4 Lecture 4. Misc topics on MP and the proof of the MP

In the previous lecture, we studied the maximum principle. The main result is the following theorem.

Theorem 1. *Consider the system $\dot{x} = f(x,u)$ with cost function*

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

and boundary constraint

$$x(t_f) \in M \subseteq \mathbb{R}^n$$

Assume f, φ and L are C^1 in x. Let $(x^*(\cdot), u^*(\cdot))$ correspond to the optimal solution to the minimization problem

$$\min_{u \in \mathcal{U}_{\mathrm{ad}}} J(u)$$

in which $\mathcal{U}_{\mathrm{ad}} = \{u : [0,t_f] \to U \subseteq \mathbb{R}^m\}$. Define the Hamiltonian function

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

Then there exists a function $p^* : [0,t_f] \to \mathbb{R}^n$ and a constant $p_0^* \in \{0,-1\}$, satisfying $(p_0^*, p^*(t)) \not\equiv (0,0)$ such that

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

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

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

2) The transversality condition holds:

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

3) The maximum principle holds:

$$H(x^*(t), u^*(t), p^*(t), p_0^*) = \max_{u \in U \subseteq \mathbb{R}^m} H(x^*(t), u, p^*(t), p_0^*) = \text{constant} \tag{1}$$

for all $t \in [0,t_f]$. In particular, this constant is zero if t_f is free.

We ended with the lunar lander example. But in fact, there was another important example that I wanted to show you, which is the example of time optimal control.

4.1 Time optimal control

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

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

We impose a terminal constraint $x(t_f) \in S$. This is an optimal control problem with free terminal time t_f. We focus on the case when S is a singleton. The general case is essentially the same.

It is worth mentioning that the time optimal control problem is closely related to controllability and stabilizability. Loosely speaking, the system is controllable (to the target) if and only if $\min J < \infty$. Thus optimal control is an important tool in studying controllability and stabilization. Now we consider the following system which is affine in the input:

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

where $u \in \mathbb{R}^m$. The constraint for u is $|u_i| \leq 1$ for all $i = 1, \cdots m$. This model is quite general and includes most of the control systems that we meet and is thus reasonable to work with.

The Hamiltonian for the system is

$$H = p^\top (f(x) + g(x)u) + p_0 = p^\top f(x) + \sum_{i=1}^{m} u_i (p^\top g_i(x)) + p_0$$

The costate equation is

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

And there's no transversality condition. Good thing is that we don't need to discuss abnormal extremals: H depends trivially on p_0.

Recall that $|u_i| \leq 1$, the maximum principle can be seen as a linear programming on a convex set since u is affine in this case. In our case, (in general, u can be constrained in a polytope)

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

So typically, the optimal control switches between 1 and -1, except at those time instants that $p^\top(t) g_i(x^*(t)) = 0$. Such control is named *bang-bang control* (a control whose components are either 1 or -1). A critical issue here is that we don't know the optimal control when $p^\top(t) g_i(x^*(t)) = 0$. If the function $\gamma(t) = p^\top(t) g_i(x^*(t))$ has only finite zeros, that is fine, the input at those finite points can be taken arbitrarily. In this case, we say that the optimal solution is *normal*. But if γ is zero on some interval $[t_1, t_2]$, then you cannot choose the control arbitrarily on that interval. Such control on $[t_1, t_2]$ is called *singular*, and the corresponding trajectory $x^*|_{[t_1,t_2]}$ is called a *singular arc*.

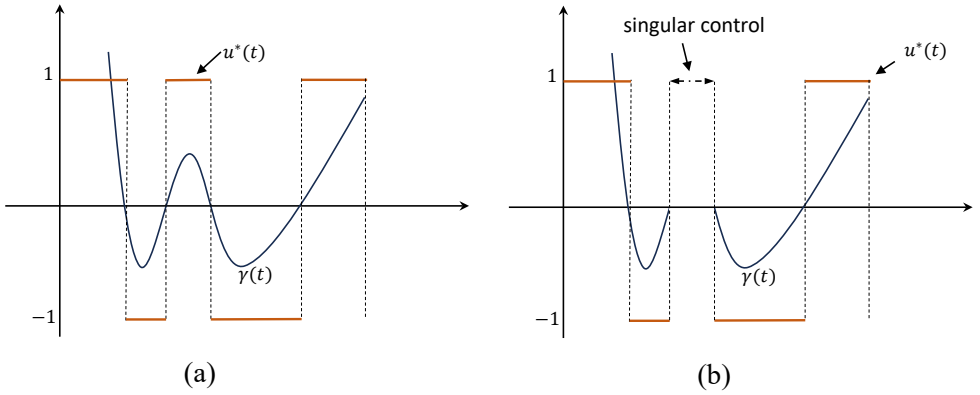

Figure 1: Normal control and singular control

The existence of singular arcs is a bit unpleasant because in this situation, the maximum principle does not tell us the information of the optimal control. Therefore, it would be nice to have some prior test to exclude the existence of singular arcs. Fortunately, for most control systems, singular arcs do not exist and there are simple criteria to check this. For example, we consider the linear system

$$\dot{x} = Ax + Bu, \quad u \in \mathbb{R}^m, \ |u_i| \leq 1,$$

For this system $g_i = b_i$ and the Hamiltonian is $H(x, u, p, p_0) = p^\top (Ax + Bu) + p_0$. The costate equation reads

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

The condition under which singularity happens is when

$$p^\top(t) b_i(x^*(t)) = 0, \quad \forall t \in [t_1, t_2]. \tag{2}$$

Note that if (2) is to be satisfied, its derivatives of any order should also be zero. For instance,

$$\begin{aligned} p^\top b_i &= 0, \\ p^\top A b_i &= 0, \\ &\vdots \\ p^\top A^k b_i &= 0 \end{aligned}$$

or

$$p^\top \begin{bmatrix} b_i \\ Ab_i \\ \vdots \\ A^{n-1}b_i \end{bmatrix} = 0.$$

Thus if the system pair (A, b_i) is controllable, $p(t) \equiv 0$, thus $p_0 = -1$ and $0 = H(x^*(t), u^*(t), p^*(t), -1) = p_0 = -1$, since t_f is free. This is a contradiction. Thus there is no singular arc.

Claim 1. For the linear system, if (A, b_i) is controllable for all i, then there is no singular arc. We call such system a *normal system.*

Example 1 (Double integrator, normal system)**.** Consider a double integrator

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= u \end{aligned}$$

with initial condition (ξ, η). The objective is to drive the initial condition to the origin $x(0) = (0, 0)$ in minimal time under the constraint $|u| \le 1$. The first observation is that the system is a single input system which is controllable. Thus there exists no singular arc.

The Hamiltonian is $H = p_1 x_2 + p_2 u + p_0$. The costate equation reads

$$\begin{aligned} \dot{p}_1 &= 0 \\ \dot{p}_2 &= -p_1 \end{aligned}$$

Thus $p_1 = c_1$, $p_2 = -c_1 t + c_2$ for some constants c_1 and c_2, and

$$u^*(t) = \operatorname{sign}(p_2(t)).$$

There are four possibilities for the control sequence: 1) $(-1, 1)$; 2) $(1, -1)$; 3) (1); 4) (-1). The last two are special cases of the first two and it suffices to discuss the first two cases.

Case 1): In this case, $c_1 < 0$. At the switching time t_s, $-c_1 t_s + c_2 = 0$, hence $t_s = c_2/c_1$, from which we see that $c_2 > 0$. Now integrate the system from the initial state (ξ_1, ξ_2) to get

$$\begin{cases} x_2(t) = \xi_2 - t, \\ x_1(t) = \xi_1 + \xi_2 t - \frac{1}{2}t^2 \end{cases}, \ t \in [0, t_s]$$

and

$$\begin{cases} x_2(t) = \xi_2 - 2t_s + t, \\ x_1(t) = \xi_1 + \xi_2 t_s - \frac{1}{2}t_s^2 + (\xi_2 - 2t_s)(t - t_s) + \frac{1}{2}(t^2 - t_s^2) \end{cases}, \ t \in (t_s, t_f]$$

In view of the terminal condition $x_1(t_f) = x_2(t_f) = 0$, we get an equation

$$\begin{cases} \xi_2 - 2t_s + t_f = 0, \\ \xi_1 + \xi_2 t_s - \frac{1}{2}t_s^2 + (\xi_2 - 2t_s)(t_f - t_s) + \frac{1}{2}(t_f^2 - t_s^2) = 0 \end{cases}$$

from which we can solve

$$t_s = \xi_2 \pm \sqrt{\xi_1 + \frac{\xi_2^2}{2}}$$

provided that

$$\xi_1 + \frac{\xi_2^2}{2} > 0.$$

When $\xi_1 + \frac{\xi_2^2}{2} = 0$, ($\xi_2$ must be negative) then there is no switching and the control sequence is (-1) for all $t \geq 0$. Otherwise, there is one switch. For $\xi_2 < 0$, in order that $t_s > 0$, we must have $t_s = \xi_2 + \sqrt{\xi_1 + \frac{\xi_2^2}{2}}$ and $\xi_2^2 < \xi_1 + \frac{\xi_2^2}{2}$, or $\xi_1 > \frac{\xi_2^2}{2}$.

When $t \in [0, t_s)$, we see

$$x_1 = -\frac{1}{2}x_2^2 + \text{const},$$

therefore, we can draw the phase plot as in Figure 2.

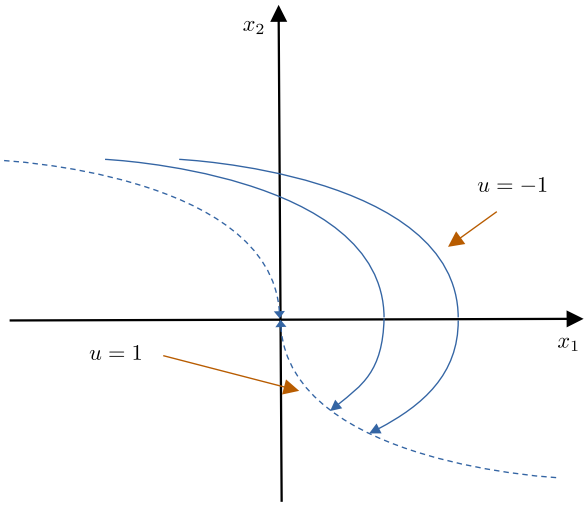

Figure 2: Minimal time double integrator

Case 2) is done in exactly the same way.

Ok, this is the unfinished work of the last lecture. Now, I want to discuss some extensions of the MP and some theoretical aspects that are worth mentioning.

4.2 Some extensions of the maximum principle

4.2.1 Time varying cost

Up until now, we have always been focused on time invariant systems. Let's see if our MP can also be extended to time varying systems with possibly time varying costs. For that, let us consider the system

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

with cost function

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

If we introduce a new variable $x_{n+1} := t$, then the original system can be written as

$$\begin{bmatrix} \dot{x} \\ \dot{x}_{n+1} \end{bmatrix} = \begin{bmatrix} f(x_{n+1}, x, u) \\ 1 \end{bmatrix}$$

which turns into time invariant. The cost functional becomes

$$J(u) = \varphi(x_{n+1}, x) + \int_0^T L(x_{n+1}, x, u)\mathrm{d}t$$

which is too, time invariant. The constraints, initial conditions etc., shall also be put into the new coordinate, e.g., the initial condition

$$\begin{bmatrix} x(0) \\ x_{n+1}(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ 0 \end{bmatrix}$$

and the terminal constraint $x_{n+1}(T) > 0$. After this, it is enough to apply the MP for time invariant system. For example, in this case

$$H(x, x_{n+1}, u, p, p_{n+1}, p_0) = p^\top f(x_{n+1}, x, u) + p_{n+1} + p_0 L(x, x_{n+1}, u).$$

The costate equation is

$$\dot{p} = -H_x^\top$$
$$\dot{p}_{n+1} = -H_{x_{n+1}}$$

and the maximum principle

$$H(x^*(t), t, u^*(t), p^*(t), p^*_{n+1}(t), p^*_0) = \max_u H(x^*(t), t, u, p^*(t), p^*_{n+1}(t), p^*_0).$$

Example 2 (Paper mill production [5])**.** The paper mill production system (suggested by Karl Johan) can be described by a linear system of the form

$$\dot{x} = Cv(t) + Bu$$

where $x(t) \in \mathbb{R}^n$ represents the tank levels in the system, i.e., the state that needs to be controlled, $v(t) \in \mathbb{R}^m$ is the given paper production – a time varying signal and u is the control input. The constraint for control input is $u(t) \in \Omega_u$ for some convex polytope Ω_u. The control objective is to minimize

$$J = \int_0^1 \sum_{i=1}^m |u_i(t) - a_i(t)| \mathrm{d}t$$

for some given time varying signal $a(\cdot)$. As pointed out in [5], due to the structure of the cost functional, this problem can be turned into a LP.

4.2.2 State constraints

In the present form of the maximum principle, there does not involve any state constraint. However, state constraint is extremely important in practical problems. Fortunately, there is one type of state constraints that can be easily handled, namely, the mixed input and state constraint. In this setting, the constraint is described by variable control region, say $u(t) \in U(x(t))$ for all t.

Fortunately, the maximum principle also holds for variable control region in the sense that if the input space U at each moment is a function of the state, say $U(x) \subseteq \mathbb{R}^m$, the MP still holds by merely changing the maximum principle to

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

and leaving the rest unchanged.

Example 3 (Rayleigh problem)**.** Consider minimizing

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

subject to (the controlled van de Pol oscillator):

$$\dot{x}_1 = x_2,$$
$$\dot{x}_2 = -x_1 + x_2(1.4 - 0.14x_2^2) + 4u$$

with initial condition $(x_1(0), x_2(0)) = (-5, -5)$, $t_f = 4.5$ and a mixed input and state constraint:

$$-1 \le u(t) + \frac{x_1(t)}{6} \le 0.$$

The Hamiltonian (assume there's no abnormal extremals) for this problem is

$$H(x,u,p) = p_1 x_2 + p_2(-x_1 + x_2(1.4 - 0.14x_2^2) + 4u) - (u^2 + x_1^2)$$

and the costate equation is

$$\begin{aligned}\dot{p}_1 &= p_2 + 2x_1\\ \dot{p}_2 &= -p_1 - p_2(1.4 - 0.42x_2^2)\end{aligned}$$

with boundary condition $p_1(t_f) = p_2(t_f) = 0$. The maximum principle says

$$u^*(t) = \arg \max_{-1-\leq u+\frac{x_1^*(t)}{6} \leq 0} (-u^2 + 4p_2 u).$$

Thus in principle one can solve this problem by solving the canonical equation together with the NLP. Since in principle, you can solve $u^*(t)$ as a function of p_2, then we substitute u^* into the system equation. At last the canonical equation becomes autonomous with know initial or terminal condition which can be solved numerically.

There is also another type of constraints which is much harder and admits only limited solution: pure state constraint

$$g_i(x(t)) \leq 0,\ i = 1, \cdots, k.$$

Unlike the mixed input state case, this situation is quite hard to solve in general. Normally, this is solved by discretization and then solving the problem as a nonlinear programming problem. There do exist some interesting results regarding MP with pure state constraints, interested readers may refer to a survey paper [3] or a more recent online tutorial [4]. The basic idea is to take the differential of g_i until the input appears explicitly in the expression an then treats the constraint as a mixed state input constraint.

4.2.3 Existence of optimal control

There is still the problem of existence of optimal controller that we haven't answered yet. You remember the contradiction that Karl Johan posed in the first lecture, which says that if there does not exist an optimal controller, then the necessary condition may not make any sense. Fortunately, in practice, for most of the time, it's quite safe to apply the MP directly without checking the existence of an optimal controller. This is because we don't need very strong assumptions to guarantee the existence of an optimal controller. The following condition is an easily understandable one.

Assumption (H). The set $U \subseteq \mathbb{R}^m$ is compact, $f : \Omega \times U \to \mathbb{R}^n$ is continuous in (x,u) and C^1 in x. And that f has linear growth bound on x:

$$|f(x,u)| \leq C(1 + |x|)$$

on $\Omega \times U$ for some constant C. (Recall that linear growth bound guarantees forward completeness)

Now consider the optimal control problem:

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

with admissible input set

$$\mathcal{U}_{\text{ad}} = \{u : \mathbb{R}_{\geq 0} \to U\}$$

and terminal constraint $x(t_f) \in S \subseteq \mathbb{R}^n$. We have the following useful theorem:

Theorem 2. *Let the assumption **(H)** hold. Assume S is closed and reachable*[1]*, φ and L are continuous. If the sets*

$$F(x) := \{(y_0, y) \in \mathbb{R}^{n+1} : y_0 \geq L(x,u),\ y = f(x,u) \text{ for some } u \in U\}$$

are convex, then the above problem has an optimal solution.

[1]There exists at least one controller in $\mathcal{U}_{\text{ad}}$ which drives the initial condition to the set S.

Example 4 (Linear system)**.** Consider the linear system

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

with costs

$$J_1 = L_f^\top x(t_f) + \int_0^{t_f} L^\top x + S^\top u \mathrm{d}t$$

and

$$J_2 = x^\top(t_f) Q_f x(t_f) + \int_0^{t_f} x^\top Q x + u^\top R u \mathrm{d}t$$

for some $R > 0$. The input is constrained on a polytope $|u_i| \leq 1$ for all i. Then the assumptions of Theorem 2 are satisfied for both cost functions.

Proof strategy:

Step 1: Let $(x^{(k)}(\cdot), u^{(k)}(\cdot))$ be a minimizing sequence, i.e., $J(u^{(k)}) \to \inf_u J(u)$.

Step 2: Extract a sub-sequence $(x^{(k_j)}(\cdot), u^{(k_j)}(\cdot))$ such that $x^{(k_j)}(\cdot)$ converges uniformly to some $x^*(\cdot)$ on $[0, t_f]$. Then show $x^*(\cdot)$ is an admissible trajectory – this is where the convexity of $F(x)$ is used.

To readers can refer to [2] for a complete proof.

4.2.4 Sufficient condition

One should always bear in mind that like the EL equation, the maximum principle only provides sufficient condition for a minimizer. But you have noticed that in all the examples, we never checked whether the solution obtained by the MP is a true minimizer. The reason behind this is quite simple. Again, we consider the optimal control of the system

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

with $x \in \Omega \subseteq \mathbb{R}^n$ and $u \in U \subseteq \mathbb{R}^m$, terminal constraint $x(t_f) \in S$, and cost function

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

Assumption (H'). The set $U \subseteq \mathbb{R}^m$ of control values is compact. The vector filed $f, L : \Omega \times U \to \mathbb{R}^n$ are continuous on $\Omega \times U$ for some open set $\Omega \subset \mathbb{R}^n$ and are C^1 w.r.t. x.

Under this technical assumption, we have the following theorem.

Theorem 3. *Let assumption **(H')** hold. Assume that an optimal solution exists, and that there are only finite admissible control functions that satisfy the maximum principle. Then the one which yields the lowest value of the cost is optimal. In particular, if there is only one solution to the maximum principle, then that solution is automatically optimal.*

Proof. Obvious. □

Very nice! In practice, we have noticed that in most problems, the solution to the maximum principle admits unique solution, therefore by the theorem, that solution is optimal.

4.3 Proof of the maximum principle

Now, let's make a brief summary. First, we studied the CoV, and then we tried to use CoV to study the optimal control problem. But there we could only make some conjectures because of the essential limitations of CoV. To cope with that, we introduced the maximum principle, which turned out to be quite successful and not so complicated to use. But still, we haven't yet proved the maximum principle. In the rest of the time, we will start proving the maximum principle. Before that, I would like to cite a paragraph written by L.C. Young in his famous book *Lectures on the Calculus of Variations and Optimal Control (1969),* a mathematician famous for his work in calculus of variation and optimal control:

> The proof of the maximum principle, given in the book of Pontryagin, Boltyanskii, Gamkrelidze and Mischenko... represents, in a sense, the culmination of the efforts of mathematicians, for considerably more than a century, to rectify the Lagrange multiplier rule.

From this paragraph you can already see that the proof of the MP must not be trivial. But you will soon see that the ideas of the proof are rather simple. However, these are really great ideas and are well worth learning.

Remember that we mentioned at least two essential difficulties in CoV. One is that its inefficiency to handle constraints. The other is that it requires certain smoothness conditions which are too restrictive in practice. Thus in order to prove the maximum principle, we will need some new tools for doing nonsmooth analysis. Fortunately, only one non-trivial tool will be sufficient for our purpose, i.e., the separability of tents.

Tent method

To motivate the idea, we consider a static nonlinear optimization problem:

$$\begin{aligned} &\min\ g_0(x) \\ &\text{subject to } g_i(x) \le 0, \quad i = 1, \cdots, m \end{aligned} \tag{LM}$$

in which $\{g_i\}_{i=0}^m \in C^1(\mathbb{R}^n; \mathbb{R})$. Suppose that the problem is feasible, i.e., there exists an admissible x_* which minimizes $g_0(x)$.

To solve this optimization problem, it is standard practice to use the so called *Lagrangian multiplier* method. Or you can also use CoV that we have introduced previously.

Exercise. Derive the first order necessary condition of the (LM) problem using calculus of variation.

The *method of tent* is a totally different approach which is very powerful and which works for non-smooth problems. This method was introduced by Boltyanskii (student of Pontrayagen) and his colleagues when proving the maximum principle. Let's see how the method works.

Define the following sets:

$$\Omega_i = \{x \in \mathbb{R}^n : g_i(x) \le 0\}, \quad i = 1, \cdots, m$$

So the constraints can also be expressed as

$$x \in \Omega_1 \cap \cdots \cap \Omega_m.$$

And for $x_1 \in \mathbb{R}^n$, let

$$\Omega_0 = \{x : g_0(x) < g_0(x_1)\} \cup \{x_1\}.$$

Take the intersection of all these sets

$$\Sigma := \Omega_0 \cap \Omega_1 \cap \cdots \cap \Omega_m.$$

We claim that x_1 is a minimizer *if and only if* $\Sigma = \{x_1\}$. To see this, suppose x_1 is a minimizer, then $g_i(x_1) \le 0$ for $i \ge 1$ and $g_0(x_1) \le g_0(x)$ for all feasible x. Thus $x_1 \in \Sigma$. If there is another point $x_2 \in \Sigma$, then x_2 is feasible and $g_0(x_2) < g_0(x_1)$, a contradiction, thus if x_1 is a minimizer, there must hold $\Sigma = \{x_1\}$. Conversely, suppose that $\Sigma = \{x_1\}$, if x_1 is not a minimizer, then either x_1 is not feasible or there exists $x_2 \ne x_1$, both feasible such that $g_0(x_2) < g_0(x_1)$. For the first case, $x_1 \notin \Omega_1 \cap \cdots \cap \Omega_m$, thus $x_1 \notin \Sigma$, a contradiction. For the second case, $\{x_1, x_2\} \subseteq \Sigma$, a contradiction.

As an example, let $m = 1$ and Figure 3 show two sets Ω_0 and Ω_1 on a plane. In this figure, Ω_1 and Ω_0 intersects only at the point x_1. So x_1 is a minimizer.

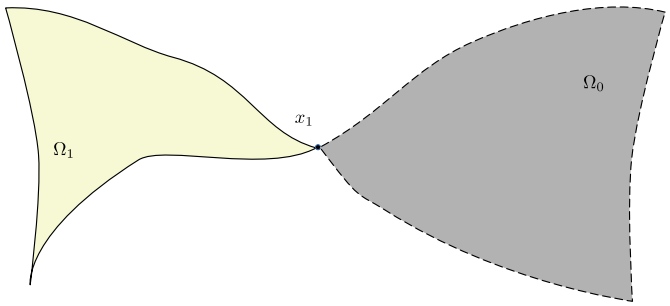

Figure 3: Separating 2-dim tents.

However, in Figure 4, the the intersection of the two sets contains also some other points. Thus x_1 is not a minimizer.

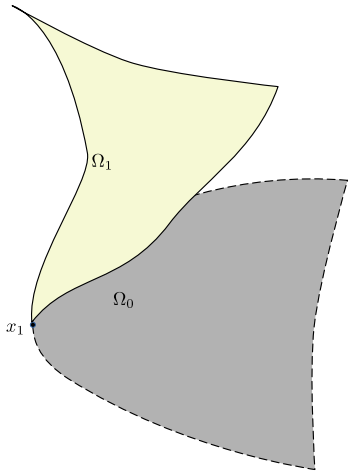

Figure 4: Separating 2-dim tents.

Next, we need the notion of tangent cone. Given a set $\Omega \subseteq \mathbb{R}^n$ (may be non-convex), the *tangent cone* at $x \in \Omega$ is defined as

$$T_x\Omega := \left\{ v \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \exists \{x_i\}_{i=1}^{\infty} \subseteq \Omega,\ \exists \{t_i\}_{i=1}^{\top} \subseteq \mathbb{R}_{>0}, \text{ s.t.} \\ t_i \downarrow 0,\ x_i \to x, \text{ and } (x_i - x)/t_i \to v \end{array} \right\}$$

see Figure 5.

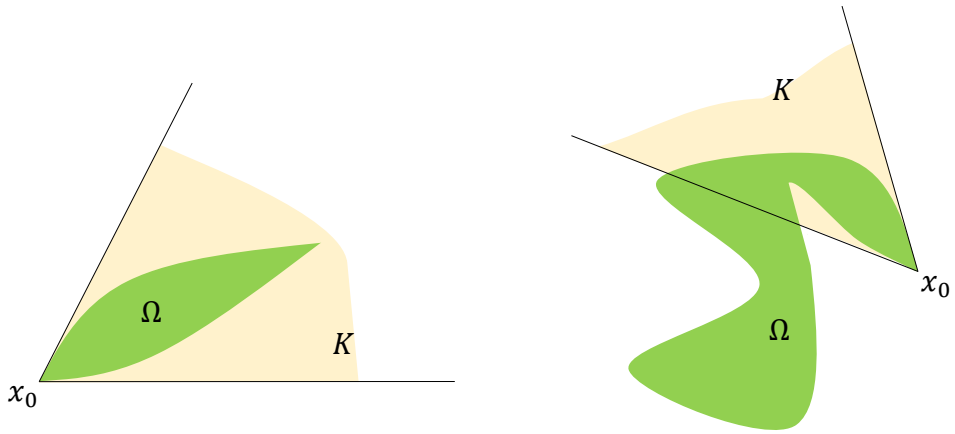

Figure 5: Tangent cones of convex and non-convex sets Ω.

In particular, when Ω is a smooth sub-manifold – think of a smooth surface in $\mathbb{R}^n$ – then $T_x\Omega$ is the tangent space of Ω at x. Hence the notation coincides with tangent space in the smooth case.

Definition 1 (Tent)**.** Given a set Ω and its tangent cone $T_x\Omega$ at x, a *tent* is a convex convex cone $K \subseteq T_x\Omega$ with apex x.

The reason why we need to use the notion of tent instead of a tangent cone is that a tangent cone of a set may be non-convex and that non-convex objects are hard to work with. aIn Figure 6, K_0 represents the tangent cones while K_1 some tents.

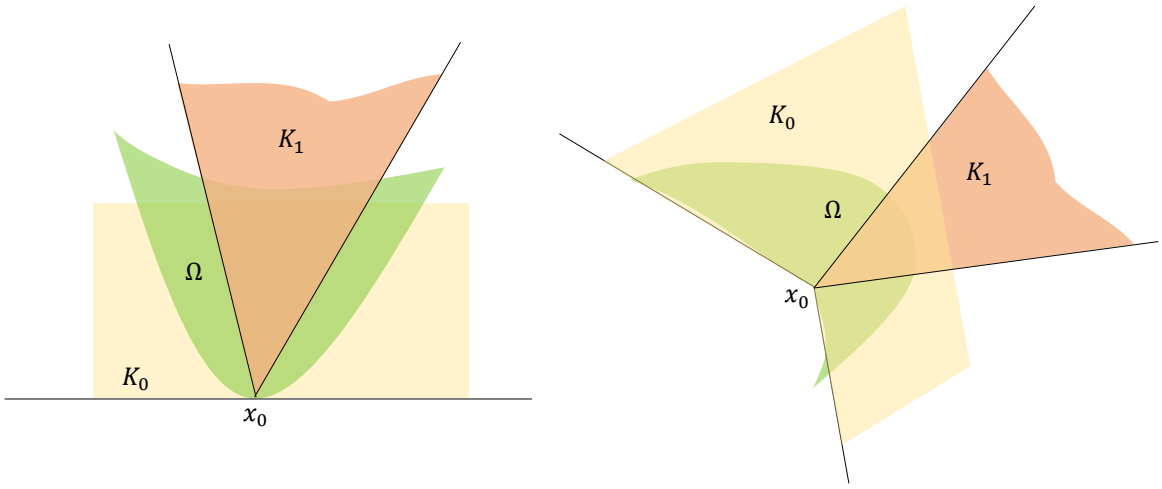

Figure 6: Tents.

Intuitively, to be able to "separate" Ω_0 and Ω_1, the tents of the two sets at the intersecting point should also be separable in the sense that they intersect only at the apex. Or equivalently, there is a hyperplane passing through x_1 which separates $T_{x_1}\Omega_0$ and $T_{x_1}\Omega_1$, see Figure 7.

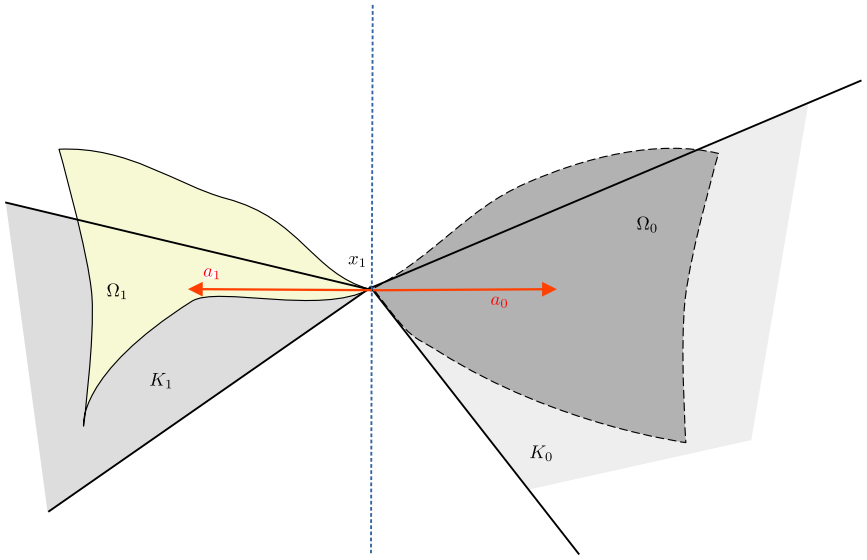

Figure 7: Separating 2-dim tents.

In Figure 7, let us choose two arbitrary nonzero vectors a_0 and a_1 perpendicular to the separating hyperplane such that

$$a_0 + a_1 = 0 \tag{3}$$

Furthermore, we see that

$$a_i^\top (x - x_1) \geq 0, \quad \forall x \in K_i,\ i = 0, 1. \tag{4}$$

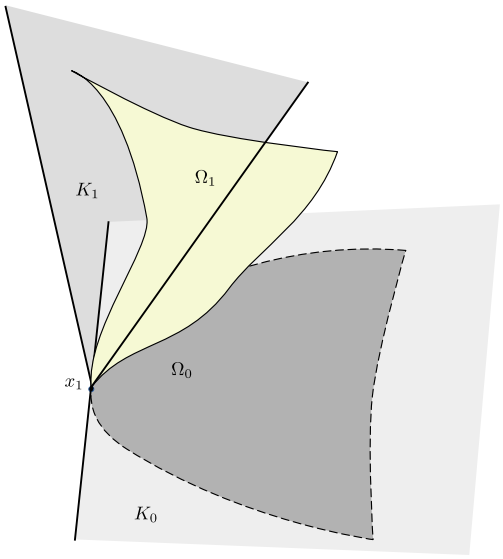

Figure 8: Tents not separable.

Thus if we can find K_0 and K_1, we can obtain a necessary condition based on the relation (4). For problem (LM), this is easy since g_0 and g_1 are smooth:

$$K_i = \{x : \nabla g_i(x_1)(x - x_1) \leq 0\}, \quad i = 0, 1$$

That is, K_i are half spaces, see Figure 9.

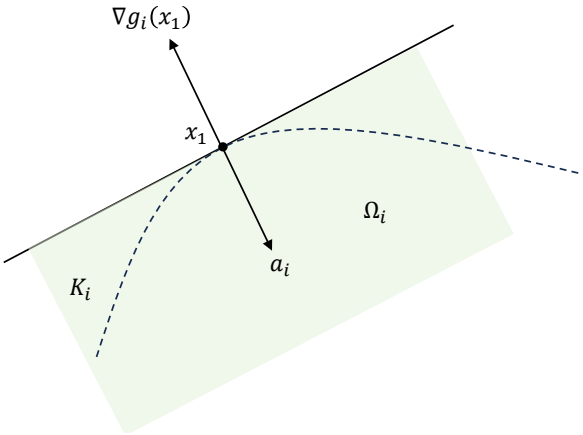

Figure 9: The tents are half spaces.

Therefore, a_i must be of the form

$$a_i = \lambda_i \nabla g_i(x_1)$$

for $\lambda_i \leq 0$. Since λ_i cannot be zero at the same time, $\lambda_i < 0$ for $i = 0, 1$. Thus the relation (3) becomes

$$\lambda_0 \nabla g_0(x_1) + \lambda_1 \nabla g_1(x_1) = 0$$

for $\lambda_0, \lambda_1 < 0$ or

$$\nabla g_0(x_1) + \lambda \nabla g_1(x_1) = 0$$

for some $\lambda > 0$. This is a special case of the famous KKT (Karush-Kuhn-Tucker) condition which we will be able to prove once we have generalize the above reasoning.

The separability of tents

We generalize our previous discussions to arbitrary finite many tents.

Definition 2 (Separability)**.** Let $K_0, \cdots, K_p$ be some closed, convex cones with a common apex x in $\mathbb{R}^n$. They are said to be *separable* if there exists a hyper plane Γ through x that separates one of the cones from the intersection of the others.

Lemma 1. *Let $K_0, \cdots, K_p$ be some closed, convex cones with a common apex x in $\mathbb{R}^n$. Then they are separable if and only if there exist dual vectors $a_i,\ i = 0, 1, \cdots, p$ fulfilling*[2]

$$a_i^\top (y - x) \leq 0, \quad \forall y \in K_i$$

and at least one of which is not zero and such that

$$a_0 + \cdots + a_p = 0.$$

Lemma 2. *Let $\Omega_0, \cdots, \Omega_p$ be sets in $\mathbb{R}^n$ satisfying*

$$\Omega_0 \cap \cdots \cap \Omega_p = \{x\},$$

and $K_0, \cdots, K_p$ be tents of these sets at x. If all the tents are convex and that at least one of the tents is distinct from its affine hull. Then $K_0, \cdots, K_p$ is separable.

The proofs of the above two results are quite technical and are hence omitted. Interested readers are referred to [1].

Problem statement

We start by introducing the optimal control problem under fixed terminal time. First, let us recall our optimal control problem. We focus on time-invariant control systems:

$$\dot{x} = f(x, u), \tag{5}$$

where $x(t) \in \mathbb{R}^n,\ u(t) \in U \subset \mathbb{R}^m$ for all $t \in [0, t_f]$, the initial condition $x(0) = x_0$ is assumed to be fixed. The cost function is

$$J(u(\cdot)) = \varphi(x(t_f)) + \int_0^{t_f} L(x(s), u(s))ds,$$

where $\varphi(\cdot),\ f(\cdot, u),\ L(\cdot, u)$ are continuously differentiable for all u. The optimal control problem amounts to finding a process $u_*(t),\ x_*(t),\ 0 \leq t \leq t_f$, with a (measurable) controller $u_*(t)$ such that $x_*(t_f) \in M$ for some manifold M, and $J(u_*(\cdot))$ attains a minimum. We say that the problem is in 1) *Mayer form* if $L = 0$; 2) *Lagrange form* if $\varphi = 0$; 3) *Bolza form* if neither L nor φ is zero.

We claim that the preceding three types of optimal control problems can all be reduced to Mayer form. In fact, let

$$x_{n+1}(t) = \int_0^t L(x(s), u(s))ds$$

the system becomes

$$\begin{cases} \dot{x} = f(x, u) \\ \dot{x}_{n+1} = L(x, u) \end{cases} \tag{6}$$

\with initial condition $(x_0, 0)$, and the cost function becomes

$$J = \varphi(x(t_f)) + x_{n+1}(t_f). \tag{7}$$

This is an optimal control problem of the Mayer form of a time-invariant system. Due to this reason, it suffices to study the optimal control problem with cost function:

$$J = \varphi(x(t_f)).$$

[2]Note that we can also use $a_i^\top (y - x) \geq 0$ by reversing the sign of a_i, see (4).

Introduce the following notations which will be used in the proof:

$$\begin{aligned} &x_1 := x_*(t_f) \\ &\Omega_0 = \{x_1\} \cup \{x : \varphi(x) < \varphi(x_1)\} \\ &\Omega_1 : \text{ reachability set from } x_0 \\ &\Omega_2 = M\text{: the terminal manifold} \end{aligned}$$

Let $u_*(t)$, $x_*(t)$, $0 \le t \le t_f$ be an optimal process. Then it is easily seen that

$$\Omega_0 \cap \Omega_1 \cap \Omega_2 = \{x_1\}. \tag{8}$$

The reader should immediately realize that such type of condition implies separability of tents of the three sets, this is the content of Lemma 2. Denote K_i the tent of Ω_i at x_1. It thus remains to find the tents K_i. The tents K_0 and K_2 can be easily computed:

$$\begin{aligned} K_0 &= \{x \in \mathbb{R}^n : \nabla\varphi(x_1)(x - x_1) \le 0\} \\ K_2 &= T_{x_1}\Omega_2 \end{aligned}$$

(note that Ω_2 is a fixed manifold).

Therefore, our problem boils down to calculating the tent of Ω_1 at x_1: K_1. By definition, a tent is only a convex subcone of the tangent cone of Ω_1 at x_0, we should however, try to find a tent as big as possible, since the bigger the tent, the more necessary information it conveys. This is the main non-trivial step in proving the maximum principle (if we already know Lemma 1, 2) and was first achieved by Boltyanskii and his colleagues using the so called needle variation.

Needle variation

Suppose at the moment that the optimal control $u_* : [0, t_f] \to U$ is continuous. Fix $\tau \in (0, t_f]$ and consider the following *needle shaped variation* of u_* for small $\varepsilon > 0$:

$$u_\varepsilon(t) = \begin{cases} w, & t \in (\tau - \varepsilon, \tau] \\ u_*(t), & \text{otherwise} \end{cases}$$

where $w \in U$ is some constant, see Figure 10.

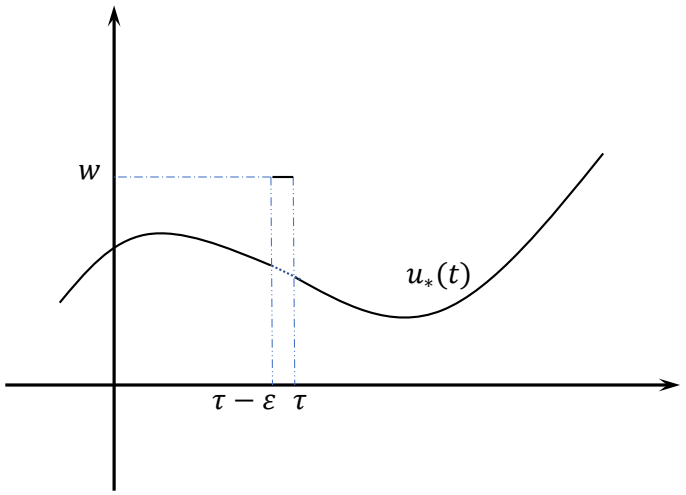

Figure 10: Needle variation.

Denote $t \mapsto x_\varepsilon(t)$ the solution to $\dot{x} = f(x, u_\varepsilon)$. Obviously, $u_\epsilon(\cdot)$ is admissible, thus $x_\epsilon(t_f)$ belongs to the reachable set at t_f, i.e., $x_\epsilon(t_f) \in \Omega_1$ for all ϵ chosen above. Thus by definition, $\left.\frac{\partial x_\varepsilon(t_f)}{\partial \varepsilon}\right|_{\varepsilon=0+}$

must belong to the tangent cone of Ω_1. Denote

$$v(t) = \left.\frac{\partial x_\varepsilon(t)}{\partial \varepsilon}\right|_{\varepsilon=0+}, \quad t \in [\tau, t_f]$$

then it remains to find $v(t_f)$. We call $v(t_f)$ a *deviation vector*. To find the deviation vector, first we need to characterize $x_\epsilon(t)$. Denote $v_\epsilon(t) = \frac{\partial x_\varepsilon(t)}{\partial \varepsilon}$, since $u_\epsilon(t) = u_*(t)$ for $t \geq \tau$, it follows that

$$\begin{aligned}\frac{dv_\epsilon(t)}{dt} &= \frac{\partial}{\partial \epsilon} f(x_\epsilon(t), u_*(t)) = \frac{\partial f}{\partial x}(x_\epsilon(t), u_*(t)) \frac{\partial x_\epsilon(t)}{\partial t} \\ &= \frac{\partial f}{\partial x}(x_\epsilon(t), u_*(t)) v_\epsilon(t), \quad \forall t \in (\tau, t_f]\end{aligned}$$

Evaluating at $\epsilon = 0+$, we get $\dot{v}(t) = \frac{\partial f}{\partial x}(x_*(t), u_*(t))v(t)$. That is, $v(t)$ satisfies a linear ODE. It still remains to find the initial condition $v(\tau)$. Note that

$$\begin{aligned}x_\varepsilon(\tau) &= x_*(\tau - \epsilon) + \int_{\tau-\epsilon}^{\tau} f(x_\epsilon(s), w)\mathrm{d}s, \\ &= x_*(\tau - \epsilon) + \int_{\tau-\epsilon}^{\tau} f(x_*(s), u_*(s))\mathrm{d}s + \int_{\tau-\epsilon}^{\tau} [f(x_\epsilon(s), w) - f(x_*(s), u_*(s))]\mathrm{d}s \\ &= x_*(\tau) + \int_{\tau-\epsilon}^{\tau} [f(x_\epsilon(s), w) - f(x_*(s), u_*(s))]\mathrm{d}s\end{aligned}$$

thus

$$\begin{aligned}v(\tau) &= \lim_{\varepsilon \to 0+} \frac{x_\varepsilon(\tau) - x_*(\tau)}{\varepsilon} \\ &= \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} \left[\int_{\tau-\varepsilon}^{\tau} f(x_\varepsilon(t), w)dt - f(x_*(t), u_*(t))\mathrm{d}t \right] \\ &= f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)).\end{aligned} \tag{9}$$

To summarize, $v(\cdot)$ is the solution to the following Cauchy problem

$$\begin{cases} \dot{v} = \dfrac{\partial f}{\partial x}(x_*(t), u_*(t))v, \quad \forall t \in [\tau, t_f] \\ v(\tau) = f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)). \end{cases}$$

To construct more deviation vectors, let $v_1(t_f), \cdots, v_r(t_f)$ be some different deviation vectors obtained as above corresponding to some distinct time instants $\tau_1 < \cdots < \tau_r$ and constant inputs $w_1, \cdots, w_r$. Consider the combined needle variation

$$u_{\varepsilon,k}(t) = \begin{cases} w_i, & t \in (\tau_i - k_i\varepsilon, \tau_i] \text{ for some } i \in \{1, \cdots, r\} \\ u_*(t), & \text{otherwise} \end{cases}$$

where k_i are non-negative constants satisfying $\sum_{i=1}^r k_i = 1$. One can show that

$$\sum_{i=1}^{r} k_i v_i(t_f) = \left.\frac{\partial x(t_f, u_{\varepsilon,k})}{\partial \varepsilon}\right|_{\varepsilon=0+}$$

which implies that $\sum_{i=1}^r k_i v_i(t_f)$ are again in $T_{x_1}\Omega_1$. Still call these vectors deviation vectors and define K_1 to be the set of all deviation vectors, i.e.,

$$K_1 = \left\{ \sum_{i=1}^{r} k_i v_i(t_f) \;\middle|\; \begin{array}{l} \exists r \in \mathbb{Z}_+,\ \tau_i \in [0, t_f),\ w_i \in U,\ k_i \geq 0,\ \sum_{i=1}^r k_i = 1, \\ v_i(t_f) \text{ the deviation vector obtained from needle} \\ \text{variation at} \tau_i \text{ with spike } w_i \end{array} \right\}$$

Then K_1 is a tent of Ω_1 at x_1.

Final step: the costate equation and the maximum principle

Condition (8) implies that K_0, K_1, K_2 are separable. Invoking Lemma 1 and Lemma 2, we deduce that there exist three vectors a_i, at least one of which is nonzero, satisfying

$$a_i^\top v \le 0, \quad v \in K_i,\ i = 0, 1, 2 \tag{10}$$

and

$$a_0 + a_1 + a_2 = 0. \tag{11}$$

In particular, $a_1^\top v(t_f) \le 0$ for any deviation vector $v(t_f)$. Now we introduce a small trick: if we are able to construct some function $p : [0, t_f] \to \mathbb{R}^n$ such that $p(t)^\top v(t) \equiv$ constant with $p(t_f) = a_1$, then we obtain immediately $p(t)^\top v(t) = a_1^\top v(t_f) \le 0$ for all $t \in [0, t_f]$. In particular, if v is the deviation vector obtained by needle variation at time τ with spike w, then $v(\tau) = f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau)$. Thus at $t = \tau$, $p(\tau)^\top [f(x_*(\tau), w) - f(x_*(\tau), u_*(\tau))] \le 0$ or

$$p(\tau)^\top f(x_*(\tau), u_*(\tau)) \ge p(\tau)^\top f(x_*(\tau), w)) \tag{12}$$

For convenience, define

$$H(x, u, p) := p^\top f(x, u)$$

which is the Hamiltonian associated with the system . Now that the spike can be any $w \in U$ and $t \in [0, t_f)$, it follows from (12) that

$$H(x_*(t), u_*(t), p(t)) = \max_{u \in U} H(x_*(t), u, p(t)) = \text{constant}, \quad \forall t \in [0, t_f). \tag{13}$$

This is the maximum principle that we have been looking for! Except two things: the interval $[0, t_f)$ doesn't include the endpoint t_f and the function p hasn't been determined yet. The first issue can be fixed if everything is continuous in the above formula, which is indeed true as long as we have shown p is, since f, x_* and u_* are continuous as assumed. For the second issue, let us recall the following simple fact:

Lemma 3. *Consider two linear ODE*

$$\dot{x} = A(t)x$$
$$\dot{p} = -A(t)^\top p$$

where $x, p \in \mathbb{R}^n$. Then $p(t)^\top x(t) = p(t')^\top x(t')$ for any $t, t' \in \mathbb{R}$.

With this lemma, we can now construct p to be the solution of the following ODE

$$\begin{aligned} \dot{p} &= -\left[\frac{\partial f}{\partial x}(x_*(t), u_*(t))\right]^\top p \\ &= -H_x^\top (x_*, u_*, p) \end{aligned} \tag{14}$$

with terminal state $p(t_f) = a_1$ (note that this is exactly the costate equation).

Recall that

$$K_0 = \{x \in \mathbb{R}^n : \nabla\varphi(x_1)(x - x_1) \le 0\}$$
$$K_2 = T_{x_1}\Omega_2$$

For a_0, since K_0 is a half space, $a_0^\top v \le 0$ for $v \in K_0$ implies $a_0 = \lambda \nabla\varphi(x_1)^\top$ for some constant $\lambda \ge 0$. For a_2, since K_2 is a sub-manifold, $a_2 \perp K_2$. It follows from (11) that (recall $a_1 = p(t_f)$):

$$\lambda \nabla\varphi(x_*(t_f))^\top + p(t_f) \perp \Omega_2 \tag{15}$$

for some constant $\lambda \ge 0$.

Up to now, we have prove the maximum principle for the Mayer problem under the assumption that u_* is continuous.

For u not continuous, only the condition (13) needs to be modified by noticing that the limits in (9) exist for almost all $t \in [0, t_f]$. Summarizing, we have proved the following.

References

[1] Vladimir Grigorevich Boltyanskii. The method of tents in the theory of extremal problems. *Russian Mathematical Surveys*, 30(3):1, 1975.

[2] Alberto Bressan and Benedetto Piccoli. *Introduction to the mathematical theory of control*, volume 1. American institute of mathematical sciences Springfield, 2007.

[3] Richard F Hartl, Suresh P Sethi, and Raymond G Vickson. A survey of the maximum principles for optimal control problems with state constraints. *SIAM review*, 37(2):181–218, 1995.

[4] Helmut Maurer. Tutorial on control and state constrained optimal control problems. In *SADCO Summer School 2011-Optimal Control*, 2011.

[5] Bengt Pettersson. *Production Control of a Pulp and Paper Mill*. PhD thesis, Lund Institute of Technology, 1970.