Course Notes for *Optimization*

Fei Li*

*Email: fei.li.best@gmail.com.

Contents

1 Banach Spaces

1.1 Basic Definitions and Properties

Definition 1.1 (Normed Vector Space)**.** A *norm* $\|\cdot\| : V \to \mathbb{R}$ defined on a vector space V is a real-valued function such that

(1) $\|v\| \geq 0$ for all $v \in V$ with $\|v\| = 0$ if and only if $v = 0$.

(2) $\|\alpha v\| = |\alpha|\|v\|$ for any scalar α.

(3) $\|u + v\| \leq \|u\| + \|v\|$ for all $u, v \in V$.

A *normed vector space* is a pair $(V, \|\cdot\|)$ where V is a vector space and $\|\cdot\|$ is a norm defined on V.

Sometimes when the norm is clear from the context we will simply say V is a normed vector space.

Definition 1.2 (Cauchy Sequence)**.** A sequence $\{x_i\}_{i=1}^{\infty}$ in a normed vector space $(X, \|\cdot\|)$ is called a *Cauchy sequence* if for every real number $\epsilon > 0$ there is a positive integer N such that

$$\|x_m - x_n\| < \epsilon, \qquad \forall m, n \geq N.$$

X is called *complete* if every Cauchy sequence in X converges.

Note that

- A convergent sequence is a Cauchy sequence. If $x_n \to x$, then for every $\epsilon > 0$ there is $N > 0$ such that $\|x_n - x\| < \epsilon/2$ for all $n \geq N$. Hence by the triangle inequality for $m, n \geq N$

$$\|x_m - x_n\| \leq \|x_m - x\| + \|x - x_n\| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

  The converse is not true in general. Take $A = (0, 1]$ and the sequence $\{1/n\}_{n=1}^{\infty}$ for example. It converges to 0, which does not belong to the set A.

- A Cauchy sequence must be bounded: there is an $N \geq 1$ such that $\|x_m - x_n\| < 1$ for all $m, n \geq N$. If we take $B = \{x_n : n \geq N\}$ then $\operatorname{diam} B \leq 1$ so that B is bounded. On the other hand the set of finite points $A = \{x_1, \ldots, x_N\}$ is bounded, so that $\{x_n\}_{n=1}^{\infty} = A \cup B$ is bounded.

Definition 1.3. A ***Banach space*** is a complete normed vector space.

Proposition 1.4. A closed subspace of a Banach space is again a Banach space.

Proof. Let M be a closed subspace of a Banach space X, and let $\{x_n\} \subset M \subseteq X$ be a Cauchy sequence in M. Since X is complete, the sequence converges to a point $x \in X$. Since M is closed, $x \in M$. □

Let $\mathcal{B}[a, b]$ be the space of all bounded real-valued functions on $[a, b]$. Define a norm $\|\cdot\|_\infty$ on $\mathcal{B}[a, b]$ by

$$\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|. \tag{1}$$

Then it is easy to see that $\|\cdot\|_\infty$ is a norm on $\mathcal{B}[a, b]$. It is called the *supremum norm* or *uniform norm* on $\mathcal{B}[a, b]$. Recall that for a sequence of real-valued functions $\{f_n\}_{n=1}^{\infty}$ defined on $[a, b]$, we say $\{f_n\}_{n=1}^{\infty}$ *converges uniformly* to f if for every $\epsilon > 0$, there is $N > 0$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n \geq N$ and *all* $x \in [a, b]$. It is easy to see that $\{f_n\}_{n=1}^{\infty}$ converges uniformly to f if and only if $\lim_{n\to\infty} \|f_n - f\|_\infty = 0$, i.e. f_n converges to f in $\mathcal{B}[a, b]$ with respect to the uniform norm. This justifies the name "uniform norm". We next show that $\mathcal{B}[a, b]$ equipped with the uniform norm is a Banach space.

Proposition 1.5. $(\mathcal{B}[a,b], \|\cdot\|_\infty)$ is a Banach space.

Proof. We need to show that every Cauchy sequence in $\mathcal{B}[a,b]$ converges with respect to $\|\cdot\|_\infty$. So let $\{f_n\}_{n=1}^\infty$ be a Cauchy sequence in $\mathcal{B}[a,b]$. This means that given $\epsilon > 0$, there is $N > 0$ such that $\|f_n - f_m\|_\infty < \epsilon$ for all $n, m > N$. Hence for all $x \in [a,b]$,

$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty < \epsilon.$$

This implies that for each $x \in [a,b]$, the sequence $\{f_n(x)\}_{n=1}^\infty$ is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, the sequence is convergent. Define

$$f(x) := \lim_{n\to\infty} f_n(x)$$

for each $x \in [a,b]$. We want to show that $\{f_n\}_{n=1}^\infty$ converges to f.

First of all, is $f \in \mathcal{B}[a,b]$ at all? The Cauchy sequence $\{f_n\}_{n=1}^\infty$ is bounded, so that there exists $M > 0$ for which $\|f_n\|_\infty \leq M$ for all $n \in \mathbb{N}$. By the definition of the uniform norm, we have $|f_n(x)| \leq \|f_n\|_\infty \leq M$ for all $x \in [a,b]$ and all $x \in \mathbb{N}$. Taking $n \to \infty$ we find $|f(x)| = \lim_{n\to\infty} |f_n(x)| \leq M$, so that f is indeed bounded.

For any $\epsilon > 0$ and for each $x \in X$, there is $N > 0$ so that for all $n \geq N$,

$$|f_n(x) - f(x)| = \lim_{m\to\infty} |f_n(x) - f_m(x)| \leq \epsilon.$$

This implies that for all $n \geq N$, $\|f_n - f\|_\infty \leq \epsilon$. This proves the convergence of the sequence $\{f_n\}_{n=1}^\infty$ to f. □

Let $C^0[a,b]$ denote the space of continuous real-valued functions on $[a,b]$. If we equip it with the uniform norm Eq. (1), then thanks to Proposition 1.6, it is a closed subspace of $\mathcal{B}[a,b]$. Thus according to Proposition 1.4 $(C^0[a,b], \|\cdot\|_\infty)$ is a Banach space.

Proposition 1.6. Let $\{f_n\}_{n=1}^\infty$ be a sequence in $C^0[a,b]$. If $\|f_n - f\|_\infty \to 0$ as $n \to \infty$ (i.e. $\{f_n\}_{n=1}^\infty$ converges to f uniformly), then f is also continuous.

Proof. Fix $\epsilon > 0$ and $x' \in [a,b]$. To prove that f is continuous at x', we need to find a $\delta > 0$ such that $|x - x'| < \delta \Rightarrow |f(x) - f(x')| < \epsilon$. The idea is to approximate $|f(x) - f(x')|$ by the three segments in the triangle inequality below:

$$|f(x) - f(x')| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(x')| + |f_n(x') - f(x')|.$$

Each of the three terms on the right side can be made small:

- we can choose $N > 0$ such that $|f(x) - f_n(x)| < \epsilon/3$ and $|f_n(x') - f(x')| < \epsilon/3$ for all $n \geq N$, by uniform convergence;
- by continuity of f_n, we can find some $\delta > 0$ such that $|x - x'| < \delta$ implies that $|f_n(x) - f_n(x')| < \epsilon/3$.

Then for this $|x - x'| < \delta$ we have

$$|f(x) - f(x')| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

This proves that f is indeed continuous. □

Besides the supremum norm, we can also define other norms on $C^0[a,b]$, in particular the p-norms, which are generalizations of the p-norms defined on finite dimensional Euclidean space.

Proposition 1.7. $\|\cdot\|_1 : C^0[a,b] \to \mathbb{R}$ defined by

$$\|f\|_1 = \int_a^b |f(x)|dx$$

is a norm on $C^0[a,b]$.

Proof. We need to verify the three properties in Definition 1.1.

(1) We need to prove that for $f \in C^0[a,b]$, $\int_a^b |f(x)|dx \equiv 0 \Rightarrow |f| \equiv 0$ on $[a,b]$ (so that $f \equiv 0$ on $[a,b]$). Suppose to the contrary, $|f(x')| \neq 0$ for some $x' \in [a,b]$, say $f(x') > 0$. Then since f is continuous on $[a,b]$, $f(x) > 0$ for all x sufficiently close to x'. To be rigorous choose $0 < \epsilon < f(x')$. By continuity there exists $\delta > 0$ such that for all $x \in [x' - \delta, x + \delta]$ we have $|f(x) - f(x')| < \epsilon$, so that $f(x) > f(x') - \epsilon > 0$. Now

$$\int_a^b |f(x)|dx \geq \int_{x'-\delta}^{x'+\delta} |f(x)|dx > 0,$$

a contradiction.

(2) Obvious.

(3) For the triangle inequality, since $|f(x) + g(x)| \leq |f(x)| + |g(x)|$, we have

$$\int_a^b |f(x) + g(x)|dx \leq \int_a^b (|f(x)| + |g(x)|)dx = \int_a^b |f(x)|dx + \int_a^b |g(x)|dx.$$

□

For $1 \leq p < \infty$, the p-norm $\|\cdot\|_p$ on $C^0[a,b]$ is defined by

$$\|f\|_p = \left(\int_a^b |f(x)|^p dx\right)^{1/p}.$$

The triangle inequality for the p-norm

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

is also called the Minkowski inequality. $C^0[a,b]$ equipped with the p-norm is not complete. Take $p = 1$ and $C^0[0,1]$ for example, and consider the sequence of functions in Fig. 1a, i.e.

$$f_n(x) = \begin{cases} 0 & \text{if } x \in [0, 1/2]; \\ \text{linear} & \text{if } x \in [1/2, a_n]; \\ 1 & \text{if } x \in [a_n, 1] \end{cases}$$

with $a_n = 1/2 + 1/n$. The distance between two elements f_n and f_m under the 1-norm is the area between them, as shown in Fig. 1b. The area can be made arbitrarily small for large n or m, so $\{f_n\}$ is a Cauchy sequence. However, the sequence converges to the function

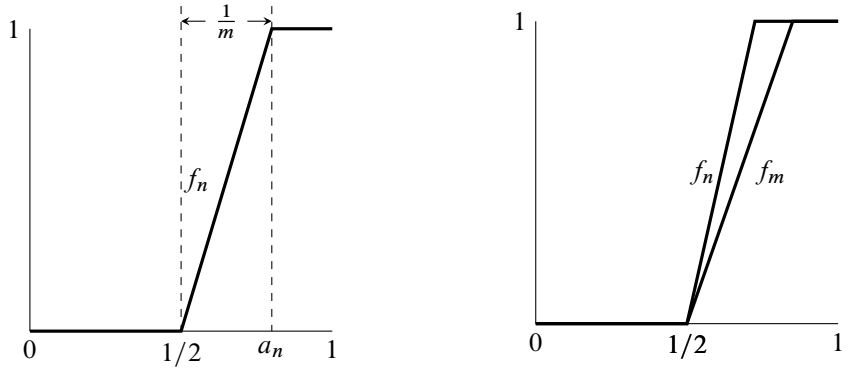
$$f(x) = \begin{cases} 0 & \text{if } x \in [0, 1/2] \\ 1 & \text{if } x \in (1/2, 1] \end{cases},$$

which is not continuous and does not belong to $C^0[0,1]$.

Definition 1.8. A linear map $L : X \to Y$ between two Banach spaces is called *bounded* if there is a constant $K > 0$ such that

$$\|Lx\| \leq K\|x\| \qquad \text{for all } x \in X.$$

Theorem 1.9. A linear map $L : X \to Y$ between two Banach spaces is bounded if and only if it is continuous.



(a) The sequence $\{f_n\}$.

(b) $\|f_n - f_m\|_1$ can be made small.

Figure 1: An example showing $C^0[0, 1]$ with 1-norm is not complete.

Proof. If L is bounded, then for $x_n \to x$ we have $\|Lx_n - Lx\| = \|L(x_n - x)\| \leq K\|x_n - x\|$ converges to 0, so that it is continuous. If L is continuous, then in particular it is continuous at 0, and so for *any* sequence $\{x_n\}_{n=1}^{\infty} \subset X$ for which $x_n \to 0$ in X, we have $Lx_n \to 0$ in Y. Suppose L is not bounded, then

$$\frac{\|Lx_n\|}{\|x_n\|} \to \infty$$
$$\Downarrow$$
$$\left\| L\frac{x_n}{\|x_n\|} \right\| := a_n \to \infty$$
$$\Downarrow$$
$$\left\| L\frac{x_n}{\|x_n\| \cdot \sqrt{a_n}} \right\| = \left\| L\frac{x_n}{\|x_n\|} \right\| \Big/ \sqrt{a_n} = \sqrt{a_n} \to \infty.$$

We arrived at a contradiction, since $y_n = x_n / \left(\|x_n\| \cdot \sqrt{a_n}\right)$ tends to 0. □

The space of linear continuous maps between X and Y is denoted by $\mathcal{L}(X, Y)$. We can define a norm on it by $\|L\| := \sup_{\|x\|=1} \|Lx\|$. The space $\mathcal{L}(X, Y)$ equipped with this norm is again a Banach space.

Not all linear maps are bounded. An example is the map $L : \ell^2 \to \mathbb{R}^{\infty}$ with $Le_k = ke_k$, where $\{e_k\}$ is an orthonormal basis in ℓ^2. We can imagine that it is an infinite matrix with $1, 2, 3, 4, \ldots$ on the diagonal...

1.2 Banach Fixed Point Theorem

Definition 1.10 (Contraction)**.** Let $(X, \|\cdot\|)$ be a Banach space. A mapping $T : X \to X$ is called a *contraction* on X if there is $\alpha \in (0, 1)$ such that

$$\|Tx - Ty\| \leq \alpha\|x - y\|$$

for all $x, y \in X$.

Theorem 1.11 (Banach Fixed Point Theorem)**.** Let $T : X \to X$ be a contraction on a Banach space X. Then T has a unique fixed point.

Proof. First, for a contraction T, its fixed point is necessarily unique. For suppose $x = Tx$ and $x' = Tx'$ are two fixed points of T. Then

$$\|x - x'\| = \|Tx - Tx'\| \leq \alpha\|x - x'\|.$$

Since $\alpha < 1$, we have $\|x - x'\| = 0$ and consequently $x = x'$. Pick an arbitrary point $x_0 \in X$ and define a sequence $\{x_n\}$ by

$$\begin{aligned} x_1 &= Tx_0; \\ x_2 &= Tx_1 = T^2x_0; \\ x_3 &= Tx_2 = T^3x_0; \\ &\vdots \\ x_n &= Tx_{n-1} = T^nx_0; \\ &\vdots \end{aligned}$$

From Definition 1.10, a contraction is continuous. If our $\{x_n\}$ converges to some $x \in X$, then from $x_n = Tx_{n-1}$, we have

$$x = \lim_{n\to\infty} x_n = \lim_{n\to\infty} Tx_{n-1} = T\left(\lim_{n\to\infty} x_{n-1}\right) = Tx,$$

so that $x \in X$ will be a fixed point of T. Our next task is to show that $\{x_n\}$ is Cauchy, so by completeness of X the convergence is guaranteed, and from which the existence of fixed point will be established. Now,

$$\begin{aligned} \|x_{m+1} - x_m\| &= \|Tx_m - Tx_{m-1}\| \\ &\le \alpha\|x_m - x_{m-1}\| \\ &= \alpha\|Tx_{m-1} - Tx_{m-2}\| \\ &\le \alpha^2\|x_{m-1} - x_{m-2}\| \\ &\vdots \\ &\le \alpha^m\|x_1 - x_0\|. \end{aligned}$$

Then by the triangle inequality, we have

$$\begin{aligned} \|x_m - x_n\| &\le \|x_m - x_{m+1}\| + \|x_{m+1} - x_{m+2}\| + \cdots + \|x_{n-1} - x_n\| \\ &\le (\alpha^m + \alpha^{m-1} + \cdots + \alpha^{n-1})\|x_1 - x_0\| \\ &= \frac{\alpha^m}{1-\alpha}(1 - \alpha^{n-m})\|x_1 - x_0\| \\ &\le \frac{\alpha^m}{1-\alpha}\|x_1 - x_0\| \to 0 \quad \text{as } m \to \infty. \end{aligned}$$

This proves $\{x_n\}$ is Cauchy. Consequently, $\{x_n\}$ converges to some $x \in X$ and this point is a fixed point of X. □

1.2.1 Counter-examples of the Banach Fixed Point Theorem

Note the Banach fixed point theorem remains true if we restrict to $T : M \to M$ for some closed subset M of X.

(1) (M not closed) Consider $M = (0, 1)$ and $f(x) = x/2$. In this case fixed point does not exist.

(2) (T not contraction) Consider $T : [0, 1] \to [0, 1]$ with $T(x) = x^2$. We have $T(0) = 0$ and $T(1) = 1$, so that in this case the fixed point is not unique.

(3) (T not contraction) For $T : \mathbb{R} \to \mathbb{R}$, T admits a fixed point if the graph of T intersects with the line $y = x$. Consider

$$Tx = \begin{cases} x + \frac{1}{x+1} & \text{if } x \ge 0 \\ 1 & \text{if } x \le 0 \end{cases}$$

For $x \geq 0$ the graph asymptotically approaches the line $y = x$ as $x \to \infty$, but it never intersects with $y = x$. The derivative of T is

$$T'(x) = \begin{cases} 1 - \frac{1}{(x+1)^2} & \text{if } x \geq 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

so $|T'(x)| < 1$ for all $x \in \mathbb{R}$, but $|T'(x)| \to 1$ as $x \to \infty$. We have $|Tx - Ty| < |x - y|$ for any $x, y \in \mathbb{R}$, but there does not exist a constant $\alpha \in (0, 1)$ such that $|Tx - Ty| \leq \alpha|x - y|$ for *any* $x, y \in \mathbb{R}$, since the derivative tends to 1. We see that the condition "$|Tx - Ty| < |x - y|$ for any $x, y \in \mathbb{R}$" is not enough.

(4) (Contraction and stability) Consider the affine transformation $T : \mathbb{R} \to \mathbb{R}$ where $T(x) = ax + b$. If $a \neq 1$ then we have a unique fixed point. However, the situation is different for $a > 1$ and $a < 1$. For $a < 1$, the fixed point is *stable*: we can start with any point x_0 on the real line and use the iterative procedure of the Banach fixed point theorem to converge to the fixed point. For $a > 1$ the fixed point is *unstable*: apply the iterative procedure for any starting point x_0 not equal to the fixed point itself will lead us far away to infinity.

Another example is $T(x) = x^2$. If we start with $|x_0| > 1$ then the sequence $\{x_n\}$ ($x_n = Tx_{n-1}$) will diverge to infinity; if we start with $|x_0| = \pm 1$ then we will stay at the fixed point, but if we start with any $|x_0| < 1$ then the sequence $\{x_n\}$ will converge to 0. We see that the fixed point $x = 0$ is stable, while the point $x = 1$ is unstable. Indeed, if we take $M = [-1/2 + \epsilon, 1/2 + \epsilon]$ then $\max |T'(x)| = |\frac{1}{2} - \epsilon| \cdot 2 = 1 - 2\epsilon < 1$, so T restricted to M is a contraction.

More examples:

- $x_{n+1} = \sqrt{x_n + 2}$,
- $x_{n+1} = e^{-x_n^2}$,
- $x_{n+1} = 2 \arctan x_n$,
- $x_{n+1} = (1/2) \arctan x_n$,
- $x_{n+1} = 2 \log(1 + x_n)$.

(5) For affine transformation $Tx = Ax + b$, we want $Ax + b = x \Rightarrow x = (I - A)^{-1}b$. We then want

$$\det(I - A) \neq 0 \iff \lambda = 1 \text{ not an eigenvalue of } A.$$

For T to be contraction we want $\max |\lambda_i| < 1$ where λ_is are eigenvalues of A, or $\|A\| \leq 1$ in general.

1.2.2 Application of the Banach Fixed Point Theorem

The standard application of the Banach fixed point theorem is Picard's existence theorem.

Theorem 1.12 (Picard's existence and uniqueness theorem)**.** Let $F(t, y)$ be a continuous function of (t, y) on a strip

$$\mathcal{S} = \{(t, y) : a \leq t \leq b, -\infty < y < \infty\}$$

and suppose further it is Lipschitz continuous with respect to the second argument, i.e. there is a constant $K > 0$ such that

$$|F(t, y_1) - F(t, y_2)| \leq K|y_1 - y_2| \qquad \forall (t, y_1), (t, y_2) \in \mathcal{S}.$$

If (t_0, y_0) is an interior point of $\mathcal{S}$, then there exists a unique solution $y(t)$ of

$$y' = F(t, y), \qquad y(t_0) = y_0, \tag{2}$$

on some interval $[a', b']$ with $t_0 \in [a', b'] \subset [a, b]$.

Proof. First note that y is a solution to Eq. (2) if and only if it is a solution to

$$y(t) = y_0 + \int_{t_0}^{t} F(u, y(u))du. \tag{3}$$

Define the operator T as

$$(Ty)(t) = y_0 + \int_{t_0}^{t} F(u, y(u))du.$$

y is a solution to Eq. (3) iff $Ty = y$, i.e. y is a fixed point of T. Our aim is to choose an appropriate interval $[a', b']$, and prove that $T : C^1[a', b'] \to C^1[a', b']$ is a contraction (with ∞-norm), so that we can apply Banach fixed point theorem to assert that there is a unique fixed point of T:

$$\begin{aligned}
\|Ty_1 - Ty_2\| &= \left\| \int_{t_0}^{t} F(u, y_1(u))du - \int_{t_0}^{t} F(u, y_2(u))du \right\| \\
&= \left| \int_{t_0}^{t^*} \left(F(u, y_1(u)) - F(u, y_2(u))\right) du \right| \\
&\leq \int_{t_0}^{t^*} |F(u, y_1(u)) - F(u, y_2(u))| \, du \\
&\leq \int_{t_0}^{t^*} K \, |y_1(u) - y_2(u)| \, du \\
&\leq \int_{t_0}^{t^*} K \, \|y_1 - y_2\| \, du \\
&= (t^* - t_0) K \cdot \|y_1 - y_2\|
\end{aligned}$$

where $t^* = \arg\max_{t \in [a,b]} |(Ty_1)(t) - (Ty_2)(t)|$. To make T a contraction ,we want

$$(t^* - t_0)K < 1,$$

i.e. $t^* < t_0 + 1/K$. So if we choose some $0 < \delta < 1/K$ and let $a' = t_0 - \delta, b' = t_0 + \delta$ we are done.

□

1.3 Compact Operators

Definition 1.13. A set $A \subset X$ is *relatively* compact if its closure in X is compact.

Definition 1.14. A linear operator $L : X \to Y$ is called *compact* if one of the following equivalent statements holds:

- ◇ the image of the unit ball of X under T is relatively compact in Y;
- ◇ the image of any bounded subset of X under T is relatively compact in Y;
- ◇ there exists a neighborhood U of 0 in X and a compact subset $V \subset Y$ such that $T(U) \subset V$;
- ◇ for any bounded sequence $(x_n)_{n \in \mathbb{N}}$ in X, the sequence $(Tx_n)_{n \in \mathbb{N}}$ contains a converging subsequence.

An example of compact operator is $T : C^0[a, b] \to C^0[a, b]$ with

$$(Ty)(t) = \int_{a}^{t} f(u, y(u))du.$$

1.4 Schauder's Fixed Point Theorem

Schauder's fixed point theorem is the infinite dimensional version of the Brouwer fixed point theorem:

Theorem 1.15 (Schauder's Fixed Point Theorem)**.** Every continuous function $T : K \to K$ from a nonempty, compact and convex subset K of a *Banach space* X to itself has a fixed point.[1]

Proof. The idea of the proof is to approximate $T : K \to K$ by some finite dimensional maps $T_n : K_n \to K_n$ and make use of the Brouwer's fixed point theorem (Theorem 1.16).

Since K is compact, for every $\epsilon > 0$ there are N points $\{x_1, \dots, x_N\} \subset K$ such that $K \subset \bigcup_{i=1}^N B(x_i, \epsilon)$. Let $\text{conv}(\{x_1, \dots, x_N\})$ denote the convex hull of $\{x_1, \dots, x_N\}$. Since it is formed by finite number of points, it is homeomorphic to some compact and convex subset in some finite Euclidean space. We can then define a projection $P_N : K \to \text{conv}(\{x_1, \dots, x_N\})$ such that $\|P_N(x) - x\| < \epsilon$ for any $x \in K$. For example, we can take a partition of unity $\{\psi_i\}_{i=1}^N$ of K subordinate to $\{B(x_i, \epsilon)\}_{i=1}^N$ and define P_N as

$$P_N(x) = \frac{\psi_1(x)}{\psi(x)} x_1 + \frac{\psi_2(x)}{\psi(x)} x_2 + \cdots + \frac{\psi_N(x)}{\psi(x)} x_N$$

where $\psi = \sum_{i=1}^N \psi_i$. Indeed,

$$\begin{aligned} \|P_N(x) - x\| &= \left\| \sum_{i=1}^N \frac{\psi_i(x)}{\psi(x)} x_i - \sum_{i=1}^N \frac{\psi_i(x)}{\psi(x)} x \right\| \\ &= \left\| \sum_{i=1}^N \frac{\psi_i(x)}{\psi(x)} (x_i - x) \right\| \\ &\le \sum_{i=1}^N \frac{\psi_i(x)}{\psi(x)} \|x_i - x\| \\ &< \sum_{i=1}^N \frac{\psi_i(x)}{\psi(x)} \epsilon = \epsilon \end{aligned}$$

where in the last inequality we used the fact that $\psi_i(x) = 0$ for $x \notin B(x_i, \epsilon)$. Now we are able to define $T_N : \text{conv}(\{x_1, \dots, x_N\}) \to \text{conv}(\{x_1, \dots, x_N\})$ as $T_N := P_N \circ T|_{\text{conv}(\{x_1, \dots, x_N\})}$. By Theorem 1.16 the function has a fixed point x_N^*.

Let $\{x_{N_k}^*\}$ be a sequence of fixed points in K defined above. By compactness of K it has a convergent subsequence. Let us denote this subsequence by $\{x_m^*\}$ and its limit by x^*, i.e. $\lim_{m\to\infty} x_m^* = x^*$.

We would like to argue that the sequence $\{x_m^*\}$ also converges to $T(x^*)$, i.e. $\lim_{m\to\infty} x_m^* = T(x^*)$. Then since the Banach space X is Hausdorff we would have $T(x^*) = x^*$. To see this, we use the triangle inequality

$$\|x_m^* - T(x^*)\| \le \|x_m^* - T(x_m^*)\| + \|T(x_m^*) - T(x^*)\|.$$

The first term on the right hand side goes to zero as $m \to \infty$ since $x_m^* = T_m(x_m^*) = P_m \circ T(x_m^*)$ and we have $\|P_m \circ T(x_m^*) - T(x_m^*)\| < \epsilon \to 0$ by construction of P_m. The second term goes to zero as $x_m^* \to x^*$ since T is continuous. □

Theorem 1.16 (Brouwer's fixed point theorem)**.** On any nonempty compact and convex subset $K \subseteq \mathbb{R}^n$, every continuous function $f : K \to K$ has a fixed point.

[1] Another statement is that if $T : K \to K$ is a compact operator where K is nonempty, convex and closed, then T has a fixed point.

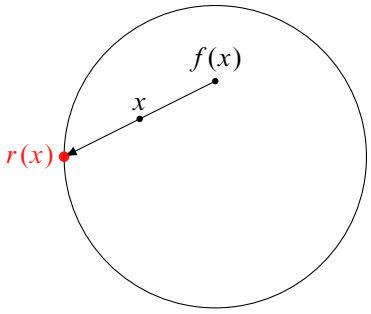


Figure 2: Construction of retraction in the proof of Brouwer's fixed point theorem.

Proof. We prove the theorem for $K = D^n$, the closed unit ball, since every nonempty compact and convex subset in $\mathbb{R}^n$ is homeomorphic to D^m for some suitable dimension $m \leq n$. We discuss the proof for $n = 1$, $n = 2$, and $n > 2$ separately below.

$n = 1$ In this case let's take $K = [0, 1]$. The proof is simple: if for $f : [0, 1] \to [0, 1]$ we have $f(0) = 0$ or $f(1) = 1$, then we are done. Otherwise, if $f(0) \neq 0$ and $f(1) \neq 1$ then we must have $f(0) > 0$ and $f(1) < 1$. Then for the function $g(x) = f(x) - x$ defined on $[0, 1]$ we have $g(0) > 0$ and $g(1) < 0$, so by the intermediate value theorem in calculus there must be some point $x^* \in [0, 1]$ such that $g(x^*) = 0$, i.e. $f(x^*) = x^*$.

$n = 2$ Suppose $f : D^2 \to D^2$ does not have a fixed point. Then for every $x \in D^2$ we have $f(x) \neq x$, so we for every $x \in D^2$ we can draw a ray starting at $f(x)$ and passing through x. Call the intersection of the ray with the boundary $r(x)$ (See Fig. 2 for illustration). We constructed a function $r : D^2 \to S^1$ that is a *retraction* of D^2 to S^1:

- $r(x) = x$ for $x \in S^1$, i.e. r restricted to $S^1 \subset D^2$ is the constant map;
- r is continuous: a small perturbation in x would result in a small perturbation in $f(x)$ and consequently $r(x)$.

This implies that $r \circ \iota = \mathbb{1}$ on S^1, where $\iota : S^1 \to D^2$ is the inclusion map. The induced homomorphisms between fundamental groups of S^1 and D^2 would have the relation $(r \circ \iota)_* = r_* \iota_* = \mathbb{1}_*$:

$$\pi_1(S^1) \xrightarrow{r_*} \pi_1(D^2) \xrightarrow{\iota_*} \pi_1(S^1)$$

which is impossible, since $\pi_1(S^1) \simeq \mathbb{Z}$ but $\pi_1(D^2)$ is trivial.

The fundamental group of a space is the group of equivalent classes of loops in the space, where two loops are equivalent if they are homotopic. It records information about "holes" in the space. Since D^2 is convex, you can continuously shrink every loop in D^2 to a point on the loop, but you cannot do so for a loop in S^1 without tearing apart the loop. So in S^1 a loop around the circle and a point on the circle is not (homotopic) equivalent.

$n > 2$ To prove for the general case, we go from homotopy to homology. If f has no fixed point, then we can construct a retraction $r : D^n \to S^{n-1}$ as before. We would then have $(r \circ \iota)_* = r_* \iota_* = \mathbb{1}_*$ for

$$H_{n-1}(S^{n-1}) \xrightarrow{r_*} H_{n-1}(D^n) \xrightarrow{\iota_*} H_{n-1}(S^{n-1})$$

which is again impossible since $H_{n-1}(S^{n-1}) \simeq \mathbb{Z}$ while $H_{n-1}(D^n)$ is trivial. □

Schauder's fixed point theorem can be used to prove the Peano's existence theorem:

Theorem 1.17 (Peano's existence theorem)**.** Let $F(t, y)$ be a continuous function of (t, y) on a rectangle

$$\mathcal{R} = \{(t, y) : a \leq t \leq b, c \leq y \leq d\}.$$

If (t_0, y_0) is an interior point of $\mathcal{R}$, then there exists a solution $y(t)$ of

$$y' = F(t, y), \qquad y(t_0) = y_0,$$

on some interval $[a', b']$ with $t_0 \in [a', b'] \subset [a, b]$.

2 Hilbert Spaces

2.1 Basic Definitions and Properties

Definition 2.1. A ***Hilbert space*** is an inner product space $(H, \langle\cdot,\cdot\rangle)$ that is also complete w.r.t. the norm induced by the inner product. The induced norm is $\|x\| = \langle x, x\rangle^{1/2}$.

Proposition 2.2 (Cauchy-Schwarz inequality)**.** Let H be a Hilbert space. For any $u, v \in H$, we have $\langle u, v\rangle \leq \|u\|\|v\|$.

Proof. For $t \in \mathbb{R}$, we have $\langle u + tv, u + tv\rangle \geq 0$ (positivity). Expand this out we have $t^2\|v\|^2 + 2t\langle u, v\rangle + \|u\|^2 \geq 0$ for any $t \in \mathbb{R}$. This is a quadratic function in t that is always larger than or equal to 0. Thus the discriminant is non-positive:

$$\Delta = 4\langle u, v\rangle^2 - 4\|u\|^2\|v\|^2 \leq 0 \quad \Rightarrow \quad \langle u, v\rangle \leq \|u\|\|v\|.$$

□

A useful property of the Hilbert space that is relevant for optimization is the following:

Proposition 2.3. Let $K \subseteq H$ be a nonempty closed and convex subset of H. Then for every $x \in H$ there exists $y^* \in K$ such that

$$\|x - y^*\| = \min_{z\in K} \|x - z\|.$$

Moreover, we have $\langle x - y^*, z - y^*\rangle \leq 0$ for every $z \in K$.

What are examples of infinite dimensional Hilbert spaces? $\mathbb{R}^\infty$ is not. $\ell^\infty = \{a \in \mathbb{R}^\infty : \|a\|_\infty < \infty\}$ is a Banach space but not a Hilbert space, but $\ell^p = \{a \in \mathbb{R}^\infty : \|a\|_p < \infty\}$ is also *not* a Hilbert space, since although there is the p-norm in it, we cannot define an inner product that can induce the p-norm (note that when $p < q$ we have $\ell^p \subset \ell^q$). We only have a Hilbert space when $p = 2$, i.e.

$$\ell^2 = \left\{a \in \mathbb{R}^\infty : \sum_{i=1}^{\infty} |a_i|^2 < \infty\right\}.$$

The inner product is defined, since if $a, b \in \ell^2$, then

$$\begin{aligned}
|\langle a, b\rangle| = \left|\sum_i a_i b_i\right| &\leq \sum_i |a_i b_i| \\
&\leq \sum_i \left(\frac{a_i^2}{2} + \frac{b_i^2}{2}\right) = \sum_i \frac{a_i^2}{2} + \sum_i \frac{b_i^2}{2} \\
&= (\|a\|^2 + \|b\|^2)/2 \\
&< \infty.
\end{aligned}$$

Another example of Hilbert space would be $L^2[a, b]$, i.e. square integrable functions defined on a closed interval, or more generally any reasonable subset Ω of $\mathbb{R}^n$. Are there other infinite dimensional Hilbert spaces? In fact, in some sense there isn't: every (infinite dimensional) separable Hilbert space is isomorphic to ℓ^2. Recall that a space is separable if it contains a countable dense subset. In reality most Hilbert spaces we encountered are separable, and we may never have to deal with in-separable spaces. The proof is straightforward once we have the following theorem:

Theorem 2.4. A Hilbert space H is separable if and only if it has a countable orthonormal basis (in the sense of a Schauder basis, i.e. there is $\{x_n\}_{n=1}^\infty$ such that every $x \in H$ can be represented uniquely as $x = \sum_{n=1}^\infty a_n x_n$ in the sense of the induced metric)

Proof. If H is separable, then we would like to have any orthonormal basis $\{e_i\}_{i\in I}$ to be countable. Assume the contrary that $\{e_i\}_{i\in I}$ is uncountable. Note that each element in $\{e_i\}_{i\in I}$ is of distance $\sqrt{2}$ apart, like in the Euclidean case: $\|e_i - e_j\|^2 = \|e_i\|^2 + \|e_j\|^2 = 2$. Thus we can have a small open ball round each point in $\{e_i\}_{i\in I}$. Those balls are disjoint, and we see that it is no longer possible for a dense set to be countable.

On the other hand, if H has a countable orthonormal basis $\{e_i\}_{i=1}^{\infty}$, then let

$$A_n = \{q_1 e_1 + \cdots + q_n e_n : q_i \in \mathbb{Q} \text{ for } i = 1, \ldots, n\}$$

and let $A = \bigcup_{n=1}^{\infty} A_n$. Then A is the countable dense subset. We can prove that every open ball around any point $x \in H$ contains an element from A by the triangular inequality. □

Then from Theorem 2.4 it is not hard to show that every (infinite dimensional) separable Hilbert space H is isomorphic to ℓ^2: pick a countable orthonormal basis $\{f_n\}_{n=1}^{\infty}$ and define $T : H \to \ell^2$ by

$$T(f) = (\langle f, f_n\rangle)_{n=1}^{\infty}.$$

One can show that T is indeed an (isometric) isomorphism. So for example $L^2[a,b]$ is isomorphic to ℓ^2.[2] The idea is that we can represent any function f in $L^2[a,b]$ by countably many functions.

For example, $\{1/\sqrt{2}, \cos x, \sin x, \cos 2x, \sin 2x, \ldots\}$ is an orthonormal basis for $L^2(-\pi,\pi)$ with inner product $\langle f, g\rangle = (1/\pi)\int_{-\pi}^{\pi} fg$. For any $u \in L^2(-\pi,\pi)$ we have

$$u(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \tag{4}$$

and

$$\|u\|^2 = \frac{1}{\pi}\int_{-\pi}^{\pi} u^2 = \frac{a_0^2}{2} + \sum_{n=1}^{\infty} \left(a_n^2 + b_n^2\right).$$

This is the Parseval's identity, which is just the infinite dimensional Pythagorean theorem. Several remarks:

(1) Odd functions are orthogonal to even functions (if u is odd and v is even, then uv is odd, so that $\int uv = 0$.) If u is odd then $a_n = 0$ for any $n \geq 1$, and if u is even then $b_n = 0$ for $n \geq 1$. In general,

$$u(x) = \frac{u(x) + u(-x)}{2} + \frac{u(x) - u(-x)}{2}$$

where the first term is even and the second term is odd.

(2) For the Heaviside function $H(x)$, since it is odd we have $a_n = 0$ and

$$\begin{aligned} b_n &= \frac{2}{\pi}\int_0^{\pi} \sin nx\, dx = \frac{2}{\pi}\left[-\frac{\cos(nx)}{n}\right]_0^{\pi} \\ &= \frac{2}{\pi}\left\{\frac{2}{1}, \frac{0}{2}, \frac{2}{3}, \frac{0}{4}, \frac{2}{5}, \frac{0}{6}, \frac{2}{7}, \cdots\right\} \end{aligned}$$

so that

$$H(x) = \frac{4}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \frac{\sin 7x}{7} + \cdots\right].$$

At $x = \pi/2$ we have

$$1 = \frac{4}{\pi}\left[\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right] \quad \text{so that} \quad \pi = 4\left[\frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right].$$

[2] The statement $L^2[a,b]$ is a separable Hilbert space is a very strong one and the proof is very involved.

(3) If u is continuous and u' is bounded, then Eq. (4) holds point-wise.

(4) For the function $u(x) = |x|$, we have $b_n = 0$ and $a_0 = \pi/2$ and

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} |x|\cos nx dx = \frac{2}{\pi}\int_0^{\pi} x\cos nx dx = \frac{2}{\pi n^2}\left[(-1)^n - 1\right]$$

So

$$u(x) = \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{\cos(2k+1)x}{(2k+1)^2}.$$

In particular setting $x = \pi$ (or $x = 0$) we get

$$\sum_{k=0}^{\infty}\frac{1}{(2k+1)^2} = \frac{\pi^2}{8}.$$

(5) Parseval's identity:

$$\frac{1}{\pi}\int_{-\pi}^{\pi} x^2 dx = \frac{\pi^2}{2} + \sum_{k=0}^{\infty}\frac{16}{\pi^2(2k+1)^4}$$

and we get

$$\sum_{k=0}^{\infty}\frac{1}{(2k+1)^4} = \frac{\pi^4}{96}.$$

(6) Next we take $u(x) = x$, for $x \in (-\pi, \pi)$. Since we require $u(x)$ to be 2π-periodic we remove the end points $x = \pi$ and $x = -\pi$. It is odd so that $a_n = 0$ and

$$b_n = \frac{2}{\pi}\int_0^{\pi} x\sin nx dx = \frac{2}{n}(-1)^{n+1}.$$

We obtain

$$u(x) = 2\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin nx.$$

Parseval's identity yields:

$$\frac{1}{\pi}\int_{-\pi}^{\pi} x^2 dx = \sum_{n=1}^{\infty}\frac{4}{n^2} \Longrightarrow \sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}.$$

(7) Note that if we differentiate u in Eq. (4) we get worse function. Regularity of u is about fast decay of a_n and b_n.

2.2 Riesz Representation Theorem

The Riesz representation theorem implies that H and H^* are actually isomorphic.

Theorem 2.5 (Riesz Representation Theorem)**.** Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space. Every element $\varphi : H \to \mathbb{R}$ in the dual space H^* of H can be represented as $\varphi(x) = \langle x, y\rangle$ for some $y \in H$. Moreover $\|\varphi\| = \|y\|$.

Proof. Suppose $\varphi \neq 0$ and let $M = \ker\varphi \neq H$ and $M^\perp \neq \{0\}$. Take *any* $w \in M^\perp$, $w \neq 0$ and let

$$y = \frac{\varphi(w)}{\|w\|}\frac{w}{\|w\|} \in M^\perp.$$

The norm of y is $\|y\| = |\varphi(w)|/\|w\|$, and $\varphi(y) = \varphi(w)^2/\|w\|^2$, so $\varphi(y) = \|y\|^2 = \langle y, y\rangle$. For any $x \in H$ let

$$x' = x - \frac{\varphi(x)}{\|y\|}\frac{y}{\|y\|}.$$

Then

$$\varphi(x') = \varphi(x) - \frac{\varphi(x)}{\|y\|^2}\cancel{\varphi(y)} = 0.$$

This implies that $x' \in \ker\varphi = M$. Then

$$0 = \langle x', y\rangle = \langle x, y\rangle - \frac{\varphi(x)}{\|y\|^2}\cancel{\langle y, y\rangle}$$

$$\Downarrow$$

$$\varphi(x) = \langle x, y\rangle.$$

We also have

$$\|\varphi\| = \sup_{\|x\|=1} \varphi(x) = \sup_{\|x\|=1} \langle x, y\rangle \leq \sup_{\|x\|=1} \|x\|\|y\| = \|y\|.$$

On the other hand we can choose $x = y/\|y\|$ and we have

$$\|\varphi\| \geq \varphi\left(\frac{y}{\|y\|}\right) = \frac{\varphi(y)}{\|y\|} = \|y\|.$$

This proved $\|\varphi\| = \|y\|$. To prove uniqueness, suppose $\varphi(x) = \langle x, y_1\rangle = \langle x, y_2\rangle$. Then we would have $\langle x, y_1 - y_2\rangle = 0$ for *any* $x \in H$. This implies $y_1 - y_2 = 0$ and consequently $y_1 = y_2$. □

2.3 Topology of infinite dimensional Hilbert space

OK, now we know that every infinite dimensional separable Hilbert space is like ℓ^2. Let's investigate the natural topology (i.e. the topology induced by the norm) on ℓ^2. Is the closed unit ball $\bar{B} = \{a \in \ell^2 : \|a\|_2 \leq 1\}$ compact? We can easily see that for the sequence $\{f_n\} \subset \bar{B}$ where $f_n = (0, \ldots, 1, \ldots, 0, \ldots)$ we have $\|f_n\| = 1$ for every $n \in \mathbb{N}$, so it does not have any convergent subsequence in this strong topology, so the closed and bounded unit ball is not compact. The sequence has a 1 in it that escapes to infinity. Not good! The topology is too "fine" in this infinite dimensional space.

From another perspective, for any $g \in \ell^2$, we have

$$\lim_{n\to\infty} \int f_n g = \lim_{n\to\infty} g(n) = 0 = \int 0 \cdot g$$

but it is not true that $f_n \to 0$, since $\|f_n\| = 1 \quad \forall n \in \mathbb{N}$. In other words, from the Reisz representation theorem $\varphi(f_n) \to \varphi(f)$ for every $\varphi \in H^*$ but it is not true that $f_n \to f$. We would like $f_n \to f$ to be true in some sense, so that we are able to say $f_n \to f \implies \varphi(f_n) \to \varphi(f)$, i.e. φ remains continuous. Recall we define this to be the *weak topology*, i.e. the weak topology on H is the topology generated by the linear functionals in H^*, so that it is the coarsest topology with respect to which all $\varphi \in H^*$ remain continuous. Then we can use the Banach-Alaoglu theorem to establish that *the closed unit ball is indeed compact in the weak topology*. In fact, this is true if and only if the space is reflexive (Hilbert space is reflexive. To prove use the Riesz representation theorem twice).

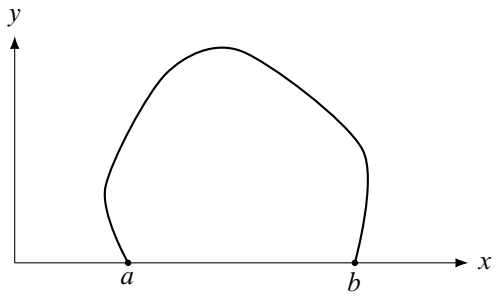


Figure 3: Dido's problem

3 Calculus of Variations

In finite dimensional optimization, our problem is $\min_{x \in A} f(x)$ for some $A \subseteq \mathbb{R}^n$. Here we are interested in

$$\min_{v \in A} J(v)$$

where A is some subset of an infinite dimensional vector space V and J is some cost functional.

3.1 Examples in Calculus of Variations

We discuss three historical examples that motivated calculus of variations.

Example 3.1 (Dido's isoperimetric problem)**.** We are given a curve of a fixed length, and two fixed points a and b. Our task is to choose the shape of the curve in order to maximize the area under the curve. See Fig. 3 for illustration. Formally, the problem is to maximize

$$J(y) = \int_a^b y(x)dx$$

for $y : [a, b] \to \mathbb{R}$ continuous and $y(a) = y(b) = 0$, subject to the constraint that

$$\int_a^b \sqrt{1 + (y'(x))^2}dx = C_0$$

for some fixed constant C_0.

The solution is easily guessed to be an arc of a circle. However, a rigorous demonstration requires tools from calculus of variations.

Example 3.2 (Catenary problem)**.** We would like to determine the shape of the rope that hangs over two points of equal height. See Fig. 4. The shape should minimize the total potential energy of the rope. If our rope has uniform density ρ, then the mass of a length ds is $dm = \rho ds$, and the potential energy of that segment is $dU = dm \cdot g \cdot y = (\rho g)y \cdot ds$. The total potential energy of the curve is thus

$$J(y) = \int_a^b dU = (\rho g) \int_a^b y ds = (\rho g) \int_a^b y(x)\sqrt{1 + (y'(x))^2}dx.$$

We would like to minimize $J(y)$ among continuous functions $y : [a, b] \to \mathbb{R}_+$ subject to the constraint that

$$\int_a^b \sqrt{1 + (y'(x))^2}dx = C_0$$

as before.

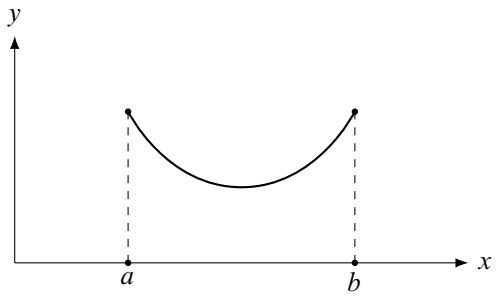


Figure 4: Catenary problem

The solution of the problem turns out to be

$$y(x) = c\cosh(x/c) \qquad \text{for some } c > 0,$$

instead of a parabola. The curve describes many things we see in real life, like freely-hanging electric power cables, spider's webs, or *simple* suspension bridges.

Example 3.3 (Brachistochrone)**.** Given two fixed points in the vertical plane, we want to find a path between them so that a particle sliding without friction along this path takes the shortest time (see Fig. 5). Note that we take the y-direction downward, so the negative direction of y corresponds to increase of potential energy. Suppose the particle is initially at rest so that the total energy is zero. By conservation of mechanical energy we have

$$\frac{1}{2}mv^2 - mgy = 0,$$

so the speed is $v = \sqrt{2gy}$. Time is distance divided by speed, so the traveling time of an arc ds is ds/v. The total traveling time is

$$J(y) = \int_a^b \frac{ds}{v} = \int_a^b \frac{\sqrt{1 + (y'(x))^2}}{\sqrt{2g \cdot y(x)}} dx.$$

We look for a continuous function $y : [a, b] \to \mathbb{R}_+$ with two fixed end points that minimizes J. The optimal curves turn out to be *cycloids*.

All examples take the following form: Among all continuous functions $y : [a, b] \to \mathbb{R}$ from $C^1[a, b]$ satisfying given boundary conditions

$$y(a) = y_0, \qquad y(b) = y_1$$

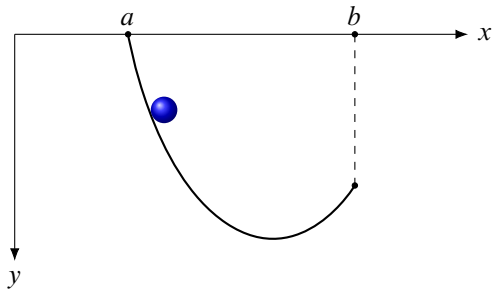


Figure 5: Brachistochrone problem

find (local) minima of the cost functional

$$J(y) = \int_a^b L(x, y(x), y'(x))dx. \tag{5}$$

$L : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called the *Lagrangian*. To minimize $J(y)$, we'd like to have its "derivative" to be zero. We next define the corresponding notion of derivatives.

3.2 Definition of Variations

The Fréchet derivative generalizes the notion of derivatives in one dimensional case.

Definition 3.4 (Fréchet derivative)**.** Let $T : X_1 \to X_2$ be a function between two Banach spaces. We say it is *Fréchet differentiable* at $x_0 \in X_1$ if there exists a continuous linear map $A : X_1 \to X_2$ such that

$$T(x_0 + \eta) = T(x_0) + A\eta + o(\|\eta\|)$$

for all $\eta \in X_1$. L is called the *Fréchet derivative* of T.

If for any $x \in M \subseteq X$, T is Fréchet-differentiable and moreover T' is continuous, then we say $T \in C^1(M)$.

Example 3.5. We discussed the following examples in class:

(1) $T : C^0[a, b] \to \mathbb{R}$ with

$$T(y) = \int_a^b (\sin^3(t) + y^2(t))dt.$$

T is convex as the function $s \mapsto s^2$ is convex.

(2) $T : C^1[a, b] \to \mathbb{R}$ with

$$T(y) = \int_a^b \rho(t)\sqrt{1 + y'(t)^2}dt$$

where $\rho \in C^0[a, b]$. The function $s \mapsto \sqrt{1 + s^2}$ is convex. T is convex if $\rho > 0$ and T is concave if $\rho < 0$. However, it is not possible for T to be strictly convex, since for example $T(y) = T(y + 1)$.

(3) $T : C^0[0, 1] \to \mathbb{R}$ with $T(y) = y(0)$. T is linear so it is convex.

(4) $T : C^0[0, 1] \to \mathbb{R}$ with $T(y) = y(0)^2 - y(1)$. T is convex. The derivative of T is $T'(y)h = 2y(0)h(0) - y(1)h(0)$.

(5) (Linear) $T : C^1[a, b] \to \mathbb{R}$ with

$$T(y) = y'\left(\frac{a + b}{2}\right).$$

(6) (Linear) $T : C^0[a, b] \to \mathbb{R}$ with

$$T(y) = \int_a^b t^2 y(t)dt.$$

(7) $T : M \to \mathbb{R}$ with $M = \{y \in C^1[0, 1] : y(0) = 0 \text{ and } y(1) = 1\}$ and

$$T(y) = \int_0^1 y'(t)^2 dt.$$

M is a closed and convex subset of the Banach space $C^1[0, 1]$. T is continuous and convex and $T(y) \geq 0$ for any $y \in M$. Suppose y^* is a minimizer. Then $T(y^*) \leq T(y^* + h)$ for any $h \in C^1[0, 1]$ such that $h(0) = (1) = 0$. We have

$$T(y^* + h) - T(y^*) = \int_0^1 [2(y^*)'(t)h'(t) + h'(t)^2]dt.$$

Let's guess $y^*(t) = t$. Then $T(y^* + h) - T(y^*) = \int_0^1 h'(t)^2 dt \geq 0$. It is strictly positive unless $h = 0$. So we have a unique minimizer.

Definition 3.6 (First variation)**.** The Fréchet derivative for the cost functional $J : V \to \mathbb{R}$ at point $y \in V$ is called the *first variation* of J and is denoted by $\delta J|_y$. Namely, it is a linear functional such that

$$J(y + \eta) = J(y) + \delta J|_y(\eta) + o(\|\eta\|). \tag{6}$$

We should be aware of another definition of first variation, namely thr Gateaux derivative

$$J(y + \alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + o(\alpha). \tag{7}$$

Equivalently, it is also

$$\delta J|_y(\eta) = \lim_{\alpha \to 0} \frac{J(y + \alpha\eta) - J(y)}{\alpha} = \frac{d}{d\alpha}\bigg|_{\alpha=0} J(y + \alpha\eta). \tag{8}$$

It is a weaker notion than the Fréchet derivative, but is also easier to work with. In the following we will often work with this weaker notion.

A real-valued functional $B : V \times V \to \mathbb{R}$ is *bilinear* if it is linear in each argument. Setting $Q(y) = B(y, y)$ we get a *quadratic form* on V.

Definition 3.7 (Second Variation)**.** A quadratic form $\delta^2 J\big|_y : V \to \mathbb{R}$ is called the *second variation* of J at y if for all $\eta \in V$ and all α we have

$$J(y + \alpha\eta) = J(y) + \delta J|_y(\eta)\alpha + \delta^2 J\big|_y(\eta)\alpha^2 + o(\alpha^2). \tag{9}$$

3.3 Euler-Lagrange Equation

Similar to the first-order necessary condition in finite dimensional optimization, if y^* is optimal we'd like to have

$$\delta J|_{y^*}(\eta) = 0 \tag{10}$$

for all admissible perturbations η (i.e.$\eta(a) = \eta(b) = 0$). We expand the Lagrangian inside $J(y + \alpha\eta)$ in the left side of Eq. (7) using Taylor expansion ($f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \nabla f \cdot \mathbf{v} + o(\mathbf{v})$) and then equate the corresponding term with $\delta J|_y(\eta) \equiv 0$:

$$\begin{aligned} J(y + \alpha\eta) &= \int_a^b L(x, y + \alpha\eta, y' + \alpha\eta') \\ &= \int_a^b \left[L(x, y, y') + L_y(x, y, y')\alpha\eta + L_{y'}(x, y, y')\alpha\eta' + o(\alpha)\right] \end{aligned}$$

so we have

$$\int_a^b \left[L_y(x, y, y')\eta + L_{y'}(x, y, y')\eta'\right] = \delta J|_y(\eta) \equiv 0. \tag{11}$$

Applying integration by parts for the second term inside the integral we get

$$\int_a^b \left[L_y(x, y, y') - \frac{d}{dx}L_{y'}(x, y, y')\right]\eta = 0 \tag{12}$$

for all $\eta \in C^1[a, b]$ such that $\eta(a) = \eta(b) = 0$. Since we required the term inside the bracket to be continuous, it must be zero on $[a, b]$. We get the renowned **Euler-Lagrange equation**:

$$L_y = \frac{d}{dx}L_{y'} \tag{13}$$

Example 3.8. This example demonstrates that we can use the Euler-Lagrange equation to solve seemingly complicated problems. Consider

$$\min J(y) = \int_0^{\log 2} e^{(-x-2y+2y')} dx$$

with $y(0) = 2$ and $y(\log 2) = 1$. Apply the Euler-Lagrange equation we get $L_y = -2L$ and $(d/dx)L_{y'} = 2(-1-2y'+2y'')L$, so $y'' - y' = 0$. The solution is $y(x) = -e^x + 3$.

Example 3.9. This example demonstrates that the solution to the Euler-Lagrange equation is not necessarily optimal. Consider

$$\min J(y) = \int_{-1}^{1} (y')^2(1-y')^2 dx = \int_{-1}^{1} (y' - y'^2)^2 dx$$

and let the function space be $M = \{y \in C^1[-1,1] : y(-1) = 0, y(1) = 1\}$. The derivative of L with respect to y' is $L_{y'} = 2(y' - y'^2)(1-2y') = 2(y' - 3y'^2 + 2y'^3)$. Since there is no y in L, we have $(d/dx)L_{y'} = 0$ and so $L_{y'}$ must be a constant. Consequently y' must be a constant, so y is linear. The solution then is $y(x) = (x+1)/2$ and $J(y) = \int_{-1}^{1} \frac{1}{4} \cdot \frac{1}{4} dx = \frac{1}{8}$.

Consider the following sequence

$$y_n(x) = \begin{cases} 0 & -1 \le x \le -\frac{1}{n}, \\ \frac{n}{4}(x + \frac{1}{n})^2 & -\frac{1}{n} \le x \le \frac{1}{n}, \\ x & \frac{1}{n} \le x \le 1. \end{cases}$$

The value of the functional is

$$\begin{aligned} J(y_n) &= \int_{-\frac{1}{n}}^{\frac{1}{n}} \left(\frac{n}{2}\left(x + \frac{1}{n}\right)\right)^2 \left(1 - \frac{n}{2}\left(x + \frac{1}{n}\right)\right)^2 dx \\ &= \frac{1}{16} \int_{-\frac{1}{n}}^{\frac{1}{n}} (nx+1)^2(1-nx)^2 dx = \frac{1}{16} \int_{-\frac{1}{n}}^{\frac{1}{n}} (1 - n^2x^2)^2 dx \\ &= \frac{1}{16} \int_{-\frac{1}{n}}^{\frac{1}{n}} (n^4x^4 - 2n^2x^2 + 1) dx \\ &= \cdots = \frac{c}{n} \end{aligned}$$

for some constant c. Hence $J(y_n) \to 0$ as $n \to \infty$. The sequence converges to

$$y(x) = \begin{cases} 0 & -1 \le x \le 0 \\ x & 0 \le x \le 1 \end{cases},$$

which does not belong to $C^1[-1,1]$. In fact, for any "stair" function y we all have $J(y) = 0$. Note also that the Lagrangian $L(s) = s^2(1-s^2)$ is not convex.

Example 3.10. Consider

$$\min J(y) = \int_0^1 \left(\frac{y'^2}{p} + 2q \cdot y \cdot y' + q' \cdot y^2\right) dx$$

with $y(0) = 0$ and $y(1) = 1$, where $p, q \in C^1[0,1]$ with $p(x) \neq 0$ on $[0,1]$. The partial derivatives of the Lagrangian with respect to y' and y are respectively

$$L_{y'} = \frac{2y'}{p} + 2q \cdot y,$$

$$L_y = 2q \cdot y' + 2q' \cdot y = \frac{d}{dx}(2qy).$$

So we have

$$\frac{d}{dx}\left(\frac{2y'}{p} + 2q \cdot y - 2q \cdot y\right) = 0$$
$$y' = C \cdot p \quad \text{for some constant } C \in \mathbb{R}$$
$$y(x) = C \int_0^x p(t)dt.$$

From $y(1) = 1$ we obtain $C = \int_0^1 p(t)dt$.

3.3.1 Special Cases

1. (no y) $L = L(x, y')$. In this case $L_{y'}(x, y') = c$. We can then infer y' and consequently y.
2. (no x) $L = L(y, y')$. Then

$$\frac{d}{dx}L_{y'}(y, y') - L_y(y, y') = 0$$
$$\Downarrow$$
$$y'\left(\frac{d}{dx}L_{y'}(y, y') - L_y(y, y')\right) = 0$$
$$\Downarrow$$
$$y'\frac{d}{dx}L_{y'}(y, y') + y''L_{y'}(y, y') - y''L_{y'}(y, y') - y'L_y(y, y') = 0$$
$$\Downarrow$$
$$\frac{d}{dx}\left[y'L_{y'}(y, y') - L(y, y')\right] = 0$$
$$\Downarrow$$
$$y'L_{y'}(y, y') - L(y, y') = c.$$

For the Lagrangian in Example 3.3, this leads to

$$\frac{y' \cdot y'}{\sqrt{1 + (y')^2}\sqrt{2gy}} - \frac{\sqrt{1 + (y')^2}}{\sqrt{2gy}} = c$$
$$\Downarrow$$
$$\frac{-1}{\sqrt{1 + (y')^2}\sqrt{2gy}} = c$$
$$\Downarrow$$
$$(1 + (y')^2)(2gy) = k > 0 \qquad \left(k = \frac{1}{c^2}\right)$$
$$(y')^2 = \frac{k}{2gy} - 1$$
$$\Downarrow$$
$$y' = \sqrt{\frac{k}{2gy} - 1},$$

a first order differential equation in y.

3. (no y') $L = L(x, y)$. $L_y(x, y) = 0$. This is not a differential equation of y. We cannot use the Euler-Lagrange equation in this case. For example
$$J(y) = \int_0^1 x^2 y^2 dx$$
on $M = \{y \in C^1[0, 1] : y(0) = 0, y(1) = 1\}$. The Euler-Lagrange equation gives $y = 0$, which is not in M. If we take $y_n(x) = x^n$ then we will have $J(y_n) \to 0$. The minimum is not attainable. As another example, take same M and
$$J(y) = \int_0^1 (y^2 - 2x \cdot y)dx.$$
The Euler-Lagrange equation gives $2y - 2x = 0 \Rightarrow y = x$ and we have $J(y) = -1/3$. Note that J is convex in y, so y is the unique minimizer.

Example 3.11. Below is a collection of exercises we did in class.

(1) Consider the cost functional $J(y) = \int_a^b \sin y'(x)dx$, with endpoints $y(a) = y(b) = 0$. The Lagrangian only depends on y', so the Euler-Lagrange equation yields
$$\frac{d}{dx} L_{y'} = \frac{d}{dx} \cos y' = 0 \Rightarrow \cos y' \text{ is constant } \Rightarrow y' \text{ is constant.}$$
Since $y(a) = y(b) = 0$, the solution is $y = 0$. However, as J is odd with $J(0) = 0$, the solution $y = 0$ is not optimal. Namely, if we take any y such that $J(y) > 0$, then $J(-y) = -J(y) < 0$.

(2) Consider the cost functional
$$J(y) = \int_a^2 x^3 y'(x)^3 dx$$
where $a \geq 0$. J is not bounded from below, as can be seen from Fig. 6. Since $x^3 \geq 0$ on $[a, 2]$ is increasing we have
$$\left| \int_a^{(a+2)/2} x^3 y'^3 dx \right| < \left| \int_{(a+2)/2}^2 x^3 y'^3 dx \right|$$
so that $J(y) < 0$. As we increase the height of the function the functional J goes to $-\infty$. However, let's still solve the Euler-Lagrange equation and see what we will get:
$$\frac{d}{dx}\left[3x^3 y'^2\right] = 0 \Rightarrow x^3 y' = c \Rightarrow y'^2 = \frac{c}{x^3} \Rightarrow y' = \pm\sqrt{\frac{c}{x^3}} = \pm\frac{k}{x^{3/2}}.$$
We considered three cases:

(i) $a = 0$, $y(0) = 0$, $y(2) = 1$. In this case we really do not have a solution for y, since $y(0)$ and $y'(0)$ are not defined. That the solution is not defined on some points is not obvious at first from the look of the functional. If it were the case that $y(2) = 0$, then we could set $k = 0$ and the solution would be $y = 0$.

(ii) $a = 1$, $y(1) = y(2) = 1$. Since J is odd, let's first take $y' = k/x^{3/2}$. Integrate and we get
$$y(x) = \frac{h}{\sqrt{x}} + b \Rightarrow \begin{cases} 1 = y(1) = h + b \\ 1 = y(2) = \frac{h}{\sqrt{2}} + b \end{cases} \Rightarrow \begin{cases} h = 0 \\ b = 1 \end{cases} \Rightarrow y = 1$$
for which $J(y) = 0$.

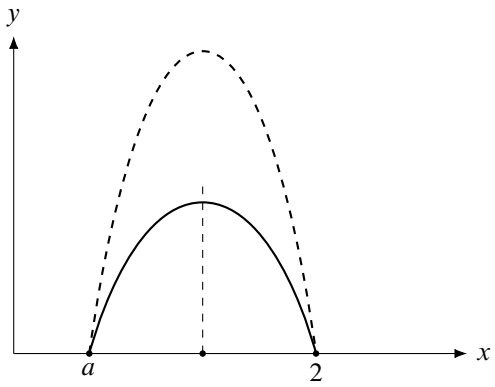


Figure 6: $J(y)$ is not bounded from below

(iii) $a = 1, y(1) = 1, y(2) = 2$. In this case the solution is

$$y(x) = \frac{h}{\sqrt{x}} + b \Rightarrow \begin{cases} 1 = y(1) = h + b \\ 2 = y(2) = \frac{h}{\sqrt{2}} + b \end{cases} \Rightarrow \begin{cases} h = -\sqrt{2}(1+\sqrt{2}) \\ b = 3 + \sqrt{2} \end{cases}$$
$$\Rightarrow y = \frac{-\sqrt{2}(1+\sqrt{2})}{\sqrt{x}} + 3 + \sqrt{2}.$$

(3) The functional is

$$J(y) = \int_1^2 x^3 y'(x)^2 dx$$

with $y(1) = 5$ and $y(2) = 2$. Now J is convex, so the Euler-Lagrange equation would yield a unique optimal solution. Let's solve it:

$$\frac{d}{dx} 2x^3 y'(x) = 0 \quad \Rightarrow \quad y'(x) = \frac{c}{x^3} \quad \Rightarrow \quad y(x) = k_1 x^{-2} + k_2$$

and from the boundary condition

$$\begin{cases} y(1) = k_1 + k_2 = 5 \\ y(2) = k_1/4 + k_2 = 2 \end{cases} \Rightarrow k_1 = 4, k_2 = 1 \quad \Rightarrow y(x) = 4x^{-2} + 1.$$

(4) The functional is

$$J(y) = \int_0^1 (2xy - y'^2 + 3y'y^2) dx$$

with $y(0) = 0, y(1) = -1$. $L_{y'} = -2y' + 3y^2$ so $(d/dx)L_{y'} = -2y'' + 6yy'$, and $L_y = 2x + 6yy'$, so

$$\frac{d}{dx} L_{y'} = L_y \Rightarrow y'' = -x$$

and consequently $y' = -(1/2)x^2 + k$ and $y = -(1/6)x^3 + k_1 x + k_2$. From the boundary condition $y(0) = k_2 = 0$ and $y(1) = -1/6 + k_1 = -1 \Rightarrow k_1 = -5/6$ so the solution is

$$y = -\frac{1}{6}x^3 - \frac{5}{6}x.$$

(5) The functional is

$$J(y) = \int_0^1 (y^3 + 3x^2 y') dx.$$

We discussed two cases

(i) $y(0) = 0, y(1) = 1$,

(ii) $y(0) = 0$ and $y(1)$ is free.

If we now solve for the Euler-Lagrange equation, then we have $L_{y'} = 3x^2$ and $L_y = 3y^2$, which gives us $y^2 = 2x \Rightarrow y = \sqrt{2x}$. This solution can not satisfy the boundary condition of Item (i). It is only feasible under Item (ii). What's going wrong? Notice that we can do integration by parts on the second part of J to get rid of y', so in this sense J does not really depend on y':

$$\int_0^1 3x^2 y' dx = \int_0^1 3x^2 dy = y \cdot 3x^2 \Big|_0^1 - \int_0^1 6xy = 3 - 6xy$$

so the functional is

$$J(y) = \int_0^1 (y^3 - 6xy) + 3.$$

(6) Functional

$$J(y) = \int_0^1 (2xy^3 + e^x \sin y + 3x^2 y^2 \cdot y' + y' e^x \cos y) dx$$

with

(i) $y(0) = 0, y(1) = 1$,

(ii) $y(0) = 0, y(1) = \sqrt{8}$.

It is easy to see that

$$J(y) = \int_0^1 \frac{d}{dx}(x^2 y^3 + e^x \sin y) dx = \left[x^2 y^3 + e^x \sin y\right]_0^1 = c$$

for any y.

3.3.2 Variable-endpoint problems

If we do not fix the right point of y, then the set of admissible perturbations will change. We still have $\eta(a) = 0$ but $\eta(b)$ does not have to be 0. Then in Eq. (11) we would have an extra term $L_{y'}(x, y, y')\eta(b) = 0$ for all admissible η. We then have an extra condition $L_{y'}(x, y, y') = 0$ at $x = b$, besides the Euler-Lagrange equation.

Example 3.12. (1)

$$J(y) = \int_0^\pi (y'(x)^2 + y(x)^2 - 2y(x) \sin x) dx$$

(2)

$$J(y) = \int_0^\pi (y^2 - y'^2) dx$$

where $y \in C^1[0, \pi]$. We place no restrictions on the two endpoints. In this case $y'' + y = 0$ and we find

$$\begin{cases} y(x) = a \sin x + b \cos x \\ y'(0) = y'(\pi) = 0 \end{cases}$$

i.e. $a = 0$. $y_b(x) = b \cos x$ are extremals. $J(y_b) = 0$ for any b. Next we require $y(0) = y(\pi) = 0$. In this case

$$\begin{cases} y(x) = a \sin x + b \cos x \\ y(0) = y(\pi) = 0 \end{cases}$$

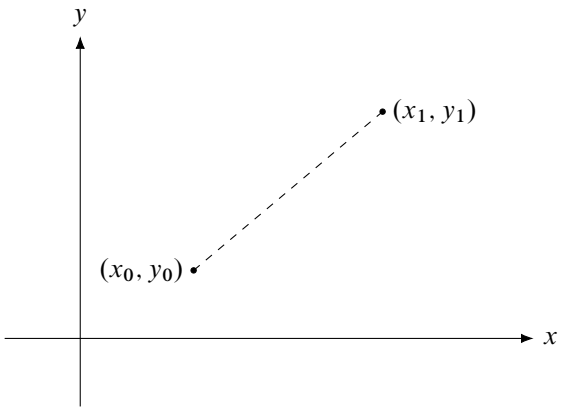


Figure 7: shortest path between two points on the plane.

We find $b = 0$, and so $y_a(x) = a \sin x$ and $J(y_a) = 0$ for any a. But if we take $y(x) = x(x - \pi)$, then $J(y) < 0$. In fact $J(\alpha y) < 0$ for any $\alpha > 0$ and if we let $\alpha \to \infty$ then $J(\alpha y) \to -\infty$. The functional J is not bounded from below. It is indeed bounded from above and the solutions come out of the Euler-Lagrange equation are maximizers.

(3) Functional

$$J(y) = \int_0^1 ((y' - x)^2 + 2xy)dx$$

with $y \in C^1[0, 1]$ and $y(0) = 1$. The Euler-Lagrange equation yields

$$\begin{aligned} 2(y'' - 1) &= 2x \\ y'' &= x + 1 \\ y' &= \frac{1}{2}x^2 + x + k \\ y &= \frac{1}{6}x^3 + \frac{1}{2}x^2 + k_1 x + k_2. \end{aligned}$$

Now $y(0) = k_2 = 1$, and we use the extra condition $L_{y'} = 2(y' - x) = 0$ at $x = 1$, so that $y'(1) = 1 \Rightarrow k_1 = -1/2$. The final solution is thus

$$y = \frac{1}{6}x^3 + \frac{1}{2}x^2 - \frac{1}{2}x + 1.$$

Since J is a convex function of y, this is the unique optimal solution.

3.3.3 Multiple Degrees of Freedom

It is straightforward to extend the derivation of the Euler-Lagrange equation in Section 3.3 to multiple degrees of freedom settings, i.e. when $y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$: simply interpret L_y and $L_{y'}$ as gradients and replace multiplications by inner products in Eqs. (11) and (12). The resulting equation is the same as Eq. (13). Write out for each coordinates

$$L_{y_i} = \frac{d}{dx} L_{y_i'} \quad \forall i = 1, \ldots, n.$$

Example 3.13. Consider the problem of finding the shortest path between two points (x_0, y_0) and (x_1, y_1) on the plane $\mathbb{R}^2$ (see Fig. 7). The problem can be formulated as finding the optimal

$$\begin{cases} x = x(t) & \text{with } x(0) = x_0,\ x(1) = x_1 \\ y = y(t) & \text{with } y(0) = y_0,\ y(1) = y_1 \end{cases}$$

for $t \in [0, 1]$ that minimizes

$$J(x, y) = \int_0^1 \sqrt{x'^2 + y'^2}dt.$$

$J(x, y)$ is convex, so we will have a unique solution. The Euler-Lagrange equations are

$$\begin{cases} \dfrac{d}{dt}\dfrac{x'}{\sqrt{x'^2 + y'^2}} = 0 \\ \dfrac{d}{dt}\dfrac{y'}{\sqrt{x'^2 + y'^2}} = 0 \end{cases}$$

so that x' is constant and y' is constant. From the boundary conditions we get $x' = (x_1 - x_0)/(1 - 0) = (x_1 - x_0)$ and similarly $y' = (y_1 - y_0)/(1 - 0) = (y_1 - y_0)$. The optimal value is

$$J^* = \int_0^1 \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}dt = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}.$$

Example 3.14. Consider

$$J(u, v) = \int_0^{\pi/2} (u'^2 + v'^2 + 2uv)dx$$

where $u, v \in C^1[0, \pi/2]$ and $u(0) = v(0) = 0$, $u(\pi/2) = v(\pi/2) = 1$. From the Euler-Lagrange equation

$$\begin{cases} \dfrac{d}{dx}2u' - 2v = 0 \\ \dfrac{d}{dx}2v' - 2u = 0 \end{cases} \Rightarrow \begin{cases} u'' - v = 0 \\ v'' - u = 0 \end{cases}$$

and so $u'''' - u = 0$ with $u(0) = 0$, $u(\pi/2) = 1$, $u''(0) = 0$, and $u''(\pi/2) = 0$. We can also see that $v = u$. The solution for u is

$$u(x) = a\cos x + b\sin x + c\cosh x + d\sinh x$$

and from the initial conditions

$$\left.\begin{aligned} 0 = u(0) &= a + c \\ 0 = u''(0) &= -a + c \end{aligned}\right\} \Rightarrow \begin{cases} a = 0 \\ c = 0 \end{cases}$$

$$\left.\begin{aligned} 1 = u(\frac{\pi}{2}) &= b + d\sinh\frac{\pi}{2} \\ 1 = u''(\frac{\pi}{2}) &= -b + d\sinh\frac{\pi}{2} \end{aligned}\right\} \Rightarrow \begin{cases} b = 0 \\ d = \dfrac{1}{\sinh(\pi/2)} \end{cases}$$

So the final solutions for u and v are

$$\begin{cases} u(x) = \dfrac{\sinh x}{\sinh \pi/2} \\ v(x) = \dfrac{\sinh x}{\sinh \pi/2}. \end{cases}$$

The value of J is

$$\begin{aligned} J(u, v) &= \int_0^{\pi/2} \left[2\left(\frac{\cosh x}{\sinh \pi/2}\right)^2 + 2\left(\frac{\sinh x}{\sinh \pi/2}\right)^2 \right] dx = 2\left(\int_0^{\pi/2} \cosh 2x dx\right) / (\sinh^2 \pi/2) \\ &= \sinh \pi/(\sinh^2 \pi/2) \approx 2.18. \end{aligned}$$

What if we take the straight line between $(0,0)$ and $(1,\pi/2)$? Let $u(x) = v(x) = (\pi/2)x$, then $u' = v' = \pi/2$ so

$$J(u,v) = \int_0^{\pi/2} \left(\frac{\pi^2}{2} + \frac{\pi^2}{2}x^2\right) dx = \frac{\pi^2}{2}\cdot\frac{\pi}{2} + \frac{\pi^2}{2}\cdot\frac{1}{3}\frac{\pi^3}{8} = \frac{\pi^3}{4} + \frac{\pi^5}{48} > 7.$$

3.4 Variational Problems With Constraints

3.4.1 Integral Constraints

Example 3.1 and Example 3.2 are all calculus of variations problems with integral constraints. Let's derive necessary conditions for optimality for such problems. Suppose we want to minimize

$$J(y) = \int_a^b L(x,y,y')dx$$

with usual boundary condition and an additional constraint that

$$C(y) = \int_a^b M(x,y,y')dx = C_0, \tag{14}$$

where M is a function from the same class as L, and C_0 is a given constant. Suppose y is optimal. Then for $y + \alpha\eta$ to be admissible, the perturbation η must preserve the constraint (in addition to vanishing at the endpoints as before). In other words, we must have $C(y+\alpha\eta) = C_0$ for all α sufficiently close to 0. In terms of the first variation of C, this property is easily seen to imply that

$$\delta C|_y(\eta) = 0.$$

Repeating the same calculation as in our original derivation of the Euler-Lagrange equation, we obtain

$$\int_a^b \left[M_y(x,y,y') - \frac{d}{dx}M_{y'}(x,y,y')\right]\eta = 0 \tag{15}$$

as Eq. (12). For every η satisfying Eq. (15), we should have

$$\delta J|_y(\eta) = \int_a^b \left[L_y(x,y,y') - \frac{d}{dx}L_{y'}(x,y,y')\right]\eta = 0.$$

In summary,

$$\int_a^b \left[L_y - \frac{d}{dx}L_{y'}\right]\eta = 0 \quad \forall \eta \text{ such that } \int_a^b \left[M_y - \frac{d}{dx}M_{y'}\right]\eta = 0.$$

Similar to the situation in finite-dimensional optimization, we conclude that there exists a constant λ (a Lagrange multiplier) such that

$$\left(L_y - \frac{d}{dx}L_{y'}\right) + \lambda\left(M_y - \frac{d}{dx}M_{y'}\right) = 0$$

for all $x \in [a,b]$. Rearrange the terms we get

$$(L+\lambda M)_y = \frac{d}{dx}(L+\lambda M)_{y'},$$

i.e. the Euler-Lagrange equation holds for the augmented Lagrangian $L + \lambda M$. In other words, y is an extremal of the augmented cost functional

$$(J+\lambda C)(y) = \int_a^b \left[L(x,y,y') + \lambda M(x,y,y')\right]dx. \tag{16}$$

There is a catch: if y is an extremal of the constraint functional C, i.e. it satisfies the Euler-Lagrange equation for M, then all of its nearby curves would violate the constraint. For example, consider the length constraint $C(y) = \int_0^1 \sqrt{1+(y')^2}dx$ together with the boundary conditions $y(0) = y(1) = 0$. Clearly $y \equiv 0$ is the only admissible curve (it is the unique global minimizer of the constraint functional), so no matter what J is, the solution is $y \equiv 0$, but it is not an extremal for any J.

To summarize, we used heuristics and derived the following first-order necessary condition for constrained optimality: if y is an extremum for the constrained problem and is not an extremal for the constraint functional C (i.e. does not satisfy the Euler-Lagrange equation for M), then it is an extremal of the augmented cost functional Eq. (16) for some $\lambda \in \mathbb{R}$. To prove it rigorously, we can consider a two parameter family of perturbed curves $y + \alpha_1\eta_1 + \alpha_2\eta_2$ and use the inverse function theorem.

3.4.2 Non-integral Constraints

If instead of an integral constraint, we have an equality constraint which must hold point-wise:

$$M(x, y(x), y'(x)) = 0 \tag{17}$$

for all $x \in [a, b]$, then the first-order necessary condition for optimality is similar to that for integral constraint, but the Lagrange multiplier is now a function of x. In other words, the Euler-Lagrange equation must hold for the augmented Lagrangian

$$L + \lambda(x)M$$

for some function $\lambda : [a, b] \to \mathbb{R}$. An additional assumption to rule out degenerate cases is that there are at least two degrees of freedom and that everywhere along the curve we have $M_{y'} \neq 0$, or if y' does not appear in Eq. (17), $M_y \neq 0$. The integral constraint Eq. (14) is global, in the sense that it applies to the entire curve. In contrast, the non-integral constraint Eq. (17) is local, i.e., applies to each point on the curve. Locally around each point, there is no essential difference between the two. This suggests that for each x there should exist a Lagrange multiplier, and these can be pieced together to give the desired function $\lambda = \lambda(x)$.

Example 3.15. What is the shortest path between two points on earth? I.e. we want

$$\min \int_0^1 \sqrt{x^2 + y^2 + z^2}dt$$

subject to the constraint that

$$x^2(t) + y^2(t) + z^2(t) = R^2$$

for each $t \in [0, 1]$. Geometrically, the solution can be obtained as follows: take the two points and the center of the mass, which defines a plane that cuts the sphere. The smaller of the two arcs is the shortest path. This is called the *geodesics*.

Example 3.16. There can be many solutions to the constrained-version Euler-Lagrange equation and not all of them are minimizers. Suppose we want

$$\min J(y) = \int_a^b (y')^2 dx \quad s.t. \quad C(y) = \int_a^b y^2 dx = 1$$

and with boundary conditions $y'(a) = y'(b) = 0$. Note that $J(y) \geq 0$ for any y. The Euler-Lagrange equation is

$$\frac{d}{dx}(2y' - 0) = (0 - 2\lambda y) \Rightarrow y'' + \lambda y = 0$$

If $\lambda = 0$, then y is constant and from which $y = 1/\sqrt{b-a}$. For this y we have $J(y) = 0$ so y is the unique minimizer of J under the constraint. In general, $\lambda \geq 0$. To see this, note

$$y'' + \lambda y = 0 \Rightarrow y'' \cdot y = -\lambda y^2 \Rightarrow \int_a^b y'' \cdot y = -\lambda \int_a^b y^2 \Rightarrow -\int_a^b (y')^2 = -\lambda \Rightarrow \lambda \geq 0.$$

Another way to see $\lambda \geq 0$ is that, if $\lambda = -\alpha^2 < 0$, then the solution would be $y(x) = c_1 e^{\alpha x} + c_2 e^{-\alpha x}$. The boundary conditions yields $y = 0$, which does not satisfy the constraint.

For $\lambda = \alpha^2 > 0$, the solution is $y(x) = c_1 \cos \alpha x + c_2 \sin \alpha x$. The boundary condition yields $\sin \alpha(a-b) = 0$. This implies that $\alpha(b-a) = k\pi \Rightarrow \lambda_k = k^2\pi^2/(b-a)^2$. This gives us the solutions

$$y_k(x) = A_k \cos \frac{k\pi(x-a)}{b-a}.$$

We then have $C(y) = 1 \Rightarrow A_k^2 = 2/(b-a)$. The solutions are

$$y_k(x) = \sqrt{\frac{2}{b-a}} \cos \frac{k\pi(x-a)}{b-a}$$

and $J(y) = k^2\pi^2/(b-a)^2$. The Lagrangian multiplier method gives us many critical points.

Example 3.17. Consider the problem

$$\min J(y) = \int_0^\pi \left(2 \sin x \cdot y + y'^2\right) dx$$

with $y \in C^1[0, \pi]$, boundary conditions $y(0) = y(\pi) = 0$, and an integral constraint

$$C(y) = \int_0^\pi y dx = 1.$$

The Euler-Lagrange equation yields

$$\begin{aligned} \frac{d}{dx}(2y') - 2\sin x - \lambda &= 0 \\ y''(x) &= \sin x + \frac{\lambda}{2} \\ y(x) &= -\sin x + \frac{\lambda}{4}x^2 + c_1 x + c_2. \end{aligned}$$

From the boundary conditions

$$\left.\begin{aligned} 0 = y(0) = c_2 \Rightarrow c_2 = 0 \\ 0 = y(\pi) = \frac{\lambda}{4}\pi^2 + c_1\pi \Rightarrow c_1 = -\frac{\lambda\pi}{4} \end{aligned}\right\} \Rightarrow y(x) = -\sin x + \frac{\lambda}{4}x^2 - \frac{\lambda\pi}{4}x.$$

From the constraint $C(y) = 1$ we get $\lambda = -72/\pi^3$. So the solution is

$$y(x) = -\sin x - \frac{18}{\pi^3}x^2 + \frac{18}{\pi^2}x.$$

Since J is convex, the solution is the unique minimizer.

Example 3.18. We revisit Example 3.1, in a more general setting. We consider a closed curve described by

$$\begin{cases} x = x(t) \\ y = y(t) \end{cases}$$

with $t \in [0,1]$ and $x(0) = x(1)$, $y(0) = y(1)$. The area enclosed by the curve is

$$A(x,y) = \frac{1}{2}\int_0^1 (xy' - x'y)dt.$$

The problem is

$$\max_{x,y} A(x,y) \quad s.t. \quad \int_0^1 \sqrt{x'^2 + y'^2}dt = C_0.$$

We solve for the Euler-Lagrange equation:

$$L = \frac{(xy' - x'y)}{2} \Rightarrow \quad L_{x'} = -\frac{y}{2}, \quad L_{y'} = \frac{x}{2}, \quad L_x = \frac{y'}{2}, \quad L_y = -\frac{x'}{2}.$$
$$M = \sqrt{x'^2 + y'^2} \Rightarrow \quad M_{x'} = \frac{x'}{M}, \quad M_{y'} = \frac{y'}{M}, \quad M_x = M_y = 0.$$

So

$$\begin{cases} \frac{d}{dx}\left(-\frac{y}{2} + \lambda\frac{x'}{M}\right) = \frac{y'}{2} \\ \frac{d}{dx}\left(\frac{x}{2} + \lambda\frac{y'}{M}\right) = -\frac{x'}{2} \end{cases} \Rightarrow \begin{cases} y' = \lambda\frac{d}{dt}\left(\frac{x'}{M}\right) \\ x' = -\lambda\frac{d}{dt}\left(\frac{y'}{M}\right) \end{cases} \Rightarrow \begin{cases} y = \lambda\frac{x'}{M} + c_2 \\ x = -\lambda\frac{y'}{M} + c_1. \end{cases}$$

This implies that

$$(x - c_1)^2 + (y - c_2)^2 = \lambda^2,$$

so the solution is a circle, with radius $|\lambda|$. From the constraint $2\pi|\lambda| = C_0$, we get

$$|\lambda| = \frac{C_0}{2\pi} \quad \Rightarrow \quad \lambda = \pm\frac{C_0}{2\pi}.$$

Also note that $C_0^2 = 4\pi^2|\lambda|^2 = 4\pi \cdot (\pi|\lambda|^2) = 4\pi \cdot A_{\max}$, so we have the inequality

$$C_0^2 \geq 4\pi A.$$

4 Introduction to Optimal Controls

The control system is

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

Here $x \in \mathbb{R}^n$ is the state, $u \in U \subseteq \mathbb{R}^m$ is the control, $t \in \mathbb{R}$ is time, and x_0 is the initial state. Note the state has dimension n and the control has dimension m. The control u can affect the evolution of the state through the system. We want to find a control from

$$\mathcal{U} = \{u : [0, T] \to U \mid u(\cdot) \text{ measurable}\}$$

so as to *minimize* some objective function

$$J(u) := \int_0^T L(t, x(t), u(t))dt + K(T, x(T)) \tag{18}$$

where $L : \mathbb{R} \times \mathbb{R}^n \times U \to \mathbb{R}$ is the *Lagrangian*, and $K : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ is the *terminal cost*.

4.1 Examples of Control Problems

Here we present several examples of control problems, mostly from Evans 2005.

Example 4.1 (Control of Production and Consumption)**.** Suppose $x(t)$ is the output of an economy at time $t \geq 0$. The output in each period is either reinvested or consumed. Let $u(t) \in [0, 1]$ be the fraction of output reinvested at time t, so that

$$x(t) = (1 - u(t))x(t) + u(t)x(t) := C(t) + I(t).$$

The dynamic is

$$\begin{cases} \dot{x}(t) = kI(t) = ku(t)x(t) \\ x(0) = x_0, \end{cases}$$

for some $k > 0$. Namely the growth rate of the output is proportional to the investment. We want to maximize the total consumption in $[0, T]$, so we formulate our problem as

$$\min_u J(u) = \int_0^T -C(t)dt = \int_0^T -(1 - u(t))x(t)dt.$$

To see the problem in extreme cases, note that if we take $u \equiv 0$, i.e. we consume all output produced in each period, then the system becomes

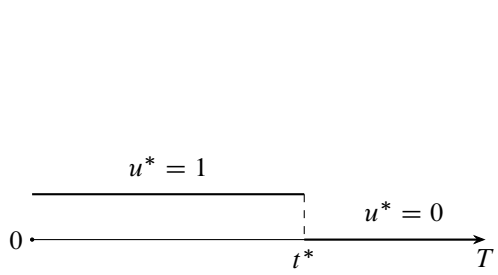
$$\begin{cases} \dot{x}(t) = 0 \\ x(0) = x_0 \end{cases}$$

so that $x(t) = x_0$ for all t. The total output of the economy does not grow and stay constant over time. The total consumption is then $\int_0^T 1 \cdot x_0 dt = Tx_0$. If we take $u \equiv 1$, i.e. we invest all the output and do not consume for all the period, then obviously our total consumption will be 0.

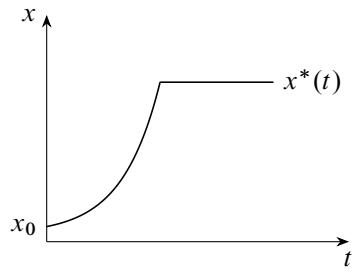
We shall see that the optimal solution is a *bang-bang* control (Fig. 8a)

$$u^*(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq t^*, \\ 0 & \text{if } t^* < t \leq T \end{cases}$$

for an appropriate *switching time* $t^* \in [0, T]$. In other words, we should invest all output up until time t^*, after which we only make consumption but no investment.



(a) The bang-bang control of the production and consumption problem.



(b) The optimal trajectory corresponding to the bang-bang-control.

Figure 8: The optimal control and the optimal trajectory in Example 4.1.

Given this information, let's find this optimal switching time. When $u^*(t) = 1$, the dynamic and the corresponding solution are

$$\begin{cases} \dot{x}(t) = kx(t), \quad 0 \le t \le t^* \\ x(0) = x_0 \end{cases} \quad \Rightarrow \quad x(t) = x_0 e^{kt}$$

so that at time t^* the output is $x(t^*) = x_0 e^{kt^*}$. When $u^*(t) = 0$ the dynamic and the solution are

$$\begin{cases} \dot{x}(t) = 0, \quad t^* \le t \le T \\ x(t^*) = x_0 e^{kt^*} \end{cases} \quad \Rightarrow \quad x(t) = x_0 e^{kt^*}.$$

The optimal trajectory is thus like Fig. 8b. The optimal consumption corresponding to the bang-bang control is then

$$\int_0^{t^*} 0dt + \int_{t^*}^{T} x_0 e^{kt^*} dt = (T - t^*) x_0 e^{kt^*}.$$

We want to find the maximizer of the function $\phi(s) = (T - s)e^{ks}$ so we set its derivative to zero

$$\phi'(s) = -e^{ks} + (T - s)ke^{ks} = 0$$
$$\Downarrow$$
$$(T - s) \cdot k = 1 \quad \Rightarrow \quad s = T - \frac{1}{k}$$

so the optimal switching time is $t^* = T - \frac{1}{k}$. The larger k, the later we can do the switch, while smaller k implies that we should do the switch earlier. This makes intuitive sense: if investments have high return, then it is best to constrain the consumption at present and invest more, so that we can have greater accumulation of wealth in the future; on the other hand if investments have low return, then we do not have incentives to commit to long-term investments, but should instead enter into consumption early on.

Example 4.2 (Pendulum)**.** We want to apply forces to a swinging pendulum so as to bring it to stop in minimum time. Let $\theta(t)$ denote the angle of the pendulum at time t. Recall The equation of motion with damping and with small angle approximation is

$$\begin{cases} \ddot{\theta}(t) + \lambda\dot{\theta}(t) + \omega^2\theta(t) = 0 \\ \theta(0) = \theta_1, \dot{\theta}(0) = \theta_2. \end{cases}$$

Let $u(\cdot)$ denote the magnitude of the torque applied to the object around the pivot, with direction perpendicular to the plane of motion. We require $|u| \le 1$. The dynamics now become

$$\begin{cases} \ddot{\theta}(t) + \lambda\dot{\theta}(t) + \omega^2\theta(t) = u(t) \\ \theta(0) = \theta_1, \dot{\theta}(0) = \theta_2. \end{cases}$$

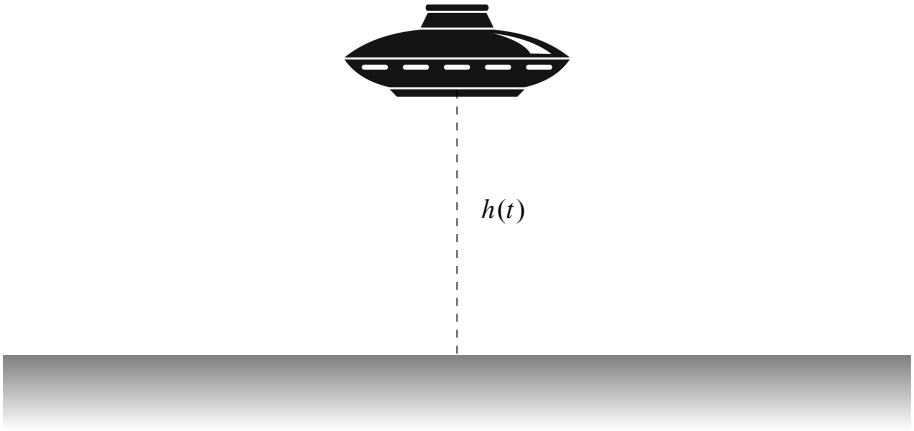


Figure 9: Illustration of the moon lander problem.

If we let $x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} := \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}$, then we can write the dynamics as

$$\dot{x}(t) = \begin{pmatrix} \dot{\theta} \\ \ddot{\theta} \end{pmatrix} = \begin{pmatrix} x_2 \\ -\lambda x_2 - \omega^2 x_1 + u(t) \end{pmatrix} = f(x,u).$$

The objective is to minimize

$$J(u) = \tau(u)$$

where $\tau(u)$ is the first time that $x(t) = 0$ (i.e. $\theta(t) = \dot{\theta}(t) = 0$). We see that this is a *free-time, fixed-endpoint* problem.

Example 4.3 (Moon Lander)**.** How do we land a spacecraft on the moon surface, so as to use the least amount of fuel (Fig. 9)? To model this problem, we consider these variables:

- $h(t)$, the height of the spacecraft at time t;
- $v(t) = \dot{h}(t)$, the velocity of the spacecraft;
- $m(t)$, the mass of the spacecraft;
- $u(t)$, the thrust at time t.

$u(t)$ is our control, and we assume $0 \le u(t) \le 1$. If $u(t) = 0$, then we turn off the engine and let the spacecraft do free fall, and $u(t) = 1$ means we apply the maximal thrust against gravity. As the we burn the fuel, the mass $m(t)$ of the spacecraft will change over time. We assume the rate of change is inversely proportional to $u(t)$. The motion of the spacecraft according to Newton's second law is (take the upward direction as the positive direction)

$$m(t)\ddot{h}(t) = -gm(t) + u(t).$$

We can write the dynamics as

$$\begin{cases} \dot{v}(t) = -g + \dfrac{u(t)}{m(t)} \\ \dot{h}(t) = v(t) \\ \dot{m}(t) = -ku(t). \end{cases}$$

Let $x(t) = (v(t), h(t), m(t))^T$. We can summarize the dynamics as $\dot{x}(t) = f(x(t), u(t))$, with $x(0) = x_0$ given. Moreover, we have additional physical constraints: $h(t) \geq 0$ and $m(t) \geq 0$ for all t.

The objective is to minimize the amount of fuel used, or in other word maximize the remaining fuel when we landed. So we want to minimize

$$J(u) = -m(\tau)$$

where τ is the first time that $h(t) = v(t) = 0$.

Example 4.4 (Rocket Car)**.** Imagine we have a railroad rocket car with engines on both sides (Fig. 10). We introduce the variables

- $q(t)$, the car's position at time t;
- $v(t) = \dot{q}(t)$, the car's velocity at time t;
- $u(t)$, the thrust from the rockets,

where $-1 \leq u(t) \leq 1$.

Given initial position and velocity, we want to figure out how to fire the rocket so as to bring the car to the origin with zero velocity in a minimal amount of time. Assuming the car has mass m, the law of motion is

$$m\ddot{q}(t) = u(t).$$

We let $x(t)$ denote $(q(t), v(t))^T$, and normalize m to 1, so that we can write the dynamics as

$$\begin{cases} \dot{x}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t) \\ x(0) = x_0 = (q_0, v_0)^T. \end{cases}$$

Our objective is to minimize the functional

$$J(u) = \tau(u)$$

where $\tau(u)$ is the first time that $q(t) = v(t) = 0$.

For this linear system, the optimal control will be a bang-bang control, explained in Section 4.3. Meanwhile, let's try to see what the dynamics are when $u = 1$ or $u = -1$.

- When $u \equiv 1$, the dynamics become

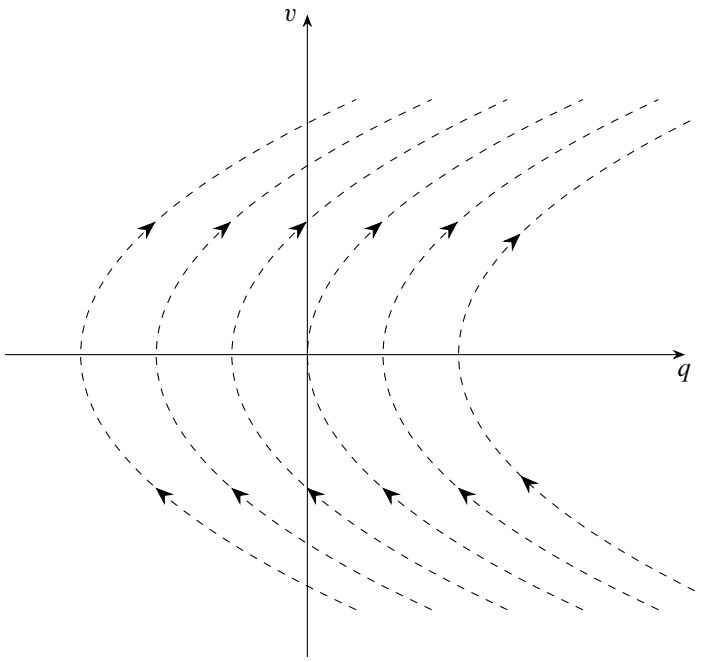
  $$\begin{cases} \dot{q} = v \\ \dot{v} = 1. \end{cases}$$

  We then have

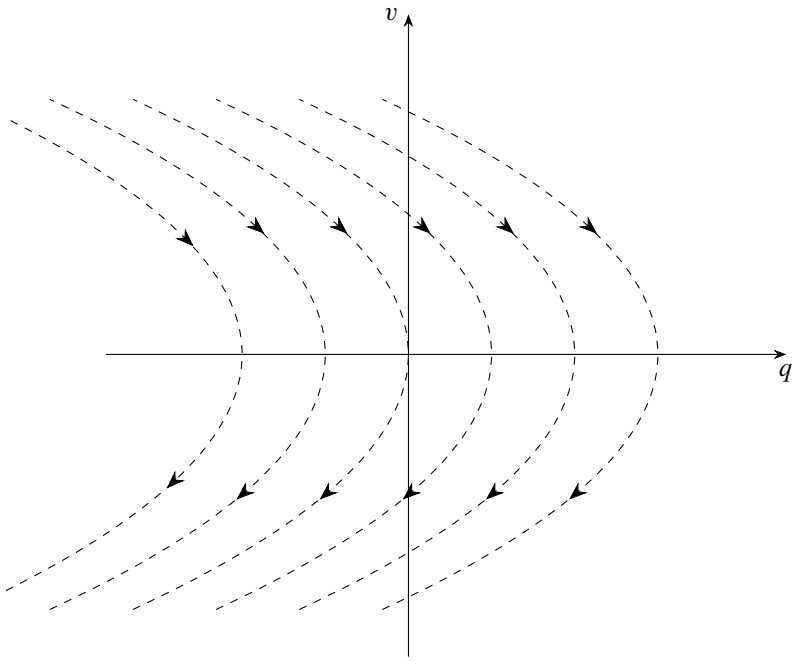
  $$v\dot{v} = \dot{q} \quad \Rightarrow \quad \frac{1}{2}(v^2)' = q' \quad \Rightarrow \quad \frac{1}{2}v^2(t) - \frac{1}{2}v(t_0) = q(t) - q(t_0)$$

  for some t_0 for which $u(t_0) = 1$. We see that as long as the control is set for $u \equiv 1$, the trajectory of the state $x = (q, v)^T$ stays on the curve $v^2 = 2q + b$ for some constant b. See Fig. 11a.

Figure 10: The rocket car problem.



(a) Illustration of the rocket car problem. When the control is $u \equiv 1$, the state trajectory $x = (q, v)^T$ stays on the curve $v^2 = 2q + b$ for some constant b. Since we go toward the positive direction, q and v increase toward the positive direction.



(b) Illustration of the rocket car problem. When the control is $u \equiv -1$, the state trajectory $x = (q, v)^T$ stays on the curve $v^2 = -2q + c$ for some constant c. Since we go toward the negative direction, q and v increase toward the negative direction.

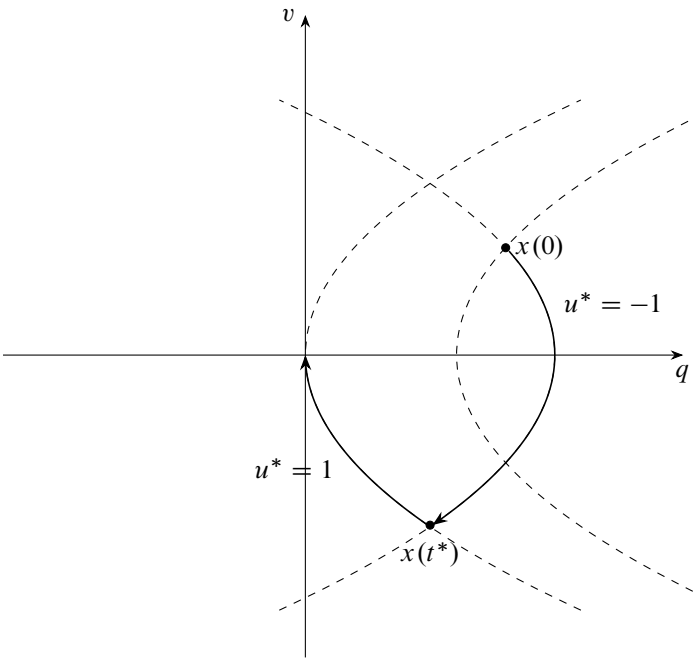


Figure 12: Geometric solution of the rocket car problem.

- When $u \equiv -1$, we have
$$\begin{cases} \dot{q} = v \\ \dot{v} = -1. \end{cases}$$
and hence $\frac{1}{2}(v^2)' = -q'$. Let t_1 belong to an interval where $u \equiv -1$ and integrate, we can
$$v^2(t) = -2q(t) + (2q(t_1) - v^2(t_1)).$$
Thus the state trajectory follows the curve $v^2 = -2q + c$ for some constant c. See Fig. 11b.

Suppose we start at $x(0)$ in Fig. 12, so that the car initially has positive velocity and position. We first firing the engine backward, to steer it toward point $x(t^*)$. Then at time t^* we give a positive force on the car, to stop it at the origin.

4.2 Hamiltonian Mechanics

Calculus of variations and optimal control theory have their connections with Hailtonian mechanics and classical mechanics. We define the *momentum* as $p := L_{\dot{x}}(t, x, \dot{x})$ and we define the *Hamiltonian* as

$$H(t, x, \dot{x}, p) := p \cdot \dot{x} - L(t, x, \dot{x}).$$

Note that

$$\dot{x}(t) = H_p(t, x, \dot{x}, p), \tag{19}$$
$$\dot{p}(t) = \frac{d}{dt} L_{\dot{x}}(t, x, \dot{x}) = L_x(t, x, \dot{x}) = -H_x(t, x, \dot{x}). \tag{20}$$

These two equations

$$\boxed{\dot{x} = H_p, \quad \dot{p} = -H_x} \tag{21}$$

are known as *Hamilton's canonical equations*. Note that the partial derivative of H with respect to $\dot{x}$ is zero:

$$H_{\dot{x}} = p - L_{\dot{x}}(t, x, \dot{x}) = 0.$$

To see the meaning of the Hamiltonian in physics, L=let $x(t) = (x_1(t), x_2(t), x_3(t))^T \in \mathbb{R}^3$ be the position of a particle, $\dot{x}(t)$ the velocity and $U = U(x)$ the potential energy, so that $-U_x$ is the force. Let

$$L(t, x, \dot{x}) = \frac{1}{2} m \|\dot{x}\|^2 - U(x)$$

be the difference between kinetic energy and potential energy. Note that $p = L_{\dot{x}} = m\dot{x}$ is the momentum. The Euler-Lagrange equation yields

$$\frac{d}{dt} L_{\dot{x}} = L_x \quad \Rightarrow \quad \frac{d}{dt}(m\dot{x}) = -U_x.$$

We have recovered Newton's second law: the derivative of the momentum is the force.

Hamilton's principle of least action states that trajectories $x(t)$ of mechanical system are extremals of the functional

$$\int_{t_0}^{t_1} \left(\frac{1}{2} m \|\dot{x}\|^2 - U(x) \right) dt,$$

which is called the *action integral*. If the potential is 0, then the trajectories are extremals of the functional $\int_{t_0}^{t_1} \left(\frac{1}{2} m \|\dot{x}\|^2 \right) dt$, which are straight lines.

Newton's second law, the derivative of momentum equals to the force, is a differential statement that holds pointwise in time, while the principle of least action is a statement about the entire trajectory. They are equivalent: if the action integral is minimized, then every small piece of the trajectory must also deliver minimal action. In the limit as the length approaches zero, we recover the differential statement.

The Hamiltonian is

$$H = p \cdot \dot{x} - L(t, x, \dot{x}) = \frac{1}{2} m \|\dot{x}\|^2 + U(x) = \text{ kinetic energy } + \text{ potential energy},$$

which is the total mechanical energy.

4.3 Bang-bang Principle

We address the following basic question: *given an initial point x_0 and a target set $S \subset \mathbb{R}^n$, does there exist a control steering the system to S in finite time?* Here we will not concern ourselves with any payoff function. We take $S = \{0\}$ and we only investigate linear systems

$$\begin{cases} \dot{x}(t) = Mx(t) + Nu(t) \\ x(0) = x_0 \end{cases}$$

where the state $x(t)$ is n-dimensional, the control $u(t)$ is m-dimensional, M is an $(n \times n)$ matrix and N is an $(n \times m)$ matrix. We assume the control lies in $U = [-1, 1]^m$.

Recall how we solve for the ODE

$$\begin{cases} \dot{x}(t) = a(t)x(t) + b(t) \\ x(0) = x_0 \end{cases}$$

where $x(t)$ is a real-valued function. We let $A(t) = \int_0^t a(s)ds$, and then multiply both sides by $e^{-A(t)}$,

$$\begin{aligned}
\dot{x}(t) - a(t)x(t) &= b(t) \\
e^{-A(t)}\dot{x}(t) - a(t)e^{-A(t)}x(t) &= e^{-A(t)}b(t) \\
\frac{d}{dt}\left[e^{-A(t)}x(t)\right] &= e^{-A(t)}b(t) \\
e^{-A(t)}x(t) &= \int_0^t e^{-A(s)}b(s)ds + x_0 \\
&\Downarrow \\
x(t) &= e^{A(t)}x_0 + e^{A(t)}\int_0^t e^{-A(s)}b(s)ds.
\end{aligned}$$

In an entirely similar way, when $x : \mathbb{R} \to \mathbb{R}^n$ is multi-dimensional, the solution to the linear control system

$$\begin{cases} \dot{x}(t) = Mx(t) + Nu(t) \\ x(0) = x_0 \end{cases}$$

is

$$x(t) = e^{tM}x_0 + e^{tM}\int_0^t e^{-sM}Nu(s)ds \tag{22}$$

where e^{tM} is defined as $\sum_{k=0}^{\infty} t^k M^k/k!$.

Definition 4.5 (Reachable set)**.** We define the *reachable set for time t* to be the set $\mathcal{C}(t)$ that consists of initial points x_0 for which there exists a control such that $x(t) = 0$. We define the *reachable set* $\mathcal{C}$ to be

$$\mathcal{C} = \bigcup_{t \geq 0} \mathcal{C}(t).$$

Observe that $x_0 \in \mathcal{C}(t)$ if and only if there is a control $u \in \mathcal{U}$ such that $x(t) = 0$, if and only if

$$0 = x(t) = e^{tM}x_0 + e^{tM}\int_0^t e^{-sM}Nu(s)ds$$

for some control $u \in \mathcal{U}$, if and only if

$$x_0 = -\int_0^t e^{-sM}Nu(s)ds \tag{23}$$

for some control $u \in \mathcal{U}$. From this we can get some information about the geometry of the reachable set from the geometry of the control set U. Recall a set A is symmetric if $x \in A \Rightarrow -x \in A$, and convex if for all $\lambda \in [0,1]$, $x, x' \in A \Rightarrow \lambda x + (1-\lambda)x' \in A$.

Proposition 4.6. The reachable set $\mathcal{C}$ is symmetric and convex. Also, if $x_0 \in \mathcal{C}(t')$, then $x_0 \in \mathcal{C}(t)$ for all $t \geq t'$.

Proof. Symmetry is obvious: if $x_0 \in \mathcal{C}(t)$, then according to Eq. (23), $x_0 = -\int_0^t e^{-sM}Nu(s)ds$ for some control $u(\cdot) \in \mathcal{U}$. Therefore $-x_0 = -\int_0^t e^{-tM}N(-u(s))ds$ and $-u \in \mathcal{U}$ since the set U is symmetric, so that $x_0 \in \mathcal{C}(t)$.

If $S = \{0\}$ can be reached in time t' via some control u', then it can certainly be reached at any later time $t \geq t'$: just set the control u to be equal to u' before time t' and zero after time t'. Convexity of $\mathcal{C}$ follows from convexity of U. □

Recall our control set is $U = [-1,1]^m$. A bang-bang control is a control that only take extreme values (1 or -1).

Definition 4.7. A control $u = (u_1, \ldots, u_m)^T \in \mathcal{U}$ is called *bang-bang* if $u_i(t) = 1$ or -1 for all t and all $i = 1, \ldots, m$

The bang-bang principle (Theorem 4.9) states that, if there is some control that steers the initial state x_0 of a linear system to origin, there must exist a bang-bang control that can also steers x_0 to origin. Thus, to search for the desired control, we can first search for a bang-bang control. If there does not exist a bang-bang control, then there is no hope for finding other ones. To prove it we first need the Krein-Milman theorem (Theorem 4.8) below.

Theorem 4.8 (Krein-Milman Theorem)**.** Let X be a locally convex topological vector space, K a nonempty, compact, convex subset of X. Then

- K has at least one extreme point.
- K is the closure of the convex hull of its extreme points.

Proof. See page 125 of Lax 2002. □

Theorem 4.9 (Bang-bang principle)**.** Let $t > 0$ and suppose $x_0 \in \mathcal{C}(t)$ for the system

$$\begin{cases} \dot{x}(t) = Mx(t) + Nu(t) \\ x(0) = x_0. \end{cases}$$

Then there exists a bang-bang control u which steers x_0 to $S = \{0\}$ at time t.

Proof. Define

$$\mathbb{K} = \{u \in \mathcal{U} \mid u \text{ steers } x_0 \text{ to } 0 \text{ at time } t\}.$$

$\mathbb{K}$ is nonempty since $x_0 \in \mathcal{C}(t)$. It is convex since $U = [-1, 1]^m$ is convex. Since $\mathcal{U}$ is weak* compact according to the Banach-Alaoglu theorem, we prove $\mathbb{K}$ is closed in $\mathcal{U}$ so that it is also weak* compact. Let $\{u_n\}_{n=1}^{\infty} \subset \mathbb{K}$ be a sequence such that $u_n \to u \in \mathcal{U}$. We need to show $u \in \mathbb{K}$. From $u_n \in \mathbb{K}$ we have

$$x_0 = -\int_0^t e^{-tM} N u_n(s) ds \quad \forall n \in \mathbb{N}$$

by Eq. (23) and according to the definition for weak* convergence,

$$-\int_0^t e^{-tM} N u_n(s) ds \quad \to \quad -\int_0^t e^{-sM} N u(s) ds \qquad \text{as } n \to \infty.$$

Thus we have

$$x_0 = \lim_{n\to\infty} -\int_0^t e^{-tM} N u_n(s) ds = -\int_0^t e^{-sM} N u(s) ds$$

and so $u \in \mathbb{K}$. This proves that $\mathbb{K}$ is convex and compact. Then we can apply Theorem 4.8 to conclude that there exists an extreme point $u^* \in \mathbb{K}$. We next show that this extreme point is indeed a bang-bang control.

We must show $|u_i^*(s)| = 1, \forall i = 1, \ldots, m$ for $0 \le s \le t$ almost everywhere. Suppose to the contrary that for some index i and a subset $E \subseteq [0, t]$ of positive measure such that $|u_i^*(s)| < 1$ for $s \in E$. In fact, there exist a number $\epsilon > 0$ and a subset $F \subseteq E$ of positive measure such that

$$|u_i^*(s)| \le 1 - \epsilon$$

for $s \in F$. Choose any control $v \in \mathcal{U}, v \not\equiv 0$ such that $v_j = 0$ for any $j \neq i$ and $\int_F e^{-tM} N v(s) ds = 0$. Define

$$u_1 := u^* + \epsilon v$$
$$u_2 := u^* - \epsilon v.$$

We claim that $u_1, u_2 \in \mathbb{K}$. To see this, first observe

$$\int_0^t e^{-tM} N u_1(s) ds = -\int_0^t e^{-tM} N u^*(s) ds - \epsilon \int_0^t e^{-tM} N v(s) ds = x_0 - 0 = x_0,$$
$$\int_0^t e^{-tM} N u_2(s) ds = -\int_0^t e^{-tM} N u^*(s) ds + \epsilon \int_0^t e^{-tM} N v(s) ds = x_0 + 0 = x_0.$$

Also, on set $[0,t] \setminus F$ we have $u_1 = u^*$ and on F we have $|u_1| \le |u^*| + \epsilon|v| \le 1 - \epsilon + \epsilon = 1$, and similarly for u_2. Hence $u_1, u_2 \in \mathbb{K}$. We found two points in $\mathbb{K}$, distinct from u^*, such that

$$u^* = \frac{1}{2} u_1 + \frac{1}{2} u_2,$$

a contradiction to u^* being an extreme point in $\mathbb{K}$. This concludes the proof that u^* is indeed a bang-bang control. □

4.4 Linear Time-Optimal Control

In this subsection, we consider a specific problem within the setting of linear systems: time-optimal control. The dynamics are as before

$$\begin{cases} \dot{x}(t) = Mx(t) + Nu(t) \\ x(0) = x_0. \end{cases} \tag{24}$$

Recall in previous subsection we did not specify the objective function. Here we investigate minimization of

$$J(u) = \tau(u) \tag{25}$$

where $\tau(u)$ is the first time that the dynamic system hits $S = \{0\}$. This is free-time, fixed-endpoint problem. We first show the existence of optimal bang-bang control (if there is optimal control at all), then we state the maximum principle (Theorem 4.11) for the problem.

Theorem 4.10. Let $x_0 \in \mathbb{R}^n$. There exists an optimal bang-bang control u^*.

Proof. Let $\tau^* = \inf_{x_0 \in \mathcal{C}(t)} t$. We want to show that $x_0 \in \mathcal{C}(\tau^*)$, namely there exists an optimal control u^* steering x_0 to 0 at time τ^*.

Choose a decreasing sequence $t_1 \ge t_2 \ge \cdots$ such that $x_0 \in \mathcal{C}(t_n)$ and $t_n \to \tau^*$. Since $x_0 \in \mathcal{C}(t_n)$, by Eq. (23) there exists a control u_n such that

$$x_0 = -\int_0^{t_n} e^{-tM} N u_n(s) ds.$$

If necessary, redefine u_n to be 0 for $t_n \le s \le t_1$. Since $\mathcal{U}$ is weak* compact, in the sequence $\{u_n\}$ there exists a subsequence $\{u_{n_k}\}$ and a control $u^* \in \mathcal{U}$ such that $u_{n_k} \to u^*$ as $n_k \to \infty$. We assert that u^* is an optimal control. We have $u^*(s) = 0$ for $s \ge \tau^*$, and

$$x_0 = -\int_0^{t_{n_k}} e^{-tM} N u_{n_k}(s) ds = -\int_0^{t_1} e^{-tM} N u_{n_k}(s) ds$$

since $u_{n_k} = 0$ for $s \ge t_{n_k}$. Let $n_k \to \infty$:

$$x_0 = \lim_{n_k \to \infty} \left(-\int_0^{t_1} e^{-tM} N u_{n_k}(s) ds \right) = -\int_0^{t_1} e^{-tM} N u^*(s) ds = -\int_0^{\tau^*} e^{-tM} N u^*(s) ds$$

since $u^*(s) = 0$ for $s \ge \tau^*$. Hence $x_0 \in \mathcal{C}(\tau^*)$, and therefore u^* is optimal. By Theorem 4.9 there in fact exists an optimal bang-bang control. □

Theorem 4.11 (Potryagin Maximum Principle for Linear Time-Optimal Control)**.** Let $u^* \in \mathcal{U}$ be the optimal control for the problem Eqs. (24) and (25). Then there exists an $n \times 1$ nonzero vector h such that

$$h^T e^{-tM} N u^*(t) = \max_{u \in U} \{h^T e^{-tM} N u\} \qquad \text{for all } 0 \le t \le \tau^*. \tag{26}$$

Proof. We define $K(t, x_0)$ to be the set of states that can be reached at time t via some control with initial condition $x(0) = x_0$. By Eq. (22), $x \in K(t, x_0)$ if and only if

$$x = e^{tM} x_0 + e^{tM} \int_0^t e^{-sM} N u(s) ds \tag{27}$$

for some control $u \in \mathcal{U}$. It is easy to deduce convexity and closedness of $K(t, x_0)$ from convexity and weak* compactness of $\mathcal{U}$.

Note that for $t_1 \le t_2$ we have $K(t_1, x_0) \subseteq K(t_2, x_0)$, namely if we can reach some $x \in K(t_1, x_0)$, in time t_1, then we can certainly reach x using more time. Let $\tau^* = \min J(u)$ denote the minimum time it takes to steer the initial state x_0 to 0, using the optimal control u^*. Then $0 \in \partial K(\tau^*, x_0)$, i.e. the final state 0 must lie on the boundary of the set $K(\tau^*, x_0)$. To see this, note that if 0 lies outside of $K(\tau^*, x_0)$, then this means we cannot reach 0 at time τ^*; if it lies in the interior then we can reach 0 in a shorter time.

Since $0 \in \partial K(\tau^*, x_0)$ and $K(\tau^*, x_0)$ is convex, by the supporting hyperplane theorem there exists some $g \neq 0$ such that

$$g \cdot x \le 0 \quad \text{for all } x \in K(\tau^*, x_0).$$

From Eq. (27), for $0 \in K(\tau^*, x_0)$ and an arbitrary $x \in K(\tau^*, x_0)$ we have

$$x = e^{\tau^* M} x_0 + e^{\tau^* M} \int_0^{\tau^*} e^{-sM} N u(s) ds$$
$$0 = e^{\tau^* M} x_0 + e^{\tau^* M} \int_0^{\tau^*} e^{-sM} N u^*(s) ds$$

for some control $u \in \mathcal{U}$. Then

$$g^T \left(e^{\tau^* M} x_0 + e^{\tau^* M} \int_0^{\tau^*} e^{-sM} N u(s) ds \right) \le 0 = g^T \left(e^{\tau^* M} x_0 + e^{\tau^* M} \int_0^{\tau^*} e^{-sM} N u^*(s) ds \right).$$

Define $h^T := g^T e^{\tau^* M}$. Then

$$\int_0^{\tau^*} h^T e^{-sM} N u(s) ds \le \int_0^{\tau^*} h^T e^{-sM} N u^*(s) ds \tag{28}$$

for some control $u \in \mathcal{U}$. This must imply Eq. (26), for if

$$h^T e^{-sM} N u^*(s) < \max_{u \in U} \{h^T e^{-sM} N u\}$$

for $s \in E$ where $E \subset [0, \tau^*]$ has positive measure, then we can design a new control

$$\hat{u}(s) = \begin{cases} u^*(s) & s \notin E \\ u(s) & s \in E \end{cases}$$

where $u(s)$ is defined so that

$$h^T e^{-sM} N u(s) = \max_{u \in U} \{h^T e^{-sM} N u\}.$$

We then have

$$\int_0^{\tau^*} h^T e^{-sM} N \hat{u}(s) ds > \int_0^{\tau^*} h^T e^{-sM} N u^*(s) ds,$$

a contradiction to Eq. (28). □

To see how Theorem 4.11 is related to the general maximum principle (Theorem 5.1), the Hamiltonian is effectively $H(x, p, u) = (Mx + Nu) \cdot p$. We can define the costate as $p^*(t)^T = h^T e^{-tM}$. Then from Theorem 4.11 we have

$$\begin{aligned} H(x^*(t), p^*(t), u^*(t)) &= p^*(t)^T \left(Mx^*(t) + Nu^*(t)\right) \\ &= \max_{u \in U} \left\{p^*(t)^T \left(Mx^*(t) + Nu^*(t)\right)\right\} = \max_{u \in U} H\left(x^*(t), p^*(t), u\right). \end{aligned}$$

The maximum principle offers us a *necessary condition* for finding the optimal solution. We may not be able to immediately find the optimal control by solving for Eq. (26), but it can give use useful information for the solution. We next use some examples to illustrate how the maximum principle can be applied.

Example 4.12 (Rocket Car)**.** Recall the rocket car problem from Example 4.4. This is a free time, fixed endpoint problem where we want to steer the car to the origin using shortest possible time. The dynamic is

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t)$$

for $x(t) = (x_1(t), x_2(t))^T$, $U = [-1, 1]$. According to the maximum principle, there exists $h \neq 0$ such that

$$h^T e^{-tM} N u^*(t) = \max_{u \in U} \{h^T e^{-tM} N u\}$$

where $M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $N = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$. To use the maximum principle, we meed to compute e^{-tM}. In this case it is simple:

$$M^0 = I, \quad M = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad M^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and so $M^k = 0$ for all $k \geq 2$. Consequently

$$\begin{aligned} e^{-tM} &= I - tM = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix}, \\ e^{-tM} N &= \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -t \\ 1 \end{pmatrix}, \\ h^T e^{-tM} N &= \begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} -t \\ 1 \end{pmatrix} = -th_1 + h_2. \end{aligned}$$

The maximum principle says

$$(-th_1 + h_2)u^*(t) = \max_{|u| \leq 1} \{(-th_1 + h_2)u\}$$

so

$$u^*(t) = \text{sign}(-th_1 + h_2).$$

Therefore the optimal control switches at most once. Thus, the geometric solution we obtained earlier is indeed optimal.

Example 4.13 (Control of a Vibrating Spring)**.** Consider a ball with unit mass hanging from a spring (Fig. 13). Our goal is to apply a control $u(t) \in [-1, 1]$ so as to bring the ball to stop in minimal time. This is again a free time, fixed endpoint problem. The dynamic is

$$\ddot{x} + x = u,$$

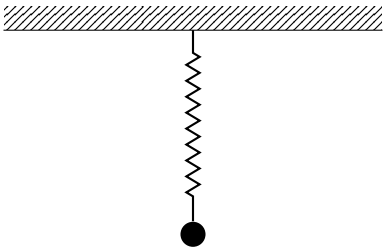


Figure 13: Illustration of the vibrating spring problem.

namely the forces it experiences (besides gravity and the hanging force) are the spring force and the control. We write the dynamic as

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t) =: Mx(t) + Nu(t).$$

where $x(t) = (x_1(t), x_2(t))^T$. Again we need to compute the matrix exponential e^{-tM}. Observe that

$$\begin{aligned} M^0 &= I, \\ M^1 &= M, \\ M^2 &= -I, \\ M^3 &= -M, \\ M^4 &= I, \end{aligned}$$

so if k is even then M^k oscillates between I and $-I$, while if k is odd then M^k oscillates between M and $-M$. We thus have

$$\begin{aligned} e^{tM} &= I + tM + \frac{t^2}{2!}M^2 + \cdots \\ &= (1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \cdots)I + (t - \frac{t^3}{3!} + \frac{t^5}{5!} - \cdots)M \\ &= \cos t I + \sin t M \\ &= \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}. \end{aligned}$$

So

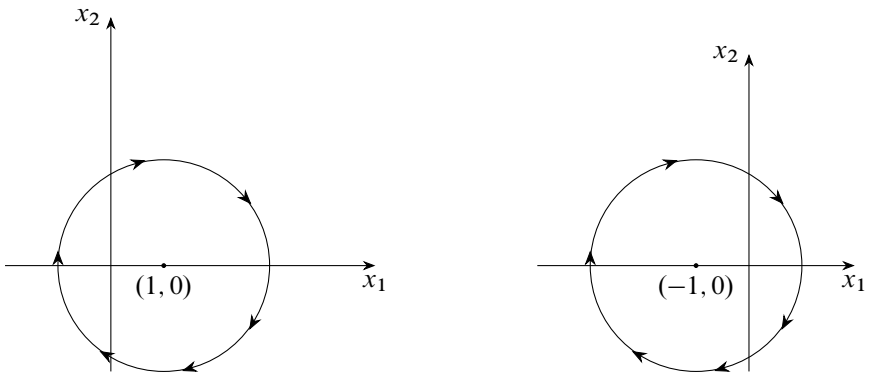
$$\begin{aligned} e^{-tM} &= \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}, \\ e^{-tM}N &= \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix}, \\ h^T e^{-tM} N &= \begin{pmatrix} h_1 & h_2 \end{pmatrix} \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} = -h_1 \sin t + h_2 \cos t. \end{aligned}$$

From Eq. (26), for each time t we have

$$(-h_1 \sin t + h_2 \cos t)u^*(t) = \max_{|u| \le 1} \{(-h_1 \sin t + h_2 \cos t)u\}$$

and so

$$u^*(t) = \text{sign}(-h_1 \sin t + h_2 \cos t).$$



(a) When $u \equiv 1$.

(b) When $u \equiv -1$.

Figure 14: State trajectory of the vibrating spring problem under the control $u \equiv 1$ and $u \equiv -1$ respectively.

If we let $\|h\| = 1$, and choose δ such that $-h_1 = \cos\delta, h_2 = \sin\delta$, then

$$u^*(t) = \text{sign}(\cos\delta \sin t + \sin\delta \cos t) = \text{sign}(\sin(t + \delta)).$$

We see that u^* switches between -1 and 1 every π units of time.

A geometric interpretation of the solution is also possible. When $u \equiv 1$, the dynamic is

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x_2} = -x_1 + 1. \end{cases}$$

The trajectory of the state $(x_1, x_2)^T$ follows a circle with center $(1, 0)$:

$$\frac{d}{dt}\left[(x_1(t) - 1)^2 + x_2^2(t)\right] = 2(x_1(t) - 1)\dot{x}_1(t) + 2x_2(t)\dot{x}_2(t) = 0.$$

See Fig. 14a. Similarly, when $u \equiv -1$ the trajectory follows a circle with center $(-1, 0)$ (Fig. 14b).

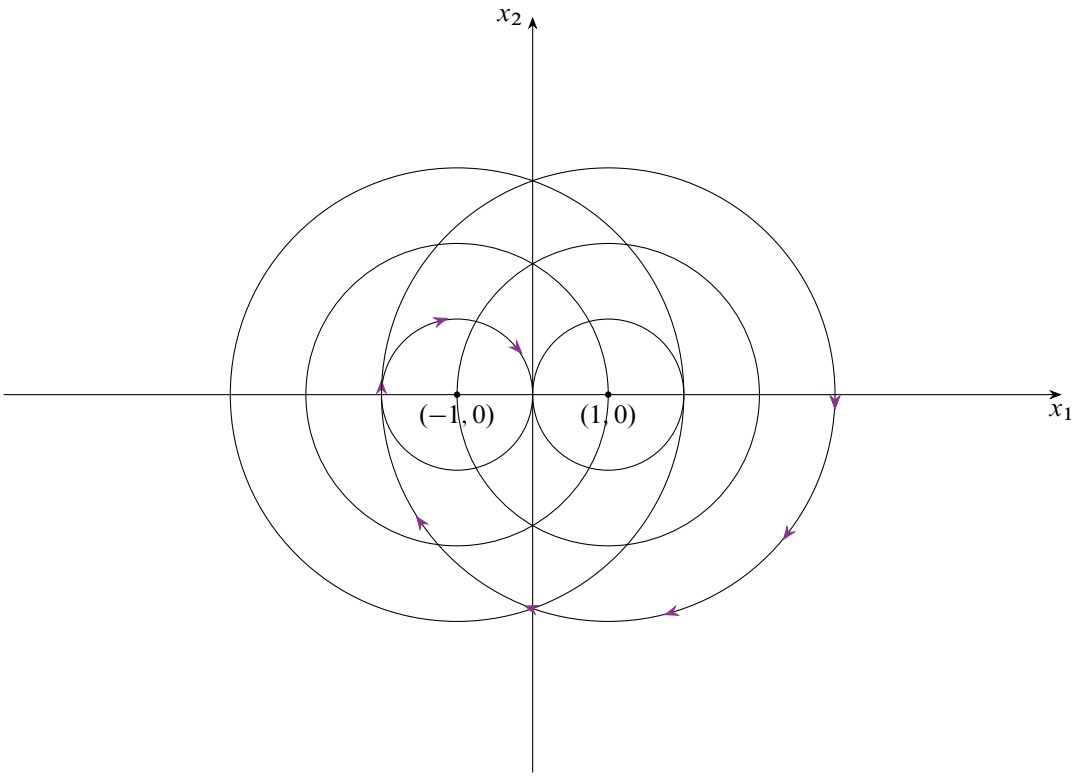


Figure 15: Geometric solution of the vibrating spring problem. Given an initial state $x(0) = (x_1(0), x_2(0))^T$, we want to steer it toward the origin. The optimal control corresponds to switch from one circle to the other.

5 The Pontraygin Maximum Principle

In this section we introduce the Pontraygin Maximum Principle. It provides us *necessary* conditions for finding optimal controls.

We define the Hamiltonian as

$$H(x, u, p) = p \cdot f(x, u) - L(x, u).$$

5.1 Free Time, Fixed Endpoint Problem

The system is

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

The target is $S = \{x_1\}$ with $x_1 \in \mathbb{R}^n$. Objective:

$$J(u) := \int_0^\tau L(x(t), u(t))dt$$

where $L : \mathbb{R}^n \times U \to \mathbb{R}$ is the running cost, and $\tau = \tau(u) \leq \infty$ is the first time the state hits the target point x_1.

Theorem 5.1 (Pontryagin Maximum Principle for Free Time, Fixed Endpoint Problem)**.** Assume u^* is the optimal control and x^* is the corresponding trajectory. Then there exists a function $p^* : [0, \tau^*] \to \mathbb{R}^n$, called the *costate*, such that

$$\dot{x}^*(t) = \nabla_p H(x^*(t), p^*(t), u^*(t)), \tag{29}$$

$$\dot{p}^*(t) = -\nabla_x H(x^*(t), p^*(t), u^*(t)), \tag{30}$$

and

$$H(x^*(t), p^*(t), u^*(t)) = \max_{u \in U} H\left(x^*(t), p^*(t), u\right), \quad \forall t \in [0, \tau^*]. \tag{31}$$

Also

$$H(x^*(t), p^*(t), u^*(t)) \equiv 0. \tag{32}$$

Here τ^* denote the first time the state $x^*(t)$ hits the target point x_1.

5.2 Fixed Time, Free Endpoint Problem

The system is

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

Objective:

$$J(u) := \int_0^T L(x(t), u(t))dt + K(x(T)).$$

Theorem 5.2 (Pontryagin Maximum Principle for Fixed Time, Free Endpoint Problem)**.** Assume u^* is the optimal control and x^* is the corresponding trajectory. Then there exists a function $p^* : [0, T] \to \mathbb{R}^n$, called the *costate*, such that

$$\dot{x}^*(t) = \nabla_p H(x^*(t), p^*(t), u^*(t)), \tag{33}$$

$$\dot{p}^*(t) = -\nabla_x H(x^*(t), p^*(t), u^*(t)), \tag{34}$$

and

$$H(x^*(t), p^*(t), u^*(t)) = \max_{u \in U} H\left(x^*(t), p^*(t), u\right), \quad \forall t \in [0, T]. \tag{35}$$

In addition the mapping

$$t \to H(x^*(t), p^*(t), u^*(t)) \tag{36}$$

is constant. Finally we also have the terminal condition

$$p^*(T) = \nabla K(x^*(T)).$$

5.3 Applications of the Maximum Principle

Example 5.3 (Control of Production and Consumption)**.** Let's see how we can apply the maximum principle to the production and consumption problem (Example 4.1). For simplicity we take $k = 1$. This is a fixed time, free endpoint problem. The dynamic is

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

and the functional we want to minimize is

$$J(u) = \int_0^T -C(t)dt = \int_0^T -(1 - u(t))x(t)dt.$$

The Hamiltonian is then

$$H(x, p, u) = p \cdot f(x, u) - L(x, u) = p \cdot xu + (1 - u)x = x + ux(p - 1).$$

The two canonical equations are

$$\begin{aligned} \dot{x} &= H_p = ux, \\ \dot{p} &= -H_x = -1 - u(p - 1). \end{aligned}$$

Since there is no terminal cost, $p(T) = 0$. Finally, according to the maximum principle we need to solve the following equation

$$H(x(t), p(t), u(t)) = \max_{0 \le u \le 1} \{x(t) + ux(t)(p(t) - 1)\}.$$

If follows that the optimal control takes the form

$$u(t) = \begin{cases} 1 & \text{if } p(t) > 1, \\ 0 & \text{if } p(t) \le 1. \end{cases}$$

To solve for the optimal control, we need to solve for p first:

$$\begin{cases} \dot{p}(t) = -1 - u(t)(p(t) - 1), \\ p(T) = 0. \end{cases}$$

If $p(t) \le 1$, then $u(t) = 0$, so the equation becomes $\dot{p}(t) = -1$. The solution is then $p(t) = T - t$. Consequently

$$T - t \le 1 \quad \Rightarrow \quad t \le T - 1.$$

If $p(t) \ge 1$, then $u(t) = 1$. The ODE is then

$$\begin{cases} \dot{p}(t) = -p(t), \\ p(T - 1) = 1. \end{cases}$$

The solution is $p(t) = e^{T-1-t}$. Consequently

$$p(t) = e^{T-1-t} > 1 \quad \Rightarrow \quad t \leq T-1.$$

Thus the optimal control is

$$u^*(t) = \begin{cases} 1 & \text{if } 0 \leq t \leq T-1, \\ 0 & \text{if } T-1 \leq t \leq T. \end{cases}$$

This confirms our guess in Example 4.1.

Example 5.4 (Linear-Quadratic Regulator)**.** Let's see how we can use the maximum principle to solve for the LQ problem. For simplicity we take $n = m = 1$. The dynamic is linear:

$$\begin{cases} \dot{x}(t) = x(t) + u(t) \\ x(0) = x_0 \end{cases}$$

and we want to minimize a quadratic cost functional

$$J(u) = \int_0^T x^2(t) + u^2(t)dt.$$

For this problem the values of the controls are not constrained, i.e. $U = \mathbb{R}$. The Hamiltonian is

$$H(x, p, u) = p \cdot f(x, u) - L(x, u) = p \cdot (x + u) - (x^2 + u^2).$$

Thus we need to solve for

$$H(x(t), p(t), u(t)) = \max_{u \in \mathbb{R}} \{p(t) \cdot (x(t) + u) - (x^2(t) + u^2)\}.$$

Again, since there is no terminal cost, $p(T) = 0$. The right hand side is a quadratic function of u, so setting $H_u = 0$ we get

$$H_u = -2u + p = 0 \quad \Rightarrow \quad u(t) = \frac{p(t)}{2}.$$

Thus to solve for the optimal control, we first need to solve for the canonical equations

$$\begin{cases} \dot{x}(t) = x(t) + \dfrac{p(t)}{2}, \\ \dot{p}(t) = 2x(t) - p(t), \end{cases} \tag{37}$$

with $x(0) = x_0$ and $p(T) = 0$. Write it in matrix form this is

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 1/2 \\ 2 & -1 \end{pmatrix}}_{M} \begin{pmatrix} x \\ p \end{pmatrix}.$$

The general solution is

$$\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = e^{tM} \begin{pmatrix} x_0 \\ p_0 \end{pmatrix}.$$

However, here we shall avoid calculating e^{tM}, since it is a very complicated matrix. Instead, we look for a *feedback* control of the form

$$u(t) = c(t)x(t).$$

We have

$$u(t) = c(t)x(t) = \frac{p(t)}{2} \quad \Rightarrow \quad c(t) = \frac{p(t)}{2x(t)}.$$

Define $d(t) := p(t)/x(t)$ so that $c(t) = d(t)/2$. We seek an ODE that d satisfies. Compute

$$\dot{d} = \frac{\dot{p}}{x} - \frac{p\dot{x}}{x^2}$$

and substitute Eq. (37) into above, we get

$$\dot{d} = \frac{2x - p}{x} - \frac{p}{x^2}\left(x + \frac{p}{2}\right) = 2 - d - d(1 + \frac{d}{2}) = 2 - 2d - \frac{d^2}{2}.$$

Since $p(T) = 0$, the terminal condition is $d(T) = 0$. So far we obtained a (nonlinear) first-order ODE for d:

$$\begin{cases} \dot{d} = 2 - 2d - \frac{1}{2}d^2 \quad (0 \le t \le T) \\ d(T) = 0. \end{cases}$$

This is called the *Riccati equation*. If we can solve for this equation, then we will obtain the solution for the optimal control as

$$u(t) = \frac{1}{2}d(t)x(t).$$

To solve for the Ricatti equation, a trick is to write

$$d(t) = \frac{2\dot{b}(t)}{b(t)}$$

for some function $b(\cdot)$, and then compute

$$\dot{d} = \frac{2\ddot{b}}{b} - \frac{2(\dot{b})^2}{b^2} = \frac{2\ddot{b}}{b} - \frac{d^2}{2}.$$

The Riccati equation becomes

$$\frac{2\dot{b}}{b} = \dot{d} + \frac{d^2}{2} = 2 - 2d = 2 - 2\frac{2\dot{b}}{b}$$

and consequently

$$\begin{cases} \ddot{b} = b - 2\dot{b}, \quad (0 \le t \le T) \\ \dot{b}(T) = 0, b(T) = 1. \end{cases}$$

This is a seconder oder linear ODE, which we can solve by standard techniques. We can then set $d = 2\dot{b}/b$ and derive the solution of the Riccati equation.

6 Dynamic Programming

Fix a terminal time $T > 0$ and consider the dynamics

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

Suppose our objective is to *minimize* the objective function (Eq. (18))

$$J(u) := \int_0^T L(t, x(t), u(t))dt + K(T, x(T)).$$

The idea of dynamic programming is to consider the family of minimization problems associated with the cost functionals

$$J(t, x, u) = \int_t^T L(s, x(s), u(s))ds + K(x(T))$$

where $t \in [0, T)$ and $x \in \mathbb{R}^n$. $J(t, x, u)$ is the cost if we start from time t, with state x at time t, and use the control u. We define the *value function* on $[0, T] \times \mathbb{R}^n$ as

$$V(t, x) := \inf_{u|_{[t,T]}} J(t, x, u). \tag{38}$$

The notation $u|_{[t,T]}$ means the restriction of u to the time interval $[t, T]$. This value function is the optimal *cost-to-go* from the point (t, x). In some applications we are interested in finding the optimal control, but in other situations our first goal is to obtain $V(0, x_0)$, the optimal value at the starting point. Dynamic programming aims to solve for the *whole* value function $V(t, x)$ for any t and x. It is clear that the value function must satisfy the boundary condition

$$V(T, x) = K(x) \quad \forall x \in \mathbb{R}^n.$$

In particular, if there is no terminal cost, then we should have $V(T, x) = 0$ for all $x \in \mathbb{R}^n$. The ***principle of optimality*** states that, for every $(t, x) \in [0, T) \times \mathbb{R}^n$ and every $\Delta t \in (0, T - t]$, the value function must satisfy

$$V(t, x) = \inf_{u|_{[t,t+\Delta t]}} \left\{ \int_t^{t+\Delta t} L(s, x(s), u(s))ds + V(t + \Delta t, x(t + \Delta t)) \right\} \tag{39}$$

where $x(\cdot)$ on the right-hand side is the state trajectory corresponding to the control $u|_{[t,t+\Delta t]}$ and statisfying $x(t) = x$. Namely, the optimal cost at any time t and state x, should be the infimum of the combination of cost over an interval Δt and the optimal cost at $t + \Delta t$. This equation gives us a practical guidance of finding the optimal value at the initial position $(0, x_0)$: we know the value at the end. We can then take a small step Δt back and find the value there, for which Eq. (39) is supposed to be relatively easy to solve. Having found $V(T - \Delta t, x(T - \Delta t))$, we can again take a small step back, and solve for Eq. (39) again.

Since larger problem depends on smaller problems, the philosophy of dynamic programming is to obtain complete solutions for all smaller problems and store them, so that we can make use of them to solve for the larger problem. In solving the original problem, we solve all subproblems smaller than the original problem, so we get much more than what we want to find.

We next justify Eq. (39) rigorously. Let $\overline{V}(t, x)$ denote the right side of Eq. (39). By the definition of $V(t, x)$, Eq. (38), for every $\epsilon > 0$ there exists a control u_ϵ on $[t, T]$ such that

$$J(t, x, u_\epsilon) \leq V(t, x) + \epsilon.$$

Let x_ϵ be the corresponding trajectory. We have

$$\begin{aligned} J(t,x,u_\epsilon) &= \int_t^{t+\Delta t} L(s,x_\epsilon(s),u_\epsilon(s))ds + J(t+\Delta t, x_\epsilon(t+\Delta t), u_\epsilon) \\ &\geq \int_t^{t+\Delta t} L(s,x_\epsilon(s),u_\epsilon(s))ds + V(t+\Delta t, x_\epsilon(t+\Delta t)) \geq \overline{V}(t,x). \end{aligned}$$

Thus

$$\overline{V}(t,x) \leq J(t,x,u_\epsilon) \leq V(t,x) + \epsilon$$

for arbitrary ϵ. This proves $V(t,x) \geq \overline{V}(t,x)$. On the other hand, since $V(t,x)$ is the optimal cost-to-go at time t and state x, we should have

$$V(t,x) \leq \int_t^{t+\Delta t} L(s,x(s),u(s))ds + V(t+\Delta t, x(t+\Delta t))$$

for any control u on $[t, t+\Delta t]$. Take the infimum on the right side we get $V(t,x) \leq \overline{V}(t,x)$. This proves Eq. (39).

6.1 The Hamilton-Jacobi-Bellman Equation

Let's re-write Eq. (39) in a way that is more convenient to work with for the continuous case. We first approximate the term $V(t+\Delta t, x(t+\Delta t))$ using Taylor expansion. Let $g(t) = V(t, x(t))$, then

$$g'(t) = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\frac{dx}{dt}$$

so that

$$\begin{aligned} V(t+\Delta t, x(t+\Delta t)) &= g(t+\Delta t) = g(t) + g'(t)\Delta t + o(\Delta t) \\ &= V(t,x) + V_t(t,x)\Delta t + f(t,x,u(t)) \cdot V_x(t,x)\Delta t + o(\Delta t). \end{aligned} \tag{40}$$

We also have

$$\int_t^{t+\Delta t} L(s,x(s),u(s))ds = L(t,x,u(t))\Delta t + o(\Delta t) \tag{41}$$

Substituting Eq. (40) and Eq. (41) into Eq. (39) we obtain

$$V(t,x) = \inf_{u|_{[t,t+\Delta t]}} \{L(t,x,u(t))\Delta t + V(t,x) + V_t(t,x)\Delta t + V_x(t,x) \cdot f(t,x,u(t))\Delta t + o(\Delta t)\}.$$

Cancel $V(t,x)$ on both sides, divide by Δt, and take $\Delta t \to 0$, we obtain the **Hamilton-Jacobi-Bellman (HJB) equation**

$$-V_t(t,x) = \inf_{u \in U} \{L(t,x,u) + V_x(t,x) \cdot f(t,x,u)\}. \tag{42}$$

for all $t \in [0,T)$ and all $x \in \mathbb{R}^n$. This is a partial differential equation (PDE) of V. If we can solve this partial differential equation, then we could know V, and knowledge of V may provide us with insights about the optimal control.

It is easy to check that Eq. (42) is equivalent to

$$V_t(t,x) = \sup_{u \in U} \{-V_x(t,x) \cdot f(t,x,u) - L(t,x,u)\}.$$

Recalling the definition of Hamiltonian $H(t,x,u,p) = p \cdot f(t,x,u) - L(t,x,u)$, we see this is

$$V_t(t,x) = \sup_{u \in U} H(t,x,u,-V_x(t,x)). \tag{43}$$

If u^* is optimal, and then the infimum in Eq. (42) becomes minimum, and we will have

$$\begin{aligned} -V_t(t,x^*) &= \min_{u\in U}\{L(t,x^*,u) + V_x(t,x^*)\cdot f(t,x^*,u)\} \\ &= L(t,x^*,u^*) + V_x(t,x^*)\cdot f(t,x^*,u^*) \end{aligned} \tag{44}$$

where x^* is the trajectory corresponding to u^*. Eq. (43) becomes

$$V_t(t,x^*) = \max_{u\in U} H(t,x^*,u,-V_x(t,x^*)) = H(t,x^*,u^*,-V_x(t,x^*)),$$

so we arrived at a condition analogous to the maximum principle:

$$H(t,x^*,u^*,-V_x(t,x^*)) = \max_{u\in U} H(t,x^*,u,-V_x(t,x^*)). \tag{45}$$

So far what we derived (Eqs. (42), (43) and (45)) are *necessary* conditions for optimality. Indeed, defining V to be the value function, we showed that it must satisfy the HJB equation. Assuming further that an optimal control exists, we showed that it must maximize the Hamiltonian along the optimal trajectory. It turns out that these conditions are also *sufficient* for optimality. Let $\widehat{V} : [0,T]\times\mathbb{R}^n \to \mathbb{R}$ be a C^1 function with boundary condition $\widehat{V}(T,x) = K(x)$ that satisfies Eq. (42), and suppose there is a control $\hat{u} : [0,T] \to U$ and a corresponding trajectory $\hat{x} : [0,T] \to \mathbb{R}^n$ with initial condition $\hat{x}(0) = x_0$ satisfy Eq. (45), i.e.

$$H(t,\hat{x},\hat{u},-\widehat{V}_x(t,\hat{x})) = \max_{u\in U} H(t,\hat{x},u,-\widehat{V}_x(t,\hat{x})),$$

then $\widehat{V}(t_0,x_0)$ is the optimal cost and $\hat{u}$ is the optimal control. To see this, from the above equation we have

$$-\widehat{V}_t(t,\hat{x}) = L(t,\hat{x},\hat{u}) + \widehat{V}_x(t,\hat{x})\cdot f(t,\hat{x},\hat{u}).$$

Moving the left side to the right, it is

$$0 = L(t,\hat{x},\hat{u}) + \frac{d}{dt}\widehat{V}(t,\hat{x}).$$

Integrating this equality with respect to t from 0 to T, we have

$$0 = \int_0^T L(t,\hat{x},\hat{u})dt + \widehat{V}(T,\hat{x}(T)) - \widehat{V}(0,x_0)$$

so that

$$\widehat{V}(0,x_0) = \int_0^T L(t,\hat{x},\hat{u})dt + K(\hat{x}(T)) = J(0,x_0,\hat{u}). \tag{46}$$

Let x be another trajectory with the same initial condition $x(0) = x_0$ corresponding to an arbitrary control. Since $\widehat{V}$ satisfies the HJB equation Eq. (42), we have

$$-\widehat{V}_t(t,x) \le L(t,x,u) + \widehat{V}_x(t,x)\cdot f(t,x,u)$$

or

$$0 \le L(t,x,u) + \frac{d}{dt}\widehat{V}(t,x).$$

Integrating over $[0,T]$ we have

$$0 \le \int_0^T L(t,x,u)dt + \widehat{V}(T,x(T)) - \widehat{V}(0,x_0)$$

or

$$\widehat{V}(0,x_0) \le \int_0^T L(t,x,u)dt + K(x(T)) = J(0,x_0,u). \tag{47}$$

Eq. (46) and Eq. (47) show that control $\hat{u}$ has the cost $\widehat{V}(0, x_0)$ while no other control u can produce a smaller cost.e Thus $\widehat{V}$ is the optimal cost and $\hat{u}$ is the optimal control.

To further see the distinctions between the maximum principle and dynamic programming, assume for simplicity that the system and cost is time-invariant (so we drop "t" in f, L and H). The maximum principle is formulated in terms of the canonical equations

$$\dot{x}^* = H_p\big|_*, \quad \dot{p}^* = -H_x|_*$$

and says that at each time t the value $u^*(t)$ of the optimal control must maximize $H(x^*(t), u, p^*(t))$ with respect to u:

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

This is an *open-loop* specification, because $u^*(t)$ depends not only on the state $x^*(t)$ but also on the costate $p^*(t)$ which has to be computed from the adjoint differential equation. On the other hand, the HJB equation says that the optimal control must satisfy

$$u^* = \arg\max_{u \in U} H(x^*(t), u, -V_x(t, x^*(t))).$$

This is a *closed-loop (feedback)* specification. After we solved the value function $V(t, x)$ everywhere, $u^*(t)$ is completely determined by the current state $x^*(t)$.

The ability to generate an optimal control policy in the form of a state feedback law is an important feature of the dynamic programming approach. Clearly, we cannot implement this feedback law unless we can first find the value function by solving the HJB partial differential equation, and this is in general a very difficult task. Therefore, from the computational point of view the maximum principle has an advantage in that it involves only ordinary and not partial differential equations. The dynamic programming approach provides more information (including sufficiency), but in reality, the maximum principle is often easier to use and allows one to solve many optimal control problems for which the HJB equation is analytically intractable.

6.2 Applications of the HJB Equation

Example 6.1 (Rocket Car)**.** In Example 4.12 we saw how the maximum principle can be applied to solve for the rocket car problem (Example 4.4). Here let's see how dynamic programming works. Recall the dynamic is

$$\dot{x}(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(t)$$

for $x(t) = (x_1(t), x_2(t))^T$ and $u \in [-1, 1]$. Note that

$$f(x, u) = \begin{pmatrix} 0 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ u \end{pmatrix} = \begin{pmatrix} x_2 \\ u \end{pmatrix}$$

and the cost functional is

$$J(u) = \int_0^\tau 1 dt = \tau$$

so $L(x, u) = 1$. Let $V(t, x)$ be the value function. The HJB equation (Eq. (42)) for this problem is

$$-V_t(t, x) = \inf_{u \in [-1,1]} \left\{1 + V_{x_1}(t, x)x_2 + V_{x_2}(t, x)u\right\}. \tag{48}$$

Note that this is a free-time, fixed-endpoint problem, for which the boundary condition takes the form $V(t, 0) = 0$ for all t, and Eq. (48) is valid away from $x = 0$. The infimum on the right-hand side of Eq. (48) is achieved by setting

$$u = -\operatorname{sign}(V_{x_2}(t, x)) = \begin{cases} 1 & \text{if } V_{x_2}(t, x) < 0 \\ -1 & \text{if } V_{x_2}(t, x) > 0. \end{cases}$$

Substituting it into Eq. (48), we arrive at a simplified HJB equation

$$-V_t(t,x) = 1 + V_{x_1}(t,x)x_2 - |V_{x_2}(t,x)|.$$

The optimal control is given by the *feedback* law

$$u^* = -\operatorname{sign}(V_{x_2}(t,x^*(t))),$$

whose implementation of course hinges on our ability to solve for V.

Example 6.2 (General Linear Quadratic Regulator)**.** In (finite horizon) *Linear Quadratic regulator (LQR)* problem, the control system is a linear time-varying system

$$\begin{cases} \dot{x} = A(t)x + B(t)u, \\ x(0) = x_0 \end{cases}$$

with $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. The target set is $S = \{T\} \times \mathbb{R}^n$, so this is a fixed-time, free-endpoint problem. The cost functional is

$$J(u) = \int_0^T \left[x^T(t)Q(t)x(t) + u^T(t)R(t)u(t)\right] dt + x^T(T)Mx(T)$$

where Q, R and M are symmetric and positive semidefinite matrices, and R is a symmetric and positive *definite* matrix, so that it has an inverse. The HJB equation (Eq. (42)) for this problem is

$$-V_t(t,x) = \inf_{u\in\mathbb{R}^m} \left\{x^T Q(t)x + u^T R(t)u + V_x(t,x)\cdot(A(t)x + B(t)u)\right\} \tag{49}$$

with boundary condition

$$V(T,x) = x^T Mx. \tag{50}$$

The minimizer of the quadratic function in Eq. (49) is

$$u = -\frac{1}{2}R^{-1}(t)B^T(t)V_x(t,x). \tag{51}$$

Substituting Eq. (51) into Eq. (49), the HJB equation becomes

$$-V_t(t,x) = x^T Q(t)x + V_x(t,x)^T A(t)x - \frac{1}{4}V_x(t,x)^T B(t)R^{-1}(t)B^T(t)V_x(t,x). \tag{52}$$

To obtain the optimal control, we need to solve for V. Let's *guess* a form of V and check that it indeed works. Inspired by the terminal condition Eq. (50), let's guess that V has the form

$$V(t,x) = x^T P(t)x$$

for some symmetric matrix $P(t)$. Then $V_x(t,x) = 2P(t)x$ and $V_t(t,x) = x^T\dot{P}(t)x$. Plugging these two expressions into Eq. (52), we obtain

$$-x^T\dot{P}(t)x = x^T\left[Q(t) + 2P(t)A(t) - P(t)B(t)R^{-1}(t)B^T(t)P(t)\right]x$$
$$\Downarrow$$
$$\dot{P} = Q(t) + 2P(t)A(t) - P(t)B(t)R^{-1}(t)B^T(t)P(t).$$

This is the matrix *Riccati equation*. If we can solve for this, then we can obtain the optimal control law

$$u^*(t) = -R^{-1}(t)B^T(t)P(t)x^*(t)$$

in feedback form.

References

Evans, L. C. (2005). "An introduction to mathematical optimal control theory". *Lecture Notes, University of California, Department of Mathematics, Berkeley*.

Lax, P. D. (2002). *Functional Analysis*. New York: John Wiley & Sons.