Curve minimization; old and recent results

H. J. Sussmann

Department of Mathematics
Rutgers University
Piscataway, NJ 08854, USA
sussmann@math.rutgers.edu

PENN STATE UNIVERSITY
February 22, 2007

These slides are available at http://www.math.rutgers.edu/~sussmann/slides.html

This work is partially supported by grant NSF-DMS 05-09930. The author is grateful to NSF for the support.

In ancient Egypt and India the art of *rope-stretching* was practiced in order to produce segments and other geometrical figures, showing knowledge of the fact that a segment solves the problem of maximizing the distance between the endpoints of an arc given the length of the arc.

In Egypt, according to Democritus, geometric constructions were carried out with the help of specialized workers—whom Democritus describes as experts in "composing lines," and calls by the Greek word "harpenodaptai," which means "rope-strethcers"—by means of pegs and cords. Their skills were used to build altars and temples, where it was deemed necessary to produce certain geometric shapes that would obey very precise specifications, such as being made of perfectly straight segments or perfect circles.

THE EULER-LAGRANGE EQUATION

Problem:

minimize $I = \int_a^b L(\xi(t), \dot{\xi}(t), t)\, dt$

in the space $\mathcal{W}$ of "all curves" $\xi : [a, b] \mapsto \Omega$ such that $\xi(a) = \bar{x}$ and $\xi(b) = \hat{x}$. (Here Ω is a given open subset of $\mathbb{R}^n$, and $(x, \dot{x}, t) \mapsto L(x, \dot{x}, t) \in \mathbb{R}$ is a given function on $\Omega \times \mathbb{R}^n \times [a, b]$, known as the Lagrangian.)

Necessary condition for a solution:

Suppose $\xi_* : [a, b] \mapsto \Omega$ solves the problem Let $\pi : [a, b] \mapsto \mathbb{R}_n$ be the momentum along ξ_* given by $\boxed{\pi(t) = \frac{\partial L}{\partial \dot{x}}(\xi_*(t), \dot{\xi}_*(t), t)\,.}$

Then $\boxed{\boxed{\dot{\pi}(t) = \frac{\partial L}{\partial x}(\xi_*(t), \dot{\xi}_*(t), t)\,.\ \text{for a.e. } t.}}$ (More precisely, π is absolutely continuous and the double-boxed condition holds.)

The words "Calculus of Variations" were introduced by Euler in 1754, when he got a letter from an 18-year youngster from Turin, called Joseph-Louis Lagrange, who suggested a method for proving results such as the one of the previous slide.

Laqrange's idea: make "variations" of ξ_*, replacing ξ_* by $\xi_* + \delta\xi$, where $\delta\xi$ is an "infinitesimal variation" of ξ_*, chosen in such a way that $\delta\xi(a) = \delta\xi(b) = 0$. Then, if $\Xi_*(t) = (\xi_*(t), \dot{\xi}_*(t), t)$,

$$\begin{aligned}
\delta I &= \int_a^b \Big(L(\xi_*(t) + \delta\xi(t), \dot{\xi}_*(t) + \delta\dot{\xi}(t), t) - L(\xi_*(t), \dot{\xi}_*(t), t) \Big)\, dt \\
&= \int_a^b \Big(\frac{\partial L}{\partial x}(\Xi_*(t)) \cdot \delta\xi(t) + \frac{\partial L}{\partial \dot{x}}(\Xi_*(t)) \cdot \delta\dot{\xi}(t) \Big)\, dt \\
&= \int_a^b \Big(\frac{\partial L}{\partial x}(\Xi_*(t)) - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}}(\Xi_*(t)) \Big) \cdot \delta\xi(t) \Big)\, dt \\
&= \int_a^b \Big(\frac{\partial L}{\partial x}(\Xi_*(t)) - \dot{\pi}(t) \Big) \cdot \delta\xi(t) \Big)\, dt\,.
\end{aligned}$$

But $\delta I = 0$ for all $\delta\xi$. Since $\delta\xi$ is arbitrary, we get $\boxed{\boxed{\frac{\partial L}{\partial x}(\Xi_*(t)) = \dot{\pi}(t).}}$

Euler liked Lagrange's method of "variations" so much that he named the study of these problems "Calculus of Variations."

The above argument can be made completely rigorous by just writing $\xi_* + \varepsilon\zeta$ instead of $\xi_* + \delta\xi$, and differentiating the integral with respect to ε at $\varepsilon = 0$.

Technical conditions are needed: for example, the proof works if the function $(x, \dot{x}, t) \mapsto L(x, \dot{x}, t)$ is of class C^1, and "all curves" means "all maps in $W^{1,\infty}([a, b], \Omega)$" (i.e., all Lipschitz curves).

But it does not work, even if L is very smooth (for example, a polynomial) if "all curves" is taken to mean "all maps in $W^{1,1}([a, b], \Omega)$" (i.e., all absolutely continuous curves), because of the "Lavrentiev phenomenon."

Adrien-Marie Legendre (1752-1833) found in 1786 an additional necessary condition for a minimum.

His condition, derived by him for the scalar case, is

$$\boxed{\frac{\partial^2 L}{\partial \dot{x}^2}(\Xi_*(t)) \geq 0\,.}$$

With an appropriate reinterpretation, Legendre's condition is also necessary in the vector case: all we have to do is read it as asserting that the Hessian matrix $\{\frac{\partial^2 L}{\partial \dot{x}^i \partial \dot{x}^j}(\Xi_*(t))\}_{1\leq i,j\leq n}$ is nonnegative definite.

Here are our three conditions together:

$$\boxed{\pi(t) = \frac{\partial L}{\partial \dot{x}}(\Xi_*(t))} \qquad \boxed{\dot{\pi}(t) = \frac{\partial L}{\partial x}(\Xi_*(t))} \qquad \boxed{\frac{\partial^2 L}{\partial \dot{x}^2}(\Xi_*(t)) \geq 0\,.}$$

DO YOU SEE ANYTHING HERE?

Define a function $H(x, p, \dot{x}, t)$ of the three sets of variables x, p, u, and of $t \in \mathbb{R}$, by letting

$$\boxed{H(x, p, \dot{x}, t) = \langle p, \dot{x} \rangle - L(x, \dot{x}, t)\,.}$$

(Really, H is defined on the fiber product of the tangent and cotangent bundles of the configuration space Ω.)

Then our conditions so far say, if $\Theta_*(t) = (\xi_*(t), \pi(t), \dot{\xi}_*(t), t)$, that, along Θ_*,

$$\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0, \qquad \dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$\frac{\partial^2 H}{\partial \dot{x}^2}(\Theta_*(t)) \leq 0\,.$$

The first system of three equations, usually written as

$$\boxed{\frac{dx}{dt} = \frac{\partial H}{\partial p}\,, \qquad \frac{dp}{dt} = -\frac{\partial H}{\partial x}\,, \qquad \frac{\partial H}{\partial \dot{x}} = 0\,,}$$

is exactly equivalent to Euler-Lagrange. And, of course, the fourth inequality is Legendre's condition.

$$\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0, \qquad \dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$\frac{\partial^2 H}{\partial \dot{x}^2}(\Theta_*(t)) \leq 0\,.$$

DO YOU SEE ANYTHING NOW?

$$\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0, \qquad \dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$\frac{\partial^2 H}{\partial \dot{x}^2}(\Theta_*(t)) \leq 0\,.$$

Clearly, what must be going on is that the Hamiltonian $H(\xi_*(t), \pi(t), \dot{x}, t)$ is being maximized, as a function of $\dot{x}$, by the value $\dot{x} = \dot{\xi}_*(t)$.

This is indeed true! And it could have been discovered, or at least guessed, in the 1830s. But it took until the 1870s for Weierstrass to discover it. And he didn't quite discover it all the way.

Weierstrass introduced the "excess function"

$$\mathcal{E}(x,u,v,t) = L(x,v,t) - \frac{\partial L}{\partial v}(x,u,t)\cdot v - (L(x,u,t) - \frac{\partial L}{\partial v}(x,u,t)\cdot u)\,,$$

depending on three sets of independent variables x, u and v. He then proved his *side condition*:

$$\mathcal{E}(\xi_*(t),\dot\xi_*(t),v,t) \geq 0 \quad \text{for all } v\,.$$

If we plug $x = \xi_*(t)$, $u = \dot\xi_*(t)$, in the formula for $\mathcal{E}$, we find

$$\begin{aligned}
&\mathcal{E}(\xi_*(t),\dot\xi_*(t),v,t)\\
&\qquad = L(\xi_*(t),v,t) - \pi(t)\cdot v - (L(\xi_*(t),\dot\xi_*(t),t) - \pi(t)\cdot\dot\xi_*(t))\\
&\qquad = H(\xi_*(t),\pi(t),\dot\xi_*(t),t) - H(\xi_*(t),\pi(t),v,t)\,,
\end{aligned}$$

so the Weierstrass condition says, precisely, that $H(\xi_*(t),\pi(t),\dot x,t)$ is maximized, as a function of $\dot x$, by the value $\dot x = \dot\xi_*(t)$.

But the Weierstrass condition only amounts to saying that the Hamiltonian is maximized if one substitutes $\pi(t)$ for $\frac{\partial L}{\partial \dot{x}}(\xi_*(t), \dot{\xi}_*(t), t)$.

This substitution, however, is redundant, because the equation $\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0$ already implies that $\pi(t) = \frac{\partial L}{\partial \dot{x}}(\xi_*(t), \dot{\xi}_*(t), t)$. (Recall: $H = p \cdot \dot{x} - L$.) And Hamiltonian maximization implies $\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0$.

Furthermore, the condition $\frac{\partial^2 H}{\partial \dot{x}^2}(\Theta_*(t)) \leq 0$ is also implied by Hamiltonian maximization.

So we can drop the conditions $\frac{\partial H}{\partial \dot{x}}(\Theta_*(t)) = 0$ and $\frac{\partial^2 H}{\partial \dot{x}^2}(\Theta_*(t)) \leq 0$, and just write our full system of conditions as

$$\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$H(\xi_*(t), \pi(t), \dot{\xi}_*(t), t) = \max\{H(\xi_*(t), \pi(t), u, t) : v \in U\}$$

where U is the set of all possible velocity values (which in Weierstrass' setting would just be $\mathbb{R}^n$).

The system of conditions

$$\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$H(\xi_*(t), \pi(t), \dot{\xi}_*(t), t) = \max\{H(\xi_*(t), \pi(t), u, t) : u \in U\}$$

is exactly equivalent to ELLW (Euler+Lagrange+Legendre+Weierstrass) under Weierstrass' assumptions, namely:

1. $U = \mathbb{R}^n$, or at least U open,
2. L is differentiable with respect to $\dot{x}$.

But now the conditions make sense even without these assumptions! (For example, they contain no derivatives of L with respect to $\dot{x}$.)

Therefore it is reasonable to guess that the conditions are still necessary in this more general situation.

And that is indeed true. (Almost.)

Tentative necessary conditions for $\xi_* : [a,b] \mapsto \Omega$ to be a minimizer of the cost $\int_a^b L(\xi(t), \dot{\xi}(t), t)\, dt$ in the class of all absolutely continuous curves $\xi : [a,b] \mapsto \Omega$ such that $\dot{\xi}(t) \in U$ for a. e. t and $\xi(a) = \bar{x}$, $\xi(b) = \widehat{x}$:

For some absolutely continuous $\pi : [a,b] \mapsto \mathbb{R}_n$:

$$\dot{\xi}_*(t) = \frac{\partial H}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H}{\partial x}(\Theta_*(t)).$$

$$H(\xi_*(t), \pi(t), \dot{\xi}_*(t), t) = \max\{H(\xi_*(t), \pi(t), u, t) : u \in U\}$$

Recall: $H(x,p,u,t) = p \cdot u - L(x,u,t)$.

This is very nice, but it doesn't quite work.

We need the abnormal multiplier.

That is, we must redefine the Hamiltonian as

$$H_{p_0}(x, p, u, t) = p \cdot u - p_0 L(x, u, t) \,,$$

and allow $p_0 \geq 0$ but not necessarily $= 1$.

REMARK: "Abnormal multipliers" occur in very old problems (e.g., the catenary), even though Bolza around 1910 seems to have been the first one to notice them.

Necessary conditions for $\xi_* : [a,b] \mapsto \Omega$ to be a minimizer of the cost $\int_a^b L(\xi(t), \dot{\xi}(t), t)\, dt$ in the class of all absolutely continuous curves $\xi : [a,b] \mapsto \Omega$ such that $\dot{\xi}(t) \in U$ for a. e. t and $\xi(a) = \bar{x}$, $\xi(b) = \hat{x}$:

For some absolutely continuous $\pi : [a,b] \mapsto \mathbb{R}_n$ and some $\pi_0 \geq 0$ such that $(\pi_0, \pi(t)) \neq (0,0)$:

$$\dot{\xi}_*(t) = \frac{\partial H_{\pi_0}}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H_{\pi_0}}{\partial x}(\Theta_*(t)).$$

$$H_{\pi_0}(\xi_*(t), \pi(t), \dot{\xi}_*(t), t) = \max\{H_{\pi_0}(\xi_*(t), \pi(t), u, t) : u \in U\}$$

Recall: $H_{p_0}(x, p, u, t) = p \cdot u - p_0 L(x, u, t)$.

Technical assumptions:

1. U is a subset of $\mathbb{R}^n$,
2. $x \mapsto L(x,u,t)$ is of class C^1 for each t,u,
3. $t \mapsto L(x,u,t)$ is measurable for each x,u,
4. for each $u \in U$ there exist $\delta_u > 0$, $k_u : [a,b] \mapsto [0,+\infty]$ integrable, such that $|L(x,u,t)| + \|\nabla_x L(x,u,t)\| \leq k_u(t)$ for $\|x - \xi_*(t)\| \leq \delta_u$, $a \leq t \leq b$,
5. there exist $\delta_* > 0$, $k_* : [a,b] \mapsto [0,+\infty]$ integrable, such that $|L(x,\dot{\xi}_*(t),t)| + \|\nabla_x L(x,\dot{\xi}_*(t),t)\| \leq k_*(t)$ for $\|x - \xi_*(t)\| \leq \delta_*$, $a \leq t \leq b$.

To see the difference between the ELLW conditions and our new form, in which the momentum is a completely independent variable, consider these examples:

1. Minimize (that is, find all the minimizers) $\int_0^1 \|\dot{\xi}(t)\|\, dt$ in the class of all Lipschitz curves $\xi : [0,1] \mapsto \mathbb{R}^n$ such that $\xi(0) = A$ and $\xi(1) = B$. Here $\|\cdot\|$ is a general (not necessarily smooth) norm on $\mathbb{R}^n$.

2. Minimize $\int_0^{10} \xi(t)^2\, dt$ in the class of all Lipschitz functions $\xi : [0,1] \mapsto \mathbb{R}$ with Lipschitz constant ≤ 1, such that $\xi(0) = 1$ and $\xi(10) = 1$.

Notice that for Problem 2 Euler-Lagrange would give $\frac{\partial L}{\partial x} = 0$, and then $\xi(t) \equiv 0$, which is of course impossible.

Furthermore, we can now take the velocity to be much more general, $f(x,u)$ rather than u, and we get the Pontryagin Maximum Principle:

Necessary conditions for $\xi_* : [a,b] \mapsto \Omega$, $\eta_* : [a,b] \mapsto U$, to be a minimizer of the cost $\int_a^b L(\xi(t),\eta(t),t)\,dt$ in the class of all pairs (ξ,η) such that $\xi : [a,b] \mapsto \Omega$ is absolutely continuous, $\eta : [a,b] \mapsto U$, $\dot{\xi}(t) = f(\xi(t),\eta(t),t)$ for a.e. t, and $\xi(a) = \bar{x}$, $\xi(b) = \hat{x}$:

For some absolutely continuous $\pi : [a,b] \mapsto \mathbb{R}_n$ and some $\pi_0 \geq 0$ such that $(\pi_0, \pi(t)) \neq (0,0)$:

$$\dot{\xi}_*(t) = \frac{\partial H_{\pi_0}}{\partial p}(\Theta_*(t)), \qquad \dot{\pi}(t) = -\frac{\partial H_{\pi_0}}{\partial x}(\Theta_*(t)).$$

$$H_{\pi_0}(\xi_*(t),\pi(t),\eta_*(t),t) = \max\{H_{\pi_0}(\xi_*(t),\pi(t),u,t) : u \in U\}$$

Now: $H_{p_0}(x,p,u,t) = p \cdot f(x,u,t) - p_0 L(x,u,t)$.

Hamiltonian = momentum times velocity minus abnormal multiplier times Lagrangian.

Technical assumptions:

(Letting $\mathbf{f}(x,u,t) = (L(x,u,t), f(x,u,t))$.)

1. U is a set,
2. $x \mapsto \mathbf{f}(x,u,t)$ is of class C^1 for each t, u,
3. $t \mapsto \mathbf{f}(x,u,t)$ is measurable for each x, u,
4. for each $u \in U$ there exist $\delta_u > 0$, $k_u : [a,b] \mapsto [0,+\infty]$ integrable, such that $\|\mathbf{f}(x,u,t)\| + \|\frac{\partial \mathbf{f}}{\partial x}(x,u,t)\| \leq k_u(t)$ for $\|x - \xi_*(t)\| \leq \delta_u$, $a \leq t \leq b$,
5. there exist $\delta_* > 0$, $k_* : [a,b] \mapsto [0,+\infty]$ integrable, such that $\|\mathbf{f}(x,\eta_*(t),t)\| + \|\frac{\partial \mathbf{f}}{\partial x}(x,\eta_*(t),t)\| \leq k_*(t)$ for $\|x - \xi_*(t)\| \leq \delta_*$, $a \leq t \leq b$.

THE TRANSVERSALITY CONDITION

Suppose that, instead of fixing the initial and terminal conditions $\xi(a) = \bar{x}$, $\xi(b) = \hat{x}$, we impose conditions

$\xi(a) = \bar{x}$, $\xi(b) \in S$, where S is some given set.

Suppose C is a tangent cone to S at $\xi_*(b)$. (For example, S is, up to a C^1 diffeomorphism, a closed convex set near $\xi_*(b)$, and C is the tangent cone in the obvious sense.)

Then we get the extra necessary condition:

$$-\pi(b) \in C^\perp$$

where $C^\perp = \{p : p \cdot c \leq 0 \quad \text{for all } c \in C\}$, i.e., $C^\perp$ is the polar cone of C.

THE MEANING OF THE MOMENTUM

If the momentum need no longer be equal to $\frac{\partial L}{\partial \dot{x}}$, then what does it mean?

ANSWER: Suppose we want to end up in a "target set" S at time T. Define the "value function" $V(x,t)$ by

$$V(x,t) = \inf\Big\{\int_t^T L(\xi(s),\dot{\xi}(s),s)\,ds : \xi \text{ such that } \xi(t)=x, \xi(T)\in S\Big\}.$$

Then (modulo lots of technical conditions) $\boxed{\pi(t) = -\nabla_x V(\xi(t),t)}$.

And one can think of $\pi(t)$ as a "shadow price" for your control:

$$\text{Maximizing} \qquad H(x,p,u,t) = p\cdot f(x,u,t) - L(x,u,t)$$

(ignoring the abnormal multiplier) means: if you are paid $p\cdot v\,dt$ for moving in the $v = f(x,u,t)$ direction during time dt, and in addition you have to pay $L(x,u,t)\,dt$, then $H(x,p,u,t)$ tells you how much you should value choosing the control u, and Hamiltonian maximization just says that you should choose the control that gives you the best value.

PROVING AND GENERALIZING THE MAXIMUM PRINCIPLE

The method of proof introduced by Pontryagin *et al.* in their 1962 book is based on four key ideas:

1. SYSTEM AUGMENTATION

2. NEEDLE VARIATIONS

3. PROPAGATION

4. SEPARATION

I. AUGMENTATION

We augment our controlled dynamics by adding the running cost as a new state variable. Now the state $\mathbf{x} = (x_0, x)$ evolves in $\mathbb{R} \times \Omega$. The dynamical equations are

$$\dot{x}_0 = L(x, u, t)\,, \qquad \dot{x} = f(x, u, t)\,,$$

that is $\boxed{\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, u, t)}$, where $\mathbf{f} = (f_0, f)$, and $\mathbf{f}(\mathbf{x}, u, t)$ only depends on x, that is, does not depend on x_0.

Let $\hat{\mathcal{R}}$ be the set of all pairs (x_0, x) that are reachable from $(0, \bar{x})$ over the time-interval $[a, b]$ by means of a trajectory of the new system.

Then $(x_0, x) \in \hat{\mathcal{R}}$ if and only if x can be reachable from $\bar{x}$ over $[a, b]$ with cost c.

Let $\xi_{0,*}(t) = \int_a^t L(\xi_*(s), \eta_*(s), s)\, ds$, $c_* = \xi_*(b)$, $\Xi_* = (\xi_{0,*}, \xi_*)$, so $\Xi_*(b) = (c_*, \xi_*(b))$.

Hence (ξ_*, η_*) is optimal if and only if $\hat{\mathcal{R}} \cap \hat{S} = \{\Xi_*(b)\}$, where $S = \{(x_0, x) : x \in S \wedge x_0 \leq c_* - \|x - \xi_*(b)\|^2\}$.

So we now have a set separation problem.

Any general necessary condition for two sets A and B to be separated at a point q (meaning "$A \cap B \subseteq \{q\}$") will give rise to a necessary condition for optimal control by applying it to our sets $\mathcal{R}$, $\widehat{S}$.

Presumably, a necessary condition for separation of two sets at a point q should involve their "tangent approximations" at q, whatever those may be.

EXAMPLE. Suppose S_1 and S_2 are two submanifolds of class C^1 of $\mathbb{R}^n$, and $q \in S_1 \cap S_2$. Then a necessary condition for S_1 and S_2 to be separated at q is that the following should not be true:

$$\boxed{T_q S_1 + T_q S_2 = \mathbb{R}^n \text{ and } T_q S_1 \cap T_q S_2 \neq \{0\}.}$$

One has to generalize this to more irregular sets, having "tangent cones" instead of "tangent subspaces."

From now on, we look for conditions for the separation problem for a system

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

We fix $\bar{x} = \xi(a)$, and let $\mathcal{R}$ be the reachable set from $\bar{x}$ over $[a, b]$.

We let S be some other given set.

We look for necessary conditions for $\mathcal{R} \cap S = \{\xi_*(b)\}$.

It is then trivial to apply the result to an augmented problem and get a necessary condition for optimal control.

II. NEEDLE VARIATIONS

We make needle variations of the reference control at various times. (This idea comes from Weierstrass.)

These variations depend on a parameter ε, and have the effect of moving us away from the reference trajectory in certain directions by a certain amount, so that the effect at time t of a variation is, to first order, εv, where v is a certain vector.

The vectors v are propagated from time t to time b by means of the reference flow, to vectors $P_{b,t}(v)$.

We then get a set of tangent vectors at the terminal point $\xi_*(b)$ These vectors form a "tangent cone" C to the reachable set at $\xi_*(b)$.

A needle variation of the reference control at time τ is a 1-parameter family of controls $\{\eta_\varepsilon\}_{0\le\varepsilon\le\bar{\varepsilon}}$ given by

$$\begin{aligned}\eta_\varepsilon(t) &= \eta_*(t) \quad \text{if} \quad t \notin [\tau, \tau+\varepsilon] \\ \eta_\varepsilon(t) &= u \quad \text{if} \quad t \in [\tau, \tau+\varepsilon]\,.\end{aligned}$$

Let $v_{u,\tau} = f(\xi_*(\tau), u, \tau) - f(\xi_*(\tau), \eta_*(\tau), \tau)$.

Then the effect of the variation at time τ is $\boxed{\boxed{\varepsilon v_{u,\tau} + o(\varepsilon)}}$.

III. PROPAGATION

The reference vector field $(x,t) \mapsto f(x,\eta_*(t),t)$ has flow maps $\Phi^*_{t,s}$.

These flow maps have differentials $D\Phi^*_{t,s}(\xi_*(s)) : T_{\xi_*(s)}\Omega \mapsto T_{\xi_*(t)}\Omega$.

Write $P_{t,s} = D\Phi^*_{t,s}(\xi_*(s))$. Then the $P_{t,s}$ can be used to propagate tangent vectors from $T_{\xi_*(s)}\Omega$ to $T_{\xi_*(t)}\Omega$.

In particular, if we propagate all the vectors $v_{u,\tau}$ from $T_{\xi_*(\tau)}\Omega$.to $T_{\xi_*(b)}\Omega$, we get a huge set $\mathcal{V}$ of tangent vectors to the reachable set at $\xi_*(b)$.

Let $\mathcal{P}$ be the convex cone generated by the vectors $P_{b,\tau}v_{u,\tau}$.

Then $\mathcal{P}$ is a "tangent cone" to $\mathcal{R}$.

IV. SEPARATION

Now we have a "tangent cone" $\mathcal{P}$ to $\mathcal{R}$ at $\xi_*(b)$, and a tangent cone C to S at $\xi_*(b)$.

It is natural to guess that if the sets are separated then the cones are linearly separated, that is, there is a nonzero covector $\bar{\pi} \in T_{\xi_*(b)}\Omega$ such that

$$\pi \cdot w \leq 0 \quad \text{whenever} \quad w \in \mathcal{P}\,,$$

$$\pi \cdot c \geq 0 \quad \text{whenever} \quad c \in C\,,$$

Now let $\pi(t) = \bar{\pi} \circ P_{b,t}$. Then $\pi(t) \cdot v_{u,t} \leq 0$, so

$$\pi(t) \cdot f(\xi_*(t), u, t) \leq \pi(t) \cdot f(\xi_*(t), \eta_*(t), t)\,.$$

And $\pi(b) \in C^{\perp}$.

How does one translate all the above heuristics into rigorous arguments?

Pontryagin *et al.*, in their 1962 book, did it in one way (Type T), for systems of class C^1 in x.

Later, Clarke 1972 did it in a different way (Type L), for systems Lipschitz in x.

In 1991, S. Lojasievicz Jr. discovered a way to go beyond the Clarke hypotheses via Type T methods. (Technically, Lojasiewicz was able to include systems that were only continuous with respect to the state x, exceot that the rerefence vector field had to be Lipschitz.)

My own work comes after this, extending the use of Type T methods as well as the understanding of when they work and when they do not, and reaching the conclusion that there are two non-comparable proof techniques that cannot be combined.

V. A ROUGH CLASIFICATION OF VERSIONS OF THE FDPMP (FINITE-DIMENSIONAL PONTRYAGIN MAXIMUM PRINCIPLE)

Every known version of the FDPMP is of one of the following two types:

- Type T. (The "T" stands for "topological.")
- Type L. (The "L" stands for "limiting.")

In the transversality condition:

- Type T versions involve some kind of Boltyanskii tangent cone to the terminal set.
- Type L versions involve the Clarke tangent cone to the terminal set, or the Mordukhovich normal cone

The proofs of Type T versions typically use a topological separation argument, based on the Brouwer fixed point theorem or some variant thereof.

All versions of the finite-dimensional Pontryagin maximum principle with high-order conditions (Knobloch, Krener, Bianchini-Stefani, Agrachev, Sarychev, Gamkrelidze, and many others) appear to be Type T.

The finite dimensionality comes in where the Brouwer fixed-point theorem is used, since that theorem depends esssentially on being in a finite-dimensional space.

The proofs of Type L versions usually produce a sequence $\{\bar{p}_k\}_{k\in\mathbb{N}}$ of "approximate terminal adjoint covectors" (using, for example, the Ekeland variational principle) and then extract a convergent (or weakly convergent) subsequence whose limit $\bar{p}_\infty$ is the terminal value of the adjoint covector.

The finite dimensionality comes in when one tries to establish that $\bar{p}_\infty \neq 0$. The $\bar{p}_k$ can be normalized so that $\|\bar{p}_k\| = 1$, and the existence of a weak*-convergent subsequence (if, say, we are working on a Hilbert space) follows from the weak*-compactness of the closed unit ball, but in infinite dimensions one cannot prove in general that $\bar{p}_\infty \neq 0$, since the unit sphere is not weak*-compact.

NATURAL QUESTIONS:

- Is it possible to unify all these versions, and their proofs, into a single general theorem?

- If so, would that theorem be Type T, Type L, or of some new type, involving techniques that somehow combine or go beyond those of the two basic types?

- In particular, is there a FDPMP with high-order conditions that would apply to a dynamical law that in some portion of the reference trajectory is only Lipschitz, and with a transversality condition involving a Clarke or Mordukhovich cone?

It turns out that the key issue is whether the following property is true:

The Transversal Intersection Property (TIP)

If two subsets S_1, S_2 of $\mathbb{R}^n$ have tangent cones C_1, C_2 at a point $p \in \mathbb{R}^n$, and the cones C_1, C_2 are strongly transversal, then $S_1 \cap S_2$ contains a sequence of points p_j converging to p and $\neq p$.

VI. CONES

A. Definition

A *cone* in a real linear space X is a subset C of X which is nonempty, and closed under multiplication by nonnegative scalars. (In particular, if C is a cone then necessarily $0 \in C$.)

B. Definition

The *polar* of a cone C in a real linear normed space X is the set $C^\perp$ of all $w \in X^\dagger$ such that $\langle w, c \rangle \leq 0$ for all $c \in C$. Clearly, $C^\perp$ is always a closed convex cone. If X is finite-dimensional (so $X \sim X^{\dagger\dagger}$ canonically), then $C^{\perp\perp}$ is the smallest closed convex cone containing C, from which it follows in particular that $C^{\perp\perp} = C$ if and only if C is closed and convex.

REMARK: $X^\dagger$ is the dual of X.

C. Definition

Assume that $S \subseteq \mathbb{R}^n$ and $p \in S$. The *Bouligand tangent cone* to S at p is the set of all vectors $v \in \mathbb{R}^n$ such that there exist

(i) a sequence $\{p_j\}_{j\in\mathbb{N}}$ of points of S converging to p,

(ii) a sequence $\{h_j\}_{j\in\mathbb{N}}$ of positive real numbers converging to 0,

such that

$$v = \lim_{j\to\infty} \frac{p_j - p}{h_j}.$$

D. Notation

We use $T_p^B S$ to denote the Bouligand tangent cone to S at p. (It is then clear that $T_p^B S$ is always a closed cone.)

E. Definition

Assume that $S \subseteq \mathbb{R}^n$ and $p \in S$. A *Boltyanskii approximating cone* to S at p is a convex cone C in $\mathbb{R}^n$ having the property that there exist

(i) a nonnegative integer m,
(ii) a closed convex cone D in $\mathbb{R}^m$,
(iii) a neighborhood U of 0 in $\mathbb{R}^m$,
(iv) a continuous map $F : U \cap D \mapsto S$,
(v) a linear map $L : \mathbb{R}^m \mapsto \mathbb{R}^n$,

such that

$$F(x) = p + Lx + o(\|x\|) \quad \text{as} \quad x \to 0\,,\ x \in D\,,$$

and $LD = C$.

F. Definition

Assume that $S \subseteq \mathbb{R}^n$, S is closed, and $p \in S$. The *Clarke tangent cone* to S at p is the set of all vectors $v \in \mathbb{R}^n$ such that, whenever $\{p_j\}_{j \in \mathbb{N}}$ is a sequence of points of S converging to p, it follows that there exist Bouligand tangent vectors $v_j \in T_{p_j}^B S$ such that $\lim_{j \to \infty} v_j = v$.

G. Notation

We use $T_p^C S$ to denote the Clarke tangent cone to S at p. Then $T_p^C S$ is a closed convex cone.

VII. TRANSVERSALITY

A. Definition

Two convex cones C_1, C_2 in $\mathbb{R}^n$ are *transversal* if

$$C_1 - C_2 = \mathbb{R}^n \,,$$

i.e., if for every $x \in \mathbb{R}^n$ there exist $c_1 \in C_1$, $c_2 \in C_2$, such that $x = c_1 - c_2$.

B. Remark

This is a very natural generalization to cones of the ordinary notion of transversality of linear subspaces. For subspaces S_1, S_2, it is customary to require that $S_1 + S_2 = \mathbb{R}^n$, but it would make no difference if we required $S_1 - S_2 = \mathbb{R}^n$ instead.

C. Intuition

The basic idea of transversality is that, if two objects O_1, O_2 have first-order approximations A_1, A_2 near a point p, and A_1 and A_2 are transversal, then $O_1 \cap O_2$ looks, near p, like $A_1 \cap A_2$.

VIII. NON-TRANSVERSALITY = LINEAR SEPARATION

Suppose C_1, C_2 are convex cones in $\mathbb{R}^n$. Then the following conditions are equivalent:

- C_1 and C_2 are not transversal,
- $C_1^{\perp} \cap (-C_2)^{\perp} \neq \{0\}$,
- there exists a nonzero linear functional $\bar{p} : \mathbb{R}^n \mapsto \mathbb{R}$ such that

$$\langle \bar{p}, c_1 \rangle \leq 0 \quad \text{for all} \quad c_1 \in C_1 \,,$$

and

$$\langle \bar{p}, c_2 \rangle \geq 0 \quad \text{for all} \quad c_2 \in C_2 \,.$$

IX. STRONG TRANSVERSALITY

A. Definition

Two convex cones C_1, C_2 in $\mathbb{R}^n$ are *strongly transversal* if they are transversal and in addition $\boxed{C_1 \cap C_2 \neq \{0\}}$.

B. Intuition:

If two sets S_1, S_2 have first-order approximations C_1, C_2 near a point p, and the cones C_1, C_2 are strongly transversal, it should follow that $S_1 \cap S_2$ contains points p_j converging to p and $\neq p$.

Reason:

Near p, $S_1 \cap S_2$ should look like $C_1 \cap C_2$, because C_1 and C_2 are transversal.

Since $C_1 \cap C_2$ contains a full half-line through 0, $S_1 \cap S_2$ should also contains a nontrivial curve through p.

C. An important caveat:

The above intuition is, of course, not a proof, and when one does things carefully, it turns out that, for very reasonable notions of "first-order approximation," all one can prove is that $S_1 \cap S_2$ must contain a nontrivial connected set through p, but this set could fail to be path-connected. And for other reasonable notions one can prove even less. (For example, that $S_1 \cap S_2$ contains a sequence of points $p_j \neq p$ that converges to p.)

The following lemma says that transversality and strong transversality are almost equivalent.

More precisely, the only gap between the two conditions occurs when the cones C_1 and C_2 are linear subspaces such that $C_1 \oplus C_2 = \mathbb{R}^n$, in which case C_1 and C_2 are transversal but not strongly transversal.

D. Lemma

If C_1, C_2 are convex cones in $\mathbb{R}^n$, then C_1 and C_2 are transversal if and only if either

(i) C_1 and C_2 are strongly transversal,

or

(ii) C_1 and C_2 are linear subspaces and $C_1 \oplus C_2 = \mathbb{R}^n$.

PROOF.

It suffices to assume that C_1 and C_2 are transversal but not strongly transversal and show that (ii) holds. (Recall that (ii) says: "C_1 and C_2 are linear subspaces and $C_1 \oplus C_2 = \mathbb{R}^n$.")

Let us prove that C_1 is a linear subspace. Pick $c \in C_1$. Using the transversality of C_1 and C_2 write

$$-c = c_1 - c_2, \quad c_1 \in C_1\,, \quad c_2 \in C_2\,.$$

Then $c_1 + c = c_2$. But $c_1 + c \in C_1$ and $c_2 \in C_2$. So $c_1 + c \in C_1 \cap C_2$, and then $c_1 + c = 0$, since C_1 and C_2 are not strongly transversal. Therefore $-c = c_1$, so $-c \in C_1$. This shows that $c \in C_1 \Rightarrow -c \in C_1$. So C_1 is a linear subspace. A similar argument shows that C_2 is a linear subspace. Then the transversality of C_1 and C_2 implies that $C_1 + C_2 = \mathbb{R}^n$, and the fact that they are not strongly transversal implies that $C_1 \cap C_2 = \{0\}$. Hence $C_1 \oplus C_2 = \mathbb{R}^n$. *END OF PROOF.*

X. Set separation

Two subsets S_1, S_2 of a Hausdorff topological space space T are locally separated at a point $p \in T$ if there exists a neighborhood U of p in T such that

$$S_1 \cap S_2 \cap U \subseteq \{p\}\,.$$

XI. The Transversal Intersection Property

If two subsets S_1, S_2 of $\mathbb{R}^n$ have tangent cones C_1, C_2 at a point p, and the cones C_1, C_2 are strongly transversal, then S_1 and S_2 are not locally separated at p.

The statement that "S_1 and S_2 are not locally separated at p" means the following:

$S_1 \cap S_2$ contains a sequence of points p_j converging to p and $\neq p$.

A. Remark. This is exactly the "intuition" discussed earlier.

B. Question. For what notions of "tangent cone to a set at a point" is the TIP (Transversal Intersection Property) true?

XII. How the TIP is applied to prove versions of the FDPMP

To apply the TIP to prove a version of the FDPMP for optimal control, one carries out the following steps:

St 1. Reduce the optimal control problem to a separation problem in which, for a dynamics $\boxed{\dot{x} = f(x, u, t)}$, and an interval $[a, b]$, it is required that the reachable set $\mathcal{R}(f, [a, b], \bar{x}_{in})$ be locally separated from some other given set S. (This reduction is well known. It amounts to "augmenting the system by adding the cost as a new dynamical variable".)

St 2. Construct a "tangent cone" C_1 to $\mathcal{R}(f, [a, b], \bar{x})$ at the terminal point $\bar{x}_{term}$ of the reference trajectory.

NOTE: $\mathcal{R}(f, [a, b], \bar{x})$ is the set of all points reachable from the initial point $\bar{x}_{in}$ over the interval $[a, b]$ for the dynamics f.

St 3. Compare C_1, the tangent cone to $\mathcal{R}(f, [a, b], \bar{x})$ at $\bar{x}_{term}$, to C_2, the tangent cone to S at $\bar{x}_{term}$.

St 4. Use the TIP to conclude that C_1 and C_2 cannot be strongly transversal, because $\mathcal{R}(f, [a, b], \bar{x}_{in})$ and S are locally separated at $\bar{x}_{term}$.

St 5. If we can go from "not strongly transversal" to "not transversal," then the non-transversality is exactly the existence of a nontrivial covector linearly separating C_1 and C_2, and this yields the desired "adjoint vector" of the Maximum Principle.

St 6. How do we go from "not strongly transversal" to "not transversal"? In optimal control this is easy, because the cone C_2 is, typically, the product of a tangent cone to the set of admissible terminal states times a half-line, so it is never a linear subspace.

Naturally, for all this to work one needs the notion of "tangent cone" used in the above steps to be such that the TIP is true.

THEOREM: The TIP is true if "tangent cone" is taken to mean "Boltyanskii approximating cone." (The proof of this is Type T.)

THEOREM: The TIP is true if "tangent cone" is interpreted to mean "Clarke tangent cone." (The proof of this is Type L.)

The first TIP result leads to a number of versions of the FDPMP with a Boltyanskii or Boltyanskii-like tangent cones in the transversality condition. In these versions, high-order conditions can easily be included. (Classical work by Pontryagin et al., work by Knobloch, Krener, Agrachev, Sarychev, Gamkrelidze, Bianchini, Stefani, HJS, and lots of others.) These results are all proved using the TIP for Boltyanskii cones or for some generalization of them, such as the "approximating multicones" used by HJS.

The second TIP result leads to a number of versions of the FDPMP with a Clarke or Mordukhovich normal cone in the transversality condition. (Work by Clarke, Vinter, Rockafellar, Ioffe, Mordukhovich, Loewen, da Pinho, Franskowska, and lots of others.) In these versions, it does not seem that high-order conditions can be incorporated. Most of these results are not proved by explicitly using the TIP for Clarke cones or for some generalization thereof, but work is now in progress by HJS which, it is hoped, will show that they can be proved that way.

It may seem natural to expect that a more general TIP might be true, containing both results. I conjectured (and even briefly believed I had proved) about 10 years ago that such a result was true.

The problem was solved in January, 2006, by Alberto Bressan, who proved the following:

XIII. Bressan's Theorem

There exist two closed subsets S_1, S_2 of $\mathbb{R}^4$, and two closed convex cones C_1, C_2 in $\mathbb{R}^4$, such that

- C_1 is a Boltyanskii approximating cone to S_1 at 0;
- C_2 is the Clarke tangent cone to S_2 at 0;
- C_1, C_2 are strongly transversal;
- $S_1 \cap S_2 = \{0\}$.

Using Bressan's example, one can construct an example of a Lagrange optimal control problem in $\mathbb{R}^8$ with a terminal state constraint, and an optimal trajectory-control pair (ξ_*, η_*), defined on an interval $[a_*, b_*]$, such that

- the dynamics and Lagrangian satisfy conditions that lend themselves to Type T arguments,
- the terminal set S has a Clarke tangent cone C at the terminal point of $\xi_*(b)$,
- there does not exist a nontrivial multiplier $(\pi(\cdot), \pi_0)$ (consisting of an adjoint covector $\pi(\cdot)$ and "abnormal multiplier" π_0) that satisfies the adjoint equation, the Hamiltonian maximization condition, and the transversality condition $-\pi(b_*) \in C^\perp$.

The actual construction is done in complete detail in my 2006 CDC paper, and it's sort of technical.

Remark: In this particular example, the usual nonsmooth "adjoint differential inclusion" is actually a true "adjoint differential equation."

A lot remains to be done. For example,

- find a good counterexample as above, with a very smooth optimal control problem, for which one can get lots of high-order necessary conditions for optimality involving high-order variations in the direction of Lie brackets, but for which the terminal condition on the state involves a set with a Clarke tangent cone.

- carry out the program of proving all Type L versions of the FDPMP using the "Type L" TIP. A first step in that direction was my paper in the Sevilla CDC, where I introduced a concept of "approximating multicones" (called "Mordukhovich-Warga approximating multicones") adapted to Type L arguments, and prove the TIP.

- find a good example of failure of the FDPMP for which the dynamics are appropriate to Type L arguments, but the terminal condition on the state involves a set with a Boltyanskii approximating cone. (Conjecture: this will probably happen for some problem which is governed by a differential inclusion $\dot{x} \in F(x,t)$, and whose adjoint equation is the "intrinsic adjoint equation" involving a partial convexification of the Mordukhovich normal cone to the graph of F.)

Argument for the conjecture: I have tried and tried to derive the intrinsic equation in the Type T setting and wasn't able to. This suggest to me that perhaps the intrinsic equation can only be derived with Type L methods, in which case it is reasonable to expect that it will not "go well" with a Boltyanskii tangent cone to the terminal set.