } P \geq 0 \\ +\infty, & \text{else} \end{cases}$$

Therefore

$$\inf_P \{\Theta(P) + \Xi(P)\} = \inf_{P \in \Pi(a,b)} \langle P, C\rangle$$

By Fenchel duality, we now deduce (4.20).

Let $\mathscr{M}(X \times Y)$ be the space of Borel measures on $X \times Y$. Mimicking the above reasoning, we define two functionals $\Theta, \Xi : \mathscr{M}(X \times Y) \to \mathbb{R} \cup \{+\infty\}$:

$$\Theta(\pi) := \begin{cases} 0, & \text{if } \pi \in \Gamma(\mu, \nu) \\ +\infty, & \text{else.} \end{cases}, \quad \Xi(\pi) = \begin{cases} \int c \mathrm{d}\pi, & \text{if } \pi \geq 0 \\ +\infty, & \text{else} \end{cases}$$

To apply Fenchel duality theorem, one has to first determine the Legendre transforms of Θ and Ξ. But the topological dual of $\mathscr{M}(X \times Y)$ is not analytically convenient to work with. Thus we go from the other direction, i.e., from the right to the left of (4.21). In this case, we need to change the sup on the right to inf instead:

$$\begin{aligned} &\sup_{\phi,\psi} \left\{ \int_X \phi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) : \phi(x) + \psi(y) \leq c(x,y) \right\} \\ =& - \inf_{\phi,\psi} \left\{ \int_X \phi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) : \phi(x) + \psi(y) \geq -c(x,y) \right\} \end{aligned}$$

Now

$$\inf_{\phi,\psi} \left\{ \int_X \phi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y) : \phi(x) + \psi(y) \geq -c(x,y) \right\} = \inf_u \Xi(u) + \Theta(u)$$

where Θ, Ξ are defined as

$$\Theta(u) = \begin{cases} 0, & \text{if } u(x,y) \geq -c(x,y) \\ +\infty, & \text{else} \end{cases}, \quad \Xi(u) = \begin{cases} \int_X \phi(x) \mathrm{d}\mu(x) + \int_Y \psi(y) \mathrm{d}\nu(y), & \text{if } \phi(x) + \psi(y) = u(x,y) \\ +\infty, & \text{else} \end{cases}$$

where the living space for u is yet to be determined. It should be chosen in a way that its topological dual is rich enough and easy to work with. A good candidate is $C_b(X \times Y)$ when $X \times Y$ is compact, since it is well-known that the topological dual of $C_b(X \times Y)$ is $\mathscr{M}(X \times Y)$.

Let's assume $X \times Y$ is compact and c is lower semi-continuous. The Legendre transforms of $\Theta, \Xi : C_b(X \times Y) \to \mathbb{R} \cup \{+\infty\}$ are

$$\begin{aligned} \Theta^*(-\pi) &= \sup_{u \in C_b} -\langle \pi, u\rangle - \Theta(u) \\ &= \sup_{u \in C_b} \left\{ -\int u \mathrm{d}\pi : u \geq -c \right\} \\ &= \sup_{u \in C_b} \left\{ \int u \mathrm{d}\pi : u \leq c \right\} \\ &= \begin{cases} \int c \mathrm{d}\pi & \text{if } \pi \geq 0 \\ +\infty & \text{else.} \end{cases} \end{aligned}$$

and

$$\begin{aligned}\Xi^*(\pi) &= \sup_{u\in C_b}\left\{\langle\pi,u\rangle - \int_X \phi(x)\mathrm{d}\mu(x) - \int_Y \psi(y)\mathrm{d}\nu(y) : \phi+\psi = u\right\}\\ &= \sup_{u\in C_b}\left\{\int \phi(x)+\psi(y)\mathrm{d}\pi(x,y) - \int \phi\mathrm{d}\mu - \int \psi\mathrm{d}\nu : \phi+\psi = u\right\}\\ &= \begin{cases} 0 & \text{if } (p_X)_\#\pi = \mu,\ (p_Y)_\#\pi = \nu\\ +\infty & \text{else.}\end{cases}\end{aligned}$$

respectively. By Fenchel duality,

$$\max_\pi -\Theta^*(-\pi) - \Xi^*(\pi) = \max_{\pi\in\Gamma(\mu,\nu)} -\int c\mathrm{d}\pi = -\min_{\pi\in\Gamma(\mu,\nu)}\int c\mathrm{d}\pi$$

from which it follows that

$$\begin{aligned}\text{RHS of (4.20)} &= -\inf_u \Theta(u) + \Xi(u) = -\max_\pi -\Theta^*(-\pi) - \Xi^*(\pi) = \min_{\pi\in\Gamma(\mu,\nu)}\int c\mathrm{d}\pi\\ &= \text{LHS of (4.20)}.\end{aligned}$$

as desired. In conclusion, we have proved the duality relation (4.21) when $X\times Y$ is compact. The following general result shows that compactness is not essential though:

Proposition 4.6 (Duality for compact $X\times Y$)**.** *Let $X\times Y$ be compact and $c : X\times Y\to[0,\infty]$ l.sc., then the Kantorovich-Rubinstein duality* (4.21) *holds.*

In order to extend to non-compact case, we need some convex analysis tools, which are important also for further understanding the structures of optimal transport plans and maps. In particular, the notion *c-cyclical monotonicity* of the supports of optimal plans will be crucial to us.

4.2.3 c-cyclical monotonicity

Convex analysis recalled

Let X be a complete metric space, recall that for a convex functional $f : X\to(-\infty,\infty]$, the *subdifferential* of f at x is defined as

$$\partial f(x) := \{x^*\in X^* : \langle x^*, y-x\rangle \le f(y)-f(x),\quad \forall\, y\in X\} \tag{4.22}$$

where X^* is as usual the topological dual of X, and $\langle,\rangle$is the paring on $X^*\times X$. The following are some well-known properties of the subdifferential (the reader is invited to verify these properties) of a convex function:

1) $\partial f(x)$ is a convex closed (possibly empty) subset of X^*.

2) $x^*\in\partial f(x)$ if and only if $f(x)+f^*(x^*) = \langle x^*, x\rangle$, where f^* is the Legendre transform of f, i.e., $f^*(x^*) = \sup_{x\in X}\{\langle x^*,x\rangle - f(x)\}$.

3) When f is differentiable [6] at x, then $\partial f(x) = \{\nabla f(x)\}$.

4) ∂f is a monotone operator: for $x_1^*\in\partial f(x_1)$, $x_2^*\in\partial f(x_2)$,

$$\langle x_2^* - x_1^*, x_2 - x_1\rangle \ge 0.$$

[6] In this course, a function f is said to be differentiable at x if $r\mapsto f(x+ry)$ is differentiable at $r=0$ for all $y\in X$.

5) Another remarkable property regarding subdifferential is cyclical monotonicity of its graph. Since $x \mapsto \partial f(x)$ is a set valued map, its graph is well defined:

$$\mathrm{Gr}(\partial f) := \{(x, x^*) \in X \times X^* : x^* \in \partial f(x)\}.$$

For a set of points $(x_1, x_1^*), \cdots, (x_N, x_N^*)$ on the graph and a permutation σ on $\{1, \cdots, N\}$, we have

$$\left\langle x_i^*, x_{\sigma(i)} - x_i \right\rangle \leq f(x_{\sigma(i)}) - f(x_i)$$

adding up together, we get

$$\sum_{i=1}^{N} \left\langle x_i^*, x_{\sigma(i)} - x_i \right\rangle \leq 0. \tag{4.23}$$

A graph $\Gamma \subseteq X \times X^*$ satisfying (4.23) is said to be *cyclically monotone*. Notice that there also holds

$$\sum_{i=1}^{N} \left\langle x_{\sigma(i)}^* - x_i^*, x_i \right\rangle \leq 0.$$

c-cyclical monotonicity

Cyclical monotonicity is a special case of a more general notion, namely, c-cyclical monotonicity, which plays a fundamental role in optimal transport. Our final goal in this subsection is to show that the support of an optimal transport plan is c-cyclically monotone.

Definition 4.6 (c-cyclical monotonicity)**.** A set $\Gamma \subseteq X \times Y$ is *c-cyclically monotone* if

$$\sum_{i=1}^{N} c(x_i, y_i) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}) \tag{4.24}$$

for every $N \geq 1$, permutation σ of $\{1, \cdots, N\}$ and $(x_i, y_i) \in \Gamma$ for $i = 1, \cdots, N$.

Example 4.11. Consider the set

$$I_c = \{(\phi, \psi) \in C_b(X) \times C_b(Y) : \phi(x) + \psi(y) \leq c(x, y),\ \forall x, y \in X \times Y\}.$$

For $\phi, \psi \in I_c$, call

$$\Gamma(\phi, \psi) := \{(x, y) \in X \times Y : \phi(x) + \psi(y) = c(x, y)\}$$

the *contact set* of the pair (ϕ, ψ). Then $\Gamma(\phi, \psi)$ is c-cyclically monotone. Indeed, for $(x_1, y_1), \cdots, (x_N, y_N)$ and any permutation σ of $\{1, \cdots, N\}$, we have

$$\sum_{i=1}^{N} c(x_i, y_i) = \sum_{i=1}^{N} \phi(x_i) + \sum_{i=1}^{N} \psi(y_{\sigma(i)}) \leq \sum_{i=1}^{N} c(x_i, y_{\sigma(i)}).$$

To see that c-cyclical monotonicity is a generalization of cyclical monotonicity, let $Y = X^*$ and $c(x, x^*) = -\langle x^*, x \rangle$ in (4.24), we immediately recover (4.23).

Similarly, we can generalize the Legendre transform:

Definition 4.7 (c-transform)**.** Given $c(x, y) : X \times Y \to \mathbb{R} \cup \{+\infty\}$, $\phi : X \to [-\infty, \infty)$, $\psi : Y \to [-\infty, \infty)$, define

$$\phi^c(y) := \inf_{x \in X} \{c(x, y) - \phi(x)\}$$
$$\psi^c(x) := \inf_{y \in Y} \{c(x, y) - \psi(y)\}$$

and say that ϕ^c (resp. ψ^c) the *c-conjugate* of ϕ (resp. ψ).

Definition 4.8 (c-concavity)**.** A function $\phi : X \to [-\infty, \infty)$ (resp. $\psi : Y \to [-\infty, \infty)$) is said to be *$c$-concave* if it is the infimum of a family of c-affine functions $c(\cdot, y) + \alpha$ (resp. $c(x, \cdot) + \beta$), i.e.,

$$\phi(x) = \inf_{y \in \mathscr{A}} c(x, y) + \alpha_y$$

for some index set $\mathscr{A}$.

c-concave function has the following important properties whose proof is left as an exercise:

- ϕ is c-concave if and only if it is the c transform of a function ψ, i.e., $\phi = \psi^c$;
- $\phi^{cc} \geq \phi$, with equality if and only if ϕ is c-concave.

The following is the last definition we will need:

Definition 4.9 (c-superdifferential)**.** Given $c(x, y) : X \times Y \to \mathbb{R} \cup \{+\infty\}$, $\phi : X \to [-\infty, \infty)$, the *$c$-superdifferential* of ϕ is a set valued map

$$\partial^c \phi(x) := \{y \in Y : c(x, y) - c(x', y) \leq \phi(x) - \phi(x') \text{ for all } x' \in X\}$$

If we specify the definition to $c(x, x^*) = -\langle x^*, x \rangle$, then we see that $\partial^c(-f)(x) = \{x^* \in X^* : \langle x^*, y - x \rangle \leq f(y) - f(x)\} = \partial f(x)$ in which the inequality is in opposite direction compared to (4.22). That is the reason why we should not call $\partial^c \phi$ c-subdifferential.

Exercise 4.3. Show that $y \in \partial^c \phi(x)$ if and only if $\phi(x) + \phi^c(y) = c(x, y)$ and deduce from this relation that the graph of $\partial^c \phi$ is c-cyclically monotone. *Hint:* $\mathrm{Gr}(\partial^c \phi)$ is a contact set.

The following result says that a c-cyclically monotone set is always contained in contact sets, or more precisely, the graph of some c-superdifferential $\mathrm{Gr}(\partial^c \phi)$.

Theorem 4.8 (Rockafellar)**.** *Assume $c : X \times Y \to \mathbb{R}$ and that $\Gamma \subseteq X \times Y$ is c-cyclically monotone. Then there exists a c-concave function $\phi : X \to [-\infty, \infty)$, $\phi \not\equiv -\infty$, such that $\Gamma \subseteq \mathrm{Gr}(\partial^c \phi)$. If c is bounded Lipschitz, i.e., $c \in \mathrm{Lip}_b(X \times Y)$, then ϕ can be chosen such that $(\phi, \phi^c) \in \mathrm{Lip}_b(X) \times \mathrm{Lip}_b(Y)$.*

Proof. We construct ϕ with the desired properties. For any $(x_N, y_N) \in \Gamma$, ϕ should be such that

$$\phi(x) \leq c(x, y_N) - c(x_N, y_N) + \phi(x_N), \quad \forall x \in X.$$

We can continue for $(x_{N-1}, y_{N-1}), \cdots, (x_0, y_0) \in \Gamma$,

$$\begin{aligned} \phi(x) &\leq c(x, y_N) - c(x_N, y_N) + c(x_N, y_{N-1}) - c(x_{N-1}, y_{N-1}) + \phi(x_{N-1}) \\ &\leq c(x, y_N) - c(x_N, y_N) + c(x_N, y_{N-1}) - c(x_{N-1}, y_{N-1}) + \cdots + c(x_1, y_0) - c(x_0, y_0) + \phi(x_0). \end{aligned}$$

If $\phi(x_0) = 0$, the above formula suggests defining

$$\phi(x) := \inf\{c(x, y_N) - c(x_N, y_N) + c(x_N, y_{N-1}) - c(x_{N-1}, y_{N-1}) + \cdots + c(x_1, y_0) - c(x_0, y_0)\}$$

where the infimum is taken over all finite set of points on Γ. We need to verify that 1) ϕ is c-concave; 2) $\phi(x_0) = 0$. For 1), it is obvious. It remains to show 2). Take $N = 1$ and $(x_1, y_1) = (x_0, y_0)$, we get $\phi(x_0) \leq 0$. Thus we need only show that $\phi(x_0) \geq 0$, which is obvious due to c-cyclical monotonicity of Γ.

It remains to show that ϕ and ϕ^c are bounded Lipschitz whenever $c \in \mathrm{Lip}_b(X \times Y)$. The Lipschitz continuity is obvious since ϕ is an infimum of a family of Lipschitz functions, see Exercise 4.2. Note that

$$\phi^c(y) \le c(x_0, y) - \phi(x_0) \le \sup c < \infty$$

Since ϕ is c-concave,

$$\inf \phi(x) = \inf \phi^{cc}(x) = \inf_x \inf_y c(x, y) - \phi^c(y) \ge \inf c - \sup \phi^c > -\infty$$

Similarly, one can show that $\sup \phi(x) < +\infty$. Thus ϕ is bounded on X. One can prove for ϕ^c analogously. □

Now, reconsider the discrete Kantorovich problem. Suppose that $\pi = \sum_{i,j} P_{ij} \delta_{(x_i, y_j)}$ is an optimal plan. If the support of π is not c-cyclically monotone, we can find a set of points $(x_{i_1}, y_{j_1}), \cdots, (x_{i_N}, y_{j_N})$ with $P_{i_k j_k} > 0$ for $k = 1, \cdots, N$ and a permutation σ of $\{j_1, \cdots, j_N\}$ such that

$$\sum_{k=1}^{N} c(x_{i_k}, y_{j_k}) > \sum_{k=1}^{N} c(x_{i_k}, y_{\sigma(j_k)}). \tag{4.25}$$

We use these information to construct a better plan:

$$\tilde{\pi} = \pi - \epsilon \sum_{k=1}^{N} \delta_{(x_{i_k}, y_{j_k})} + \epsilon \sum_{k=1}^{N} \delta_{(x_{i_k}, y_{\sigma(j_k)})}$$

Since $P_{i_k j_k}$ is strictly positive, $\tilde{\pi}$ is non-negative for small $\epsilon > 0$. Next, notice that

$$(p_X)_\# \tilde{\pi} = (p_X)_\# \pi - \epsilon \sum_{k=1}^{N} \delta_{x_{i_k}} + \epsilon \sum_{k=1}^{N} \delta_{x_{i_k}} = (p_X)_\# \pi$$
$$(p_Y)_\# \tilde{\pi} = (p_Y)_\# \pi - \epsilon \sum_{k=1}^{N} \delta_{y_{j_k}} + \epsilon \sum_{k=1}^{N} \delta_{y_{\sigma(j_k)}} = (p_Y)_\# \pi$$

since σ is a permutation. Thus $\tilde{\pi}$ is a *bona fide* optimal plan. It remains to show that $\tilde{\pi}$ performs better than π. Indeed, invoking (4.25),

$$\int c \mathrm{d}\tilde{\pi} - \int c \mathrm{d}\pi = -\epsilon \left(\sum_{k=1}^{N} c(x_{i_k}, y_{j_k}) - \sum_{k=1}^{N} c(x_{i_k}, y_{\sigma(j_k)}) \right) < 0$$

as desired. Thus we have shown that the support of the optimal plan of discrete Kantorovich problem is c-cyclically monotone. The same reasoning also holds for general Kantorovich problem:

Theorem 4.9 (c-cyclical monotonicity of the support of optimal plan)**.** *Assume that $c : X \times Y \to [0, \infty)$ is continuous and that $\pi \in \Gamma(\mu, \nu)$ is optimal with $\int c \mathrm{d}\pi < \infty$. Then* $\mathrm{supp}\pi$ *is c-cyclically monotone.*

Exercise 4.4. Prove Theorem 4.9. The proof strategy is almost the same as that of the preceding discrete version. *Hint:* since c is continuous, one may construct some neighborhoods $U_i \times V_i$ around (x_i, y_i) such that $c(x, y) > c(x_i, y_i) - \epsilon$ on $U_i \times V_i$ and $c(x, y) < c(x_i, y_{\sigma(i)}) + \epsilon$ on $U_i \times V_{\sigma(i)}$.

4.2.4 Duality theory II: $X \times Y$ non-compact

Let us see how much possible can we extend our result to non-compact $X \times Y$. Let $\pi \in \Gamma(\mu,\nu)$ be a minimizer (whose existence is guaranteed by Proposition 4.6). When $c : X \times Y \to \mathbb{R}$ is continuous, by Theorem 4.9, the support of π is c-cyclically monotone. Thus by Theorem 4.8,

$$\mathrm{supp}\pi \subseteq \mathrm{Gr}(\partial^c \phi) = \{(x,y) \in X \times Y : \phi(x) + \phi^c(y) = c(x,y)\}$$

for some c-concave function ϕ, and $(\phi,\phi^c) \in \mathrm{Lip}_b(X) \times \mathrm{Lip}_b(Y)$ whenever $c \in \mathrm{Lip}_b(X \times Y)$. Since $c(x,y) - (\phi(x) + \phi^c(y)) = 0$ on the support of π, we have

$$\begin{aligned}\int c\mathrm{d}\pi &= \int c(x,y) - (\phi(x) + \phi^c(y))\mathrm{d}\pi + \int \phi(x) + \phi^c(y)\mathrm{d}\pi \\ &= \int \phi(x) + \phi^c(y)\mathrm{d}\pi(x,y) \\ &= \int_X \phi(x)\mathrm{d}\mu(x) + \int_Y \phi^c(y)\mathrm{d}\nu(y).\end{aligned}$$

To summarize, we have proved:

Proposition 4.7 (Bounded Lipschitz cost function)**.** *The Kantorovich-Rubinstein duality* (4.21) *holds for $c \in \mathrm{Lip}_b(X \times Y)$. In addition, the maximum on the RHS is attained at a pair $(\phi,\phi^c) \in \mathrm{Lip}_b(X) \times \mathrm{Lip}_b(Y)$, where ϕ^c is the c-transform of ϕ. The function ϕ is called the Kantorovich potential.*

Notice that in the proof of Proposition 4.7, we didn't use Fenchel duality theorem, but the result is stronger than Proposition 4.6 when c is bounded Lipschitz. In particular, 4.7 holds for c with negative parts. But Proposition 4.6 has the advantage that it holds for l.s.c. c and that it does need to be bounded. The goal of the rest of this subsection is to prove the following result for non-negative c:

Theorem 4.10 (Kantorovich-Rubinstein duality)**.** *Assume that $c : X \times Y \to [0,\infty]$ is l.s.c., then the Kantorovich-Rubinstein duality* (4.21) *holds.*

Proof. In the proof of Proposition 4.5, we have shown that c can be approximated by an increasing sequence of bounded Lipschitz functions $\{c_k\}$: $c_k \uparrow c$ as $k \to \infty$. Let π be an optimal plan, we have

$$\begin{aligned}\min_\pi \int c\mathrm{d}\pi &\geq \sup_{\phi+\psi\leq c} \int \phi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \text{ (weak duality)} \\ &\geq \sup_{\phi+\psi\leq c_k} \int \phi\mathrm{d}\mu + \int \psi\mathrm{d}\nu \\ &= \min_\pi \int c_k\mathrm{d}\pi \text{ (Proposition 4.7)}\end{aligned} \tag{4.26}$$

Thus it suffices to prove

$$\lim_{k\to\infty} \min_\pi \int c_k\mathrm{d}\pi \geq \min_\pi \int c\mathrm{d}\pi$$

which forces the inequality (4.26) to be equality. Let $\pi_k \in \arg\min_\pi \int c_k\mathrm{d}\pi$. Since $\Gamma(\mu,\nu)$ is compact, we can find a subsequence of $\{\pi_k\}$, still denoted as $\{\pi_k\}$, and some $\pi_* \in \Gamma(\mu,\nu)$, such that $\pi_k \rightharpoonup \pi_*$ as $k \to \infty$. This implies

$$\lim_{k\to\infty} \min_\pi \int c_k\mathrm{d}\pi = \lim_{k\to\infty} \int c_k\mathrm{d}\pi_k \geq \liminf_{k\to\infty} \int c_p\mathrm{d}\pi_k = \int c_p\mathrm{d}\pi_* \geq \int c_p\mathrm{d}\pi$$

for any $p \geq 1$ since c_k is an increasing sequence and π is optimal. Then letting $p \to \infty$, by monotone convergence theorem, we obtain

$$\lim_{k\to\infty} \min_{\pi} \int c_k \mathrm{d}\pi \geq \int c \mathrm{d}\pi$$

as desired. □

Remember in Theorem 4.9, we proved that when $c : X \times Y \to \mathbb{R}$ is continuous, then the support of π is c-cyclically monotone. The following rather surprising result says that the converse is also true when $x \to c(x,y)$ and $y \to c(x,y)$ are bounded by integrable functions.

Theorem 4.11. *Assume $c : X \times Y \to [0,\infty]$ is l.s.c. and there exist functions $f \in L^1(\mu)$, $g \in L^1(\nu)$ such that $c(x,y) \leq f(x) + g(y)$. Then*

1) $\pi \in \Gamma(\mu,\nu)$ is optimal if and only if suppπ *is c-cyclically monotone.*

2) there exists a c-concave function $\phi : X \to [-\infty,\infty)$ such that $\phi \in L^1(\mu)$, $\phi^c \in L^1(\nu)$ and

$$\min_{\pi} \int c \mathrm{d}\pi = \int_X \phi \mathrm{d}\mu + \int_Y \phi^c \mathrm{d}\nu.$$

4.2.5 Existence of optimal maps

We are now ready to study a bit the existence theory of minimizers of the Monge problem. Due to the highly nonlinearality of the Monge problem, general results regarding existence of minimizers are not available. However, there are at least two important situations of which strong conclusions can be made. The fist one is the quadratic cost case in Euclidean space and the other is convex cost on the real line.

Quadratic case: $c(x,y) = \frac{1}{2}|x-y|^2$

Assumption: $X = Y = \mathbb{R}^n$, $c(x,y) = \frac{1}{2}|x-y|^2$, μ, ν are probability measures with finite second moment, i.e., $\int |x|^2 \mathrm{d}\mu(x), \int |x|^2 \mathrm{d}\nu(x) < \infty$, and μ is absolutely continuous w.r.t. $\mathscr{L}^n$.

The following is the main theorem on existence of optimal maps. It reveals a rather surprising connection between Monge problem and convex analysis (be aware that the Monge problem is highly nonlinear).

Proposition 4.8. *Under the above assumption. The Monge problem has a unique solution. Further more, the optimal map is constructed from a convex function ψ differentiable μ-a.e., given by the formula $T(x) = \nabla\psi(x)$ for μ-a.e. x. Conversely, if ψ is convex, differentiable μ-a.e. with $|\nabla\psi| \in L^2(\mu)$, i.e., $\int |\nabla\psi|^2 \mathrm{d}\mu < \infty$, then $T := \nabla\psi$ is optimal from μ to $\nu := T_{\#}\mu$.*

The proof strategy is to construct the optimal map from an optimal plan, with the help of the existence theory of optimal plans that we have already studied in Section 4.2.4.

Proof. Under the assumption, the Kantorovich problem has an optimal solution π supported on $\mathrm{Gr}\partial^c\phi$ for some c-concave function ϕ. Let us take a closer look at what it means to be c-concave in this context. By definition

$$\phi(x) = \inf_{i\in\mathscr{A}} c(x,y_i) + \alpha_i = \inf_{i\in\mathscr{A}} \frac{1}{2}|x-y_i|^2 + \alpha_i = \frac{1}{2}|x|^2 + \inf_{i\in\mathscr{A}} \frac{1}{2}|y_i|^2 - y_i^\top x + \alpha_i.$$

It is immediate to see that $\phi(x) - \frac{1}{2}|x|^2$ is concave and lower-semi continuoussince it is the infimum of a family of affine functions. Obviously, the converse also holds. Thus for quadratic cost, ϕ is c-concave if

and only if $\phi(x) - \frac{1}{2}|x|^2$ is concave. Since a convex function is differentiable almost every where, so is ϕ. At point x where $\nabla\phi$ exists, we examine $\mathrm{Gr}(\partial^c\phi)$. If $(x,y) \in \mathrm{Gr}(\partial^c\phi)$, then $c(x',y) - \phi(x') = \frac{1}{2}|x'-y|^2 - \phi(x')$ is minimal at $x' = x$. Differentiating w.r.t. x', we get

$$y = x - \nabla\phi(x) = \nabla\left(\frac{1}{2}|x|^2 - \phi(x)\right) =: \nabla\psi(x)$$

where $\psi(x) = \frac{1}{2}|x|^2 - \phi(x)$ is convex and . But this implies that for any $x \in p_X(\mathrm{Gr}(\partial^c\phi))$, there is only one point corresponding to x, which is $\nabla\psi(x)$. Thus we can define a map T on $\mathbb{R}^n$ as

$$T(x) = \nabla\psi(x)$$

and π is supported on the graph of T. Consequently, $\pi = (\mathrm{id}\times\nabla\psi)_\#\mu$, and T are optimal plan and optimal map respectively invoking Pratelli Theorem 4.3. To see that T is unique up to a negligible set, suppose that T' is another optimal map. Then $\pi' = (\mathrm{id}\times T')_\#\mu$ is also an optimal plan. Moreover, $\pi'' = \frac{1}{2}(\pi+\pi')$ is also optimal, which, by similar reasoning above, is supported on a graph which is only possible when $T = T'$ μ-a.e.

To prove the converse, we utilize Theorem 4.11 invoking that $c(x,y) = \frac{1}{2}|x-y|^2 \le |x|^2 + |y|^2$. We need to show that the graph of $\nabla\psi$ is c-cyclically monotone. But since ψ is convex, by definition of the subdifferential, we have for any set of points $(x_1,\nabla\psi(x_1)),\cdots,(x_N,\nabla\psi(x_N))$ on the graph and permutation σ of $\{1,\cdots,N\}$:

$$\left\langle \nabla\psi(x_i), x_{\sigma(i)} - x_i \right\rangle \le \psi(x_{\sigma(i)}) - \psi(x_i)$$

from which it follows that

$$\sum_{i=1}^{N} \left\langle \nabla\psi(x_i), x_{\sigma(i)} - x_i \right\rangle \le 0.$$

But this is equivalent to (verify!):

$$\sum_{i=1}^{N} |\nabla\psi(x_i) - x_i|^2 \le \sum_{i=1}^{N} |\nabla\psi(x_i) - x_{\sigma(i)}|^2$$

as desired. □

Remark 4.1. As a byproduct, we also see that optimal plan for the Kantorovich problem under assumption is unique up to a $\mu\otimes\nu$ negligible set. Indeed, if π' is another optimal plan, then $\pi'' = \frac{1}{2}(\pi+\pi')$ is also optimal. But we have seen that the graph of π'' must coincide with that of π for μ-a.e. x, which implies that $\pi'' = \pi$ and consequently $\pi = \pi'$ for $\mu\otimes\nu$-a.e $x, y \in X\times Y$. More generally, whenever an optimal plan must be induced by a transport map, then we have uniqueness of both.

The real line: $c(x,y) = h(x-y)$

Let us consider the cost of the form $c(x,y) = h(x-y)$ on the real line, where $h : \mathbb{R} \to [0,\infty)$ is convex. Note carefully that h takes values in $\mathbb{R}$ instead of in $\mathbb{R}\cup\{\infty\}$, which forces c to be a continuous function on $\mathbb{R}$, a well-known fact.

Exercise 4.5. Any convex function $h : \mathbb{R}^n \to \mathbb{R}$ is continuous. Give an example that this fails to be true if we allow $+\infty$ in the range of h.

Now by Theorem 4.6, there exists an optimal plan π to the Kantorovich problem. By Theorem 4.9, if $\int c\mathrm{d}\pi < +\infty$, then $\Gamma := \mathrm{supp}\pi$ is c-cyclically monotone. Now Rockafellar's theorem 4.8 says that there exists a c-concave function $\phi : \mathbb{R} \to [-\infty, \infty)$, $\phi \not\equiv -\infty$, such that $\Gamma \subseteq \mathrm{Gr}(\partial^c \phi)$. In particular, when c is bounded Lipschitz, e.g., when c vanishes outside a compact set., ϕ can be chosen to be bounded Lipschitz, which is differentiable $\mathscr{L}^1$-almost everywhere. In this case, for any $(x, y) \in \mathrm{Gr}(\partial^c \phi)$, by definition $h(x' - y) - \phi(x')$ is minimum at $x' = x$, and thus

$$\nabla h(x - y) - \nabla \phi(x) = 0 \tag{4.27}$$

. Since h is strictly convex, ∇h is invertible, and we may further get

$$y = x - (\nabla h)^{-1} \circ \nabla \phi(x) := T(x).$$

By Remark 4.1, T is the unique optimal transport map. Notice that the formula (4.27) can be

Remember that in Example 4.9, we have shown that for $\mu \in \mathscr{P}(\mathbb{R})$ atomless, formula (4.15) provides a transport map. In particular, when $\mathrm{supp}\nu = \mathbb{R}$, then T can be explicitly written as $T = F_\nu^{-1} \circ F_\mu$ where F_μ, F_ν are the cumulative functions of the measure μ and ν respectively. To conclude, we have:

Proposition 4.9. *Suppose that $c(x, y) = h(x - y)$ for some strictly convex function $h : \mathbb{R} \to [0, \infty)$, μ is atomless and c is bounded Lipschitz. Then the optimal transport plan π is induced by the unique optimal transport map, i.e., $\pi = (\mathrm{id} \times T)_\# \mu$.*

In this proposition, the requirement bounded Lipschitz of c is somehow too strong. To relax this, we use the following technical lemma.

Lemma 4.2. *Assume that $\Gamma \subseteq \mathbb{R} \times \mathbb{R}$ is c-cyclically monotone. Then Γ is a monotone graph in the sense that whenever $(x, y), (x', y') \in \Gamma$ and $x < x'$, one has $y \leq y'$.*

Now starting from the fact that $\mathrm{supp}\pi$ is c-cyclically monotone, $\mathrm{supp}\pi$ is concentrated on a monotone graph by the above lemma. However, a monotone function can have at most countably many discontinuity points and all of them are of the first kind. Thus through the optimal plan π, a transport map can be constructed, which is uniquely determined up to $\mathscr{L}^1$-negligible sets, and as before, this is the unique optimal transport map.

We left the proof of Lemma 4.2 as an exercise.

Proposition 4.10. *Suppose that $c(x, y) = h(|x - y|)$ for some convex nondecreasing function $h : \mathbb{R} \to [0, \infty)$, μ is atomless. Then there exists an optimal transport map T (possibly non-unique) if $\int c(x, T(x))\mathrm{d}\mu(x) < \infty$.*

Example 4.12. Let $\mu = \mathscr{L}^1|_{[0,1]}$, $\nu = \mathscr{L}^1|_{[1,2]}$, and $c(x, y) = |x - y|$. To obtain calculate the optimal value, we can use the results in 4.2.5 since $c(x, y) = h(x - y)$, with $h(x) = |x|$ convex. Then $T(x) = F_\nu^{-1} \circ F_\mu(x)$ is an optimal map, which results in $\min_{(\mathrm{M})} = 1$. On the other, $T_1(x) = x + 1$ and $T_2(x) = 2 - x$ are also optimal.

Example 4.13 (Histogram equalization)**.** Histogram equalization is a common operation used to increase the global contrast of an image. This is applied for example when the intensity of pixels of the image lie in a narrow range in the histogram. Suppose that the histogram of the original image is represented by a vector of dimension N, or a probability measure (after normalization) $\mu = \sum_{i=0}^N a_i \delta_i$. For

grayscale images, N is usually 255. Histogram equalization is to find a map which transforms the histogram μ into a uniform distribution ν, still supported on $\{1,\cdots,N\}$. The problem is easier to solve if we view μ as an atomless continuous distribution, for example, by approximation. Then $T(m) = F_\nu^{-1} \circ F_\mu(m)$ (see Example 4.9) is the unique optimal transport map for any cost $c(x,y) = h(x-y)$ with h strictly convex. Since ν is uniform, T has a simple formula

$$T(m) = (N+1)\sum_{i=1}^{m} a_i.$$

This formula says that, for pixels of intensity m, it should be mapped to intensity $T(m)$.

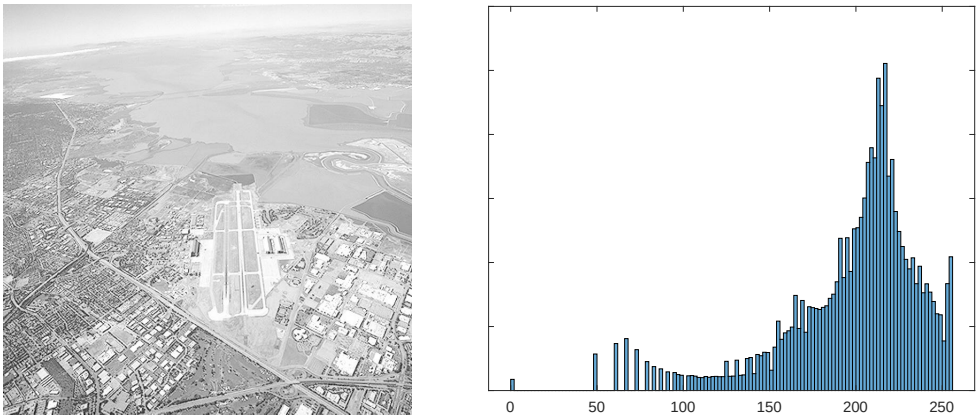

Figure 4.7: Original image and histogram

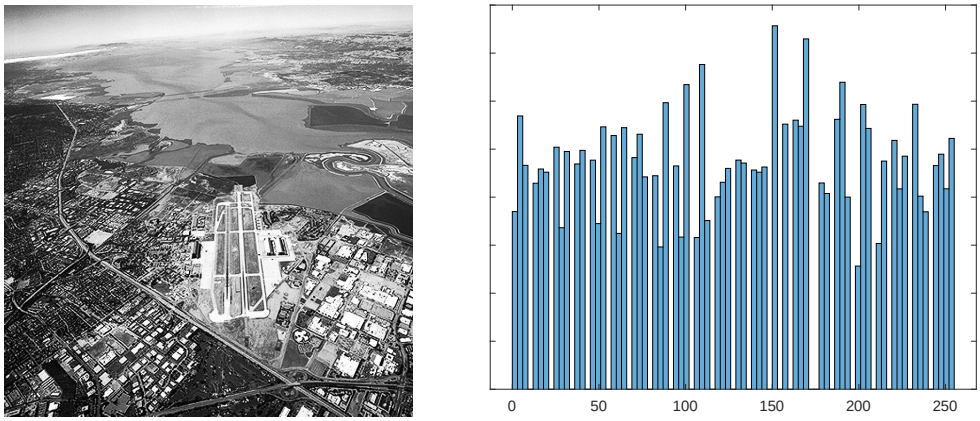

Figure 4.8: After histogram equalization

Dido's problem revisited

We now solve Dido's problem using a totally different approach. In Figure 4.9, we draw two regions in yellow color. On the right is a semi-circle with radius r and whose center is at the origin. The set on the left is a region enclosed by a curve γ and the x-axis which has the same area as the semi-circle, i.e. $\text{area}(C) = \text{area}(H) = \frac{1}{2}\pi r^2$.

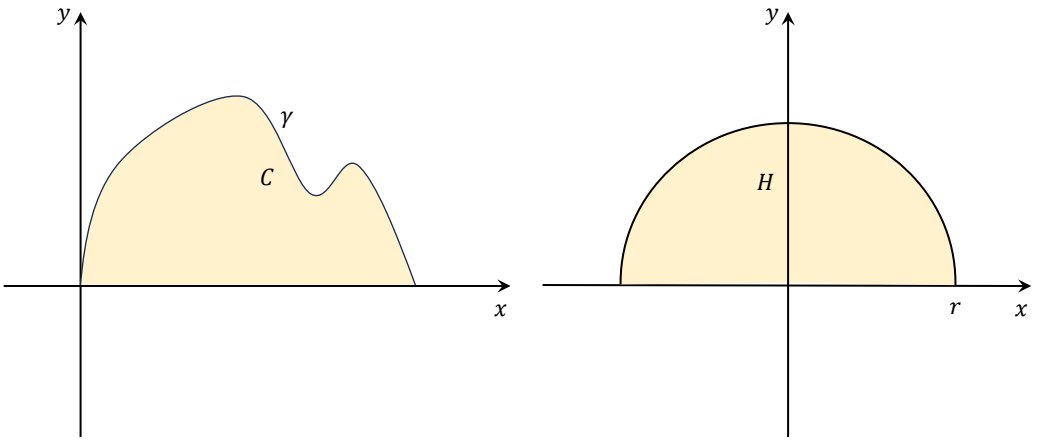

Figure 4.9: Dido's problem and optimal transport

We want to show that the length of the curve γ is less than the perimeter of the half circle, i.e., πr. If we reflect the two sets along the x-axis, then we obtain two new sets, still denoted as C and H, as in Figure 4.10.

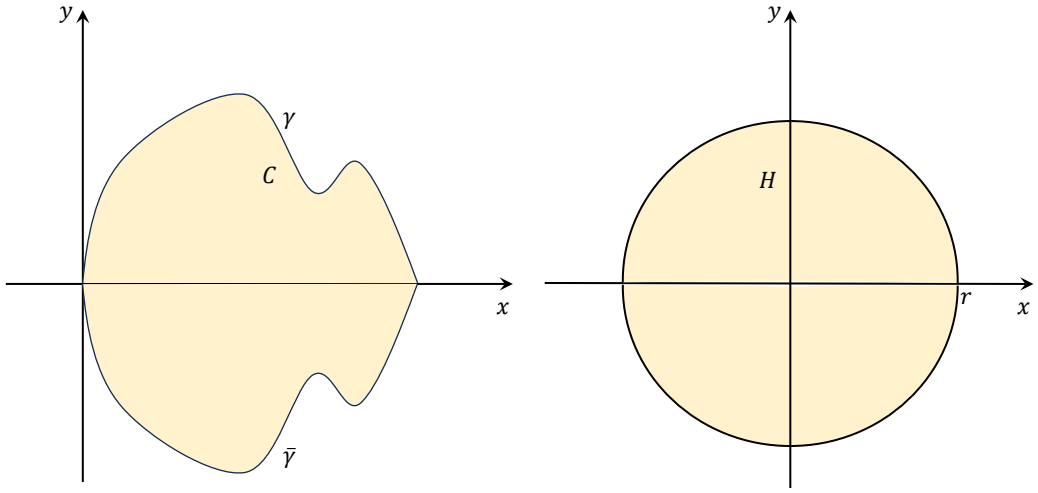

Figure 4.10: Dido's problem and optimal transport

Define two probability measures that represent the two sets:

$$\mu := \frac{1}{\pi r^2}\mathscr{L}^2|_C, \quad \nu := \frac{1}{\pi r^2}\mathscr{L}^2|_H$$

where $\mathscr{L}^2$ is the Lebesgue measure on the plane. Now since μ has density, under quadratic cost c, there exists an optimal transport map T on $\mathbb{R}^2$ such that $T_\#\mu = \nu$ by Proposition 4.8. Assume that T is C^1 (the regularity is a subtle issue), then by mass preservation, T must map the points in C onto H. Again by Proposition 4.8, T is the gradient of a convex function ϕ, i.e., $T = \nabla\phi$. By Monge-Ampere equation (4.9),

we have

$$\det \nabla T(x) = \frac{\rho_C(x)}{\rho_E(T(x))} = 1, \quad \forall x \in C,$$

where ρ_C and ρ_E are the densities of μ and ν respectively. Since $\nabla T = \nabla^2 \phi$, ∇T is symmetric and has only real eigenvalues, say $\{\lambda_1, \lambda_2\}$. Thus

$$1 = \sqrt{\det \nabla T(x)} = \sqrt{\lambda_1 \lambda_2} \leq \frac{\lambda_1 + \lambda_2}{2} = \frac{1}{2}\mathrm{tr}(\nabla T(x)) = \frac{1}{2}\mathrm{div} T(x)$$

and

$$\int_C \mathrm{div} T(x) \mathrm{d}x \geq 2\mathrm{vol}(C) = 2\pi r^2$$

On the other hand, let ν be the outward pointing normal field on ∂C, then by divergence theorem,

$$\int_C \mathrm{div} T(x) \mathrm{d}x = \int_{\partial C} \langle T, \nu \rangle \, \mathrm{d}l \leq \int_{\partial C} r \mathrm{d}l = 2r\ell(\gamma)$$

since $T(C) \subseteq B$, from which we obtain

$$\ell(\gamma) \geq \pi r$$

as desired.

It is easy to see that the above reasoning also holds in higher dimension. More precisely, let $\mathscr{L}^n$ and σ^{n-1} denote the Lebesgue volume and surface measures in $\mathbb{R}^n$. We have the following generalized result of the Dido's problem, which is called the *isopermetric inequality:*

Proposition 4.11 (Isoperimetric inequality)**.** *Let $E \subseteq \mathbb{R}^n$ be a bounded open set with C^1 boundary. Let $B \subseteq \mathbb{R}^n$ be the ball with $\mathscr{L}^n(E) = \mathscr{L}^n(B)$. Then $\sigma^{n-1}(\partial E) \geq \sigma^{n-1}(\partial B)$.*

Proof. It is sufficient to note the following inequalities:

$$1 = (\det \nabla T)^{1/n} \leq \frac{1}{n}\mathrm{div} T$$

and

$$\sigma^{n-1}(\partial B) = n\mathscr{L}^n(B) = n\mathscr{L}^n(E) \leq \int_E \mathrm{div} T \mathrm{d}x = \int_{\partial E} \langle T, \nu \rangle \mathrm{d}\sigma^{n-1} \leq \sigma^{n-1}(\partial E).$$

□

4.3 Metric properties of optimal transport

4.3.1 Wasserstein spaces

Optimal transport provides a way of measuring the difference/distance between different measures. This shall be clear once we have introduced the metric side of optimal transport.

Given a metric space (X, d), set

$$\mathscr{P}_p(X) := \left\{ \mu \in \mathscr{P}(X) : \int d(x, x_0)^p \mathrm{d}\mu(x) < \infty \text{ for some } x_0 (\text{hence for all}) \ x_0 \in X \right\}$$

for $p \in [1, \infty]$ and define the Wasserstein distance on $\mathscr{P}_p(X)$ as the optimal value of the Kantorovich problem (with $c(x, y) = d(x, y)^p$):

$$W_p^p(\mu, \nu) := \min_{\pi \in \Gamma(\mu, \nu)} \int d(x, y)^p \mathrm{d}\pi(x, y).$$

In particular, for $\mu,\nu\in\mathscr{P}_p(\mathbb{R}^n)$, $W_p^p(\mu,\nu)=\min_{\pi\in\Gamma(\mu,\nu)}\int|x-y|^p\mathrm{d}\pi(x,y)$. As in functional analysis, $p=1,2,+\infty$ are the most interesting cases. We prove that W_p defines a metric or distance on $\mathscr{P}_p(X)$, i.e., we need check for $\mu,\nu,\sigma\in\mathscr{P}_p(X)$, the three properties:

1) $W_p(\mu,\nu)\geq 0$ and $W_p(\mu,\nu)=0$ iff $\mu=\nu$;

2) $W_p(\mu,\nu)=W_p(\nu,\mu)$;

3) $W_p(\mu,\sigma)\leq W_p(\mu,\nu)+W_p(\nu,\sigma)$.

First we prove 2). Let $i:Y\times X\to X\times Y$ be defined by $i(x,y)=(y,x)$, then

$$\int d(x,y)\mathrm{d}\pi(x,y)=\int_{Y\times X}d\circ i(y,x)\mathrm{d}(i^{-1})_{\#}\pi(y,x)$$

and $\pi\in\Gamma(\mu,\nu)$ iff $(i^{-1})_{\#}\pi\in\Gamma(\nu,\mu)$. Thus

$$\min_{\pi\in\Gamma(\mu,\nu)}\int_{X\times Y}d(x,y)\mathrm{d}\pi(x,y)=\min_{\pi\in\Gamma(\nu,\mu)}\int_{Y\times X}d(y,x)\mathrm{d}\pi(y,x).$$

For 1), the only nontrivial claim we need to prove is $W_p(\mu,\nu)=0$ implies $\mu=\nu$. Suppose $W_p(\mu,\nu)=0$, then $x=y$ for π-a.e. $(x,y)\in X\times Y$. Then for any $f\in C_b(X)$,

$$\int f\mathrm{d}\mu=\int f(x)\mathrm{d}\mu(x)=\int f(x)\mathrm{d}\pi(x,y)=\int f(x)\mathrm{d}\pi(x,x)=\int f(y)\mathrm{d}\pi(y,x)=\int f\mathrm{d}\nu$$

Thus $\mu=\nu$.

It remains to prove 3). We need the following technical lemma.

Lemma 4.3 (Dudley)**.** *Let (X_i,μ_i), $i=1,2,3$ be Polish spaces, $\pi_{12}\in\Gamma(\mu_1,\mu_2)$ and $\pi_{23}\in\Gamma(\mu_2,\mu_3)$. Then there exists $\pi\in\mathscr{P}(X_1\times X_2\times X_3)$ such that*

$$p^{12}_{\#}(\pi)=\pi_{12},\quad p^{23}_{\#}(\pi)=\pi_{23}$$

where $p^{ij}(x_1,x_2,x_3)=(x_i,x_j)$.

Now let π_{12} and π_{23} be optimal plans between μ,ν and ν,σ respectively. Let π be such that $p^{12}_{\#}(\pi)=\pi_{12}$ and $p^{23}_{\#}(\pi)=\pi_{23}$. Since $p^{13}_{\#}\pi\in\Gamma(\mu,\sigma)$,

$$\begin{aligned}
W_p(\mu,\sigma)^p&\leq\int_{X^2}d(x_1,x_3)^p\mathrm{d}p^{13}_{\#}\pi(x_1,x_3)\\
&=\int_{X^3}d(x_1,x_3)^p\mathrm{d}\pi(x_1,x_2,x_3)\ \text{(change of measure)}\\
&\leq\int_{X^3}[d(x_1,x_2)+d(x_2,x_3)]^p\mathrm{d}\pi(x_1,x_2,x_3)\\
&\leq\left\{\left(\int_{X^3}d(x_1,x_2)^p\mathrm{d}\pi(x_1,x_2,x_3)\right)^{1/p}+\left(\int_{X^3}d(x_2,x_3)^p\mathrm{d}\pi(x_1,x_2,x_3)\right)^{1/p}\right\}^p\ \text{(Hölder)}\\
&=\left\{\left(\int_{X^2}d(x_1,x_2)^p\mathrm{d}\pi^{12}(x_1,x_2)\right)^{1/p}+\left(\int_{X^2}d(x_2,x_3)^p\mathrm{d}\pi^{23}(x_2,x_3)\right)^{1/p}\right\}^p\\
&=(W_p(\mu,\nu)+W_p(\nu,\sigma))^p
\end{aligned}$$

which is the desired triangle inequality.

In fact, we can say more:

Proposition 4.12. *If (X,d) is a complete metric space, then $(\mathscr{P}_p(X),W_p)$ is also complete.*

The proof of this proposition relies on a generalization of Dudley's lemma. Since it is only of theoretical interest to us, we omit the proof.

4.3.2 Geodesic structure

Given two probability measures μ and ν, suppose that we are interested not only in the initial/final destination, but also in the "path" used to move the mass in between. An application that motivates this problem can be interpolating two given images of a growing tumor at different times. This is analogous to finding the geodesic joining two points on a manifold, i.e., we care not only of the distance between points, but also of the shortest path (geodesic) joining the two points.

Let (X,d) be a complete metric space, and consider the Wasserstein space $(\mathscr{P}_2(X), W_2)$. A path joining the initial and final measures μ and ν is a map $\mu_\cdot : [0,1] \to \mathscr{P}_2(X)$ such that $\mu_0 = \mu$ and $\mu_1 = \nu$. Our objective is to characterize the path between two points (measures) with the shortest length, but the length of a curve in $\mathscr{P}_2(X)$ is yet to be defined. Remember that for a curve in Euclidean space $\gamma : [0,1] \to \mathbb{R}^n$, the length of γ is defined as $\int_0^1 |\gamma'(t)| \mathrm{d}t$. That is, to define the length of a curve, one needs the definition of derivative, or velocity of the curve, and that the derivative is defined only for functions which are differentiable almost everywhere. A wide class of curves in $\mathbb{R}^n$ that are differentiable almost everywhere are *absolutely continuous* curves. However, we will not need this and instead define absolute continuity according to the following more useful form:

Definition 4.10 (Metric derivative)**.** Let (X,d) be a metric space. We say that a curve $\gamma : [a,b] \to X$ is *absolutely continuous* and we write $\gamma \in AC([a,b];X)$ if there exists $g \in L^1(a,b)$ such that

$$d(\gamma(x),\gamma(y)) \le \int_x^y g(t)\,dt, \ \forall\, a \le x \le y \le b. \tag{4.28}$$

And the *metric derivative* of $\gamma \in AC([a,b];X)$, denoted $|\gamma'(t)|$, is the limit (when exists, otherwise set to ∞)

$$|\gamma'(t)| := \lim_{h\to 0} \frac{d(\gamma(t),\gamma(t+h))}{|h|}. \tag{4.29}$$

The following result justifies our definition:

Proposition 4.13. *For any $\gamma \in AC([a,b];X)$, the lower limit* (4.29) *is a limit which exists for $\mathscr{L}^1$-a.e. $t \in [a,b]$ and $|\gamma'(\cdot)|$ is the minimal g, up to $\mathscr{L}^1$-negligible sets that satisfies* (4.28)*.*

With this proposition at hand, we can finally define the length of a curve in metric space.

Definition 4.11 (Length)**.** Given a curve $\gamma \in AC([a,b];X)$, the *length* of curve is defined as

$$\ell(\gamma) := \int_a^b |\gamma'(t)| \mathrm{d}t.$$

As usual, the length of curve is invariant under reparametrization, i.e., if $\phi : [a,b] \to [c,d]$ is strictly increasing, then

$$\ell(\gamma) = \ell(\gamma \circ \phi).$$

Thus we can always find a reparametrization ϕ such that $\tilde{\gamma} := \gamma \circ \phi$ has constant speed, i.e., $|\tilde{\gamma}'(t)|$ is constant for $\mathscr{L}^1$-a.e. t.

Note that the length should be defined in a way that it is always larger than the distance between the two endpoints. This is indeed true since by Proposition 4.13 and formula (4.28), $d(\gamma(a),\gamma(b)) \le \int_a^b |\gamma'(t)| \mathrm{d}t$. When the inequality becomes equality, we call the curve γ a geodesic:

Definition 4.12 (Geodesic)**.** We say that $\gamma \in AC([a,b];X)$ is a *geodesic* if

$$\ell(\gamma) = d(\gamma(a),\gamma(b)).$$

An important property to notice is that the restriction of a geodesic on any interval is again geodesic (verify!).

Definition 4.13 (Geodesic space)**.** Denote $\mathrm{Geo}(X)$ the space of constant speed geodesic on $[0,1]$. We say that (X,d) is geodesic if for all $x,y \in X$, there exists $\gamma \in \mathrm{Geo}(X)$ with $\gamma(0)=x$ and $\gamma(1)=y$.

The following are some obvious properties of the space $\mathrm{Geo}(X)$:
1) For $\gamma \in \mathrm{Geo}(X)$, the length of γ is $\ell(\gamma) = d(\gamma(0),\gamma(1))$;
2) The speed of γ is $d(\gamma(0),\gamma(1))$, i.e., $|\gamma'(t)| = d(\gamma(0),\gamma(1))$ for a.e. t;
3) A continuous curve $\gamma \in C([0,1],X)$ is in $\mathrm{Geo}(X)$ if and only if

$$d(\gamma(s),\gamma(t)) = |t-s| d(\gamma(0),\gamma(1)), \quad \forall\, s,t \in [0,1]. \tag{4.30}$$

To see this, assume $\gamma \in \mathrm{Geo}(X)$, then $d(\gamma(s),\gamma(t)) = \int_s^t |\gamma'| = |t-s| d(\gamma(0),\gamma(1))$. Conversely, if $\gamma \in C([0,1],X)$ is such that the above equality holds for any $s<t\in[0,1]$, then for $s<s'<t'<t$, we have

$$\begin{aligned} d(\gamma(0),\gamma(1)) &\le d(\gamma(0),\gamma(s)) + d(\gamma(s),\gamma(t)) + d(\gamma(t),\gamma(1)) \\ &= (s + (t-s) + 1 - t) d(\gamma(0),\gamma(1)) \\ &= d(\gamma(0),\gamma(1)) \end{aligned}$$

thus the inequality is equality and (4.30) holds. This implies that γ is absolutely continuous and $|\gamma'(t)| = d(\gamma(0),\gamma(1))$. Through the proof we see that the condition (4.30) can be relaxed to

$$d(\gamma(s),\gamma(t)) \le |t-s| d(\gamma(0),\gamma(1)), \quad \forall\, s,t \in [0,1]. \tag{4.31}$$

The following is our main theorem of this subsection: it says that we $(\mathscr{P}_p(X), W_p)$ is a geodesic space whenever (X,d) is, and that we can lift the geodesic in X to $\mathscr{P}_p(X)$ in a rather simple manner.

Theorem 4.12. *If (X,d) is a geodesic space, then*

1) $(\mathscr{P}_p(X), W_p)$ is also geodesic.

2) Let $\mu_0,\mu_1 \in \mathscr{P}_p(X)$, and π an optimal plan for the cost $c(x,y) = d(x,y)^p$, $\gamma_{x,y} : [0,1] \to X$ the geodesic joining x to y, then

$$\mu_t = (\Gamma_t)_\# \pi$$

is a constant speed geodesic in $\mathscr{P}_p(X)$ connecting μ_0 and μ_1, where $\Gamma_t(x,y) = \gamma_{x,y}(t)$.

3) If μ_0 has a density, and that there exists an optimal map T, then

$$\mu_t = (T_t)_\# \mu_0$$

is a geodesic connecting μ_0 to μ_1 where $T_t(x) = \gamma_{x,T(x)}(t)$.

Proof. Let π by an optimal plan between μ_0 and μ_1. We will need to show that $t \mapsto \mu_t = (\Gamma_t)_\#\pi$ is a constant speed geodesic in $\mathscr{P}_p(X)$. By (4.31), it suffices to show

$$W_p(\mu_s,\mu_t) \le |t-s| W_p(\mu_0,\mu_1), \quad \forall\, t,s \in [0,1].$$

Indeed,

$$\begin{aligned} W_p^p(\mu_s,\mu_t) &\le \int d(x,y)^p \mathrm{d}(\Gamma_s,\Gamma_t)_{\#}\pi \\ &= \int d(\gamma_{x,y}(s),\gamma_{x,y}(t))^p \mathrm{d}\pi(x,y) \end{aligned}$$

since $(\Gamma_s,\Gamma_t)_{\#}\pi \in \Gamma(\mu_s,\mu_t)$. Now that $\gamma_{x,y}$ is a geodesic, then $d(\gamma_{x,y}(s),\gamma_{x,y}(t)) = |s-t|d(x,y)$ and

$$W_p^p(\mu_s,\mu_t) \le |s-t|^p \int d(x,y)^p \mathrm{d}\pi(x,y) = |s-t|^p W_p^p(\mu_0,\mu_1)$$

as desired. This proves 1) and 2). To prove 3), it's sufficient to note that an optimal plan is induced by an optimal map T. □

Example 4.14. If $X=\mathbb{R}^n$ and d is the usual Euclidean metric, then $\Gamma_t(x,y) = (1-t)x+ty$ and $T_t(x) = (1-t)x+tT(x) = [(1-t)\mathrm{id}+tT](x)$.

4.3.3 Benamou-Brenier formula

Let us take a look at the third item of Theorem 4.12. For convenience, we focus on Euclidean space, i.e., $X=\mathbb{R}^n$ equipped with Lebesgue measure $\mathscr{L}^n$. Suppose that

$$\mu_0 = \rho_0(x)\mathrm{d}x$$

it is then tempting to ask if the $\mu_t = (T_t)_{\#}\mu_0$ also has a density along the geodesic. Obviously, for this to hold, μ_1 should also have a density, say ρ_1. Next, assume that μ_t has a density $\rho_t(x)$, i.e.,

$$(T_t)_{\#}\mu_0(\mathrm{d}x) = \rho_t(x)\mathrm{d}x,$$

we would like to find the expression for ρ_t. Notice that the last equation implies that for all compactly supported $f:\mathbb{R}^n\to\mathbb{R}$, there holds

$$\int f(x)\mathrm{d}(T_t)_{\#}\mu_0(x) = \int f(x)\rho_t(x)\mathrm{d}x.$$

Differentiate this w.r.t. t, we get (remember that f is compactly supported):

$$\begin{aligned} \frac{d}{dt}\int f(x)\rho_t(x)\mathrm{d}x = \int f(x)\frac{\partial\rho_t}{\partial t}\mathrm{d}x &= \frac{d}{dt}\int f\circ T_t(x)\rho_0(x)\mathrm{d}x \\ &= \int \nabla f\cdot\frac{\partial T_t}{\partial t}\rho_0(x)\mathrm{d}x \\ &= -\int f(x)\mathrm{div}\left[\rho_0(x)\frac{\partial T_t(x)}{\partial t}\right]\mathrm{d}x \text{ (integration by parts)} \end{aligned}$$

By Fundamental lemma, the first and third line imply

$$\frac{\partial\rho_t}{\partial t} + \mathrm{div}\left(\rho_0 v_t\right) = 0$$

where

$$v_t(x) = (T-\mathrm{id})\circ T_t^{-1}(x) \tag{4.32}$$

since

$$\frac{\partial T_t(x)}{\partial t} = T(x)-x = (T-\mathrm{id})\circ T_t^{-1}(T_t(x)).$$

In other words, $T_t(x)$ is the solution to the following non-autonomous Cauchy problem

$$\begin{cases} \dot{\gamma}_x(t) = v_t(\gamma_x(t)) \\ \gamma_x(0) = x \end{cases} \tag{4.33}$$

Remark 4.2. On the other hand, by formula (4.14), we see

$$\rho_t(x) = \frac{\rho_0}{\det \nabla T_t} \circ (T_t)^{-1}(x).$$

The metric derivative of W_p, by definition is

$$|\mu_t'| = \lim_{h\to 0} \frac{W_p(\mu_t, \mu_{t+h})}{|h|}.$$

We want to show that for any μ_t induced by the flow of the ordinary equation (4.33), i.e., $\mu_t = (X_t)_{\#}\mu_0$, where X_t is the flow, the metric derivative of the curve $t \mapsto \mu_t$ satisfies

$$|\mu_t'|^p \le \int_{\mathbb{R}^n} |v_t(x)|^p \mathrm{d}\mu_t(x) = ||v_t(\cdot)||^p_{L^p(\mu_t)}.$$

In fact, by definition, $\mu_{t+h} = (X_{t+h} \circ X_t^{-1})_{\#}\mu_t$. Thus we have

$$\begin{aligned} W_p^p(\mu_t, \mu_{t+h}) &\le \int |x - X_{t+h} \circ X_t^{-1}(x)|^p \mathrm{d}\mu_t \\ &= \int |X_t \circ X_t^{-1}(x) - X_{t+h} \circ X_t^{-1}(x)|^p \mathrm{d}\mu_t(x) \\ &= \int \left| \int_t^{t+h} v_r(X_r \circ X_t^{-1}(x)) \mathrm{d}r \right|^p \mathrm{d}\mu_t(x) \\ &\le \left\{ \int_t^{t+h} \left[\int |v_r(X_r \circ X_t^{-1}(x)|^p \mathrm{d}\mu_t(x) \right]^{1/p} \mathrm{d}r \right\}^p \quad \text{(Minkowski inequality)} \\ &= \left\{ \int_t^{t+h} \left[\int |v_r(x)|^p \mathrm{d}\mu_r(x) \right]^{1/p} \mathrm{d}r \right\}^p \quad \text{(def. of } \mu_t) \\ &= \left\{ \int_t^{t+h} ||v_r(\cdot)||_{L^p(\mu_r)} \mathrm{d}r \right\}^p \end{aligned}$$

which implies $|\mu_t'| \le ||v_t(\cdot)||_{L^p(\mu_t)}$ for all $t \in [0,1]$.

When $t \mapsto \mu_t$ is a geodesic, then $|\mu_t'| = ||v_t(\cdot)||_{L^p(\mu_t)}$. Indeed, let T be the optimal map between μ_0 and μ_1, then

$$\begin{aligned} W_p^p(\mu_s, \mu_t) &= |s-t|^p W_p^p(\mu_0, \mu_1) \\ &= |s-t|^p \int |x - T(x)|^p \mathrm{d}\mu_0(x) \\ &= |s-t|^p \int |(\mathrm{id} - T) \circ T_t^{-1}(x)|^p \mathrm{d}(T_t)_{\#}\mu_0(x) \\ &= |s-t|^p \int |v_t(x)|^p \mathrm{d}\mu_t(x) \end{aligned}$$

as expected. Thus for $\mu_0 = \rho_0 dx$, $\mu_1 = \rho_1 dx$, we have

$$W_p^p(\mu_0, \mu_1) = \min_{\rho_t, v_t} \left\{ \int_0^1 ||v_t(\cdot)||_{L^p(\rho_t dx)} \mathrm{d}t : \frac{\partial \rho_t}{\partial t} + \mathrm{div}(\rho_t v_t) = 0 \right\}. \tag{4.34}$$

This is a special case of a more general formula, called *Benamou-Brenier formula* once its meaning is properly understood:

Theorem 4.13 (Benamou-Brenier formula)**.** *For $\mu_0, \mu_1 \in \mathscr{P}_p(\mathbb{R}^n)$, one has*

$$W_p^p(\mu_0,\mu_1) = \min\left\{\int_0^1 ||v_t(\cdot)||_{L^p(\mu_t)}\mathrm{d}t : \frac{\partial \mu_t}{\partial t} + \mathrm{div}(v_t\mu_t) = 0 \text{ in } (0,1)\times\mathbb{R}^n\right\}$$

where the minimization is among all curves $\mu_\cdot \in AC([0,1],\mathscr{P}_p(\mathbb{R}^n))$ and Borel vector field $v_t : \mathbb{R}^n \to \mathbb{R}^n$. The partial differential equation in the minimization is understood in the distributional sense, i.e.,

$$\int_0^\infty \int_{\mathbb{R}^n} \left[\frac{\partial \rho}{\partial t}(t,x) + v_t(x)\cdot\nabla\rho(t,x)\right]\mathrm{d}\mu_t(x)\mathrm{d}t = 0, \quad \forall \rho \in C_c^\infty((0,\infty)\times\mathbb{R}^n). \tag{4.35}$$

To see the motivation of (4.35), let $\mu_t = \phi(t,x)\mathrm{d}x$, then

$$\begin{aligned}
\int_0^\infty \int_{\mathbb{R}^n} \left[\frac{\partial \rho}{\partial t} + v_t\cdot\nabla\rho\right]\phi \mathrm{d}x\mathrm{d}t &= \int_{\mathbb{R}^n}\int_0^\infty \frac{\partial\rho}{\partial t}\phi\mathrm{d}t\mathrm{d}x + \int_0^\infty\int_{\mathbb{R}^n}(v_t\cdot\nabla\rho)\phi\mathrm{d}x\mathrm{d}t \\
&= -\int_{\mathbb{R}^n}\int_0^\infty \rho\frac{\partial\phi}{\partial t}\mathrm{d}t\mathrm{d}x - \int_0^\infty\int_{\mathbb{R}^n}\rho\mathrm{div}(\phi v_t)\mathrm{d}x\mathrm{d}t \\
&= -\int_0^\infty\int_{\mathbb{R}^n}\rho\left[\frac{\partial\phi}{\partial t} + \mathrm{div}(\phi v_t)\right]\mathrm{d}x\mathrm{d}t \\
&= 0
\end{aligned}$$

from which we deduce the equation in (4.34) (replacing ϕ with ρ_t).

4.4 Miscellaneous topics

4.4.1 L^1 optimal transport

In Section 4.3.2, we studied optimal transport for cost function $c(x,y) = d(x,y)^p$, $p \in (1,\infty)$ on a Polish space (X,d). In this subsection, we consider the special case $p = 1$. The reason to treat this case separately is that it has some distinguishing features from the other cases with $p > 1$. We call this type of problems L^1 optimal transport.

A first distinguishing feature of L^1 optimal transport is that c-concavity in this case is equivalent to Lipschitz continuity with Lipschitz constant 1 – denoted as $\mathrm{Lip}_1(X)$. In fact, a c-concave function has the form

$$\phi(x) = \inf_i d(x,y_i) + \alpha(y_i),$$

since $|d(x,y) - d(x',y)| \le d(x,x')$, the mapping $x \mapsto d(x,y)$ is in $\mathrm{Lip}_1(X)$, so is ϕ by Lemma 4.2. Conversely, if ϕ is in $\mathrm{Lip}_1(X)$, then $\phi(x) - \phi(y) \le d(x,y)$ for all $x, y \in X$. Thus

$$\phi(x) = \inf_y d(x,y) + \phi(y) = (-\phi)^c(x)$$

which by definition, is c-concave. To summarize:

Proposition 4.14. *Assume $c(x,y) = d(x,y)$, where d is the metric on X. Then a function $\phi : X \to [-\infty,\infty)$ is c-concave if and only if $\phi \in \mathrm{Lip}_1(X)$. The c-conjugate of ϕ is $\phi^c = -\phi$.*

Now by Proposition 4.7, strong duality holds:

$$\min_{\pi\in\Gamma(\mu,\nu)}\int_{X\times X} d(x,y)\mathrm{d}\pi(x,y) = \max_{\phi\in\mathrm{Lip}_1(X)}\int_X \phi\mathrm{d}(\mu-\nu). \tag{4.36}$$

and

$$\mathrm{supp}\pi \subseteq \{(x,y) \in X \times Y : \phi(x) - \phi(y) = d(x,y)\}$$

if ϕ is a Kantorovich potential.

Let $X = \mathbb{R}^n$ and assume μ is absolutely continuous w.r.t. $\mathscr{L}^n$. In this case, $c(x,y) = |x-y|$ where $|\cdot|$ is the Euclidean 2-norm. Now since c is differentiable almost everywhere, $\phi \in \mathrm{Lip}_1$ if and only if ϕ is continuous, differentiable a.e. and $|\nabla\phi| \leq 1$. Thus formula (4.36) can also be written as

$$\min_{\pi\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^n\times\mathbb{R}^n} d(x,y)\mathrm{d}\pi(x,y) = \max_{\phi\in \mathrm{Lip}_1(X)} \int_{\mathbb{R}^n} \phi \mathrm{d}(\mu-\nu) = \max_{|\nabla\phi|\leq 1} \int_{\mathbb{R}^n} \phi \mathrm{d}(\mu-\nu) \tag{4.37}$$

We can now proceed as in the quadratic case to arrive at the conclusion that the optimal plan π is supported on $\mathrm{Gr}(\partial^c\phi)$ for some $\phi \in \mathrm{Lip}_1$. Thus for any $(x,y) \in \mathrm{supp}\pi$, the mapping $x' \mapsto |x'-y| - \phi(x)$ achieves minimum at $x' = x$. Since ϕ is Lipschitz, it's differentiable almost everywhere. Differentiating w.r.t. x, we get

$$\nabla\phi(x) = \frac{x-y}{|x-y|}$$

from which we conclude that

$$y = x - t\nabla\phi(x)$$

for any $t > 0$, and $|\nabla\phi(x)| = 1$. Thus if an optimal transport map is to exist, at the current stage we only know the direction of the transport. This is quite different from the quadratic case. Example 4.12 shows that optimal transport maps may not be unique for L^1-optimal transport.

Let us now come back to the right most maximization of (4.37). If $d(\mu-\nu)(x)$ can be written as $\mathrm{div}(w)\mathrm{d}x$ for some function $w \in C_c^\infty(\mathbb{R}^n;\mathbb{R}^n)$, then using integration by parts, we get

$$\max_{|\nabla\phi|\leq 1} \int_{\mathbb{R}^n} \phi \mathrm{d}(\mu-\nu) = \max_{|\nabla\phi|\leq 1} \int_{\mathbb{R}^n} -\nabla\phi(x)\cdot w(x)\mathrm{d}x = \int_{\mathbb{R}^n} |w(x)|\mathrm{d}x.$$

More generally,

$$\max_{|\nabla\phi|\leq 1} \int_{\mathbb{R}^n} \phi \mathrm{d}(\mu-\nu) = \max_{\phi} \left\{ \int \phi \mathrm{d}(\mu-\nu) + \inf_w \int |w(x)| - \nabla\phi(x)\cdot w(x)\mathrm{d}x \right\} \tag{4.38}$$

since

$$\inf_w \int |w| - \nabla\phi\cdot w \mathrm{d}x = \begin{cases} 0, & \text{if } |\nabla\phi| \leq 1 \\ -\infty, & \text{else} \end{cases}.$$

If we can swap max and inf in (4.38), then we would get

$$\begin{aligned} \max_{|\nabla\phi|\leq 1} \int_{\mathbb{R}^n} \phi \mathrm{d}(\mu-\nu) &= \inf_w \int |w(x)|\mathrm{d}x + \max_\phi \left\{ \int \phi \mathrm{d}(\mu-\nu) - \int \nabla\phi\cdot w \mathrm{d}x \right\} \\ &= \inf_w \left\{ \int |w(x)|\mathrm{d}x : \int \phi \mathrm{d}(\mu-\nu) - \int \nabla\phi\cdot w\mathrm{d}x = 0, \forall \phi \right\} \end{aligned}$$

in which the constraint $\int \phi \mathrm{d}(\mu-\nu) - \int \nabla\phi\cdot w\mathrm{d}x = 0$, for all ϕ is exactly $\mathrm{div}(w)\mathrm{d}x = \mathrm{d}(\mu-\nu)(x)$ for $w \in C_c^\infty(\mathbb{R}^n;\mathbb{R}^n)$. This motivates us consider the following so called Beckmann's problem.

Beckmann's problem

Problem. Consider the minimization problem

$$\inf_w \left\{ \int |w(x)| \mathrm{d}x : w : \mathbb{R}^n \to \mathbb{R}^n, \operatorname{div} w = \mu - \nu \right\} \tag{B}$$

where the divergence $\operatorname{div} w$ is understood in the following sense

$$\int \phi(x) \operatorname{div} w(x) \mathrm{d}x = -\int \nabla \phi(x) \cdot w(x) \mathrm{d}x$$

for all $\phi \in C_c^\infty(\mathbb{R}^n)$. This problem is called Beckmann's minimization, denoted (B).

Our previous discussions are justified by the following theorem:

Theorem 4.14. *Beckmann's problem admits a minimizer. Moreover, its minimal value is equal to that of the Kantorovich problem with cost $c(x,y) = |x-y|$, i.e.,*

$$\min_w \left\{ \int |w(x)| \mathrm{d}x : \operatorname{div} w = \mu - \nu \right\} = \min_{\pi \in \Gamma(\mu,\nu)} \int |x-y| \mathrm{d}\pi(x,y).$$

Proof. First, the inequality $\min(\mathrm{K}) \le \min(\mathrm{B})$ is guaranteed by weak duality, see (4.37) and (4.38). We need only prove the reverse inequality $\min(\mathrm{B}) \le \min(\mathrm{K})$. It is sufficient to construct a solution w to the divergence equation from an optimal transport plan π such that $\int |w| \mathrm{d}x \le \int |x-y| \mathrm{d}\pi$. We provide only a formal proof. Consider the following linear operator

$$L(\xi) = \int_{\mathbb{R}^n \times \mathbb{R}^n} \int_0^1 \gamma_{xy}'(t) \cdot \xi(\gamma_{xy}(t)) \mathrm{d}t \mathrm{d}\pi(x,y)$$

on $C_0(\mathbb{R}^n)$, where $\gamma_{xy}(t) = (1-t)x + ty$. Invoking Riesz representation theorem, there exists a (vector) measure w_π, such that

$$L(\xi) = \int \xi(x) \cdot w_\pi(\mathrm{d}x).$$

To verify that w_π is indeed a solution to the divergence equation, we need to show

$$L(-\nabla \phi) = \int \phi \mathrm{d}(\mu - \nu)$$

for $\phi \in C_c^\infty(\mathbb{R}^n)$. Now by definition of $L(\cdot)$,

$$\begin{aligned} \int -\nabla \phi(x) \cdot w_\pi(\mathrm{d}x) &= \int_{\mathbb{R}^n \times \mathbb{R}^n} \int_0^1 \frac{d\phi(\gamma_{xy}(t))}{dt} \mathrm{d}t \mathrm{d}\pi(x,y) \\ &= \int_{\mathbb{R}^n \times \mathbb{R}^n} [\phi(y) - \phi(x)] \mathrm{d}\pi(x,y) \\ &= \int \phi \mathrm{d}(\mu - \nu). \end{aligned}$$

as desired. Next, we show □

Figure 4.11: From left to right: input, target and output

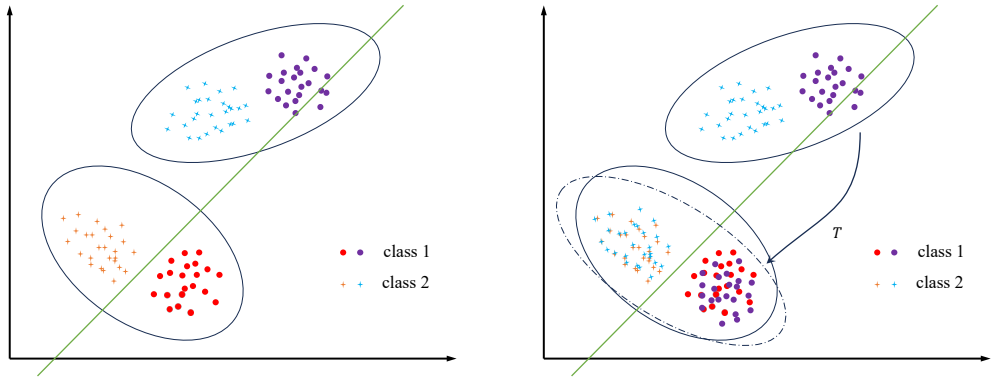

Figure 4.12: Domain adaptation

4.4.2 Image processing

Color transfer

Domain adaptation

Image interpolation

4.4.3 Control and optimal transport

Fluid dynamics viewpoint

Optimal steering

4.5 Numerical methods

Continuous methods: Brenier-Benamou formula

Discretization: Entropy regularization and matrix scaling

CHAPTER 5

APPENDIX

ODE

The solution $\phi(t;0,x)$ with initial condition $x(0)=x$ of the ODE

$$\dot{x} = f(t,x)$$

satisfies the semigroup property

$$\phi(t;s,\phi(s,0,x)) = \phi(t;0,x).$$

Proof. Let

$$\varphi(t,s) = \phi(t;s,\phi(s;0,x)), \quad t \geq s$$

We have to show

$$\varphi(t,s) = \varphi(t,0).$$

It suffices to show that

$$\frac{\partial \varphi(t,s)}{\partial s} = 0, \quad \forall s \leq t.$$

We calculate

$$\varphi(t,s) = \phi(s;0,x) + \int_s^t f(r,\varphi(r,s))\,dr$$

Then

$$\begin{aligned}
\frac{\partial \varphi(t,s)}{\partial s} &= f(s,\varphi(s,s)) - f(s,\varphi(s,s)) + \int_s^t \frac{\partial f}{\partial x}(r,\varphi(r,s))\frac{\partial \varphi(r,s)}{\partial s}\,dr \\
&= \int_s^t \frac{\partial f}{\partial x}(r,\varphi(r,s))\frac{\partial \varphi(r,s)}{\partial s}\,dr \\
\frac{d}{dt}\frac{\partial \varphi(t,s)}{\partial s} &= \frac{\partial f}{\partial x}(t,\varphi(t,s))\frac{\partial \varphi(t,s)}{\partial s}, \quad \left.\frac{\partial \varphi(t,s)}{\partial s}\right|_{t=s} = 0
\end{aligned}$$

Hence $\frac{\partial \varphi(t,s)}{\partial s} = 0$ for all $s \leq t$. □

Gaussian vectors

We gather some freqently used properties of Gaussian variables.

- Let x and y be independent Gaussian variables

  $$x \sim N(\mu_x, \Sigma_x), \quad y \sim N(\mu_y, \Sigma_y)$$

  then

  $$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \\ & \Sigma_y \end{bmatrix} \right)$$

- If $x \sim N(\mu, \Sigma)$, let $y = Ax$, then

  $$y \sim N(A\mu, A\Sigma A^T)$$

- Suppose x and y are jointly Gaussian

  $$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix} \right)$$

  then

  $$x|y \sim N\left(\mu + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y),\ \Sigma_x - \Sigma_{xy}\Sigma_y^{-1}\Sigma_{xy}^T\right)$$

  Hence

  $$E[x|y] = \mu + \Sigma_{xy}\Sigma_y^{-1}(y - \mu_y)$$

  Furthermore, y and $x - E[x|y]$ are independent.

BIBLIOGRAPHY

[1] Dimitri P. Bertsekas. Monotone mappings with application in dynamic programming. *SIAM Journal on Control and Optimization*, 15(3):438–464, 1977. 1.1.4

[2] Dimitri P Bertsekas. Value and policy iterations in optimal control and adaptive dynamic programming. *IEEE transactions on neural networks and learning systems*, 28(3):500–509, 2015. 1, 1.1.4, 1.1.4

[3] A. M. Bloch. *Nonholonomic Mechanics and Control*. Springer New York, 2003. 2.4.2

[4] Vladimir Grigorevich Boltyanskii. The method of tents in the theory of extremal problems. *Russian Mathematical Surveys*, 30(3):1, 1975. 2.3, 2.3.1

[5] Ugo Boscain and Mario Sigalotti. Introduction to controllability of nonlinear systems. *Contemporary Research in Elliptic PDEs and Related Topics*, pages 203–219, 2019. 2.2.7

[6] Alberto Bressan and Benedetto Piccoli. *Introduction to the mathematical theory of control*, volume 1. American institute of mathematical sciences Springfield, 2007. 1.2.4, 1.2.4

[7] Timothy Bretl and Zoe McCarthy. Quasi-static manipulation of a kirchhoff elastic rod based on a geometric analysis of equilibrium configurations. *The International Journal of Robotics Research*, 33(1):48–68, 2014. 2.2.2

[8] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999. 4.1.5

[9] Anthony T Fuller. Study of an optimum non-linear control system. *International Journal of Electronics*, 15(1):63–71, 1963. 2.6

[10] Gary Hewer. An iterative technique for the computation of the steady state gains for the discrete optimal regulator. *IEEE Transactions on Automatic Control*, 16(4):382–384, 1971. 1.1.4, 1.1

[11] Jean-François Le Gall et al. *Brownian motion, martingales, and stochastic calculus*, volume 274. Springer, 2016. 3.1.1

[12] Michael Margaliot and Daniel Liberzon. Lie-algebraic stability conditions for nonlinear switched systems and differential inclusions. *Systems & control letters*, 55(1):8–16, 2006. 2.2.2

[13] Jerrold E Marsden et al. *Lectures on mechanics*, volume 174. Cambridge University Press, 1992. 2.4.1

[14] John Till and D Caleb Rucker. Elastic stability of cosserat rods and parallel continuum robots. *IEEE Transactions on Robotics*, 33(3):718–733, 2017. 2.2.2