1 Lecture 1: Introduction (Karl Johan)

2 Lecture 2: Calculus of variation (CoV) and the Maximum principle

In this lecture, we are going to learn the maximum principle. The MP is a type of CoV, so we will first study the classical theory of CoV. Then we will try to move from the classical CoV theory to the optimal control setting, there we will immediately encounter some essential difficulties that the classical CoV cannot overcome. So we will need some new insights. These new insights, discovered by the Pontrayagen school finally led to the celebrated maximum principle. Without further ado, let's first take a look at the classical theory of CoV, which of course is interesting in its own right.

2.1 CoV

Behind all physical phenomena in our universe, there is almost always a principle of calculus of variation, e.g., in mechanics, thermal dynamics, electrodynamics and quantum physics, you name it. As a French mathematician once said, nature is thrifty/optimal in all its actions. In physics, they call such the principle of least action. For us, perhaps the best example for illustration is mechanics.

Example 1 (Mechanics)**.** In mechanics, a system is described by its configuration and velocity $(q, \dot{q})$ – sometimes called generalized coordinates and velocities. The most fundamental question in mechanics is to find the equation of motion of a system. It turns out there is a quite simple principle, which states that the path of a mechanical system between two fixed points q_0, q_1 without external force is the one which minimizes the following functional

$$S(q(\cdot), \dot{q}(\cdot)) := \int L(q(t), \dot{q}(t))\mathrm{d}t$$

Here L is something called the Lagrangian of the system, which is the difference between the kinetic energy and potential energy of the system:

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

If you solve this minimization problem, then you'll get the Euler-Lagragian equations of motion.

Example 2 (Optics)**.** In high school physics, we learned that the velocities of lights in different mediums are different. An interesting question is that what is the trajectory of a ray when it crosses intersection of two different mediums, see Figure 1. Fermat says that the path taken by a ray between two given points is the path that can be traveled in the least time. This again is a statement of CoV, which can help us establish the celebrated Snell's law:

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}.$$

Besides physics, there are also many other important types of CoV problems. For example, Dido's problem.

Example 3 (Dido's problem)**.** Suppose you have a curve with fixed length that you want to enclose it with a line with maximum area, see Figure 2. The question is what's the optimal shape of the curve. From calculus, we know that the area is

$$A(\gamma) := \frac{1}{2}\int_0^1 \gamma_1(s)\gamma_2'(s) - \gamma_1'(s)\gamma_2(s)\mathrm{d}s$$

and the length of the curve is

$$\int |\gamma'(s)|\mathrm{d}s = \text{constant}$$

which is an integral constraint. Due to the integral constraint, this problem is not as easy to solve if we don't resort to optimal control theory.

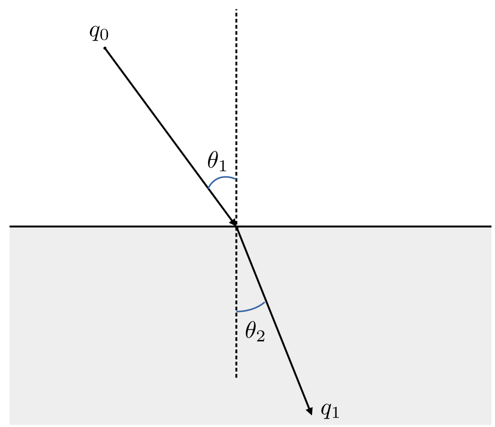

Figure 1: Snell's law

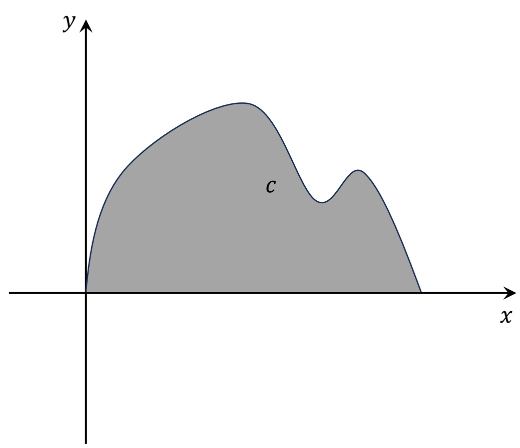

Figure 2: Dido's problem

The final example I'm giving you is from geometry.

Example 4 (Riemannian geometry)**.** On a sphere, we know that the shortest path between two points is contained in a so called great circle. This kind of path is called a geodesic in Riemannian geometry. A basic question in Riemannian geometry is how to find the geodesic between two points. That is, find $s \mapsto q(s)$ such that

$$\ell = \int_0^1 |\dot{q}(s)| \mathrm{d}s$$

is minimized with fixed endpoints with the constraint $q(0) = q_0$, $q(1) = q_1$.

Let's do a little summary. Notice that in all the previous examples, the cost functionals to be minimized have the form (with or without constraint):

$$\min_{q(\cdot), \dot{q}(\cdot)} S := \int_0^T L(q, \dot{q}) \mathrm{d}t.$$

Thus it seems a good candidate to work with. From now on, we will focus on the case that the only constraint is that the two endpoints are fixed.

The idea of CoV is quite simple. Consider the a generic minimization problem

$$\min_{u \in U} J(u)$$

where U can be arbitrarily complicated subsets of a vector space V (finite or infinite dimensional). Suppose u_* is a minimizer of the above problem, then we do an infinitesimal perturbation of u_*. More precisely, we construct an one-parameter set of u_ϵ, continuous in ϵ, such that $u_0 = u_*$ and $u_\epsilon \in U$ for all sufficiently small ϵ. In particular, if U is a vector space, then u_ϵ can be chosen as $u_\epsilon = u_* + \epsilon v$ for some $v \in U$. Conceptually, the more admissible perturbations you can find, the more info you can extract on the optimal solution. This is a quite natural idea. For example, if you want to acquire some knowledge of a linear system, then you would like to inject a signal with rich frequencies and then check its output. And we know that white noise contains all frequency components, so it's a good candidate.

Now by optimality, we must have

$$J(u_\epsilon) \geq J(u_*)$$

for all sufficiently small ϵ. But this immediately implies that

$$\epsilon \mapsto J(u_\epsilon)$$

has a minimum at $\epsilon = 0$. Thus if this function $\epsilon \mapsto J(u_\epsilon)$ is differentiable, we must have

$$\frac{\partial J}{\partial \epsilon}|_{\epsilon=0} = 0, \quad \frac{\partial^2 J}{\partial \epsilon^2}|_{\epsilon=0} \geq 0. \tag{1}$$

As a simple example, consider the minimization problem

$$\min f(x)$$
$$\text{subject to } h(x) = 0,\ x \in \mathbb{R}^n$$

where f and h are C^1. This problem can be solved using Lagrangian multiplier, but it can also be solved by CoV. Assume x_* is a minimizer. Choose any curve $\epsilon \mapsto x_\epsilon$ such that $h(x_\epsilon) = 0$ and $x_0 = x_*$. Then $\epsilon \mapsto f(x_\epsilon)$ achieves minimum at $\epsilon = 0$ for ϵ sufficiently small. This implies that

$$0 = \frac{\partial f(x_\epsilon)}{\partial \epsilon}|_{\epsilon=0} = \nabla f(x_*) \cdot \frac{\partial x_\epsilon}{\partial \epsilon}|_{\epsilon=0}$$

in which $\frac{\partial x_\epsilon}{\partial \epsilon}|_{\epsilon=0}$ is nothing but the velocity of the curve $\epsilon \mapsto x_\epsilon$ at x_*. We call such vectors tangent vectors of the set

$$M := \{x : h(x) = 0\}$$

at the point x_* and denoted it by $T_{x_*}M$. See Figure 3

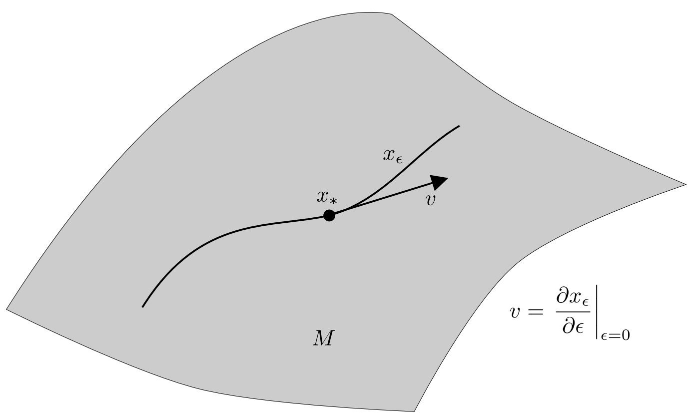

Figure 3: Tangent vectors.

The above formula says, for all tangent vectors v at the minimizer x_*, it must be perpendicular to $\nabla f(x_*)$. For second order variation.

$$\frac{\partial x_\epsilon}{\partial \epsilon}^\top f_{xx} \frac{\partial x_\epsilon}{\partial \epsilon}|_{\epsilon=0} + f_x \frac{\partial^2 x_\epsilon}{\partial \epsilon^2}|_{\epsilon=0} \geq 0$$

Choose $x_\epsilon = x_* + \epsilon v$ for v fixed locally, then $\frac{\partial^2 x_\epsilon}{\partial \epsilon^2}|_{\epsilon=0} = 0$. Thus $f_{xx}(x_*) \geq 0$.

The above calculation is a bit inefficient, for each time, you have to choose a curve x_ϵ which you haven't even used. We shall invent some more efficient notations to facilitate our computation. The δ operator does the job. For quantity F, we denote δF as the infinitesimal variation of F. Well, this is not a valid mathematical definition. But this can be made rigorous by interpreting it as a Gateaux derivative. However, we won't need the rigorous construction. The only things that we should know are the following:

P1) δ is a differential operator: it satisfies the chain rule, composition rule, etc.

P2) in our setting, δ commutes with the integration and differentiation operators, i.e., $\delta \int = \int \delta$ and $\delta \dot{x} = \frac{d}{dt}\delta x$.

P3) If a function $u \mapsto J(u)$ has a minimum at u_*, then the first variation vanishes $\delta J(u_*) = 0$ and the second variation is non-negative $\delta^2 J(u_*) \geq 0$.

Now let's go back to the Lagrangian problem and we will immediately be able to solve a number of interesting problems. Let's minimize

$$S = \int L(q, \dot{q})\mathrm{d}t.$$

Then by item P3), at the minimizer (omit the argument for ease of notation) we have

$$\delta S = 0.$$

Now using item P1 and P2, we get

$$0 = \delta S = \delta \int L\mathrm{d}t = \int \delta L\mathrm{d}t = \int \frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\delta \dot{q}\mathrm{d}t$$

i.e. the infinitesimal change of J is a result of infinitesimal changes of q and $\dot{q}$.

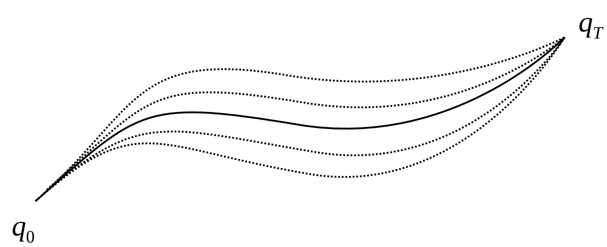

Figure 4: Variation

Using integration by parts, we get

$$\begin{aligned} 0 = \int \frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\delta \dot{q}\mathrm{d}t &= \int \frac{\partial L}{\partial q}\delta q - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\delta q\mathrm{d}t + \frac{\partial L}{\partial \dot{q}}\delta q|_0^T \\ &= \int \left(\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}}\right)\delta q\mathrm{d}t + \frac{\partial L}{\partial \dot{q}}\delta q|_0^T \end{aligned}$$

Now the boundary term vanishes if we carefully choose the variation to be such that the endpoints are fixed. Now we claim that

$$\frac{\partial L}{\partial q} - \frac{d}{dt}\frac{\partial L}{\partial \dot{q}} = 0,$$

This equation is called the Euler-Lagrange equation. To arrive at this, we need the Fundamental lemma of CoV.

Lemma 1. *If a continuous function f on on open interval (a, b) satisfies*

$$\int_a^b f(x)h(x)\mathrm{d}x = 0$$

for all $h \in C_c^\infty(a, b)$, then f is identically zero. If f is only locally integrable (in the Lebesgue sense), then f is zero almost everywhere.

Now using this lemma, we can choose δq which is compactly supported to arrive at the EL-equation (remember that L is C^1).

Let's use the EL equation to solve an example – finding the geodesic on a sphere. A point on a two sphere can be described by two angles (θ, ϕ), see Figure 5. Let $q = (\theta, \phi)$. Then a curve has the form $t \mapsto (\theta(t), \phi(t))$. Consider two points on the sphere on the xy plane. In this case, $\phi(0) = \phi(1) = \frac{\pi}{2}$. The length of the curve is calculated as

$$\int_0^1 \sqrt{R^2\dot{\theta}^2 + R^2 \sin^2\theta\dot{\phi}^2}\mathrm{d}t$$

with $\theta(0) = \theta_0$ and $\theta(1) = \theta_1$. Instead of working with this, we claim that it's equivalent to work with (only valid for Riemannian geodesic!):

$$\int_0^1 R^2\dot{\theta}^2 + R^2 \sin^2\theta\dot{\phi}^2\mathrm{d}t$$

That is, the new Lagrangian is

$$L(q, \dot{q}) = R^2\dot{\theta}^2 + R^2 \sin^2\theta\dot{\phi}^2$$

Now lets apply the EL formula. First we find

$$\frac{\partial L}{\partial q} = \begin{bmatrix} 2R^2 \sin\theta\cos\theta\dot{\phi}^2 \\ 0 \end{bmatrix}, \quad \frac{\partial L}{\partial \dot{q}} = \begin{bmatrix} 2R^2\dot{\theta} \\ 2R^2\sin^2\theta\dot{\phi} \end{bmatrix}$$

The EL equation says

$$\begin{aligned} \frac{\partial L}{\partial \theta} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}} = 0 \\ \frac{\partial L}{\partial \phi} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\phi}} = 0 \end{aligned}$$

or

$$\begin{aligned} \sin\theta\cos\theta\dot{\phi}^2 = \ddot{\theta} \\ 2\sin\theta\cos\theta\dot{\theta}\dot{\phi} + \sin^2\theta\ddot{\phi} = 0 \end{aligned}$$

It's easily seen that $\phi \equiv \frac{\pi}{2}$ is a solution, and with a bit more technique, you can show that this is the only solution. On the other hand, we have $\ddot{\theta} = 0$, hence $\theta(t) = (1-t)\theta_0 + t\theta_1$ which is an increasing function. So the geodesic lies on a great circle and there is no winding. But notice carefully that the geodesic is not necessarily minimizing! To guarantee the geodesic is minimizing, we would need some second order condition.

Remark 1. The EL equation can be easily extended to higher dimensions in the sense that we consider action of the form

$$S = \int L(q, \nabla q, x)\mathrm{d}x$$

where $x \in \mathbb{R}^n$. This is left as an exercise.

Second order variation

Recall the first variation

$$\delta S = \int \frac{\partial L}{\partial q}\delta q + \frac{\partial L}{\partial \dot{q}}\delta\dot{q}\mathrm{d}t.$$

We compute another variation:

$$\delta^2 S = \int \delta q^\top L_{qq}\delta q + \delta\dot{q}^\top L_{q\dot{q}}\delta q + L_q\delta^2 q + \delta q^\top L_{\dot{q}q}\delta\dot{q} + \delta\dot{q}^\top L_{\dot{q}\dot{q}}\delta\dot{q} + L_{\dot{q}}\delta^2\dot{q}\mathrm{d}t$$

remember that δ commutes with $\frac{d}{dt}$, we have

$$\int L_{\dot{q}}\delta^2\dot{q}\mathrm{d}t = L_{\dot{q}}\delta^2 q|_0^T - \int (\frac{d}{dt}L_{\dot{q}})\delta^2 q\mathrm{d}t.$$

Plugging this into the $\delta^2 S$ and remember the EL equation $\frac{d}{dt}L_{\dot{q}} = L_q$, we get (note $\delta^2 q = 0$ at the boundary):

$$\delta^2 S = \int \begin{bmatrix}\delta q\\ \delta\dot{q}\end{bmatrix}^\top \begin{bmatrix}L_{qq} & L_{q\dot{q}}\\ L_{\dot{q}q} & L_{\dot{q}\dot{q}}\end{bmatrix} \begin{bmatrix}\delta q\\ \delta\dot{q}\end{bmatrix} \mathrm{d}t \geq 0.$$

We assert that $L_{\dot{q}\dot{q}}(q,\dot{q})$ must be semi-positive definite, i.e.,

$$L_{\dot{q}\dot{q}} \geq 0 \tag{2}$$

along the optimal solution $(q,\dot{q})$. To see this, it is sufficient to note that there exist functions with small magnitude but with rather large derivatives; the converse is false, thus it may happen that $D_{qq}L - \frac{d}{dt}D_{q\dot{q}}L$ is non semi-positive definite (it is not even symmetric!). It is interesting to ask whether $L_{\dot{q}\dot{q}} > 0$ is sufficient as in finite dimensional optimization.

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

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

We are not going to give the precise definition of conjugate points nor are we going to prove the above result. After all, analyzing conjugate points is a delicate issue and is out of the scope of this course. It is enough to remember the sphere example to be aware of such phenomenon, see Figure 5.

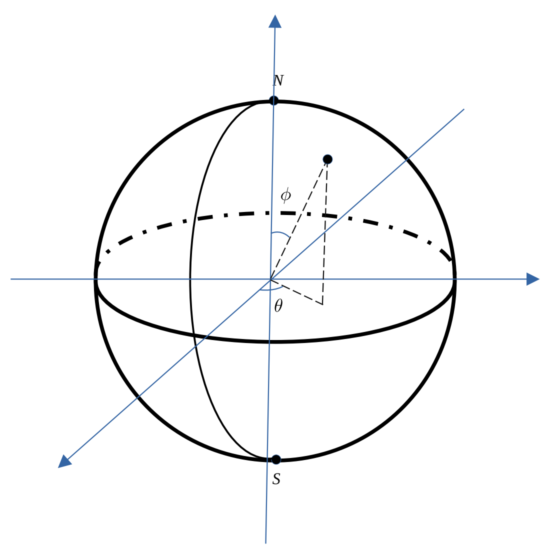

Figure 5: N and S are conjugate points.

Caveat: the EL equation does not incorporate any constraints (except on the boundary)! In this sense equation (1) is much more general and flexible. The main advantage of formula (1) is that sometimes by carefully choosing the perturbation satisfying the constraints, a lot of interesting information about the minimizer can be obtained which helps us determine the minimizer. But still, it relies on the choices of the perturbation. We do not plan to spend more time here since we will later introduce a more powerful tool that is, the maximum principle, which will ultimately help us deal with constraints in a more effective manner.

2.2 Canonical transform

Next we are going to introduce a key trick which has profound impacts on natural science and engineering. It's among the most important equations in physics and engineering. It's definitely worthwhile to put it into your knowledge box. This trick is as follows.

Consider a coordinate transform $p = \frac{\partial L}{\partial \dot{q}}$, i.e.,

$$(q, \dot{q}) \mapsto (q, p)$$

called *canonical transform* and define a function called the *Hamiltonian*

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

in which $\dot{q}$ is understood as a function of q and p. For example,

$$\frac{\partial H}{\partial p} = p^\top \frac{\partial \dot{q}}{\partial p} + \dot{q} - \frac{\partial L}{\partial \dot{q}} \frac{\partial \dot{q}}{\partial p} = \dot{q} \tag{3}$$

by definition of p. The Jocobian of the transform is

$$\begin{bmatrix} I & 0 \\ * & L_{\dot{q}\dot{q}} \end{bmatrix}$$

thus if $L_{\dot{q}\dot{q}}$ is non-singular the transform is well-defined.

Then S can be rewritten as

$$S = \int_0^T L \mathrm{d}t = \int_0^T p^\top \dot{q} - H(q, p) \mathrm{d}t$$

Let us try to calculate the first variation of S under this form:

$$\begin{aligned} \delta S &= \int_0^T p^\top \delta \dot{q} + \dot{q}^\top \delta p - H_q \delta q - H_p \delta p \mathrm{d}t \\ &= p^\top \delta q|_0^T - \int_0^T \dot{p}^\top \delta q - \dot{q}^\top \delta p + H_q \delta q + H_p \delta p \mathrm{d}t \\ &= p^\top \delta q|_0^T + \int_0^T (\dot{q}^\top - H_p) \delta p - (\dot{p}^\top + H_q) \delta q \mathrm{d}t. \end{aligned}$$

Remember that at the optimal point, the EL equation is satisfied, i.e., which results in (invoking (3)):

$$0 = \delta S = -\int_0^T (\dot{p}^\top + H_q) \delta q \mathrm{d}t$$

when the endpoints are fixed. By the same reasoning as before, we immediately get

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

This equation, together with (3), i.e.,

$$\dot{q} = H_p^\top (q, p)$$

is called the Hamiltonian equation or the canonical equation, which is extremely important in science and engineering. This equation will also appear in the maximum principle. Now we show that the Hamiltonian equation is indeed equivalent to EL-equation. Differentiating H w.r.t. q and p, we get

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

and

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

as expected.

It still remains to find the second order variation of S:

$$\begin{aligned}\delta^2 S = \delta p^\top \delta q|_0^T + p^\top \delta^2 q|_0^T + \int_0^T (\delta \dot q^\top - \delta q^\top H_p - \delta p^\top H_{pp})\delta p + (\dot q^\top - H_p)\delta^2 p \\ - (\delta \dot p^\top + \delta q^\top H_{qq} + \delta p^\top H_{qp})\delta q - (\dot p^\top + H_q)\delta^2 q \mathrm{d}t\end{aligned}$$

Invoking the first order necessary condition and remember that the boundary variation is zero, $\delta^2 S$ can be simplified to

$$\begin{aligned}\delta^2 S &= \int_0^T (\delta \dot q^\top - \delta q^\top H_{pq} - \delta p^\top H_{pp})\delta p - (\delta \dot p^\top + \delta q^\top H_{qq} + \delta p^\top H_{qp})\delta q \mathrm{d}t \\ &= \int_0^T \delta \dot q^\top \delta p + \delta \dot p^\top \delta q - \begin{bmatrix}\delta q \\ \delta p\end{bmatrix}^\top \begin{bmatrix} H_{qq} & H_{qp} \\ H_{pq} & H_{pp}\end{bmatrix}\begin{bmatrix}\delta q \\ \delta p\end{bmatrix} \mathrm{d}t \\ &= -\int_0^T \begin{bmatrix}\delta q \\ \delta p\end{bmatrix}^\top \begin{bmatrix} H_{qq} & H_{qp} \\ H_{pq} & H_{pp}\end{bmatrix}\begin{bmatrix}\delta q \\ \delta p\end{bmatrix} \mathrm{d}t.\end{aligned}$$

In order that $\delta^2 S \geq 0$ at the optimizer, it seems that we necessary have

$$\begin{bmatrix} H_{qq} & H_{qp} \\ H_{pq} & H_{pp}\end{bmatrix} \leq 0$$

since q and p seem to be independent variables and we can vary them independently. But this is **false**!

Although in the Hamiltonian $H(q, p)$, q and p can be seen as independent variables, as long as $q(t)$ and $p(t)$ are defined to be the solution to the canonical equation, they are indeed coupled.

What's worse, it's not obvious how to reason as in the Lagragian case to get a second order necessary condition: the role of variables p and q are somehow symmetric, unlike the relation between q and $\dot q$. We'll circumvent this issue later while we derive the maximum principle.

The canonical transform is nice, but it is somehow criptic. To get a better understanding of it, let's check the example of mechanics. Remember that the Lagrangian in mechanics is

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

The canonical transform is

$$p = \frac{\partial L}{\partial \dot q} = M(q)\dot q.$$

This transform is well-defined since $\frac{\partial^2 L}{\partial \dot q \partial \dot q} = M(q) > 0$. Now substitute this into

$$\begin{aligned}H(q, p) = p^\top \dot q - L(q, \dot q) &= \dot q^\top M(q)\dot q - (\frac{1}{2}\dot q^\top M(q)\dot q - V(q)) \\ &= \frac{1}{2}\dot q^\top M(q)\dot q + V(q)\end{aligned}$$

which is nothing but the mechanical energy of the system! Thus the Hamiltonian corresponds to the energy of the system. Recalling the Hamiltonian equation, we have

$$\frac{dH}{dt} = \frac{\partial H}{\partial q}\dot q + \frac{\partial H}{\partial p}\dot p = 0.$$

Thus the mechanical energy of a closed mechanical system without external force is preserved.

Another interpretation

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

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

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

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

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

2.3 Optimal control via CoV

Now we are ready to study optimal control problems using CoV and let's see how far we can get. We focus on the nonlinear system

$$\dot{x} = f(x, u), \quad x \in \mathbb{R}^n,\ u \in \mathbb{R}^m$$

with fixed initial condition x_0 and cost function

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

For the moment we impose no constraints on the input u. Let's do the first variation along the optimal solution (x_*, u_*):

$$\delta J = \nabla\varphi(x(T)) \cdot \delta x(T) + \int_0^T L_x \cdot \delta x(t) + L_u \cdot \delta u(t)\mathrm{d}t$$

Recall that the essential step in "key trick" is using integration by parts formula. But in the above we don't have a term involving $\delta\dot{x}$, instead, it's replaced by δu. However, we know

$$\delta\dot{x} = f_x\delta x + f_u\delta u$$

or

$$\delta\dot{x} - f_x\delta x - f_u\delta u = 0$$

thus, we may add this term into the above formula. However, this is a vector, so we multiply it by an arbitrary vector $p(t)$ from the left

$$p(t)^\top (\delta\dot{x} - f_x\delta x - f_u\delta u) = 0$$

and incorporate it in δJ:

$$\begin{aligned}
\delta J &= \varphi_x\delta x(T) + \int_0^T L_x\delta x + L_u\delta u + \left(p^\top (\delta\dot{x} - f_x\delta x - f_u\delta u)\right) \mathrm{d}t \\
&= \varphi_x\delta x(T) + \int_0^T \left(L_x - p^\top f_x\right)\delta x + \left(L_u - p^\top f_u\right)\delta u + p^\top\delta\dot{x}\mathrm{d}t \\
&= \varphi_x\delta x(T) + \int_0^T \frac{\partial(L - p^\top f)}{\partial x}\delta x + \frac{\partial(L - p^\top f)}{\partial u}\delta u + p^\top\delta\dot{x}\mathrm{d}t \\
&= \varphi_x\delta x(T) + \int_0^T -H_x\delta x - H_u\delta u + p^\top\delta\dot{x}\mathrm{d}t
\end{aligned}$$

for $H(x, p, u) = p^\top f(x, u) - L(x, u)$. Now we are in a position to use integration by parts formula

$$\delta J = \varphi_x\delta x(T) + \int_0^T \left(-H_x\delta x - H_u\delta u - \dot{p}^\top\delta x\right)\mathrm{d}t + p(t)^\top\delta x(t)|_0^T \tag{4}$$

Since the initial state is fixed, we have $\delta x(0) = 0$ and we arrive at

$$\delta J = -\int_0^T (\dot{p}^\top + H_x)\delta x + H_u \delta u \mathrm{d}t + (\varphi_x + p(T)^\top)\delta x(T) \tag{5}$$

Choose p such that

$$\dot{p} = -H_x^\top(x_*, u_*)$$

and

$$(\varphi_x + p(T)^\top)\delta x(T) = 0$$

then

$$\delta J = -\int_0^T H_u \delta u \mathrm{d}t = 0$$

from which it follows that

$$H_u(x_*, u_*) = 0.$$

On the other hand,

$$\dot{x}_* = H_p^\top(x_*, u_*)$$

thus we again obtain a canonical equation

$$\begin{aligned} \dot{x}_* &= H_p^\top(x_*, u_*) \\ \dot{p} &= -H_x^\top(x_*, u_*) \end{aligned}$$

If $x(T) \in M$ for some manifold, then

$$p(T) + \varphi_x^\top(x(T)) \perp T_{x_*(T)}M.$$

For example, if M is the total space, i.e., free boundary, then $p(T) = -\varphi_x^\top(x(T))$. If the boundary is fixed, then $p(T)$ is undetermined.

If $H_{uu}(x, u)$ is nonsingular, then by implicit function theorem, we can solve $u_* = u_*(x)$ from $H_u(x, u) = 0$ at least locally. With this, the Hamiltonian equation can be solved numerically.