Notes for ENEE 664: Optimal Control

André L. Tits

August 2020

Contents

Chapter 1

Motivation and Scope

1.1 Some Examples

We give some examples of design problems in engineering that can be formulated as mathematical optimization problems. Although we emphasize here engineering design, optimization is widely used in other fields such as economics or operations research. Such examples can be found, e.g., in [22].

Example 1.1 Design of an operational amplifier (opamp)
Suppose the following features (specifications) are desired

1. a large gain-bandwidth product
2. sufficient stability
3. low power dissipation

In this course, we deal with parametric optimization. This means, for this example, that we assume the topology of the circuit has already been chosen, the only freedom left being the choice of the value of a number of "design parameters" (resistors, capacitors, various transistor parameters). In real world, once the parametric optimization has been performed, the designer will possibly decide to modify the topology of his circuit, hoping to be able to achieve better performances. Another parametric optimization is then performed. This loop may be repeated many times.
To formulate the opamp design problem as an optimization problem, one has to specify one (possibly several) objective function(s) and various constraints. We decide for the following goal:

minimize	the power dissipated
subject to	gain-bandwidth product $\geq M_1$ (given)
	frequency response $\leq M_2$ at all frequencies.

The last constraint will prevent two high a "peaking" in the frequency response, thereby ensuring sufficient closed-loop stability margin. We now denote by x the vector of design parameters

$$x = (R_1, R_2, \ldots, C_1, C_2, \ldots, \alpha_i, \ldots) \in \mathbf{R}^n$$

For any given x, the circuit is now entirely specified and the various quantities mentioned above can be computed. More precisely, we can define

$P(x)$ = power dissipated
$GB(x)$ = gain-bandwidth product
$FR(x,\omega)$ = frequency response, as a function of the frequency ω.

We then write the optimization problem as

$$\min\{P(x)|GB(x) \geq M_1,\ FR(x,\omega) \leq M_2\ \forall\omega \in \Omega\} \tag{1.1}$$

where $\Omega = [\omega_1, \omega_2]$ is a range of "critical frequencies." To obtain a canonical form, we now define

$$f(x) := P(x) \tag{1.2}$$
$$g(x) := M_1 - GB(x) \tag{1.3}$$
$$\phi(x,\omega) := FR(x,\omega) - M_2 \tag{1.4}$$

and we obtain

$$\min\{f(x)|g(x) \leq 0,\ \phi(x,\omega) \leq 0\ \forall\omega \in \Omega\} \tag{1.5}$$

■

Note. We will systematically use notations such as

$$\min\{f(x)\ :\ g(x) \leq 0\}$$

not to just indicate the minimum value, but rather as a short-hand for

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0 \end{array}$$

More generally, one would have

$$\min\{f(x)|g^i(x) \leq 0,\ i = 1,2,\ldots,m,\ \phi^i(x,\omega) \leq 0,\ \forall\omega \in \Omega^i,\ i = 1,\ldots,k\}. \tag{1.6}$$

If we define $g:\ \mathbf{R}^n \to \mathbf{R}^m$ by

$$g(x) = \begin{bmatrix} g^1(x) \\ \vdots \\ g^m(x) \end{bmatrix} \tag{1.7}$$

and, assuming that all the Ω^i's are identical, if we define $\phi:\ \mathbf{R}^n \times \Omega \to \mathbf{R}^k$ by

$$\phi(x,\omega) = \begin{bmatrix} \phi^1(x,\omega) \\ \vdots \\ \phi^k(x,\omega) \end{bmatrix} \tag{1.8}$$

we obtain again

$$\min\{f(x)|g(x) \leq 0,\ \phi(x,\omega) \leq 0\ \forall\omega \in \Omega\} \tag{1.9}$$

[This is called a *semi-infinite* optimization problem: finitely many variables, infinitely many constraints.]
Note. If we define

$$\psi^i(x) \;=\; \sup_{\omega\in\Omega}\phi^i(x,\omega), \quad i=1,\ldots,k, \tag{1.10}$$

(1.9) is *equivalent* to

$$\min\{f(x)|g(x)\le 0,\ \psi(x)\le 0\} \tag{1.11}$$

(more precisely, $\{x|\phi(x,\omega)\le 0\ \forall\omega\in\Omega\} \;=\; \{x|\psi(x)\le 0\}$). Further, $\psi(x)$ can be absorbed into $g(x)$.

Exercise 1.1 *Prove the equivalence between (1.9) and (1.11). (To prove $A = B$, prove $A \subset B$ and $B \subset A$.)*

This transformation may not be advisable, for the following reasons:

(i) some potentially useful information (e.g., what are the 'critical' values of ω) is lost when replacing (1.9) by (1.11)

(ii) for given x, $\psi(x)$ may not be computable exactly in finite time (this computation involves another optimization problem)

(iii) ψ may not be smooth even when ϕ is, as shown in the exercise below. Thus (1.11) may not be solvable by classical methods.

Exercise 1.2 *Suppose that ϕ: $\mathbf{R}^n \times \Omega \to \mathbf{R}$ is continuous and that Ω is compact, so that the 'sup' in (1.10) can be written as a 'max'.*

(a) Show that ψ is continuous.

(b) By exhibiting a counterexample, show that there might not exist a continuous function $\omega(\cdot)$ such that, for all x, $\psi(x) = \phi(x,\omega(x))$.

Exercise 1.3 *Again referring to (1.10), show, by exhibiting counterexamples, that continuity of ψ is no longer guaranteed if either (i) Ω is compact but ϕ is merely continuous in each variable separately or (ii) ϕ is jointly continuous but Ω is not compact, even when the "sup" in (1.10) is achieved for all x.*

On the other hand, smoothness of ϕ and compactness of Ω do not guarantee differentiability of ψ.

Exercise 1.4 *Referring still to(1.10), exhibit an example where $\phi \in C_\infty$ (all derivatives exist and are continuous), where Ω is compact, but where ψ is not everywhere differentiable.*

In this course, we will mostly limit ourselves to classical (non semi-infinite) problems (and will generally assume continuous differentiability), i.e., to problems of the form

$$\min\{f^0(x) \ : \ f(x) \leq 0, \ g(x) \ = \ 0\} \tag{1.12}$$

where $f^0 : \ \mathbf{R}^n \to \mathbf{R}, \ f : \ \mathbf{R}^n \to \mathbf{R}^m, \ g : \mathbf{R}^n \to \mathbf{R}^\ell$, for some positive integers n, m and ℓ, are continuously differentiable.

Remark 1.1 To fit the opamp design problem into formulation (1.11) we had to pick one of the design specifications as objective (to be minimized). Intuitively more appealing would be some kind of *multiobjective* optimization problem.

Remark 1.2 Problem (1.12) is broader than it may seem at first sight. For instance, it includes 0/1 variables (in the scalar case, take $g(x) := x(x-1)$) and integer variables (in the scalar case take, e.g., $g(x) := \sin(\pi x)$.

Example 1.2 Design of a p.i.d controller (proportional - integral - derivative)
The scalar plant $G(s)$ is to be controlled by a p.i.d. controller (see Figure **1.1**). Again, the structure of the controller has already been chosen; only the values of three parameters have to be determined ($x = [x_1, x_2, x_3]^T$).

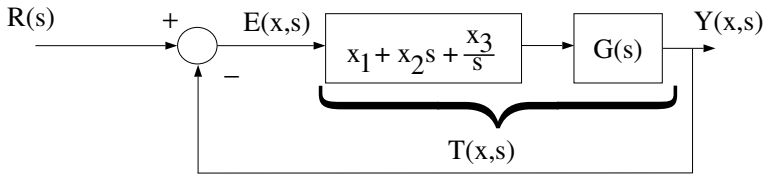


Figure **1.1**:

Suppose the specifications are as follows

- low value of the ISE for a step input (ISE = integral of the square of the difference (error) between input and output, in time domain)
- "enough stability"
- short rise time, settling time, low overshoot

We decide to minimize the ISE, while keeping the Nyquist plot of $T(x, s)$ outside some forbidden region (see Figure **1.2**) and keeping rise time, settling time, and overshoot under given values. The following constraints are also specified.

$$-10 \leq x_1 \leq 10 \quad , \quad -10 \leq x_2 \leq 10 \quad , \quad .1 \leq x_3 \leq 10$$

Exercise 1.5 *Put the p.i.d. problem in the form (1.6), i.e., specify f, g^i, ϕ^i, Ω^i.*

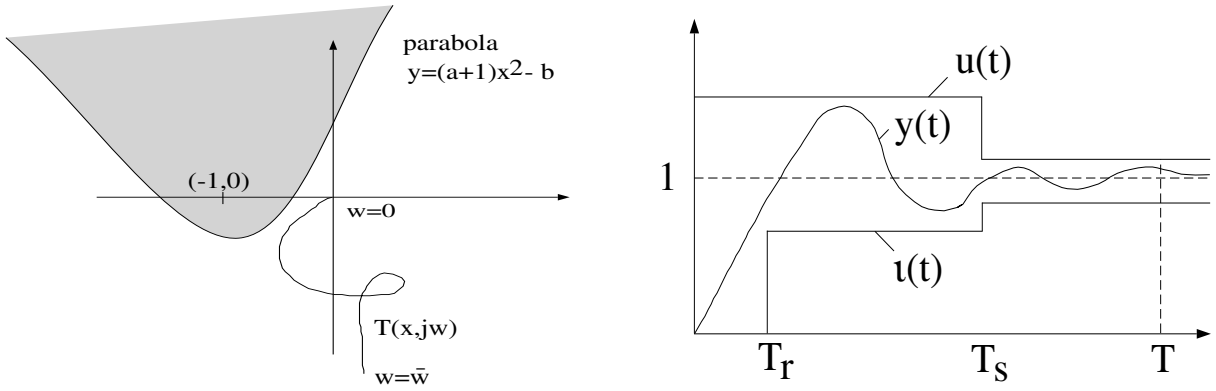


Figure **1.2**:

$T(x, j\omega)$ has to stay outside the forbidden region $\forall \omega \in [0, \bar{\omega}]$

For a step input, $y(x,t)$ is desired to remain between $l(t)$ and $u(t)$ for $t \in [0,T]$

Example 1.3 Consider again a plant, possibly nonlinear and time varying, and suppose we want to determine the best control $u(t)$ to approach a desired response.

$$\begin{aligned} \dot{x} &= F(x,u,t) \\ y &= G(x,t) \end{aligned}$$

We may want to determine $u(\cdot)$ to minimize the integral

$$J(u) \;=\; \int_0^T (y_u(t) \;-\; v(t))^2 dt$$

where $y_u(t)$ is the output corresponding to control $u(\cdot)$ and $v(\cdot)$ is some reference signal. Various features may have to be taken into account:

- Constraints on $u(\cdot)$ (for realizability)
  - piecewise continuous
  - $|u(t)| \;\leq\; u_{\max}\; \forall t$
- T may be finite or infinite
- $x(0)$, $x(T)$ may be free, fixed, constrained
- The entire state trajectory may be constrained ($x(\cdot)$), e.g., to keep the temperature reasonable
- One may require a "closed-loop" control, e.g., $u(t) \;=\; u(x(t))$. It is well known that such 'feedback' control systems are much less sensitive to perturbations and modeling errors.

■

Unlike Example 1.1 and Example 1.2, Example 1.3 is an 'optimal control' problem. Whereas discrete-time optimal control problems can be solved by classical optimization techniques, continuous-time problems involve optimization in infinite dimension spaces (a complete 'waveform' has to be determined).

1.2 Scope of the Course

To conclude this chapter we now introduce the class of problems that will be studied in this course. Consider the abstract optimization problem

$$(P) \qquad \min\{f(x) \mid x \in S\}$$

where S is a subset of a vector space X and where $f : X \to \mathbf{R}$ is the *cost* or *objective* function. S is the *feasible set*. Any x in S is a *feasible point*.

Definition 1.1 *A point $\hat{x}$ is called a (strict) global minimizer for (P) if $\hat{x} \in S$ and*

$$\begin{array}{rcll} f(\hat{x}) & \leq & f(x) & \forall x \in S \\ & (<) & & (\forall x \in S, x \neq \hat{x}) \end{array}$$

Assume now X is equipped with a norm.

Definition 1.2 *A point $\hat{x}$ is called a (strict) local minimizer for (P) if $\hat{x} \in S$ and $\exists\ \epsilon > 0$ such that*

$$\begin{array}{rcll} f(\hat{x}) & \leq & f(x) & \forall x \in S \cap B(\hat{x}, \epsilon) \\ & (<) & & (\forall x \in S \cap B(\hat{x}, \epsilon), x \neq \hat{x}) \end{array}$$

Notation.

It often helps to distinguish the scalar 0 from the original in $\mathbf{R}^n$ or in a more general vector space (see Appendix A). We will usually denote that latter by θ, sometimes specialized to θ_n in the case of $\mathbf{R}^n$ or to θ_V in the case of a vector space V.

Scope

1. Type of optimization problems considered

   (i) Finite-dimensional

   unconstrained

   equality constrained

   inequality [and equality] constrained

linear, quadratic programs, convex problems

multiobjective problems

discrete optimal control

(ii) Infinite-dimensional

calculus of variations (no "control" signal) (old: 1800)

optimal control (new: 1950's)

Note: most types in (i) can be present in (ii) as well.

2. Results sought

Essentially, solve the problem. The steps are

- conditions of optimality ("simpler" characterization of solutions)
- numerical methods: solve the problem, generally by solving some optimality condition or, at least, using the insight such conditions provide.
- sensitivity: how "good" is the solutions in the sense of "what if we didn't solve exactly the right problem?"

(– duality: some transformation of the original problem into a hopefully simpler optimization problem)

Chapter 2

Linear Optimal Control: Some Readily Solvable Instances

References: [1, 7, 21, 29].
At the outset, we consider linear time–varying models. A motivation for not starting with the time-invariant case is that while, in a finite-horizon context (which is the simpler situation as far as optimal control is concerned), allowing for time–varying models hardly complicates the analysis, linear time–varying models are of much practical importance, e.g., in the context of trajectory tracking for nonlinear (even when time-invariant) systems.
Indeed, given a nominal control signal $\hat{u}$ and corresponding state trajectory $\hat{x}$, a typical approach to synthesizing close tracking of the trajectory is to first substitute a linear model of the system, obtained by inearizing the original system around that trajectory, and then focus on keeping the state of the linearization close to the origin. E.g., given

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

(with appropriate regularity assumptions on f) and a "nominal" associated trajectory $(\hat{u}, \hat{x})$, linearization about this trajectory gives rise to the linear time–varying system

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

with $A(t) := \frac{\partial f}{\partial x}(\hat{x}(t), \hat{u}(t))$ and $B(t) := \frac{\partial f}{\partial u}(\hat{x}(t), \hat{u}(t))$.
Thus, consider the linear control system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0 \tag{2.1}$$

where $x(t) \in \mathbf{R}^n, u(t) \in \mathbf{R}^m$ for all t, and $A(\cdot)$ and $B(\cdot)$ are matrix-valued functions. Suppose $A(\cdot)$, $B(\cdot)$ and $u(\cdot)$ continuous. Then (2.1) has the unique solution

$$x(t) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, \sigma)B(\sigma)u(\sigma)d\sigma \tag{2.2}$$

where the state transition matrix Φ satisfies the homogeneous differential equation

$$\frac{\partial}{\partial t}\Phi(t, t_0) = A(t)\Phi(t, t_0)$$

with initial condition

$$\Phi(t_0, t_0) = I.$$

Further, for any t_1, t_2, the transition matrix $\Phi(t_1, t_2)$ is invertible and

$$\Phi(t_1, t_2)^{-1} = \Phi(t_2, t_1).$$

2.1 Free terminal state, unconstrained, quadratic cost: linear quadratic regulator (LQR)

A quadratic cost function of a function takes the form of an integral plus possibly terms associated with "important" time points, typically the "terminal time". The quadratic cost function might be obtained from a second order expansion of the "true" cost function around the desired trajectory.

2.1.1 Finite horizon

For simplicity, we start with the case for which the terminal state is free, and no constraint (apart from the dynamics and the initial condition) are imposed on the control and state values.
Consider the optimal control problem (see Exercise 2.3 below for a more general quadratic cost function)

$$\begin{aligned} \text{minimize} \quad & J(u) := \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} \left(x(t)^T L(t) x(t) + u(t)^T u(t)\right) dt + \frac{1}{2} x(t_\mathrm{f})^T Q x(t_\mathrm{f}) && (P) \\ \text{subject to} \quad & \dot{x}(t) = A(t)x(t) + B(t)u(t) \quad \forall t \in [t_0, t_\mathrm{f}], && (2.3) \\ & x(t_0) = x_0, \qquad u \in \mathcal{C}, && (2.4) \end{aligned}$$

where $x(t) \in \mathbf{R}^n, u(t) \in \mathbf{R}^m$ and $A(\cdot)$, $B(\cdot)$ and $L(\cdot)$ are matrix-valued functions, and $\mathcal{C}$ denotes the set of continuous mappings; minimization is with respect to u and x. (Equivalently, x can be viewed as a function x_u of u defined by the dynamics and initial condition, so the only constraint is continuity of u.) The initial and final times t_0 and t_f are given, as is the initial state x_0. The mappings $A(\cdot)$, $B(\cdot)$, and $L(\cdot)$, defined on the domain $[t_0, t_\mathrm{f}]$, are assumed to be continuous. Without loss of generality, $L(t)$ (for all t) and Q are assumed symmetric.

Remark 2.1 Clearly, inclusion of the terminal cost in (P) could equivalently be achieved by replacing $L(t)$ with $L(t)+\delta(t-t_\mathrm{f})Q$, where δ is the Dirac impulse. In these notes though, we rule out such impulses. (Controls u are functions, not distributions.)

The problem just stated is, in a sense, the simplest meaningful continuous-time optimal control problem. Indeed, the cost function is quadratic and the dynamics linear, and there are no constraints on u (except for continuity). While a linear cost function may be even simpler than a quadratic one, in the absence of (implicit or explicit) constraints on the control, such problem would have no solution (except for the trivial situation where the cost function is constant, independent of u (or x)).

In fact, the problem is simple enough that it can be solved without much advanced mathematical machinery, by simply "completing the square". Doing this of course requires that we add to and subtract from $J(u)$ a quantity that involves x and u. But doing so would likely modify our problem! The following fundamental lemma gives us a key to resolving this conundrum. The idea is that, while the integrand in the lemma involves the paths $x(t), u(t), t \in [t_0, t_\mathrm{f}]$, the integral depends only on the end points of the trajectory $x(\cdot)$, i.e., this integral is path independent.

Lemma 2.1 *(Fundamental Lemma/Path Independence Lemma) Let $A(\cdot), B(\cdot)$ be continuous matrix-value functions and $K(\cdot) = K^T(\cdot)$ be a matrix-valued function, with continuously differentiable entries (in fact, absolute continuity is sufficient) on $[t_0, t_\mathrm{f}]$. Then, if $x(t)$ and $u(t)$ are related by*

$$\dot{x}(t) = A(t)x(t) + B(t)u(t) \quad \forall t, \tag{2.5}$$

it holds

$$x(t_\mathrm{f})^T K(t_\mathrm{f})x(t_\mathrm{f}) - x(t_0)^T K(t_0)x(t_0) = \int_{t_0}^{t_\mathrm{f}} \phi(x(t), u(t), K(t), \dot{K}(t))dt \tag{2.6}$$

where

$$\phi(x(t), u(t), K(t), \dot{K}(t)) = x(t)^T(\dot{K}(t) + A^T(t)K(t) + K(t)A(t))x(t) + 2x(t)^T K(t)B(t)u(t).$$

Proof. Because $x(\cdot)^T K(\cdot)x(\cdot)$ is continuously differentiable

$$x(t_\mathrm{f})^T K(t_\mathrm{f})x(t_\mathrm{f}) - x(t_0)^T K(t_0)x(t_0) = \int_{t_0}^{t_\mathrm{f}} \frac{d}{dt} x(t)^T K(t)x(t)\ dt$$

$$= \int_{t_0}^{t_\mathrm{f}} (\dot{x}(t)^T K(t)x(t) + x(t)^T \dot{K}(t)x(t) + x(t)^T K(t)\dot{x}(t))dt$$

and the claim follows if one substitutes for $\dot{x}(t)$ the right hand side of (2.5).

■

Note that, given that $x(t_0) = x_0$ is fixed, the second term in the LHS of equality 2.6 is independent of u as long as $K(\cdot)$ is. The idea is then to add to $J(u)$ (indeed to $2J(u)$, for simplicity) the difference between the two sides (RHS−KHS) of equality 2.6, i.e., zero, while exploiting the freedom in the choice of $K(\cdot)$ to obtain a perfect square under the integral sign and cancel the terminal cost term. To that effect, (i) we select $K(t_\mathrm{f}) = Q$ (thus cancelling the terminal cost term $x(t_\mathrm{f})^T Q x(t_\mathrm{f})$), we note that the integrand becomes a perfect square provided $K(\dot{)}$ satisfies on $[t_0, t_\mathrm{f}]$ a certain differential equation that does not involve u. Now, for arbitrary $K(\cdot)$ satisfying the assumptions, we get

$$2J(u) = 2J(u) + \int_{t_0}^{t_\mathrm{f}} \phi(x(t), u(t), K(t), \dot{K}(t))dt - x(t_\mathrm{f})^T K(t_\mathrm{f})x(t_\mathrm{f}) + x_0^T K(t_0)x_0.$$

Taking K to satisfy $K(t_\mathrm{f}) = Q$ (to cancel the terminal cost), we get

$$2J(u) = \int_{t_0}^{t_\mathrm{f}} (x(t)^T L(t)x(t) + u(t)^T u(t) + \phi(x(t), u(t), K(t), \dot{K}(t))dt + x_0^T K(t_0)x_0$$

$$= \int_{t_0}^{t_\mathrm{f}} (x^T(\dot{K} + A^T K + KA + L)x + u(K^T B + B^T K)x + u^T u)dt + x_0^T K(t_0)x_0,$$

where we have removed the explicit dependence on t for notational compactness. Now recall that the above holds independently of the choice of $K(\cdot)$. If we select it to satisfy (if a solution exists on $[t_0, t_\mathrm{f}]$!) the differential equation

$$\dot{K}(t) = -A^T(t)K(t) - K(t)A(t) - L(t) + K(t)B(t)B(t)^T K(t) \tag{2.7}$$

the above simplifies to

$$2J(u) = \int_{t_0}^{t_\mathrm{f}} \left[x(t)^T (K(t)B(t)B(t)^T K(t))x(t) + 2x(t)^T K(t)B(t)u(t) + u(t)^T u(t)\right] dt + x_0^T K(t_0)x_0,$$

i.e.,

$$J(u) = \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} \|B(t)^T K(t)x(t) + u(t)\|_2^2 dt + \frac{1}{2}x_0^T K(t_0)x_0. \tag{2.8}$$

Note that, because $K(\cdot)$ is selected independently of u, the last term in (2.8) does not depend on u.

We have completed the square! Equation (2.7) is an instance of a differential Riccati equation (DRE) (after Jacopo F. Riccati, Italian mathematician, 1676–1754). Its right-hand side is quadratic in the unknown K. We postpone the discussion of existence and uniqueness of a solution to this equation and of whether such solution is symmetric (as required for the Fundamental Lemma to hold). The following scalar example shows that this is an issue indeed.

Example 2.1 (scalar Riccati equation) Consider the case of scalar, time-independent values $a = 0$, $b = 1$, $l = -1$, $q = 0$, corresponding to the optimal control problem

$$\text{minimize } \int_{t_0}^{t_\mathrm{f}} (u(t)^2 - x(t)^2)dt \quad \text{s.t.} \quad \dot{x}(t) = u(t) \;\; t \in [t_0, t_\mathrm{f}), u \in \mathcal{U}.$$

The corresponding Riccati equation is

$$\dot{k}(t) = 1 + k(t)^2, \quad k(t_\mathrm{f}) = 0$$

We get

$$\mathrm{atan}(k(t)) - \mathrm{atan}(k(t_\mathrm{f})) = t - t_\mathrm{f},$$

yielding

$$k(t) = \tan(t - t_\mathrm{f}),$$

with a finite escape time at $\hat{t} = t_{\rm f} - \frac{\pi}{2}$. (In fact, if $t_0 < t_{\rm f} - \frac{\pi}{2}$, even if we "forget" about the singularity at $\hat{t}$, $k(t) = \tan(t - t_{\rm f})$ is not the integral of its derivative (as would be required by the Fundamental Lemma): $\dot{k}(t)$ is positive everywhere, while $k(t)$ goes from positive values just before $\hat{t}$ to negative values after $\hat{t}$.) It is readily verified that in fact this optimal control problem itself has no solution if, e.g., with $x_0 = 0$, when $t_{\rm f}$ is too large. Indeed, with $x_0 = 0$, controls of the form $u(t) = \alpha t$ yields

$$J(u) = \alpha^2 \int_0^{t_{\rm f}} t^2 (1 - \frac{t^2}{4}) dt,$$

which is negative for $t_{\rm f} > \sqrt{10}$ and hence can be made arbitrary largely negative by letting $\alpha \to \infty$.

For the time being, we assume that (2.7) with terminal condition $K(t_{\rm f}) = Q$ has a unique solution exists on $[t_0, t_{\rm f}]$, and we denote its value at time t by

$$K(t) = \Pi(t, Q, t_{\rm f}),$$

so that $\Pi(t, Q, t_{\rm f})$ satisfies the DRE, more precisely,

$$\frac{\partial}{\partial t}\Pi(t, Q, t_{\rm f}) = -A^T(t)\Pi(t, Q, t_{\rm f}) - \Pi(t, Q, t_{\rm f})A(t) - L(t) + \Pi(t, Q, t_{\rm f})B(t)B(t)^T\Pi(t, Q, t_{\rm f}) \quad \forall t.$$

(Note that such solution $\Pi(\cdot, Q, t_{\rm f})$ must then be continuous on $[t_0, t_{\rm f}]$—indeed continuously differentiable on $[t_0, t_{\rm f}]$) since the (continuous) right-hand side is its derivative for all $[t_0, t_{\rm f}]$.) Accordingly, since u is unconstrained, it would seem that J is minimized by the choice

$$u(t) = -B(t)^T\Pi(t, Q, t_{\rm f})x(t), \tag{2.9}$$

and that its optimal value is $x_0^T\Pi(t_0, Q, t_{\rm f})x_0$. Equation (2.9) is a feedback law, which specifies a control signal in closed-loop form.

Exercise 2.1 *Show that this feedback law yields a well-defined, continuous control signal $\hat{u}$. (Hint. First solve for x, then obtain $\hat{u}$.)*

Remark 2.2 By "closed-loop" it is meant that the right-hand side of (2.9) does not depend on the initial state x_0 nor on the initial time t_0, but only on the current state and time. Such formulations are of major practical importance: If, for whatever reason (modeling errors, perturbations) the state at some time $\tau \in [t_0, t_{\rm f}]$ is not what it was predicted to be (when the optimal control $u^*(\cdot)$ was computed, at time t_0), the control generated by (2.9) is still optimal over $\tau \in [t_0, t_{\rm f}]$ (for the same cost function but with the integral starting at time τ)—assuming no modeling errors or perturbations between times τ and $t_{\rm f}$. This is of course not so for the open-loop optimal control obtained for the original problem, with starting time t_0.

Remark 2.3 It follows from the above that existence of a solution to the DRE over $[t_0, t_{\rm f}]$ is a sufficient condition for existence of a solution to optimal control problem (P) for every $x_0 \in \mathbf{R}^n$. Indeed, of course, the existence of a solution to (P) is not guaranteed at the outset. This was seen, e.g., in Example 2.1 above.

Let us now use a $*$ superscript to denote optimality, so (2.9) becomes

$$u^*(t) = -B(t)^T \Pi(t, Q, t_\mathrm{f}) x^*(t), \tag{2.10}$$

where x^* is the "optimal trajectory", i.e., the trajectory generated by the optimal control. As noted above, the optimal value is given by

$$V(t_0, x_0) := J(u^*) = \frac{1}{2} x_0^T \Pi(t_0, Q, t_\mathrm{f}) x_0.$$

V is known as the value function. Now suppose that, starting from x_0 at time t_0, perhaps after having undergone perturbations, the state reaches $x(\tau)$ at time $\tau \in (t_0, t_\mathrm{f})$. The remaining portion of the minimal cost, to be incurred over $[\tau, t_\mathrm{f}]$, is the minimum, over $u \in \mathcal{U}$, subject to $\dot{x} = Ax + Bu$ with $x(\tau)$ fixed, of

$$\begin{aligned} J_\tau(u) := & \quad \tfrac{1}{2} \textstyle\int_\tau^{t_\mathrm{f}} \left(x(t)^T L(t) x(t) + u(t)^T u(t) \right) dt + \tfrac{1}{2} x(t_\mathrm{f})^T Q x(t_\mathrm{f}) & (2.11) \\ = & \quad \tfrac{1}{2} \textstyle\int_\tau^{t_\mathrm{f}} \| B(t)^T \Pi(\tau, Q, t_\mathrm{f}) x(t) + u(t) \|^2 dt + \tfrac{1}{2} x(\tau)^T \Pi(\tau, Q, t_\mathrm{f}) x(\tau), & (2.12) \end{aligned}$$

where we simply have replaced, in (2.8) t_0 with τ and x_0 with $x(\tau)$. The cost-to-go is

$$J_\tau(u^*) = \frac{1}{2} x(\tau)^T \Pi(\tau, Q, t_\mathrm{f}) x(\tau).$$

Hence, the "cost-to-go" from an arbitrary time $t < t_\mathrm{f}$ and state $\xi \in \mathbf{R}^n$ is

$$V(t, \xi) = J_t(u^*) = \frac{1}{2} \xi^T \Pi(t, Q, t_\mathrm{f}) \xi. \tag{2.13}$$

Remark 2.4 We have not made any positive definiteness (or semi-definiteness) assumption on $L(t)$ or Q. The key assumption we have made is that the stated Riccati equation has a solution $\Pi(t, Q, t_\mathrm{f})$ over $[t_0, t_\mathrm{f}]$. Below we investigate conditions (in particular, on L and Q) which insure that this is the case. At this point, note that, if $L(t) \succeq 0$ for all t and $Q \succeq 0$, then $J(u) \geq 0$ for all $u \in \mathcal{U}$, and expression (2.13) of the cost-to-go implies that $\Pi(t, Q, t_\mathrm{f}) \succeq 0$ whenever it exists.

Remark 2.5 Note that (DRE) does not involve t_0 nor x_0. In particular, when a solution $\Pi(\cdot, Q, t_\mathrm{f})$ exists on $[t_0, t_\mathrm{f}]$, it yields optimal controls for every problem $\mathcal{P}_{\tau,\xi}$ with the same $A(\cdot)$, $B(\cdot)$, $L(\cdot)$, Q and t_f as in (P), but with initial state $\xi \in \mathbf{R}^n$ and $\tau \in [t_0, t_\mathrm{f}]$. In particular the value function $V : (\tau, \xi) \mapsto V(\tau, \xi)$ is that same for every problem $\mathcal{P}_{\tau,\xi}$, as long as (DRE) has a solution on $[\tau, t_\mathrm{f}]$, in particular, for every (τ, ξ) such that $\tau \in [t_0, t_\mathrm{f}]$. The notation $\mathcal{P}_{\tau,\xi}$ will be used in the remainder of this chapter.

Returning to the question of existence/uniqueness of the solution to the differential Riccati equation, first note that the right-hand side of (2.7) is locally Lipschitz-continuous (indeed, continuously differentiable) in K over $\mathbf{R}^{n \times n}$. This, together with continuity of A, B, and L, implies that, for any given Q, there exists $\tau < t_\mathrm{f}$ such that a (continuously differentiable) solution exists and is unique in $[\tau, t_\mathrm{f}]$. The alternative to existence of a unique solution on $-\infty, t_\mathrm{f})$ is existence of a finite escape time. Indeed, the following holds.

Fact. (See, e.g., [16, Chapter 1, Theorems 2.1 & 3.1].) Let $\varphi : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ be continuous, and Lipschitz-continuous in its first argument. Then for every $x_0 \in \mathbf{R}^n$ and $t_0 \in \mathbf{R}$, there exists $t_1, t_2 \in \mathbf{R}$, with $t_0 \in (t_1, t_2)$, such that the differential equation $\dot{x} = \varphi(x(t), t)$, with $x(t_0) = x_0$, has a (continuously differentiable) solution $x(t)$ in (t_1, t_2). Furthermore, this solution is unique. Finally, suppose there exists a compact set $S \subset \mathbf{R}^n$, with $x_0 \in S$, that enjoys the following property: For every t_1, t_2 such that $t_0 \in (t_1, t_2)$ and the solution $x(t)$ exists for all $t \in (t_1, t_2)$, $x(t)$ belongs to S for all $t \in (t_1, t_2)$. Then the solution $x(t)$ exists and is unique (and continuously differentiable) for all $t \in \mathbf{R}$.

Lemma 2.2 *Let $\hat{t} = \inf\{\tau : \Pi(t, Q, t_\mathrm{f})$ exists $\forall t \in [\tau, t_\mathrm{f}]\}$. If $\hat{t}$ is finite, then $\|\Pi(\cdot, Q, t_\mathrm{f})\|$ is unbounded on $(\hat{t}, t_\mathrm{f}]$.*

Proof. Let

$$\varphi(K, t) = -A(t)^T K - KA(t) + KB(t)B^T(t)K - L(t),$$

so that (2.7) can be written

$$\dot{K}(t) = \varphi(K(t), t), \quad K(t_\mathrm{f}) = Q,$$

where $\varphi \in \mathcal{C}$ is continuous and is Lipschitz-continuous in its first argument on every bounded set. Proceeding by contradiction, let S be a compact set containing $\{\Pi(t, Q, t_\mathrm{f}) \ : \ t \in (\hat{t}, t_\mathrm{f}]\}$. The claim is then an immediate consequence of the previous Fact.

■

In other words, either $\Pi(t, Q, t_\mathrm{f})$ exists and is unique over $(-\infty, t_\mathrm{f})$, or there exists a finite $\hat{t} < t_\mathrm{f}$ such that $\Pi(t, Q, t_\mathrm{f})$ is unbounded on $\hat{t}, t_\mathrm{f})$ (finite escape time). Further, since clearly the transpose of a solution to the Riccati equation is also a solution, this *unique solution must be symmetric*, as required in the path-independent lemma.

Remark 2.6 It follows that, without any further conditions, if t_0 is close enough to t_f, the optimal control problem has a (unique) solution.

Next, note that if $L(t)$ is positive semi-definite for all t, and Q is positive semi-definite, then $J(u) \geq 0$ for all $u \in \mathcal{U}$. Hence, in such case, as long as an optimal control exists,

$$V(t, \xi) \geq 0 \quad \forall t \leq t_\mathrm{f}, \ \xi \in \mathbf{R}^n.$$

It then follows from (2.13) that $\Pi(\tau, Q, t_\mathrm{f})$ is positive semidefinite for every τ such that $\Pi(t, Q, t_\mathrm{f})$ exists for all $t \in [\tau, t_\mathrm{f}]$.

Theorem 2.1 *Suppose $L(t)$ is positive semi-definite for all t and Q is positive semi-definite. Then $\Pi(t, Q, t_\mathrm{f})$ exists $\forall t \leq t_\mathrm{f}$, i.e, $\hat{t} = -\infty$.*

Proof. Again, let $\hat{t} = \inf\{\tau : \Pi(t, Q, t_\mathrm{f})$ exists $\forall t \in [\tau, t_\mathrm{f}]\}$, so that $\Pi(\cdot, Q, t_\mathrm{f})$ exists on $(\hat{t}, t_\mathrm{f}]$. Below, we show that $\|\Pi(\cdot, Q, t_\mathrm{f})\|$ is bounded by a continuous function over $(\hat{t}, t_\mathrm{f}]$. This will imply that, if $\|\Pi(\cdot, Q, t_\mathrm{f})\|$ were to be unbounded over $(\hat{t}, t_\mathrm{f}]$, we would have $\hat{t} = -\infty$; in view of Lemma 2.2, "$\hat{t}$ is finite" is ruled out, and the claim will follow.

Thus, let $\tau \in (\hat{t}, t_{\rm f}]$. For any $x \in \mathbf{R}^n$, using the positive definiteness assumption on $L(t)$ and Q, we have

$$\xi^T \Pi(\tau, Q, t_{\rm f})\xi = 2V(\tau, \xi) = \min_{u \in \mathcal{U}} \int_{\tau}^{t_{\rm f}} \big(u(t)^T u(t) + x(t)^T L(t) x(t)\big) dt + x(t_{\rm f})^T Q x(t_{\rm f}) \geq 0, \quad (2.14)$$

where $x(t)$ satisfies (2.3) with initial condition $x(\tau) = \xi$. We show that there exists a continuous function $F : (-\infty, t_{\rm f}) \to \mathbf{R}^n$ such that, if $\Pi(\cdot, Q, t_{\rm f})$ exists (and is unique and continuously differentiable) on $[\tau, t_{\rm f})$, then, for all $\xi \in \mathbf{R}^n$,

$$\xi^T \Pi(\tau, Q, t_{\rm f})\xi \leq \xi^T F(\tau)\xi, \quad (2.15)$$

implying, in view of Exercise 2.2 below, that

$$\|\Pi(\tau, Q, t_{\rm f})\|_2 \leq \|F(\tau)\|_2 \quad \forall \tau \in (\hat{t}, t_{\rm f}],$$

as claimed. To conclude the proof, we construct such F. From (2.14), we have for every $u \in \mathcal{C}$,

$$\xi^T \Pi(\tau, Q, t_{\rm f})\xi \leq \int_{\tau}^{t_{\rm f}} \big(u(t)^T u(t) + x(t)^T L(t) x(t)\big)\, dt + x(t_{\rm f})^T Q x(t_{\rm f}) \quad (2.16)$$

To obtain an upper bound quadratic in ξ, we let $\hat{u}$ be identically zero. The corresponding $\hat{x}$ then satisfies $\dot{x} = Ax$, so that $\hat{x}(t) = \Phi_A(t, \tau)\xi$ for all τ. This yields the upper bound 2.15 with

$$F(\tau) := \int_{\tau}^{t_{\rm f}} \Phi_A(t, \tau)^T L(t) \Phi_A(t, \tau) dt + \Phi_A(t_{\rm f}, \tau)^T Q \Phi_A(t_{\rm f}, \tau).$$

■

Exercise 2.2 *Prove that, if $A = A^T$ and $F = F^T$, and $0 \leq x^T A x \leq x^T F x$ for all x, then $\|A\|_2 \leq \|F\|_2$, where $\|\cdot\|_2$ denotes the spectral norm of the matrix argument. (I.e., $\|A\|_2 = \max\{\|Ax\|_2 \ : \ \|x\|_2 = 1\}$.)*

Thus, when $L(t)$ is positive semi definite for all t and Q is positive semi definite, our problem has a unique optimal control given by (2.10).

Exercise 2.3 *Investigate the case of the more general cost function*

$$J(u) = \int_{t_0}^{t_{\rm f}} \left(\frac{1}{2} x(t)^T L(t) x(t) + u(t)^T S(t) x(t) + \frac{1}{2} u(t)^T R(t) u(t)\right) dt,$$

where L, S and R are continuous on $[t_0, t_{\rm f}]$, $L(t)$ and $R(t)$ are symmetric on $[t_0, t_{\rm f}]$, and $R(t) \succ 0$ for all t. Hint: Let $v(t) = T(t)u(t) + M(t)x(t)$, where T satisfies $R(t) = T(t)^T T(t)$ for all t, and T is continuous (does such T exist?).

From here on, we assume a cost function of the form

$$J(u) = \int_{t_0}^{t_\mathrm{f}} \left(\frac{1}{2} x(t)^T L(t) x(t) + \frac{1}{2} u(t)^T R(t) u(t) \right) dt,$$

with L and R as in Exercise 2.3.

Solution to DRE
For $t \in [\tau, t_\mathrm{f}]$, define

$$p^*(t) = -\Pi(t, Q, t_\mathrm{f}) x^*(t), \tag{2.17}$$

so that the optimal control law (2.10) satisfies

$$u^*(t) = R(t)^{-1} B^T(t) p^*(t). \tag{2.18}$$

Then x^* and p^* together satisfy the the linear system

$$\begin{bmatrix} \dot{x}^*(t) \\ \dot{p}^*(t) \end{bmatrix} = \begin{bmatrix} A(t) & B(t)R(t)^{-1}B^T(t) \\ L(t) & -A^T(t) \end{bmatrix} \begin{bmatrix} x^*(t) \\ p^*(t) \end{bmatrix} \tag{2.19}$$

evolving in $\mathbf{R}^{2n}$.

Exercise 2.4 *Verify that x^* and p^* satisfy (2.19).*

Suppose $x^*(t_0) = x_0$ is fixed and note that, from (2.17), we also have the conditions $p^*(t_\mathrm{f}) = -Qx^*(t_\mathrm{f})$. Then we have a two-point boundary-value problem (TPBVP). We know that, without a positive semidefiniteness assumption on L and Q, there might not be a solution. Indeed, in contrast with initial value problems, TPBVPs do not always have a solution.
Suppose now that $\Pi(t, Q, t_\mathrm{f})$ exists and is continuous on $[t_0, t_\mathrm{f}]$. Then the optimal control problem can be solved explicitly, e.g., along the lines of the next theorem and exercise.

Theorem 2.2 *Let $X(t)$ and $P(t)$ be $n \times n$ matrices satisfying the differential equation*

$$\begin{bmatrix} \dot{X}(t) \\ \dot{P}(t) \end{bmatrix} = \begin{bmatrix} A(t) & B(t)R(t)^{-1}B^T(t) \\ L(t) & -A^T(t) \end{bmatrix} \begin{bmatrix} X(t) \\ P(t) \end{bmatrix}, \tag{2.20}$$

with $X(t_\mathrm{f}) = I$ and $P(t_\mathrm{f}) = -Q$. Then

$$P(t) = -\Pi(t, Q, t_\mathrm{f}) X(t) \tag{2.21}$$

solves (2.7) for $t \in [\tau, t_\mathrm{f}]$, for every $\tau < t_\mathrm{f}$ such that $\Pi(t, Q, t_\mathrm{f})$ exists on $[\tau, t_\mathrm{f}]$.

Proof. Just plug in. ■

Exercise 2.5 *Show that, if the DRE has a continuous solution $\Pi(t, Q, t_\mathrm{f})$ on $[t_0, t_\mathrm{f}]$ then $X(t)$ is nonsingular on $[t_0, t_\mathrm{f}]$, so that $\Pi(t, Q, t_\mathrm{f}) = -P(t)X(t)^{-1}$ for all $t \in [t_0, t_\mathrm{f}]$. [Hint: Use (2.21) to solve (2.20) for X.]*

Instance of Pontryagin's Principle (Lev S. Pontryagin, Soviet Mathematician, 1908–1988.)
In connection with cost function J with $u^T u$ generalized to $u^T R u$ (although we will still assume $R = I$ for the time being), let $\psi : \mathbf{R}^n \to \mathbf{R}$ be the "terminal cost" function, i.e.,

$$\psi(\xi) = \frac{1}{2}\xi^T Q \xi \quad \forall \xi \in \mathbf{R}^n,$$

and let $H : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ be given by

$$H(\tau, \xi, \eta, v) = -\frac{1}{2}\left(v^T R(\tau) v + \xi^T L(\tau)\xi\right) + \eta^T (A(\tau)\xi + B(\tau)v).$$

Equations (2.19), (2.18), and (2.17) then yield

$$\dot{x}^*(t) = \nabla_\eta H(t, x^*(t), p^*(t), u^*(t)), \quad x^*(t_0) = x_0, \tag{2.22}$$
$$\dot{p}^*(t) = -\nabla_\xi H(t, x^*(t), p^*(t), u^*(t)), \quad p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f})), \tag{2.23}$$

a two-point boundary-value problem. Now note that, from (2.18),

$$H(t, x^*(t), p^*(t), u^*(t)) = \max_{v \in \mathbf{R}^m} H(t, x^*(t), p^*(t), v) \ \forall t,$$

where clearly the minimum is achieved at $v = R(t)^{-1}B^T(t)p^*(t)$. Thus the following result (an instance of Pontryagin's Principle; see Chapter 6 for more details) holds.

Theorem 2.3 *$u^* : \mathbf{R} \to \mathbf{R}^m$, continuous, solves (P) if and only if*

$$H(t, x^*(t), p^*(t), u^*(t)) = \max_{v \in \mathbf{R}^m} H(t, x^*(t), p^*(t), v) \quad \forall t \in [t_0, t_\mathrm{f}],$$

where x^*, with $x^*(t_0) = x_0$, is the state trajectory generated by u^*, and where $p^* : \mathbf{R} \to \mathbf{R}^n$, continuously differentiable, satisfies

$$\dot{p}^*(t) = -\nabla_\xi H(t, x^*(t), p^*(t), u^*(t)) \ (= -A^T p^*(t) + L(t)x^*(t)), \quad \forall t \in [t_0, t_\mathrm{f}],$$
$$p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f})) \ (= -Qx^*(t_\mathrm{f})).$$

This suggests we define $\mathcal{H} : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ to take values

$$\mathcal{H}(\tau, \xi, \eta) = \max_{v \in \mathbf{R}^m} H(\tau, \xi, \eta, v),$$

Then

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \ \forall t.$$

Function H is the pre-Hamiltonian,[1] (sometimes called control Hamiltonian or pseudo-Hamiltonian) and $\mathcal{H}$ the Hamiltonian (or true Hamiltonian). (Sir William R. Hamilton, Irish mathematician, 1805–1865.) Thus

$$\begin{aligned}\mathcal{H}(t, \xi, \eta) &= -\frac{1}{2}\xi^T L(t)\xi + \eta^T A(t)\xi + \frac{1}{2}\eta^T B(t)R(t)^{-1}B^T(t)\eta \\ &= \frac{1}{2}\begin{bmatrix} \xi \\ \eta \end{bmatrix}^T \begin{bmatrix} -L(t) & A^T(t) \\ A(t) & B(t)R(t)^{-1}B^T(t) \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix}.\end{aligned}$$

[1] This terminology is borrowed from P.S. Krishnaprasad

Finally, note that the optimal cost $J(u^*)$ can be equivalently expressed as

$$J(u^*) = -\frac{1}{2}x(t_0)^T p(t_0).$$

Remark 2.7 The gradient of $\mathcal{H}$ with respect to the first $2n$ arguments is given by

$$\nabla_z \mathcal{H}(t,\xi,\eta) = \begin{bmatrix} -L(t) & A^T(t) \\ A(t) & B(t)B^T(t) \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix}.$$

so that, in view of (2.17),

$$\nabla_z \mathcal{H}(t, x^*(t), p^*(t)) = \nabla_z H(t, x^*(t), p^*(t), u^*(t)). \tag{2.24}$$

Note however that while, as we will see later, (2.22)–(2.23) hold rather generally, (2.24) no longer does when constraints are imposed on the values of control signal, as is the case later in this chapter as well as in Chapter 6. Indeed, $\mathcal{H}$ usually is non-smooth.

Remark 2.8 Because $\nabla_u H(t, x^*(t), p^*(t), u^*(t)) = 0$ and u^* is differentiable, along trajectories of (2.19), with $z^*(t) := [x^*(t); p^*(t)]$, we have

$$\begin{aligned} \frac{d}{dt}H(t, x^*(t), p^*(t), u^*(t)) &= \nabla_z H(t, x^*(t), p^*(t), u^*(t))^T \dot{z}^*(t) + \frac{\partial}{\partial t}H(t, x^*(t), p^*(t), u^*(t)) \\ &= \frac{\partial}{\partial t}H(t, x^*(t), p^*(t), u^*(t)) \end{aligned}$$

since

$$\begin{aligned} \nabla_z H(t, x^*(t), p^*(t), u^*(t))^T \dot{z}^*(t) &= \nabla_z H(t, x^*(t), p^*(t), u^*(t))^T J \nabla_z H(t, x^*(t), p^*(t), u^*(t)) \\ &= 0 \quad (\text{since } J^T = -J). \end{aligned}$$

In particular, if A, B and L do not depend on t, $H(t, x^*(t), p^*(t), u^*(t))$ is *constant* along trajectories of (2.19).

2.1.2 Infinite horizon, LTI systems

We now turn our attention to the case of infinite horizon ($t_\mathrm{f} = \infty$). Because we are mainly interested in stabilizing control laws (so that $x(t) \to 0$ as $t \to \infty$ will be guaranteed), we assume without loss of generality that $Q = 0$. To simplify the analysis, we further assume that A, B and L are constant. We also simplify the notation by translating the origin of time to the initial time t_0, i.e., by letting $t_0 = 0$. Assuming (as above) that $L = L^T \geq 0$, so that existence and uniqueness of a solution to (DRE) is guaranteed over every finite interval, we can write $L = C^T C$ for some (non-unique) C, so that the problem can be written as

$$\text{minimize } J(u) = \frac{1}{2}\int_0^\infty \big(y(t)^T y(t) + u(t)^T u(t)\big)\, dt$$

subject to

$$\begin{aligned}&\dot{x}(t) = Ax(t) + Bu(t),\ \forall t \in [t_0, t_\mathrm{f}),\\ &y(t) = Cx(t),\\ &u \in \mathcal{C}\\ &x(0) = x_0.\end{aligned}$$

Note that y is merely some linear image of x, and need not be a physical output. For example, it could be an error signal to be driven to the origin.
Consider the differential Riccati equation (2.7) with our constant A, B, and L and, in agreement with the notation used in the previous section, denote by $\Pi(t, 0, \tau)$ the value at time t of the (continuously differentiable) solution to this equation that vanishes at time τ. Since L is positive semi-definite, in view of Theorem 2.1, such solution exists for all t and τ (with $t \leq \tau$) and is unique, symmetric and positive semi-definite. Noting that $\Pi(t, 0, \tau)$ only depends on $t - \tau$ (since the Riccati equation is now time-invariant), we define

$$\Pi(\tau) := \Pi(0, 0, \tau),$$

so that

$$\Pi(t, 0, \tau) = \Pi(0, 0, \tau - t) = \Pi(\tau - t).$$

Exercise 2.6 *Formally prove that* $\Pi(t, 0, \tau) = \Pi(0, 0, \tau - t)$ *for all* $t, \tau \in \mathbf{R}$.

It is intuitively reasonable to expect that, for fixed t, as τ goes to ∞, the optimal feedback law $u(t) = -B^T\Pi(t, 0, \tau)x(t)$ for the finite-horizon problem with cost function

$$J^\tau(u) := \frac{1}{2}\int_0^\tau (x(t)^T C^T C x(t) + u(t)^T u(t))dt\ .$$

will tend to an optimal feedback law for our infinite horizon problem. Since the optimal cost-to-go is $x(t)^T\Pi(\tau - t)x(t)$, it is also tempting to guess that $\Pi(\tau - t)$ itself will converge, as $\tau \to \infty$, to some matrix Π_∞ independent of t (since the time-to-go will approach ∞ at every time t and the dynamics is time-invariant), satisfying the algebraic Riccati equation (since the derivative of Π_∞ with respect to t is zero)

$$A^T\Pi_\infty + \Pi_\infty A - \Pi_\infty B B^T \Pi_\infty + C^T C = 0, \tag{ARE}$$

with optimal control law

$$u^*(t) = -B^T\Pi_\infty x^*(t).$$

We will see that, under a mild assumption, introduced next, our intuition is correct.
When dealing with an infinite-horizon problem, it is natural (if nothing else, for practical reasons) to seek control laws that are stabilizing. Otherwise, while x^* might still remain bounded for some initial conditions, such property would not hold robustly.
Assumption. (A, B) is stabilizable.
We first show that $\Pi(t)$ converges to some limit Π_∞, then that Π_∞ satisfies (ARE) (i.e., Π_∞ is an equilibrium point for (DRE)), and finally that the control law $u(t) = -B^T\Pi_\infty x(t)$ is optimal for our infinite-horizon problem.

Lemma 2.3 *As $t \to \infty$, $\Pi(t)$ converges to some symmetric, positive semi-definite matrix Π_∞. Furthermore, Π_∞ satisfies (ARE).*

Proof. Since the optimal value for $J^\tau(u)$ is $\frac{1}{2}x_0^T\Pi(\tau)x_0$, for any $u \in \mathcal{C}$ we have, for all $\tau > 0$,

$$x_0^T\Pi(\tau)x_0 \leq \int_0^\tau x(\sigma)^T C^T C x(\sigma) + u(\sigma)^T u(\sigma) d\sigma. \tag{2.25}$$

Invoking stabilizability, let $u(t) = Fx(t)$ be a stabilizing static state feedback and let $\hat{x}$ be the corresponding solution of

$$\begin{aligned} \dot{x} &= (A+BF)x \\ x(0) &= x_0, \end{aligned}$$

i.e., $\hat{x}(t) = e^{(A+BF)t}x_0$; further, let $\hat{u} = F\hat{x}$. Then, since $A+BF$ is Hurwitz stable and the integrand in (2.25) is nonnegative, we have, for all $\tau \geq 0$,

$$x_0^T\Pi(\tau)x_0 \leq \int_0^\tau \hat{x}(\sigma)^T C^T C \hat{x}(\sigma) + \hat{u}(\sigma)^T \hat{u}(\sigma) d\sigma \leq \int_0^\infty \hat{x}(\sigma)^T C^T C \hat{x}(\sigma) + \hat{u}(\sigma)^T \hat{u}(\sigma) d\sigma = x_0^T M\ x_0,$$

where

$$M \ = \ \int_0^\infty e^{(A+BF)^T t}(C^TC + F^TF)e^{(A+BF)t}dt$$

is well defined and independent of τ. Since in view of (2.25) $x_0^T\Pi(\tau)x_0$ is nonnegative, it is bounded for every fixed x_0. Further, non-negative definiteness of C^TC implies that $x_0^T\Pi(\tau)x_0$ is monotonically non-decreasing as τ increases. Since it is bounded, $x_0^T\Pi(\tau)x_0$ must converge for every x_0.[2] Using the fact that $\Pi(\tau)$ is symmetric it is easily shown that $\Pi(\tau)$ converges (see Exercise 2.7 below), i.e., for some symmetric matrix Π_∞,

$$\lim_{\tau\to\infty} \Pi(\tau) = \Pi_\infty.$$

Symmetry and positive semi-definiteness are inherited from $\Pi(\tau)$. Finally, since[3]

$$-\dot{\Pi}(\tau) = -A^T\Pi(\tau) - \Pi(\tau)A - C^TC + \Pi(\tau)BB^T\Pi(\tau)$$

and the right-hand side converges when $\tau \to \infty$, $\dot{\Pi}(\tau)$ must also converge, and the limit must be zero (since if $\dot{\Pi}(\tau)$ were to converge to a non-zero constant, $\Pi(\tau)$ could not possibly converge). Hence Π_∞ satisfies (ARE). ■

Note: The final portion of the proof is taken from [30, p.296–297].

[2] It cannot be inferred from the mere fact that $\Pi(\tau)$ converges as $\tau \to \infty$ that its derivative converges to zero—which would imply, by taking limits in (2.7), that Π_∞ satisfies the ARE. (E.g., as $t \to \infty$, $\exp(-t)\sin(\exp(t))$ tends to a constant (zero) but its derivative does not go to zero.) In particular, Barbalat's Lemma can't be applied without first showing that $\Pi(\cdot)$ has a uniformly continuous derivative.

[3] The minus sign in the right-hand side is due to the identity (when $Q = 0$ and $t_f = 0$) $K(t) = \Pi(t,0,0) = \Pi(0,0,-t) = \Pi(-t)$.

Exercise 2.7 *Prove that convergence of $x_0^T\Pi(\tau)x_0$ for arbitrary x_0 implies convergence of $\Pi(\tau)$.*

Now note that, for any $\tilde{u} \in \mathcal{U}$ and $\tau \geq 0$, we have

$$\frac{1}{2}x_0^T\Pi(\tau)x_0 = \frac{1}{2}x_0^T\Pi(0,0,\tau)x_0 = \min_{u\in\mathcal{C}} J^\tau(u) \leq J^\tau(\tilde{u}) \leq J(\tilde{u}),$$

where $J(\tilde{u})$ might be infinite. Letting $\tau \to \infty$ we obtain

$$\frac{1}{2}x_0^T\Pi_\infty x_0 \leq J(\tilde{u}) \quad \forall \tilde{u} \in \mathcal{C},$$

implying that

$$\frac{1}{2}x_0^T\Pi_\infty x_0 \leq \inf_{u\in\mathcal{C}} J(u) \tag{2.26}$$

(where $\inf_{u\in\mathcal{U}} J(u)$ could be infinite). Finally, we construct a control u^* which attains the cost value $\frac{1}{2}x_0^T\Pi_\infty x_0$, proving optimality of u^* and equality in (2.26). For this, note that the constant matrix Π_∞, since it satisfies (is an equilibrium point for) (ARE), automatically satisfies the corresponding differential Riccati equation. Using this fact and making use of the Fundamental Lemma with $K(t) := \Pi_\infty$ for all t, we obtain, analogously to (2.8),

$$J^\tau(u) = \frac{1}{2}\left(\int_0^\tau \|B^T\Pi_\infty x(t) + u(t)\|_2^2 dt - x(\tau)^T\Pi_\infty x(\tau) + x_0^T\Pi_\infty x_0\right) \quad \forall \tau \geq 0. \tag{2.27}$$

Noting that, since $\Pi_\infty \succeq 0$, $x(\tau)^T\Pi_\infty x(\tau)$ is nonnegative, and taking the limit as $\tau \to \infty$ on both sides yields, for every $u \in \mathcal{C}[0,\infty)$,

$$J(u) \leq \frac{1}{2}\left(\int_0^\tau \|B^T\Pi_\infty x(t) + u(t)\|_2^2 dt + x_0^T\Pi_\infty x_0\right).$$

Substituting the feedback control law

$$u = -B^T\Pi_\infty x, \tag{2.28}$$

denoting by u^* the resulting control signal, and using the fact that Π_∞ is positive semi-definite yields

$$J(u^*) \leq \frac{1}{2}x_0^T\Pi_\infty x_0 \quad \forall \tau \geq 0,$$

which, together with (2.26), proves that u^* is optimal. Finally, if feedback (2.28) is stabilizing, then asymptotic stability implies that $x(\tau) \to 0$ as $\tau \to \infty$, and again taking limits on both sides of (2.27) yields

$$J(u) = \frac{1}{2}\left(\int_0^\tau \|B^T\Pi_\infty x(t) + u(t)\|_2^2 dt + x_0^T\Pi_\infty x_0\right).$$

showing that u^* is the unique optimal control.

Theorem 2.4 *Suppose (A, B) is stabilizable. Then Π_∞ solves (ARE), the control law $u = -B^T\Pi_\infty x$ is optimal, yielding u^*, and*

$$J(u^*) = \frac{1}{2}x_0^T\Pi_\infty x_0.$$

$$x^*(t) = e^{(A-BB^T\Pi_\infty)t}x_0 \quad \forall t.$$

Also, $V(t,\xi) = \frac{1}{2}\xi^T\Pi_\infty\xi$ does not depend on t. Finally, if control law (2.28) is stabilizing, then u^* is the <u>only</u> optimal control.

This solves the infinite horizon LTI problem. Note however that it is not guaranteed that the optimal control law is stabilizing. For example, consider the extreme case when $C = 0$ (in which case $\Pi_\infty = 0$) and the system is open loop unstable. It seems clear that this is due to unstable modes not being observable through C. This of course is undesirable. We now show that, indeed, under a detectability assumption, the optimal control law is stabilizing.

Theorem 2.5 *Suppose (A, B) is stabilizable and (C, A) detectable. Then, if $K \succeq 0$ solves (ARE), $A - BB^TK$ is Hurwitz stable; in particular, $A - BB^T\Pi_\infty$ is Hurwitz stable.*

Proof. (ARE) can be rewritten

$$(A - BB^TK)^TK + K(A - BB^TK) = -KBB^TK - C^TC. \tag{2.29}$$

Proceed now by contradiction. Let λ, with $\mathrm{Re}\lambda \geq 0$, and $v \neq 0$ be such that

$$(A - BB^TK)v = \lambda v. \tag{2.30}$$

Multiplying (2.29) on the left by v^* and on the right by v we get

$$2(\mathrm{Re}\lambda)(v^*Kv) = -\|B^TKv\|^2 - \|Cv\|^2.$$

Since the left-hand side is non-negative (since $K \succeq 0$) and the right-hand side non-positive, both sides must vanish. Thus (i) $Cv = 0$ and, (ii) $B^TKv = 0$ which together with (2.30), implies that $Av = \lambda v$. Since $\mathrm{Re}\lambda \geq 0$, this contradicts detectability of (C, A). ■

Corollary 2.1 *If (A, B) is stabilizable and (C, A) is detectable, then the optimal control law $u = -B^T\Pi_\infty x$ is stabilizing.*

Exercise 2.8 *Suppose (A, B) is stabilizable. Then (i) Given $x \in \mathbf{R}^n$, $\Pi_\infty x = \theta$ (the origin of $\mathbf{R}^n$) if and only if $x^T\Pi_\infty x = 0$; and (ii) If $\Pi_\infty x = \theta$ then x belongs to the unobservable subspace. In particular, if (C, A) is observable, then $\Pi_\infty > 0$. [Hint: Use the fact that $J(u^*) = x_0^T\Pi_\infty x_0$.]*

To summarize, we have the following theorem.

Theorem 2.6 *Suppose (A, B) is stabilizable. Then, as $t \to \infty$, $\Pi(t) \to \Pi_\infty$, a symmetric, positive semi-definite matrix that satisfies (ARE); and the feedback law $u(t) = -B^T\Pi_\infty x(t)$ is optimal. Further, if (C, A) is detectable, then (i) the resulting closed-loop system is asymptotically stable (i.e., Π_∞ is a "stabilizing" solution of (ARE)); and (ii) the optimal solution u^* is unique. Finally, if (C, A) is observable, then Π_∞ is positive definite.*

Remark 2.9 Some intuition concerning the solutions of (ARE) can be gained as follows. In the scalar case, if $B \neq 0$ and $C \neq 0$, the left-hand side is a downward parabola that intersects the vertical axis at C^2 (> 0); hence (ARE) has two real solutions, one positive, the other negative. When $n > 1$, the number of solutions increases, most of them being indefinite matrices (i.e., with some positive eigenvalues and some negative eigenvalues). In line with the fact that $\Pi(t) \succeq 0$ for all t, a (or the) positive semi-definite solution will be the focus of our investigation.

Finally, we discuss how (ARE) can be solved directly (without computing the limit of $\Pi(t)$). In the process, we establish that, under stabilizability of (A, B) and detectability of (C, A), Π_∞ is that unique stabilizing solution of (ARE), hence the unique symmetric positive semi-definite solution of (ARE). Also see Appendix E in [1].
Consider the Hamiltonian matrix (see (2.19))

$$H = \begin{bmatrix} A & -BB^T \\ -L & -A^T \end{bmatrix},$$

and let K^+ be a stabilizing solution to (ARE), i.e., $A - BB^TK^+$ is stable. Let

$$T = \begin{bmatrix} I & 0 \\ K^+ & I \end{bmatrix}.$$

Then

$$T^{-1} = \begin{bmatrix} I & 0 \\ -K^+ & I \end{bmatrix}.$$

Now (elementary block column and block row operations)

$$\begin{aligned} T^{-1}HT &= \begin{bmatrix} I & 0 \\ -K^+ & I \end{bmatrix} \begin{bmatrix} A & -BB^T \\ -L & -A^T \end{bmatrix} \begin{bmatrix} I & 0 \\ K^+ & I \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ -K^+ & I \end{bmatrix} \begin{bmatrix} A - BB^TK^+ & -BB^T \\ -L - A^TK^+ & -A^T \end{bmatrix} \\ &= \begin{bmatrix} A - BB^TK^+ & -BB^T \\ 0 & -(A - BB^TK^+)^T \end{bmatrix} \end{aligned}$$

since K^+ is a solution to (ARE). It follows that

$$\sigma(H) = \sigma(A - BB^TK^+) \cup \sigma(-(A - BB^TK^+)),$$

where $\sigma(\cdot)$ denotes the spectrum (set of eigenvalues). Thus, if (A, B) is stabilizable and (C, A) detectable (so such solution K^+ exists), H cannot have any imaginary eigenvalues:

It must have n eigenvalues in $\mathbf{C}^-$ and n eigenvalues in $\mathbf{C}^+$. Furthermore the first n columns of T form a basis for the stable invariant subspace of H, i.e., for the span of all generalized eigenvectors of H associated with stable eigenvalues (see, e.g., Chapter 13 of [34] for more on this).
Now let $\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix}$ be a basis for the stable invariant subspace of H, i.e., let

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$$

be any invertible matrix such that

$$S^{-1}HS = \begin{bmatrix} X & Z \\ 0 & Y \end{bmatrix}$$

for some X, Y, Z such that $\sigma(X) \subset \mathbf{C}^-$ and $\sigma(Y) \subset \mathbf{C}^+$. (Note that $\sigma(H) = \sigma(X) \cup \sigma(Y)$.) Then it must hold that, for some non-singular R',

$$\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix} = \begin{bmatrix} I \\ K^+ \end{bmatrix} R'.$$

It follows that $S_{11} = R'$ and $S_{21} = K^+R'$, thus

$$K^+ = S_{21}S_{11}^{-1},$$

which also shows that (ARE) cannot have more than one stabilizing solution. From Theorem 2.5, it also follows that, if (A, B) is stabilizable and (C, A) detectable, then (ARE) has exactly one positive semidefinite solution, namely Π_∞.
We have thus proved the following.

Theorem 2.7 *Suppose (A, B) is stabilizable and (C, A) is detectable. Then $\Pi_\infty = S_{21}S_{11}^{-1}$, where $\begin{bmatrix} S_{11} \\ S_{21} \end{bmatrix}$ is any basis for the stable invariant subspace of H. In particular, $S_{21}S_{11}^{-1}$ is symmetric and there is exactly one stabilizing solution to (ARE), namely Π_∞, which is also the only positive semi-definite solution.*

Exercise 2.9 *Given $J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$ (a unitary matrix such that $J^2 = -I$), any real matrix H that satisfies $J^{-1}HJ = -H^T$ is said to be Hamiltonian. Show that if H is Hamiltonian and λ is an eigenvalue of H, then $-\lambda$ also is.*

In summary, we have the following. If (A, B) is stabilizable then Π_∞ is well defined (as the limit of $\Pi(t)$; it is a positive semi-definite solution of (ARE); and the control law $u(t) = -B^T\Pi_\infty x(t)$ is optimal, with optimal value $\frac{1}{2}x_0^T\Pi_\infty x_0$. Further, if in addition

- (C, A) is detectable, then (i) the optimal control u^* is unique and is generated by the feedback control law $u = -B^T\Pi_\infty$; (ii) $A - BB^T\Pi_\infty$ is Hurwitz stable, i.e., the optimal control law is stabilizing; and (iii) Π_∞ is the only stabilizing solution of (ARE) and the only positive semi-definite solution of (ARE).

- (C, A) is observable, then Π_∞ is positive definite.

2.2 Fixed terminal state, unconstrained control values, quadratic cost

Question: Given $x_\mathrm{f} \in \mathbf{R}^n$, $t_\mathrm{f} > t_0$, does there exist $u \in \mathcal{U}$ such that, for system (2.1), $x(t_\mathrm{f}) = x_\mathrm{f}$? If the answer to the above is "yes", we say that x_f is *reachable* from (x_0, t_0) at time t_f. If moreover this holds for all $x_0, x_\mathrm{f} \in \mathbf{R}^n$ then we say that the *system* (2.1) is reachable on $[t_0, t_\mathrm{f}]$.
There is no loss of generality in assuming that $x_\mathrm{f} = \theta$,[4] as shown by the following exercise.

Exercise 2.10 *Define $\hat{x}(t) := x(t) - \Phi(t, t_\mathrm{f})x_\mathrm{f}$. Then $\hat{x}$ satisfies*

$$\frac{d}{dt}\hat{x}(t) = A(t)\hat{x}(t) + B(t)u(t) \quad \forall t \in [t_0, t_\mathrm{f}).$$

Conclude that, under dynamics (2.1), u steers (x_0, t_0) to $(x_\mathrm{f}, t_\mathrm{f})$ if and only if it steers (ξ, t_0) to (θ, t_f), where ξ $(= \xi(x_0, t_0)) := x_0 - \Phi(t_0, t_\mathrm{f})x_\mathrm{f}$.

Since $\Phi(t_0, t_\mathrm{f})$ is invertible, it follows that system (2.1) is reachable on $[t_0, t_\mathrm{f}]$ if and only if it is controllable on $[t_0, t_\mathrm{f}]$, i.e., if and only if, given x_0, there exists $u \in \mathcal{U}$ that steers (x_0, t_0) to (θ, t_f). [**Note.** Equivalence between reachability and controllability (to the origin) does **not** hold in the discrete-time case, where controllability is a weaker property than reachability.]
Now controllability to (θ, t_f) from (ξ, t_0), for some ξ, is equivalent to solvability of the equation (in $u \in \mathcal{U}$):

$$\Phi(t_\mathrm{f}, t_0)\xi + \int_{t_0}^{t_\mathrm{f}} \Phi(t_\mathrm{f}, \sigma)B(\sigma)u(\sigma)d\sigma = \theta.$$

Equivalently (multiplying on the left by the non-singular matrix $\Phi(t_0, t_\mathrm{f})$), (θ, t_f) can be reached from (ξ, t_0) if there exists $u \in \mathcal{U}$ such that

$$\xi = Lu := -\int_{t_0}^{t_\mathrm{f}} \Phi(t_0, \sigma)B(\sigma)u(\sigma)d\sigma,$$

where $L : \mathcal{U} \to \mathbf{R}^n$ is a linear map.
If θ is indeed reachable at time t_f from (ξ, t_0), we might want want to steer (ξ, t_0) while spending the least amount of energy, i.e., while minimizing

$$J(u) := \frac{1}{2}\langle u, u\rangle = \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} u(t)^T u(t)dt$$

subject to $\xi = Lu$, where $\langle \cdot, \cdot \rangle : \mathcal{U} \times \mathcal{U} \to \mathbf{R}$ is the L_2^m inner product and, as before, the factor of one half has been inserted for convenience. Linear map L is continuous (bounded) with respect to the norm derived from this inner product (see Exercise A.50) and the problem at

[4] Here and elsewere in these notes, θ is the origin of the vector space under consideration.

hand is a linear least-squares problem. Clearly, θ is reachable at time t_f from (ξ, t_0) if and only if $\xi \in \mathcal{R}(L)$, so that (2.1) is controllable on $[t_0, t_\mathrm{f}]$ if and only if $\mathcal{R}(L) = \mathbf{R}^n$. From Exercise A.50, L has an adjoint L^*. In view of Theorem A.5, because $\mathcal{R}(LL^*)$ (a subspace of $\mathbf{R}^n$) is closed, $\mathcal{R}(L) = \mathcal{R}(LL^*)$. It follows from Theorem A.6 that the unique optimal control is given by the unique $u \in \mathcal{R}(L^*)$ satisfying $Lu = \xi$ is

$$u = L^*\eta,$$

where η is any point in $\mathbf{R}^n$ satisfying

$$LL^*\eta = \xi$$

(and such points do exist). It is shown in Exercise A.50 of Appendix A that L^* is given by

$$(L^*\mu)(t) = -B^T(t)\Phi^T(t_0, t)\mu$$

which yields

$$\begin{aligned} LL^*\mu &= -\int_{t_0}^{t_\mathrm{f}} \Phi(t_0, t)B(t)(L^*\mu)(t)dt \\ &= \int_{t_0}^{t_\mathrm{f}} \Phi(t_0, t)B(t)B^T(t)\Phi^T(t_0, t)dt\ \mu, \quad \forall \mu \in \mathbf{R}^n, \end{aligned}$$

i.e., $LL^* : \mathbf{R}^n \to \mathbf{R}^n$ is given by

$$LL^* = \int_{t_0}^{t_\mathrm{f}} \Phi(t_0, t)B(t)B^T(t)\Phi^T(t_0, t)dt =: W(t_0, t_\mathrm{f}).$$

Since $\mathcal{R}(L) = \mathcal{R}(LL^*)$, θ is reachable at t_f from (ξ, t_0) if and only if

$$\xi \in \mathcal{R}(W(t_0, t_\mathrm{f}))\ .$$

Note that $W(t_0, t_\mathrm{f})$ has entries

$$(W(t_0, t_\mathrm{f}))_{ij} = \left\langle (\Phi_{i\cdot}(t_0, \cdot)B(\cdot))^T, (\Phi_{j\cdot}(t_0, \cdot)B(\cdot))^T \right\rangle$$

where $\langle \cdot, \cdot \rangle$ is again the L_2^m inner product (for each i and t, think of $\Phi_{i,\cdot}(t_0, t)B(t)$ as a row vector in $\mathbf{R}^m$), i.e., $W(t_0, t_\mathrm{f})$ is the Gramian matrix (or Gram matrix, or Gramian; Jørgen P. Gram, Danish mathematician, 1850–1916) associated with the vectors $(\Phi_{j\cdot}(t_0, \cdot)B(\cdot))^T$, $j = 1, \ldots, n$. It is known as the reachability Gramian. It is invertible if and only if $\mathcal{R}(L) = \mathbf{R}^n$, i.e., if and only if the system is reachable on $[t_0, t_\mathrm{f}]$.
Suppose $W(t_0, t_\mathrm{f})$ is invertible indeed and, for simplicity, assume that that target state is the origin, i.e., $x_\mathrm{f} = \theta$. The unique minimum energy control that steers (x_0, t_0) to (θ, t_f) is

then given by[5]

$$\hat{u} = L^*(LL^*)^{-1}x_0$$

i.e.

$$\hat{u}(t) = -B^T(t)\Phi^T(t_0,t)W(t_0,t_\mathrm{f})^{-1}x(t_0) \tag{2.31}$$

and the corresponding energy is given by

$$\frac{1}{2}\langle\hat{u},\hat{u}\rangle = \frac{1}{2}x(t_0)^T W(t_0,t_\mathrm{f})^{-1}x(t_0). \tag{2.32}$$

Note that, as expressed in (2.31), $\hat{u}(t)$ depends explicitly on the initial state x_0 and initial time t_0. Consequently, if between t_0 and the current time t, the state x has been affected by an external perturbation, $\hat{u}$ as expressed by (2.31) is no longer optimal (minimum energy) over the remaining time interval $[t,t_\mathrm{f}]$. Let us address this issue. (We still assume that $x_\mathrm{f} = \theta$.) At time t_0, we have

$$\begin{aligned}\hat{u}(t_0) &= -B^T(t_0)\Phi^T(t_0,t_0)W(t_0,t_\mathrm{f})^{-1}x(t_0)\\ &= -B^T(t_0)W(t_0,t_\mathrm{f})^{-1}x(t_0).\end{aligned}$$

Intuitively, this must hold independently of the value of t_0, i.e., for the problem under consideration, *Bellman's Principle of Optimality* holds (Richard E. Bellman, American mathematician, 1920–1984): Independently of the initial state (at t_0), for $\hat{u}$ to be optimal, it is necessarily the case that $\hat{u}$ applied from the current time $t \geq t_0$ up to the final time t_f, starting at the current state $x(t)$, be optimal for the remaining problem, i.e., for the objective function $\int_t^{t_\mathrm{f}} u(\tau)^T u(\tau)d\tau$. Specifically, given $x \in \mathbf{R}^n$ and $t \in [t_0,t_\mathrm{f}]$ such that x_f is reachable at time t_f from (x,t), denote by $P(x,t;x_\mathrm{f},t_\mathrm{f})$ the problem of determining the control of least energy that steers (x,t) to $(x_\mathrm{f},t_\mathrm{f})$, i.e., with (x,t) replacing (x_0,t_0). Let $x(\cdot)$ be the state trajectory that results when optimal control $\hat{u}$ is applied, starting from x_0 at time t_0. Then Bellman's Principle of Optimality asserts that, for any $t \in [t_0,t_\mathrm{f}]$, the restriction of $\hat{u}$ to $[t,t_\mathrm{f}]$ solves $P(x(t),t;x_\mathrm{f},t_\mathrm{f})$. (See Chapter 3 for a proof of Bellman's Principle in the discrete-time case.)

Exercise 2.11 *Prove that Bellman's Principle of Optimality holds for the minimum energy fixed endpoint problem. [Hint: Given an optimal control u^* for $P(x_0,t_0;x_\mathrm{f},t_\mathrm{f})$, assuming by contradiction that a lower energy control $\hat{u}$ exists for $P(x(t),t;x_\mathrm{f},t_\mathrm{f})$, construct, based on $\hat{u}$, a better control that u^* for $P(x_0,t_0;x_\mathrm{f},t_\mathrm{f})$, a contradiction. Likely you will first come up with a discontinuous control. This control then will have to be perturbed to make it continuous while keeping the objective value lower than $J(u^*)$ and the perturbation to $\hat{x}(t_\mathrm{f})$ small, then tweaked again (e.g., by adding a small enough Δu to still keep the overall objective value lower than $J(u^*)$) to insure that the prescribed terminal state x_f is recovered.]*

[5] $L^\dagger := L^*(LL^*)^{-1}$ is known as the pseudo-inverse of integral operator L; this concept was first introduced by E.I. Fredholm (Swedish mathematician, 1866–1937) in 1903. It was later further developed, in the narrower context of linear operators L from $\mathbf{R}^m$ to $\mathbf{R}^n$, by E.H. Moore (American mathematician, 1962–1932), A. Bjerhammer (Swedish geodesist, 1917–2011), and R. Penrose (English mathematician, 1931–. Also see end of Appendix A.

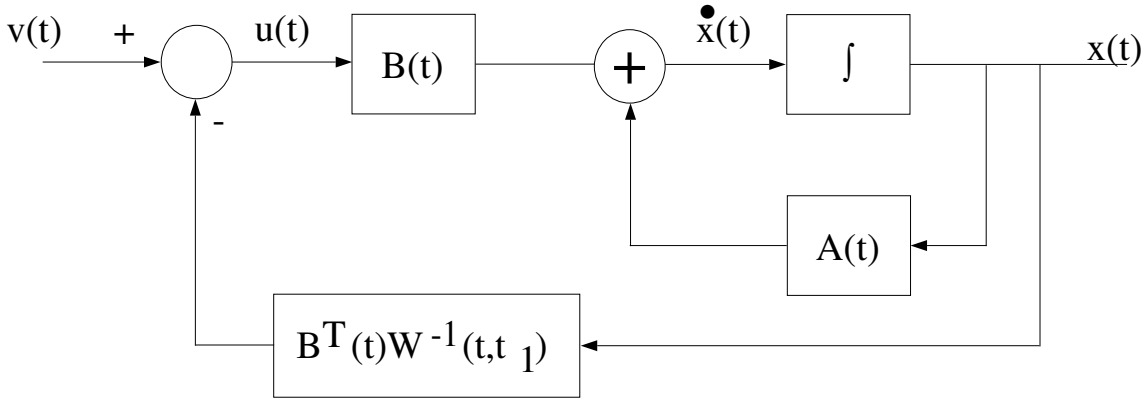


Figure **2.1**: Closed-loop implementation

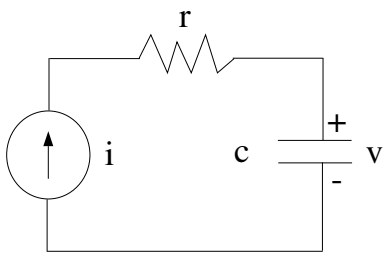


Figure **2.2**: Charging capacitor

It follows that, for any t such that $W(t, t_\mathrm{f})$ is invertible,

$$\hat{u}(t) = -B^T(t)W(t, t_\mathrm{f})^{-1}x(t) \ \forall t \tag{2.33}$$

which yields the "closed-loop" implementation, depicted in Figure 2.1; in (2.33), $v = \theta$. Further the optimal cost-to-go $V(t, \xi)$ is (from (2.32)),

$$V(t, \xi) = \frac{1}{2}\xi^T W(t, t_\mathrm{f})^{-1}\xi$$

whenever the inverse exists. Note that, with a fixed $\xi \neq \theta$, $V(t, \xi) \to \infty$ as $t \to t_\mathrm{f}$. This reflects the fact that, with $x(t_\mathrm{f}) = \theta$, for t close to t_f, very high energy must be expended to reach the origin in a very short time $t - t_\mathrm{f}$.

Exercise 2.12 *Prove (2.33) from (2.31) directly, without invoking Bellman's Principle.*

Example 2.2 (charging capacitor)

$$\frac{d}{dt}cv(t) = i(t)$$

$$\text{minimize} \int_0^{t_\mathrm{f}} ri(t)^2 dt \qquad \text{s.t.} \quad v(0) = v_0,\ v(t_\mathrm{f}) = v_1$$

We obtain

$$B(t) \equiv \frac{1}{c}, \quad A(t) \equiv 0$$

$$W(0, t_\mathrm{f}) = \int_0^{t_\mathrm{f}} \frac{1}{c^2} dt = \frac{t_\mathrm{f}}{c^2}$$

$$\eta_0 = c^2 \frac{v_0 - v_1}{t_\mathrm{f}}$$

$$i_0(t) = -\frac{1}{c} \frac{c^2(v_0 - v_1)}{t_\mathrm{f}} = \frac{c(v_1 - v_0)}{t_\mathrm{f}} = \text{ constant.}$$

The closed-loop optimal feedback law is given by

$$i_0(t) = \frac{-c}{t_\mathrm{f} - t}(v(t) - v_1).$$

Exercise 2.13 *Discuss the same optimal control problem (with fixed end points) with the objective function replaced by*

$$J(u) \equiv \frac{1}{2} \int_{t_0}^{t_\mathrm{f}} u(t)^T R(t) u(t)\ dt$$

where $R(t) = R(t)^T \succ 0$ for all $t \in [t_0, t_\mathrm{f}]$ and $R(\cdot)$ is continuous. [Hint: define a new inner product on $\mathcal{U}$.]

The controllability Gramian $W(\cdot, \cdot)$ happens to satisfy certain simple equations. Recalling that

$$W(t, t_\mathrm{f}) = \int_t^{t_\mathrm{f}} \Phi(t, \sigma) B(\sigma) B^T(\sigma) \Phi(t, \sigma)^T d\sigma$$

one easily verifies that $W(t_\mathrm{f}, t_\mathrm{f}) = 0$ and

$$\frac{d}{dt} W(t, t_\mathrm{f}) = A(t) W(t, t_\mathrm{f}) + W(t, t_\mathrm{f}) A^T(t) - B(t) B^T(t) \tag{2.34}$$

implying that, if $W(t, t_\mathrm{f})$ is invertible, it satisfies

$$\frac{d}{dt} \left(W(t, t_\mathrm{f})^{-1} \right) = -W(t, t_\mathrm{f})^{-1} A(t) - A^T(t) W(t, t_\mathrm{f})^{-1} + W(t, t_\mathrm{f})^{-1} B(t) B^T(t) W(t, t_\mathrm{f})^{-1} \tag{2.35}$$

Equation (2.34) is linear. It is a Lyapunov equation. Equation (2.35) is quadratic. It is a Riccati equation (for $W(t, t_\mathrm{f})^{-1}$).

Exercise 2.14 *Prove that if a matrix $M(t) := W(t, t_\mathrm{f})$ satisfies Lyapunov equation (2.34) then, at every $t < t_\mathrm{f}$ at which it is invertible, its inverse satisfies Riccati equation (2.35). (Hint: $\frac{d}{dt}M(t)^{-1} = -M(t)^{-1}(\frac{d}{dt}M(t))M(t)^{-1}$.)*

Exercise 2.15 *Prove that $W(\cdot,\cdot)$ also satisfies the functional equation*

$$W(t_0, t_\mathrm{f}) = W(t_0, t) + \Phi(t_0, t)W(t, t_\mathrm{f})\Phi^T(t_0, t).$$

As we have already seen, the Riccati equation plays a fundamental role in optimal control systems involving linear dynamics and quadratic cost (linear-quadratic problems). At this point, note that, if $x_\mathrm{f} = \theta$ and $W(t, t_\mathrm{f})$ is invertible, then $\hat{u}(t) = -B(t)^T P(t)x(t)$, where $P(t) = W(t, t_\mathrm{f})^{-1}$ solves Riccati equation (2.35).
We have seen that, if $W(t_0, t_\mathrm{f})$ is invertible, the optimal cost for problem (FEP) is given by

$$J(\hat{u}) = \frac{1}{2}\langle \hat{u}, \hat{u}\rangle = \frac{1}{2}x_0^T W(t_0, t_\mathrm{f})^{-1}x_0. \tag{2.36}$$

This is clearly true for any t_0, so that, from a given time $t < t_\mathrm{f}$ (such that $W(t, t_\mathrm{f})$ is invertible) the "cost-to-go" is given by

$$\frac{1}{2}x(t)^T W(t, t_\mathrm{f})^{-1}x(t)$$

I.e., the value function is given by

$$V(t, \xi) = \frac{1}{2}\xi^T W(t, t_\mathrm{f})^{-1}\xi \quad \forall \xi \in \mathbf{R}^n, t < t_\mathrm{f}.$$

This clearly bears resemblance with results we obtained for the free endpoint problem since $W(t, t_\mathrm{f})^{-1}$ satisfies the associated Riccati equation (see Exercise 2.14).
Now consider the more general quadratic cost

$$J(u) := \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} x(t)^T L(t)x(t) + u(t)^T u(t)\; dt = \int_{t_0}^{t_\mathrm{f}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} L(t) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt, \tag{2.37}$$

where $L(\cdot) = L(\cdot)^T \in \mathcal{C}$. Let $K(t) = K(t)^T$ be some continuously differentiable time-dependent matrix. Using the Fundamental Lemma we see that, since x_0 and x_i are fixed, it is equivalent to minimize (we no longer assume that $x_\mathrm{f} = \theta$)

$$\tilde{J}(u) := J(u) + \frac{1}{2}({x_\mathrm{f}}^T K(t_\mathrm{f})x_\mathrm{f} - x_0^T K(t_0)x_0)$$

$$= \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} L(t) + \dot{K}(t) + A^T(t)K(t) + K(t)A(t) & K(t)B(t) \\ B(t)^T K(t) & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt.$$

To "complete the square", suppose there exists such $K(t)$ that satisfies

$$L(t) + \dot{K}(t) + A^T(t)K(t) + K(t)A(t) = K(t)B(t)B(t)^T K(t)$$

i.e., satisfies the Riccati differential equation

$$\dot{K}(t) = -A^T(t)K(t) - K(t)A(t) + K(t)B(t)B(t)^T K(t) - L(t), \tag{2.38}$$

As we have seen when discussing the free endpoint problem, if $L(t)$ is positive semi-definite for all t, then a solution exists for every prescribed positive semi-definite "final" value $K(t_\mathrm{f})$.) Then we get

$$\tilde{J}(u) = \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^T \begin{bmatrix} K(t)B(t)B(t)^T K(t) & K(t)B(t) \\ B(t)^T K(t) & I \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt$$

$$= \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} \left\| B(t)^T K(t)x(t) + u(t) \right\|_2^2 \, dt.$$

Now, again supposing that some solution to (DRE) exists, let $K(\cdot)$ be any such solution and let

$$v(t) = B^T(t)K(t)x(t) + u(t).$$

It is readily verified that, in terms of the new control input v, the systems dynamics become

$$\dot{x}(t) = [A(t) - B(t)B^T(t)K(t)]x(t) + Bv(t) \quad t \in [t_0, t_\mathrm{f}),$$

and the cost function takes the form

$$\tilde{J}(u) = \frac{1}{2}\int_{t_0}^{t_\mathrm{f}} v(t)^T v(t) dt.$$

That is, we end up with the problem

$$\begin{array}{rlll} \text{minimize} & \displaystyle\int_{t_0}^{t_\mathrm{f}} & \langle v(t), v(t)\rangle dt & \\ \text{subject to} & \dot{x}(t) & = [A(t) - B(t)B^T(t)\Pi(t, K_\mathrm{f}, t_\mathrm{f})]x(t) + Bv(t) \quad t \in [t_0, t_\mathrm{f}) & (2.39) \\ & x(t_0) & = x_0 & \\ & x(t_\mathrm{f}) & = x_\mathrm{f} & \\ & v \in \mathcal{U} & & (2.40) \end{array}$$

where, following standard usage, we have parametrized the solutions to DRE by their value (K_f) at time t_f, and denoted them by $\Pi(t, K_\mathrm{f}, t_\mathrm{f})$. This transformed problem (parametrized by K_f) is of an identical form to that we solved earlier. Denote by $\Phi_{A-BB^T\Pi}$ and $W_{A-BB^T\Pi}$ the state transition matrix and controllability Gramian for (2.39) (for economy of notation, we have kept implicit the dependence of Π on K_f). Then, for a given K_f, we can write the optimal control v^{K_f} for the transformed problem as

$$v^{K_\mathrm{f}}(t) = -B(t)^T W^{-1}_{A-BB^T\Pi}(t, t_\mathrm{f})\xi(x(t), t),$$

where $\Pi := \Pi(t, K_\mathrm{f}, t_\mathrm{f})$, and the optimal control u^* for the original problem as

$$u^*(t) = v^{K_\mathrm{f}}(t) - B(t)^T \Pi(t, K_\mathrm{f}, t_\mathrm{f}) \xi(x(t), t) = -B(t)^T \left(\Pi(t, K_\mathrm{f}, t_\mathrm{f}) + W^{-1}_{A-BB^T\Pi}(t, t_\mathrm{f}) \right) \xi(x(t), t).$$

We also obtain, with $\xi_0 = \xi(x_0, t_0)$,

$$J(u^*) = \tilde{J}(u^*) + \xi_0^T \Pi(t_0, K_\mathrm{f}, t_\mathrm{f}) \xi_0 = \xi_0^T \left(W^{-1}_{A-BB^T\Pi}(t_0, t_\mathrm{f}) + \Pi(t_0, K_\mathrm{f}, t_\mathrm{f}) \right) \xi_0.$$

The cost-to-go at time t is

$$\xi(x(t), t)^T (W^{-1}_{A-BB^T\Pi}(t, t_\mathrm{f}) + \Pi(t, K_\mathrm{f}, t_\mathrm{f})) \xi(x(t), t)$$

Finally, if $L(t)$ is identically zero, we can pick $K(t)$ identically zero and we recover the previous result.

Exercise 2.16 *Show that reachability of (2.1) on $[t_0, t_\mathrm{f}]$ implies invertibility of $W_{A-BB^T\Pi}(t_0, t_\mathrm{f})$ and vice-versa.*

We obtain the block diagram depicted in Figure **2.3**.[6] Thus $u^*(t) = B(t)^T p(t)$, with

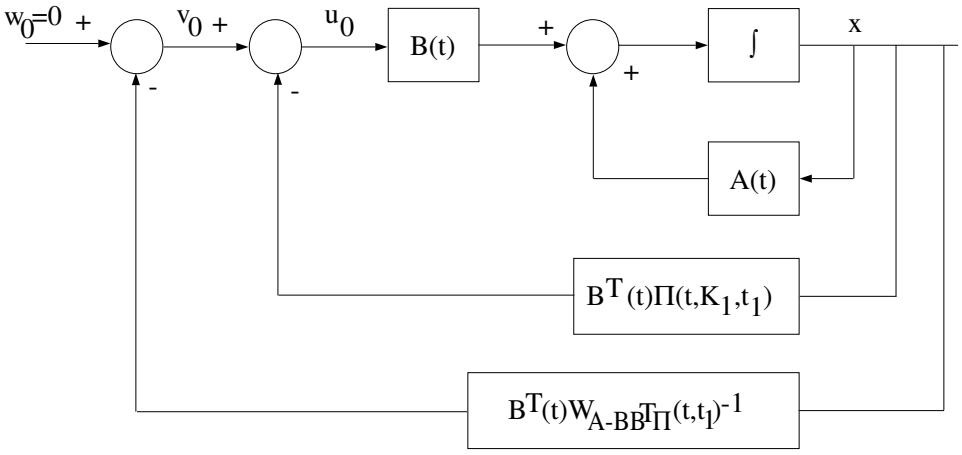


Figure **2.3**: Optimal feedback law

$$p(t) = -\left(\Pi(t, K_\mathrm{f}, t_\mathrm{f}) + W^{-1}_{A-BB^T\Pi}(t, t_\mathrm{f}) \right) x(t).$$

Note that while v_f^K clearly depends on K_f, u^* obviously cannot, since K_f is an arbitrary symmetric matrix (subject to DRE having a solution with $K(t_\mathrm{f}) = K_\mathrm{f}$). Thus we could have assigned $K_\mathrm{f} = 0$ throughout the analysis. Check the details of this.

[6] If $x_\mathrm{f} \neq \theta$, then, on Figure **2.3**, either replace $w_0 = 0$ with

$$w_0(t) = B^T(t) W^{-1}_{A-BB^T\Pi}(t, t_\mathrm{f}) \Phi_{A-BB^T\Pi}(t, t_\mathrm{f}) x_\mathrm{f},$$

or equivalently insert a summing junction immediately to the right of the bottom feedback block, adding $-\Phi_{A-BB^T\Pi}(t, t_\mathrm{f}) x_\mathrm{f}$ as external input.

The above is a valid closed-loop implementation as it does not involve the initial point (x_0, t_0) (indeed perturbations may have affected the trajectory between t_0 and the current time t). $\Pi(t, K_\mathrm{f}, t_\mathrm{f})$ can be precomputed (again, we must assume that such a solution exists.) Also note that the optimal cost $J(u^*)$ is given by (when $x_\mathrm{f} = \theta$)

$$\begin{aligned} J(u^*) &= \tilde{J}(u^*) - ({x_\mathrm{f}}^T K_\mathrm{f}, x_\mathrm{f} - x_0^T \Pi(t_0, K_\mathrm{f}, t_\mathrm{f}) x_0) \\ &= x_0^T W^{-1}_{A-BB^T\Pi}(t_0, t_\mathrm{f}) x_0 + x_0^T \Pi(t_0, K_\mathrm{f}, t_\mathrm{f}) x_0 \end{aligned}$$

and is independent of K_f.
Finally, for all optimal control problems considered so far, the optimal control can be expressed in terms of the adjoint variable (or co-state) $p(\cdot)$. More precisely, the following holds.

Exercise 2.17 *Consider the fixed terminal state problem with $x_\mathrm{f} = \theta$. Suppose that the controllability Gramian $W(t_0, t_\mathrm{f})$ is non-singular and that the relevant Riccati equation has a (unique) solution $\Pi(t, K_\mathrm{f}, t_\mathrm{f})$ on $[t_0, t_\mathrm{f}]$ with $K(t_\mathrm{f}) = K_\mathrm{f}$. Let $x(t)$ be the optimal trajectory and define $p(t)$ by*

$$p(t) = -\left(\Pi(t, K_\mathrm{f}, t_\mathrm{f}) + W^{-1}_{A-BB^T\Pi(t,K_\mathrm{f},t_\mathrm{f})}(t, t_\mathrm{f})\right) x(t)$$

so that the optimal control is given by

$$u^*(t) = B^T(t) p(t).$$

Then

$$\begin{bmatrix} \dot{x}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A(t) & B(t)B^T(t) \\ L(t) & -A^T(t) \end{bmatrix} \begin{bmatrix} x(t) \\ p(t) \end{bmatrix}$$

and the optimal cost is $-\frac{1}{2} x(t_0)^T p(t_0)$.

Exercise 2.18 *Verify that a minor modification of Theorem 2.3 holds in the present case of fixed terminal state, the only difference being that $p^*(t_\mathrm{f})$ is now free (which "compensates" for $x^*(t_\mathrm{f})$ being known).*

Exercise 2.19 *Consider an objective function of the form*

$$J(u) = \int_{t_0}^{t_\mathrm{f}} (\varphi(x(t), t) + u(t)^T u(t)) dt + \psi(x(t_\mathrm{f})),$$

for some functions φ and ψ. Show that if an optimal control exists in $\mathcal{U}$, its value $u^*(t)$ must be in the range space of B^T for all t. (Assume that B does not vary with time.)

2.3 Free terminal state, constrained control values, linear terminal cost

Consider the linear system

$$(S) \qquad \dot{x}(t) = A(t)x(t) + B(t)u(t), \quad \text{a.e. } t \in [t_0, t_f], \tag{2.41}$$

where $A(\cdot)$, $B(\cdot)$ are assumed to be continuous, and a (a.e. continuously differentiable) solution x is sought. The "a.e" (almost every) and continuity (of x) specifications are needed here because we will allow for discontinuous controls u.

Remark 2.10 When u is allowed to be discontinuous, "a.e." is clearly needed. Consider e.g., the very simple case of $\dot{x} = u$ on $[-1, 1]$, with $x(-1) = 0$, $u(t) = 0$ for $t < 0$, and $u(t) = 1$ for $t \geq 0$. The continuous "solution" is $x(t) = 0$ for $t < 0$ and $x(t) = t$ for $t \geq 0$, but such x does not satisfy the differential equation for all t: indeed, x is not differentiable at $t = 0$. It does satisfy the equation for almost every t though. As for the continuity requirement, it renders the solution unique, for given initial state $x(t_0) = x_0$. Indeed, the continuous solution is the unique solution x (for given u) of the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} (A(\tau)x(\tau) + B(\tau)u(\tau))d\tau \quad \forall t,$$

which amounts to specifying that x must be the integral of its almost-everywhere derivative. Functions that have this property are termed absolutely continuous. Absolutely continuous functions are a subset of continuously differentiable functions and a superset of almost everywhere differentiable functions. Further, if $u \in$ PC (see Definition 2.1 below) and x is continuous and satisfies (2.41) everywhere in $[t_0, t_f]$ except possibly at (finitely many) points t at which u is discontinuous, then it must be the integral of the right-handside, hence it is absolutely continuous. Hence, for those x that satisfy (2.41) with $u \in$ PC, continuity implies absolute continuity.

From this point on, we will typically merely assume that the the control function u belongs to the set PC of piecewise continuous functions, in the sense of the following definition.[7]

Definition 2.1 *A function $u : \mathbf{R} \to \mathbf{R}^m$ belongs to* PC *if it is right-continuous*[8] *and, for every (finite) $a, b \in \mathbf{R}$ with $a < b$, it is continuous on $[a, b]$ except for possibly finitely many points of discontinuity, and has (finite) left and right limits at every point.*

[7] Everything in these notes remains valid, with occasionally some minor changes, if the continuity assumption on A and B is relaxed to piecewise continuity as well.

[8] Many authors do not insist on right-continuity in the definition of piecewise continuity. The reason we do is that with such requirement Pontryagin's Maximum Principle holds for all t rather than for almost all t, and that moreover the optimal control will be unique, which is not the case without such assumption. Indeed, note that without right- (or left-) continuity requirement, changing the value of an optimal control at, say, a single time point does not affect optimality.

Throughout, given an initial time t_0 and a terminal time t_f, the set $\mathcal{U}$ of admissible controls is defined by

$$\mathcal{U} := \{u : [t_0, t_\mathrm{f}] \to \mathbf{R}^m, u \in \mathrm{PC}, u(t) \in U \ \forall t \in [t_0, t_\mathrm{f}]\}, \tag{2.42}$$

where $U \subseteq \mathbf{R}^m$ is to be specified. The reason for not requiring that admissible controls be continuous is that, in many important cases, when U is not all of $\mathbf{R}^m$, optimal controls are "naturally" discontinuous, i.e., the problem has no solution if minimization is carried out over the set of continuous function.[9]

Example 2.3 Consider the problem of bringing a point mass from rest at some point P to rest at some point Q in the least amount of time, subject to upper and lower bounds on the acceleration—equivalently on the force applied. (For example a car is stopped at a red light and the driver wants to get as early as possible to a state of rest at the next light.) Clearly, the best strategy is to use maximum acceleration up to some point, then switch instantaneously to maximum deceleration. This is an instance of a "bang-bang" control. If we insist that the acceleration must be a continuous function of time, the problem has no solution.

(For another, explicit example, see Exercise 2.23 below.)

Given the constraint on the values of u (set U), contrary to the previous section, a linear cost function can now be meaningful. Accordingly, we start with the following problem.

Let $c \in \mathbf{R}^n, c \neq 0$, $x_0 \in \mathbf{R}^n$ and let $t_\mathrm{f} \geq t_0$ be a fixed time. Find a control $u^* \in \mathcal{U}$ so as to

$$\begin{array}{ll} \text{minimize} & c^T x(t_\mathrm{f}) \text{ s.t.} \\ & \text{dynamics } \ \dot{x}(t) = A(t)x(t) + B(t)u(t) \text{ a.e. } t \in [t_0, t_\mathrm{f}], \\ & \text{initial condition } \ x(t_0) = x_0 \\ & \text{final condition } \ x(t_\mathrm{f}) \in \mathbf{R}^n \quad \text{(no constraints)} \\ & \text{control constraint } \ u \in \mathcal{U} \\ & x \text{ absolutely continuous} \end{array}$$

Definition 2.2 *The adjoint system to system*

$$\dot{x}(t) = A(t)x(t) \tag{2.43}$$

is given by[10]

$$\dot{p}(t) = -A(t)^T p(t).$$

Let $\Phi(t, \tau)$ be the *state transition matrix* for (2.43).

Exercise 2.20 *The state transition matrix $\Psi(t, \tau)$ of the adjoint system is given by*

$$\Psi(t, \tau) = \Phi(\tau, t)^T.$$

Also, if $\dot{x} = Ax$, then $x(t)^T p(t)$ is constant.

[9] This is related to the fact that the space of continuous functions is not "complete" under, say, the L_1 norm. See more on this in Appendix A.

[10] Note that for the present problem, $L = 0$ (no integral cost), and this equation is a special case of (2.19).

Exercise 2.21 *Prove that* $\frac{d}{dt}\Phi_A(t_0,t) = -\Phi_A(t_0,t)A(t)$.

Notation: For any $u \in \text{PC}$, $z \in \mathbf{R}^n$ and $t_0 \le t_1 \le t_2 \le t_\text{f}$, let $\phi(t_2,t_1,z,u)$ denote the state at time t_2 given that at time t_1 the state is z, and control u is applied, i.e., let

$$\phi(t_2,t_1,z,u) = \Phi(t_2,t_1)z + \int_{t_1}^{t_2} \Phi(t_2,\tau)B(\tau)u(\tau)d\tau.$$

Also let

$$K(t_2,t_1,z) = \{\phi(t_2,t_1,z,u) \ : \ u \in \mathcal{U}\}.$$

This set is called *reachable set at time* t_2.

Theorem 2.8 *Let* $u^* \in \mathcal{U}$ *and let*

$$x^*(t) = \phi(t,t_0,x_0,u^*),\ t_0 \le t \le t_\text{f}.$$

Let $p^*(\cdot)$ *satisfy the adjoint equation:*

$$\dot{p}^*(t) = -A^T(t)p^*(t) \quad t_0 \le t \le t_\text{f}$$

with terminal condition

$$p^*(t_\text{f}) = -c$$

Then u^* *is optimal if and only if*

$$p^*(t)^T B(t)u^*(t) = \sup\{p^*(t)^T B(t)v \ : \ v \in U\} \quad \forall t \in [t_0,t_\text{f}) \tag{2.44}$$

(implying that the "sup" is achieved) for all $t \in [t_0,t_\text{f})$. *[Note that the optimization is now over a finite dimension space.]*

Proof. u^* is optimal if and only if, for all $u \in \mathcal{U}$

$$c^T[\Phi(t_\text{f},t_0)x_0 + \int_{t_0}^{t_\text{f}} \Phi(t_\text{f},\tau)B(\tau)u^*(\tau)d\tau] \le c^T[\Phi(t_\text{f},t_0)x_0 + \int_{t_0}^{t_\text{f}} \Phi(t_\text{f},\tau)B(\tau)u(\tau)d\tau]$$

or, equivalently,

$$\int_{t_0}^{t_\text{f}} (\Phi(t_\text{f},\tau)^T c)^T B(\tau)u^*(\tau)d\tau \le \int_{t_0}^{t_\text{f}} (\Phi(t_\text{f},\tau)^T c)^T B(\tau)u(\tau)d\tau$$

As pointed out above, for $p^*(t)$ as defined,

$$p^*(t) = \Phi(t_\text{f},t)^T p^*(t_\text{f}) = -\Phi(t_\text{f},t)^T c$$

So that u^* is optimal if and only if and only if, for all $u \in \mathcal{U}$,

$$\int_{t_0}^{t_\text{f}} p^*(\tau)^T B(\tau)u^*(\tau)d\tau \ge \int_{t_0}^{t_\text{f}} p^*(\tau)^T B(\tau)u(\tau)d\tau$$

and the 'if' direction of the theorem follows immediately. Suppose now $u^* \in \mathcal{U}$ is optimal. We show that (2.44) is satisfied $\forall t \in [t_0, t_\mathrm{f})$. Indeed, if this is not the case, $\exists t^* \in [t_0, t_\mathrm{f})$, $v \in U$ s.t.

$$p^*(t^*)^T B(t^*) u^*(t^*) < p^*(t^*)^T B(t^*) v.$$

By right-continuity of u^* (and of p^* and B), there exists $\delta > 0$ such that this inequality holds for all $t \in [t^*, t^* + \delta]$. Define $\tilde{u} \in \mathcal{U}$ by

$$\tilde{u}(t) = \begin{cases} v & t^* \leq t < t^* + \delta \\ \\ u^*(t) & \text{otherwise} \end{cases}$$

Then

$$\int_{t_0}^{t_\mathrm{f}} p^*(t)^T B(t) u^*(t) dt < \int_{t_0}^{t_\mathrm{f}} p^*(t)^T B(t) \tilde{u}(t) dt$$

which contradicts optimality of u^*.

■

Corollary 2.2 *For $t_0 \leq t \leq t_\mathrm{f}$,*

$$p^*(t)^T x^*(t) \geq p^*(t)^T \xi \quad \forall \xi \in K(t, t_0, x_0). \tag{2.45}$$

Exercise 2.22 *Prove the corollary.*

Because the problem under consideration has no integral cost, the pre-Hamiltonian H defined in section 2.1.1 reduces to

$$H(\tau, \xi, \eta, v) = \eta^T (A(\tau)\xi + B(\tau)v)$$

and the Hamiltonian $\mathcal{H}$ by

$$\mathcal{H}(\tau, \xi, \eta) = \sup\{H(\tau, \xi, \eta, v) \ : \ v \in U\}.$$

Since $p^T A(t) x$ does not involve the variable u, condition (2.44) can be written as

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \quad \forall t \tag{2.46}$$

This is another instance of Pontryagin's Principle. The previous theorem states that, for linear systems with linear objective functions, Pontryagin's Principle provides a *necessary and sufficient* condition of optimality.

Remark 2.11

1. Let $\psi : \mathbf{R}^n \to \mathbf{R}$ be the "terminal cost" function, defined by $\psi(x) = c^T x$. Then $p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f}))$, just like in the case of the problem of section 2.1.

2. Linearity in u was not used. Thus the result applies to systems with dynamics of the form
$$\dot{x}(t) = A(t)x(t) + B(t, u(t))$$
where B is, say, a continuous function.

3. The optimal control for this problem is independent of x_0. I.e., the control value to be applied is independent of the current state, it only depends on the current time. (The reason is clear from the first two equations in the proof.)

Exercise 2.23 *Compute the optimal control u^* for the following time-invariant data: $t_0 = 0$, $t_f = 1$, $A = \mathrm{diag}(1, 2)$, $B = [1; 1]$, $c = [-2; 1]$, $U = [-1, 1]$. Note that u^* is not continuous! (Indeed, this is an instance of a "bang-bang" control.)*

Fact. Let A and B be constant matrices, and suppose there exists an optimal control u^*, with corresponding trajectory x^*. Then $m(t) \stackrel{\Delta}{=} \mathcal{H}(t, x^*(t), p^*(t))$ is constant (i.e., the Hamiltonian is constant along the optimal trajectory).

Exercise 2.24 *Prove the fact under the assumption that $U = [\alpha, \beta]$. (The general case will be considered in Chapter 6.)*

Exercise 2.25 *Suppose $U = [\alpha, \beta]$, so that $B(t)$ is an $n \times 1$ matrix. Suppose that $A(t) = A$ and $B(t) = B$ are constant matrices and A has n distinct real eigenvalues. Show that there is an optimal control u^* and $t_0 = \tau_0 \leq \tau_1 < \cdots \leq \tau_n = t_f$ (n = dimension of x) such that $u^*(t) = \alpha$ or β on $[\tau_i, \tau_{i+1})$, $0 \leq i \leq n-1$. [Hint: first show that $p^*(t)^T B = \gamma_1 \exp(\delta_1 t) + \cdots + \gamma_n \exp(\delta_n t)$ for some $\delta_i, \gamma_i \in \mathbf{R}$. Then use appropriate induction.]*

2.4 More general optimal control problems

We have shown how to solve optimal control problems where the dynamics are linear, the objective function is quadratic, and the constraints are of a very simple type (fixed initial point, fixed terminal point). In most problems of practical interest, though, one or more of the following features is present.

(i) nonlinear dynamics and objective function

(ii) constraints on the control or state trajectories, e.g., $u(t) \in U \ \forall\ t$, where $U \subset \mathbf{R}^m$

(iii) more general constraints on the initial and terminal state, e.g., $g(x(t_f)) \leq 0$.

To tackle such more general optimization problems, we will make use of additional mathematical machinery. We first proceed to develop such machinery.

Chapter 3

Dynamic Programming

The dynamic-programming approach to optimal control compares candidate solutions to *all* controls, yielding global solutions, even in the absence of linearity/convexity properties. This is in contrast with the Pontryagin-Principle approach, which compares them to controls that yield nearby trajectories and hence, in the absence of appropriate linearity/convexity properties, tends to yield mere local solutions. (More on this in Chapter 6.) The price for this is that dynamic programming tends to be computationally rather demanding, in many cases prohitively so. An advantage of dynamic programming is that, at an introductory level, it does not require as much mathematical machinery as Pontryagin's Principle does. This motivates its introduction at an early point in this course. We first consider discrete-time problems, then turn to continuous time.

3.1 Discrete time

See, e.g., [4, 14, 32]).

Let $X \subset \mathbf{R}^n$ and $U_i \subset \mathbf{R}^m, i = 1, \ldots, N-1$, N a positive integer; let $\psi : \mathbf{R}^n \to \mathbf{R}$ and, for $i = 1, \ldots, N-1$, let $f_0(i, \cdot, \cdot) : X \times U_i \to \mathbf{R}$, $\psi : \mathbf{R}^n \to \mathbf{R}$, and $f(i, \cdot, \cdot) : X \times U_i \to \mathbf{R}^n$ be given; and let $u := \{u_0, \ldots, u_{N-1}\} \in \mathbf{R}^{Nm}$ designate (finite) sequences of control values. No regularity is assumed. For a given $x_0 \in X$, consider the problem

$$
\begin{array}{lll}
\text{minimize}_{u \in \mathbf{R}^{Nm}} \; J(u) := & \displaystyle\sum_{i=0}^{N-1} f_0(i, x_i, u_i) + \psi(x_N) & \\
\text{s.t.} & x_{i+1} = f(i, x_i, u_i), \; i = 0, \ldots, \; N-1, \quad x_0 \text{ fixed} & \\
& u_i \in U_i, \; i = 0, \ldots, N-1, & (P) \\
& x_i \in X, \; i = 1, \ldots, N, &
\end{array}
$$

The key idea is to embed problem (P) into a family of problems, with all possible initial times $k < N-1$, and initial conditions $\xi \in X$

$$\begin{array}{lll} \text{minimize}_{u\in\mathbf{R}^{Nm}}\ J_k(u) := & \displaystyle\sum_{i=k}^{N-1} f_0(i,x_i,u_i) + \psi(x_N) & \\ \text{s.t.} & x_{i+1} = f(i,x_i,u_i),\ i=k,\ldots,\ N-1, \quad x_k = \xi, & \\ & u_i \in U_i,\ i=k,k+1,\ldots,N-1, & (P_{k,\xi}) \\ & x_i \in X,\ i=k+1,\ldots,N. & \end{array}$$

The cornerstone of dynamic programming is Bellman's Principle of Optimality, a form of which is given in the following lemma. (R.E. Bellman, 1920–1984.)

Lemma 3.1 *(Bellman's Principle of Optimality) If $u_k^*,\ldots,u_{N-1}^*$ is optimal for $(P_{k,\xi})$ with corresponding trajectory $x_k^* = \xi$, $x_{k+1}^*,\ldots,x_N^*$ and if $\ell \in \{k,\ldots,N-1\}$, then*

$$u_\ell^*, u_{\ell+1}^*, \ldots, u_{N-1}^*$$

is optimal for (P_{ℓ,x_ℓ^*}).

Proof. Suppose not. Then there exists $\hat{u} := (\hat{u}_\ell,\ldots,\hat{u}_{N-1})$, with corresponding trajectory $\hat{x}_\ell = x_\ell^*$, $\hat{x}_{\ell+1},\ldots,\hat{x}_N$, such that

$$\sum_{i=\ell}^{N-1} f_0(i,\hat{x}_i,\hat{u}_i) + \psi(\hat{x}_N) < \sum_{i=\ell}^{N-1} f_0(i,x_i^*,u_i^*) + \psi(x_N^*) \tag{3.1}$$

Consider then the control $\tilde{u} := (\tilde{u}_k,\ldots,\tilde{u}_{N-1})$ given by

$$\tilde{u}_i = \begin{cases} u_i^* & i=k,\ldots,\ell-1 \\ \hat{u}_i & i=\ell,\ldots,N-1 \end{cases}$$

and the corresponding trajectory $\tilde{x}_k = \xi, \tilde{x}_{k+1},\ldots,\tilde{x}_N$, with

$$\begin{array}{ll} \tilde{x}_i = x_i^* & i=k,\ldots,\ell \\ \hat{x}_i & i=\ell+1,\ldots,N \end{array}$$

The value of the objective function corresponding to this control is

$$\begin{aligned} J_k(\tilde{u}) = \sum_{i=k}^{N-1} & f_0(i,\tilde{x}_i,\tilde{u}_i) + \psi(\tilde{x}_N) \\ = & \sum_{i=k}^{\ell-1} f_0(i,x_i^*,u_i^*) + \underbrace{\sum_{i=\ell}^{N-1} f_0(i,\hat{x}_i,\hat{u}_i) + \psi(\hat{x}_N)} \\ < & \sum_{i=k}^{\ell-1} f_0(i,x_i^*,u_i^*) + \sum_{i=\ell}^{N-1} f_0(i,x_i^*,u_i^*) + \psi(x_N^*) \qquad (3.2) \\ = & \sum_{i=k}^{N-1} f_0(i,x_i^*,u_i^*) + \psi(x_N^*) \end{aligned}$$

where (3.1) was invoked. Hence $\tilde{u}$ yields a lower cost than u_i^*, a contradiction.

■

Remark 3.1 This result would not hold for more general objective functions, e.g., if some f_i's involve several x_j's or u_j's.

To simplify the argument, we make the following assumption. (See, e.g., [21], for a derivation without such assumption; "inf" should then be substituted for "min" in all the results.)
Assumption. $(P_{k,\xi})$ has an optimal control for all k, ξ.
Let V be again the optimal value function, i.e., for $x \in X$, $V(N,\xi) = \xi$, and for $k \in \{0,\dots,N-1\}$,

$$V(k,\xi) = \min_{u_i \in U_i \forall i} J_{k,\xi}(u)$$

$$\text{s.t. } x_{i+1} = f(i,x_i,u_i),\ i = k,\dots,N-1,\ x_i \in X, i = k+1,\dots,N,\ x_k = \xi.$$

Theorem 3.1 *(Dynamic programming)*

$$V(k,\xi) = \min\{f_0(k,\xi,v) + V(k+1, f(k,\xi,v)) | v \in U_k, f(k,\xi,v) \in X\} \qquad (3.3)$$
$$\xi \in X,\ 0 \le k \le N-1$$

and $v \in U_k$ is a minimizer for (3.3) if and only if there exists $(u_{k+1},\dots,u_{N-1})$ such that $(v,u_{k+1},\dots,u_{N-1})$ is optimal for $(P_{k,\xi})$.

Proof. Let

$$F(k,\xi,v) := f_0(k,\xi,v) + V(k+1, f(k,\xi,v)).$$

Let u_i^*, $i = k,\dots,N-1$ be optimal for $(P_{k,\xi})$, with x_i^*, $i = k+1,\dots,N$ the corresponding state trajectory. Further, let $\hat{u}_k := v \in U_k$, with $f(k,\xi,v) \in X$, be such that no control sequence of the form $(v,u_{k+1},\dots,u_{N-1})$ is optimal for $(P_{k,\xi})$, let $(\hat{u}_{k+1},\dots,\hat{u}_{N-1})$ be optimal for $(P_{k,f(k,\xi,v)})$, and let $\hat{x}_i$, $i = k+1,\dots,N$ be the state trajectory generated by $(\hat{u}_k,\dots,\hat{u}_{N-1})$. We show that $V(k,\xi) = F(k,\xi,u_k^*) < F(k,\xi,v)$, proving both claims.
Indeed,

$$V(k,\xi) = f_0(k,\xi,u_k^*) + \sum_{i=k+1}^{N-1} f_0(i,x_i^*,u_i^*) + \psi(x_N^*) = f_0(k,\xi,u_k^*) + V(k+1, f(k,\xi,u_k^*)) = F(k,\xi,u_k^*) \qquad (3.4)$$

and

$$V(k,\xi) < f_0(k,\xi,v) + \sum_{i=k+1}^{N-1} f_0(i,\hat{x}_i,\hat{u}_i) + \psi(\hat{x}_N) = f_0(k,\xi,v) + V(k+1, f(k+1,\xi,v)) = F(k,\xi,v),$$

where in both instances the penultimate equality follows from the Principle of Optimality.

■

Definition 3.1 *A function $\phi : \{0, 1, \ldots, N-1\} \times X \to \mathbf{R}^m$ is an optimal control law if, for all $k \in \{0, 1, \ldots, N-1\}$ and all $\xi \in X$, $\phi(k, \xi) \in U_k$ and*

$$V(k, \xi) = f_0(k, \xi, \phi(k, \xi)) + V(k+1, f(k, \xi, \phi(k, \xi))).$$

Corollary 3.1 *Let $\phi : \{0, 1, \ldots, N-1\} \times X \to \mathbf{R}^m$. Then ϕ is an optimal control law if and only if, for all $k \in \{0, 1, \ldots, N-1\}$ and all $\xi \in X$,*

$$f_0(k, \xi, \phi(k, \xi)) + V(k+1, f(k, \xi, \phi(k, \xi)) = \min_{v \in U_k} \{f_0(k, \xi, v) + V(k+1, f(k, \xi, v)\}. \quad (3.5)$$

How can this result be used? First note that $V(N, \xi) = \psi(\xi)$ for all ξ, then for $k = N-1, N-2, \ldots, 0$ (i.e., "backward in time") solve (3.5) to obtain, for each ξ, $\psi(k, \xi)$, then $V(k, \xi)$.

Remark 3.2

1. No regularity assumptions were made.
2. It must be stressed that, indeed, since for a given initial condition $x(0) = x_0$ the optimal trajectory is not known a priori, $V(k, \xi)$ must be computed *for all* ξ at every time $k > 0$ (or at least for ξ equal to every "possibly optimal" x_k). Thus this approach is very CPU-demanding.

Exercise 3.1 *(From [30].) Apply dynamic programming to solve the discrete-time, free terminal state linear quadratic regulator problem, with cost function*

$$\sum_{i=k}^{N-1} \left(x_i^T L_i x_i + u_i^T u_i \right) + x_N^T Q x_N,$$

where Q and L_i, $i = 0, \ldots, N-1$ are symmetric and positive semidefinite. The dynamics are

$$x_{i+1} = A_i x_i + B_i u_i, \quad i = 0, \ldots, N-1.$$

(The Riccati equation is now of course a difference equation. Its right-hand side is a bit more complicated than before.)

3.2 Continuous time

See [21, 32].
Consider the problem (P)

$$\begin{array}{ll} \text{minimize} & \displaystyle\int_{t_0}^{t_\mathrm{f}} f_0(t, x(t), u(t))dt + \psi(x(t_\mathrm{f})) \\ \text{s.t.} & \dot{x}(t) = f(t, x(t), u(t)), \text{ a.e. } t \in [t_0, t_\mathrm{f}] \\ & x(0) = x_0, \\ & u \in \mathcal{U}, \ x \text{ absolutely continuous} \end{array}$$

We impose the following regularity conditions, which are sufficient for existence and uniqueness of an absolutely continuous solution to the differential equation and well-definedness of the objective function.

(i) for each $t \in [t_0, t_1]$, $f(t, \cdot, \cdot)$ and $f_0(t, \cdot, \cdot)$ are continuously differentiable on $\mathbf{R}^n \times \mathbf{R}^m$, and ψ is continuously differentiable.

(ii) $f, \frac{\partial f}{\partial x}, \frac{\partial f}{\partial u}$ are continuous on $[t_0, t_\mathrm{f}] \times \mathbf{R}^n \times \mathbf{R}^m$.

(iii) for $\alpha \in \mathbf{R}$, $\exists \beta, \gamma$ s.t.

$$||f(t, \xi, v)|| \leq \beta + \gamma ||\xi|| \quad \forall t \in [t_0, t_1], \xi \in \mathbf{R}^n,\ v \in \mathbf{R}^m,\ ||v|| \leq \alpha.$$

Also, for every $\tau \in [0, t_\mathrm{f})$ and $x \in \mathbf{R}^n$ consider the problem $(P_{\tau,x})$

$$\begin{array}{ll} & \text{minimize} \qquad J_\tau(u) := \int_\tau^{t_\mathrm{f}} f_0(t, x(t), u(t))dt + \psi(x(t_\mathrm{f})) \\ \text{s.t.} & \dot{x}(t) = f(t, x(t), u(t)),\ \text{a.e. } t \in [\tau, t_\mathrm{f}],\ x(\tau) = x,\ u \in \mathcal{U},\ x \text{ absolutely continuous} \end{array}$$

Assumption. $(P_{\tau,\xi})$ has an optimal control for all τ, ξ.
(Again, see, e.g., [21], for a derivation without such assumption; the resulting HJB equation then involves an "inf" instead of the "min".)
As before, let $V(\tau, \xi)$ be the minimum value of the objective function, starting from state ξ at time $\tau \leq t_\mathrm{f}$; in particular $V(t_\mathrm{f}, \xi) = \psi(\xi)$. An argument similar to that used in the discrete-time case yields, for any $\Delta < t_\mathrm{f} - \tau$

$$V(\tau, \xi) = \min \left\{ \int_\tau^{\tau+\Delta} f_0(t, \tilde{x}(t), \tilde{u}(t))dt + V(\tau + \Delta, \tilde{x}(\tau + \Delta)) \mid \tilde{u} \in \mathcal{U} \right\}, \tag{3.6}$$

where $\tilde{x}(\cdot)$ is the trajectory generated by $\tilde{u}$, i.e.

$$\begin{cases} \dot{\tilde{x}}\,(t) = f(t, \tilde{x}(t), \tilde{u}(t)) \text{ a.e. } t \in [t, t_\mathrm{f}],\ x \text{ absolutely continuous} \\ \tilde{x}(\tau) = \xi \end{cases}$$

The idea is then to differentiate both sides with respect to Δ (with the left-hand side is constant), at $\Delta = 0$, thus obtaining a differential equation. (This, of course, cannot be done in the discrete-time case.) This requires a regularity assumption on $V(\cdot, \cdot)$.
Assumption. V is continuously differentiable.
Now, for given $\tau \in [t_0, t_\mathrm{f}]$, let $v \in U$, let $\tilde{u} \in \mathcal{U}$ satisfy

$$\tilde{u}(t) = v \quad \forall t \in [\tau, t_\mathrm{f}],$$

and let $\tilde{x}$ be the corresponding state trajectory. Next, motivated by (3.6), define $\varphi_v : [0, t_\mathrm{f} - \tau) \to \mathbf{R}$ by

$$\varphi_v(\Delta) := V(\tau + \Delta, \tilde{x}(\tau + \Delta)) - V(\tau, \xi) + \int_\tau^{\tau+\Delta} f_0(t, \tilde{x}(t), \tilde{u}(t))dt.$$

Then $\varphi_v(0) = 0$ and, from (3.6),

$$\varphi_v(\Delta) \geq 0 \ \forall \Delta \in [0, t_\mathrm{f} - \tau].$$

Further, φ_v is continuously differentiable on $(0, t_\mathrm{f} - \tau)$, so that its right-derivative $\varphi_v'(0)$ at $\Delta = 0$ is non-negative, i.e., for all $\xi \in \mathbf{R}^n$, $v \in U$, and all $\tau < t_\mathrm{f}$, (since $\tilde{u}(\tau) = v$),

$$\varphi_v'(0) = \frac{\partial V}{\partial t}(\tau, \xi) + \frac{\partial V}{\partial x}(\tau, \xi) f(\tau, \xi, v) + f_0(\tau, \xi, v) \geq 0. \tag{3.7}$$

Now let $u^* \in U$ be optimal for $(P_{t,\xi})$. Then, from Bellman's Principle of Optimality (u^* optimal on $(\tau, \tau + \Delta)$), for every $\Delta \neq 0$, the function $\varphi_* : [0, t_\mathrm{f} - \tau] \to \mathbf{R}$ given by

$$\varphi_*(\Delta) := V(\tau + \Delta, x^*(\tau + \Delta)) - V(\tau, \xi) + \int_\tau^{\tau + \Delta} f_0(t, x^*(t, u^*(t)))dt$$

is identically zero, so that (since u^* is right-continuous at 0), for all $\xi \in \mathbf{R}^n$, $\tau \in [t_0, t_\mathrm{f})$,

$$\varphi_*'(0) = \frac{\partial V}{\partial t}(\tau, \xi) + \frac{\partial V}{\partial x}(\tau, \xi) f(\tau, \xi, u^*(\tau)) + f_0(\tau, \xi, u^*(\tau)) = 0. \tag{3.8}$$

From (3.7) and (3.8), and since $u^*(\tau) \in U$ for all $\tau \in [t_0, t_\mathrm{f})$, we get

$$\min_{v \in U} \left\{ \frac{\partial V}{\partial t}(\tau, \xi) + \frac{\partial V}{\partial x}(\tau, \xi) f(\tau, \xi, v) + f_0(\tau, \xi, v) \right\} = 0 \ \ \forall \tau \in [t_0, t_\mathrm{f}], \ \forall \xi \in \mathbf{R}^n.$$

We summarize the above in a formal statement.

Theorem 3.2 *Suppose that the value function V is continuously differentiable and that, for every (τ, ξ), $(P_{\tau,\xi})$ has an optimal solution. Then, for all $\tau \in [t_0, t_\mathrm{f}]$, $\xi \in \mathbf{R}^n$,*

$$\left\{ \begin{array}{ll} \frac{\partial V}{\partial t}(\tau, \xi) + \min_{v \in U} \left\{ f_0(\tau, \xi, v) + \frac{\partial V}{\partial x}(\tau, \xi) f(\tau, \xi, v) \right\} = 0 & \\ V(t_\mathrm{f}, \xi) = \psi(\xi) & (HJB) \end{array} \right.$$

This partial differential equation for $V(\cdot, \cdot)$ is known as the Hamilton–Jacobi–Bellman equation (W.R. Hamilton, Irish mathematician, 1805–1865; K.G.J. Jacobi, German mathematician, 1804–1851; R.E. Bellman, American mathematician, 1920–1984). Note that minimization is now over (a subset of) $\mathbf{R}^m$ rather than over (a subset of) the function space $\mathcal{U}$.
Now, for all τ, ξ, η, v, define $H : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ and $\mathcal{H} : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ by

$$H(\tau, \xi, \eta, v) = -f_0(\tau, \xi, v) + \eta^T f(\tau, \xi, v) \tag{3.9}$$

and

$$\mathcal{H}(\tau, \xi, \eta) = \sup_{v \in U} H(\tau, \xi, \eta, v).$$

Then (HJB) can be written as

$$\left\{ \begin{array}{l} \frac{\partial V}{\partial t}(\tau, \xi) = \mathcal{H}\left(\tau, \xi, -\nabla_x V(\tau, \xi)\right) \\ V(t_\mathrm{f}, \xi) = \psi(\xi) \end{array} \right.$$

We now show that (HJB) is also a *sufficient* condition of optimality, i.e., that if $V(\cdot, \cdot)$ satisfies (HJB), it must be the value function. (In discrete time this was obvious since the solution to (3.3) was clearly unique, given the final condition $V(N, \xi) = \psi(\xi)$.) Furthermore, obtaining V by solving (HJB) yields an optimal control in feedback form.

Theorem 3.3 *Suppose there exists $V(\cdot,\cdot)$, continuously differentiable, satisfying (HJB) together with the boundary condition. Suppose there exists $\phi(\cdot,\cdot)$, with values in U, piecewise continuous in the first argument and Lipschitz in the second argument, satisfying*

$$\frac{\partial V}{\partial t}(\tau,\xi)+\frac{\partial V}{\partial x}(\tau,\xi)f(\tau,\xi,\phi(\tau,\xi))+f_0(\tau,\xi,\phi(\tau,\xi))=0 \quad \forall(\xi,\tau). \tag{3.10}$$

Then ϕ is an optimal feedback control law and V is the value function. Further, suppose that, for some $\hat{u}\in\mathcal{U}$ and the corresponding trajectory $\hat{x}$, with $\hat{x}(0)=x_0$,

$$\frac{\partial V}{\partial t}(t,\hat{x}(t))+\frac{\partial V}{\partial x}(t,\hat{x}(t))f(t,x,\hat{u}(t))+f_0(t,\hat{x}(t),\hat{u}(t))=0 \quad \forall t. \tag{3.11}$$

Then $\hat{u}$ is optimal.

Proof. (Idea: Carefully integrate back what we differentiated.) Let $\tau<t_{\rm f}$ and $\xi\in\mathbf{R}^n$ be arbitrary.
1. Let $u\in\mathcal{U}$, yielding $x(\cdot)$ with $x(\tau)=\xi$ (initial condition), i.e.,

$$\begin{cases} \dot{x}\,(t)=f(t,\tilde{x}(t),\tilde{u}(t)) \text{ a.e. } t\in[\tau,t_{\rm f}],\ x \text{ absolutely continuous} \\ x(t)=\xi \end{cases}$$

Since V satisfies (HJB) and $u(t)\in U$ for all t, we have

$$-\frac{\partial V}{\partial t}(t,x(t))\le f_0(t,x(t),u(t))+\frac{\partial V}{\partial x}(t,x(t))f(t,x(t),u(t)) \quad \forall t\in[\tau,t_{\rm f}),$$

which yields

$$\dot{V}(t,x(t))+f_0(t,x(t),u(t))\ge 0 \quad \text{a.e. } t\in[\tau,t_{\rm f}].$$

Since V is continuously differentiable and x absolutely continuous, and since $V(t_{\rm f},x(t_{\rm f}))=\psi(x(t_{\rm f}))$, integration of both sides from τ to $t_{\rm f}$ yields

$$\psi(x(t_{\rm f}))-V(\tau,\xi)+\int_\tau^{t_{\rm f}} f_0(t,x(t),u(t))dt\ge 0,$$

i.e.,

$$V(\tau,\xi)\le\int_\tau^{t_{\rm f}} f_0(t,x(t),u(t))dt+\psi(x(t_{\rm f}))\ (=J_\tau(u)). \tag{3.12}$$

Since $u\in\mathcal{U}$ is arbitrary, this implies that

$$V(\tau,\xi)\le J_\tau(u)\ \forall u\in\mathcal{U}.$$

2. To show that V is the value function and that ϕ is an optimal feedback law indeed, it now suffices to show that the control signal produced by ϕ achieves equality in (3.12). Thus let $x^*(\cdot)$ be the unique (due to the assumptions on ϕ) absolutely continuous function that satisfies

$$\begin{aligned} \dot{x}^*(t) &= f(t,x^*(t),\phi(t,x^*(t))) \text{ a.e. } t\in[\tau,t_{\rm f}] \\ x^*(\tau) &= \xi, \end{aligned}$$

and let

$$u^*(t) = \phi(t, x^*(t)).$$

Then, proceeding as above and using (3.10), we get

$$V(t,\xi) = \int_t^{t_\mathrm{f}} f_0(\tau, x^*(\tau), u^*(\tau))d\tau + \psi(x^*(t_\mathrm{f})),$$

showing optimality of control law ϕ and, by the same token, of any control signal that satisfies (3.11), and showing the V is indeed the value function. (We leave out the proof that, given the regularity assumptions on the data, ϕ satisfies the regularity conditions in the statement of the theorem.) ■

Remark 3.3 (HJB) is a partial differential equation. Thus its solution is prohibitively CPU-demanding when n is large (curse of dimensionality).

Exercise 3.2 *Solve the free end-point linear quadratic regulator problem using HJB (see Example 6.3). Is $V(t,x)$ always well-defined for all $t \in [t_0, t_\mathrm{f})$ and $x \in \mathbf{R}^n$?*

Finally, we derive an instance of Pontryagin's Principle for continuous-time problem (P), under the additional (strong) assumption that the value function V is twice continuously differentiable. We aim for a necessary condition.

Assumption. V is twice continuously differentiable.

Theorem 3.4 *Let $u^* \in \mathcal{U}$ be an optimal control for (P_{t_0,x_0}), and x^* be the corresponding state trajectory. Suppose the value function V is twice continuously differentiable, and u^* is piecewise continuous. Then Pontryagin's Principle holds, with*

$$p^*(t) := -\nabla_x V(t, x^*(t)), \tag{3.13}$$

an absolutely continuous function.

Proof. Absolute continuity (indeed, continuous differentiability) of p^* is implied by twice continuous differentiability of V. Next,

$$p^*(t_\mathrm{f}) = -\nabla_x V(t_\mathrm{f}, x^*(t_\mathrm{f})) = -\nabla\psi(t_\mathrm{f})$$

and, from (3.8) (recall that u^* is right-continuous),

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \quad \forall t.$$

It remains to show that $\dot{p}^*(t) = -\nabla_x H(t, x^*(t), p^*(t), u^*(t))$. Since V satisfies (HJB), we have, for all (t,x) and for all $v \in U$,

$$G(t,x,v) := \frac{\partial V}{\partial t}(t,x) - H(t,x,-\nabla_x V(t,x),v) = \frac{\partial V}{\partial t}(t,x) + f_0(t,x,v) + \frac{\partial V}{\partial x}(t,x) f(t,x,v) \geq 0,$$

and in particular, since $u^*(t) \in U$ for all t,

$$G(t, x, u^*(t)) \geq 0 \quad \forall t, x.$$

Now, (3.8) yields

$$G(t, x^*(t), u^*(t)) = 0 \quad \forall t$$

so that

$$G(t, x^*(t), u^*(t)) = \min_{x \in \mathbf{R}^n} G(t, x, u^*(t)).$$

In view of the twice-differentiability assumption on V, it follows that

$$\nabla_x G(t, x^*(t), u^*(t)) = 0 \quad \forall t.$$

Now, since V is twice differentiable, we have, for all (t, x, u),

$$\nabla_x G(t, x, u) = \nabla_x \frac{\partial V}{\partial t}(t, x) + \nabla_x \left(f_0(t, x, u) - \nabla_x V(t, x)^T f(t, x, u) \right)$$

$$= \nabla_x \frac{\partial V}{\partial t}(t, x) - \nabla_x H(t, x, -\nabla V(t, x), u) + \nabla^2_{xx} V(t, x) f(t, x, u).$$

i.e., using definitions (3.9) of H and (3.13) of p^*,

$$\nabla_x \frac{\partial V}{\partial t}(t, x^*(t)) - \nabla_x H(t, x^*(t), p^*(t), u^*(t)) - \nabla^2_{xx} V(t, x^*(t)) f(t, x^*(t), u^*(t)) = 0 \quad \forall t.$$

Using again definition (3.13) of p^* yields

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

This completes the proof.

■

Exercise 3.3 *Consider the optimal control problem, with scalar x and u,*

$$\text{minimize } x(1) \quad \text{s.t.} \quad \dot{x}(t) = x(t)u(t) \text{ a.e., } |u(t)| \leq 1 \ \forall t \in [0, 1],$$

where $u \in$ PC and x is absolutely continuous. Obtain (by "inspection") the value function $V(t, x)$, for $t \in [0, 1]$, $x \in \mathbf{R}$. Verify that it is not everywhere differentiable with respect to x (hence is not a solution to (HJB)). [A fortiori, V is not twice continuously differentiable, and the derivation of Pontryagin's Principle given in Theorem 3.4 is not valid. The PP itself does hold though, as we'll see down the road. This situation is typical in problems with nonlinear (in (x, u)) dynamics that is linear in u for fixed x, and with simple bounds on u, and terminal cost; such problems are common place in engineering applications.]

Chapter 4

Unconstrained Optimization

References: [22, 26].

4.1 First order condition of optimality

We consider the problem

$$\min\{f(x) \ : \ x \in V\} \tag{4.1}$$

where V is a normed vector space and $f : V \to \mathbf{R}$.

Remark 4.1 This problem is technically very similar to the problem

$$\min\{f(x) \ : \ x \in \Omega\} \tag{4.2}$$

where Ω is an open set in V, as shown in the following exercise.

Exercise 4.1 *Suppose $\Omega \subset V$ is open. Prove carefully, using the definitions given earlier, that $\hat{x}$ is a local minimizer for (4.2) if and only if $\hat{x}$ is a local minimizer for (4.1) and $\hat{x} \in \Omega$. Further, prove that the "if" direction is true for general (not necessarily open) Ω, and show by exhibiting a simple counterexample that the "only if" direction is not.*

Now suppose f is (Fréchet) differentiable (see Appendix B). (In fact, many of the results we obtain below hold under the milder assumptions.) We next obtain a first order necessary condition for optimality.

Theorem 4.1 *Let f be differentiable. Suppose $\hat{x}$ is a local minimizer for (4.1). Then $\frac{\partial f}{\partial x}(\hat{x}) = \theta$.*

Proof. Since $\hat{x}$ is a local minimizer for (4.1), there exists $\epsilon > 0$ such that

$$f(\hat{x} + h) \geq f(\hat{x}) \quad \forall h \in B(\theta, \epsilon) \tag{4.3}$$

Since f is Fréchet-differentiable, we have, for all $h \in V$,

$$f(\hat{x}+h) = f(\hat{x}) + \frac{\partial f}{\partial x}(\hat{x})h + o(h) \tag{4.4}$$

with $\frac{o(h)}{||h||} \to 0$ as $h \to \theta$. Hence, from (4.3), whenever $h \in B(\theta, \epsilon)$

$$\frac{\partial f}{\partial x}(\hat{x})h + o(h) \geq \theta \tag{4.5}$$

or, equivalently,

$$\frac{\partial f}{\partial x}(\hat{x})(td) + o(td) \geq 0 \quad \forall d \in B(\theta, 1),\ \forall t \in [0, \epsilon] \tag{4.6}$$

and, dividing by t, for $t \neq 0$

$$\frac{\partial f}{\partial x}(\hat{x})d + \frac{o(td)}{t} \geq 0 \quad \forall d \in B(\theta, 1),\ \forall t \in (0, \epsilon] \tag{4.7}$$

It is easy to show (see exercise below) that $\frac{o(th)}{t} \to 0$ as $t \to 0$. Hence, letting $t \to 0$ in (4.7), we get

$$\frac{\partial f}{\partial x}(\hat{x})d \ \geq 0 \quad \forall d \in B(\theta, \epsilon).$$

Since $d \in B(\theta, \epsilon)$ implies $-d \ \in B(\theta, \epsilon)$, we have $\frac{\partial f}{\partial x}(\hat{x})(-d) \ \geq 0$ thus

$$\frac{\partial f}{\partial x}(\hat{x})d \leq 0 \quad \forall d \in B(\theta, \epsilon).$$

Hence

$$\frac{\partial f}{\partial x}(\hat{x})d = 0 \quad \forall d \in B(\theta, \epsilon)$$

which implies (since $\frac{\partial f}{\partial x}(\hat{x})$ is linear)

$$\frac{\partial f}{\partial x}(\hat{x})d = 0 \quad \forall d \in V$$

i.e.,

$$\frac{\partial f}{\partial x}(\hat{x}) = \theta.$$

■

Remark 4.2 The same optimality condition can be established (with essentially the same proof) under less restrictive assumptions, specifically mere G-differentiability of f at $\hat{x}$ (which does not require a norm on V) and mere "weak" local optimality of $\hat{x}$: For every $h \in V$, there exists $\epsilon_h > 0$ such that $f(\hat{x}) \leq f(\hat{x} + th)$ for all $t \leq \epsilon_h$.

Exercise 4.2 *If $o(\cdot)$ is such that $\frac{o(h)}{||h||} \to \theta$ as $h \to \theta$ then, for any fixed $d \neq \theta$, $\frac{o(td)}{t} \to \theta$ as $t \to 0$ $t \in \mathbf{R}$.*

Remark 4.3 The optimality condition above, like several other conditions derived in this course, is only a *necessary* condition, i.e., a point x satisfying this condition need not be optimal, even locally. However, if there is an optimal x, it has to be among those which satisfy the optimality condition. Also it is clear that this optimality condition is also necessary for a global minimizer (since a global minimizer is also a local minimizer). Hence, if a global minimizer is known to exist, it must be, among the points satisfying the optimality condition, the one with minimum value of f.

Suppose now that we want to find a minimizer. Solving $\frac{\partial f}{\partial x}(\hat{x}) = \theta$ for x (n nonlinear equations in n unknown, when $V = \mathbf{R}^n$) is usually hard and could well yield maximizers or other stationary points instead of minimizers. A central idea is, given an "initial guess" $\tilde{x}$, to determine a "descent", i.e., a direction in which, starting from $\tilde{x}$, f decreases (at least for small enough displacement). Thus suppose $\frac{\partial f}{\partial x}(\tilde{x}) \neq \theta$ (hence $\tilde{x}$ is not a local minimizer). If h is such that $\frac{\partial f}{\tilde{x}}(\tilde{x})h < 0$ (such h exists; why?) then, for $t > 0$ small enough, we have

$$f(\tilde{x} + th) - f(\tilde{x}) = t\left\{\frac{\partial f}{\partial x}(\tilde{x})h + \frac{o(th)}{t}\right\} < 0$$

and, hence, for some $t_h > 0$,

$$f(\tilde{x} + th) < f(\tilde{x}) \qquad \forall t \in (0, t_h].$$

Such h is called a *descent direction* for f at $\tilde{x}$. The concept of descent direction is essential to numerical methods.

4.2 Steepest descent method

Suppose now that V is a Hilbert space. Then $\frac{\partial f}{\partial x}(\tilde{x})h = \langle \mathrm{grad} f(\tilde{x}), h\rangle$ and a particular descent direction is $h = -\mathrm{grad} f(\tilde{x})$. (This is so irrespective of which inner product (and associated gradient) is used. We will return to this point when studying Newton's method and variable metric methods.) This direction is known as the direction of steepest descent. The next exercise justifies this terminology.

Exercise 4.3 *Let V be a Hilbert space, with inner product $\langle\cdot,\cdot\rangle_V$ and associated gradient grad_V. Let $f : V \to \mathbf{R}$ be differentiable at $x \in V$, with gradient $\mathrm{grad}_V f(x) \neq \theta$, and let $\hat{d} := -\frac{\mathrm{grad}_V f(x)}{\|\mathrm{grad}_V f(x)\|}$. (i) Show that*

$$\mathrm{argmin}\left\{\frac{\partial f}{\partial x}(x)d \;:\; \langle d, d\rangle_V = 1\right\} = \{\hat{d}\}.$$

(Hint: Use the Cauchy-Bunyakovskii-Schwartz inequality.) (ii) Show that, given any $\tilde{d} \neq \hat{d}$ with $\langle \tilde{d}, \tilde{d}\rangle = 1$, there exists $\bar{t} > 0$ such that

$$f(x + t\hat{d}) < f(x + t\tilde{d}) \quad \forall t \in (0, \bar{t}].$$

Exercise 4.4 *Under the same assumptions as in Exercise 4.3, show that* $\hat{h} = -\text{grad}f(x)$ *is also the unique solution of*

$$\min_h f(x) + \frac{\partial f}{\partial x}(x)h + \frac{1}{2}\langle h, h\rangle,$$

i.e., the only minimizer of the second order expansion of f *about* x *when the Hessian is identity.*

In view of the above, a natural algorithm for attempting to solve (4.1) would be the following.

Algorithm SD (steepest descent with exact line search)
Data $x_0 \in H$
$i := 0$
while $\text{grad}f(x_i) \neq \theta$ do {
  pick $t_i \in \arg\min_t\{f(x_i - t\text{grad}f(x_i)) : t \geq 0\}$ (if there is no such minimizer
  the algorithm fails)
  $x_{i+1} := x_i - t_i\text{grad}f(x_i)$
  $i := i + 1$
  }
stop

Notation: Given a real-valued function ϕ, the (possibly empty) set of global minimizers for the problem

$$\text{minimize } \phi(x) \text{ s.t. } x \in S$$

is denoted by

$$\arg\min_x\{\phi(x) : \quad x \in S\}.$$

In the algorithm above, like in other algorithms we will study in this course, each iteration consists of essentially 2 operations:

- computation of a search direction (here, directly opposed to the gradient of f)
- a search along that search direction, which amounts to solving, often approximately, a minimization problem in only one variable, t. The function $\phi(t) = f(x + th)$ can be viewed as the one-dimensional section of f at x in direction h. This second operation is often also called "step-size computation" or "line search".

Before analyzing the algorithm above, we point out a practical difficulty. Computation of t_i involves an exact minimization which cannot in general be performed exactly in finite time (it requires construction of an infinite sequence). Hence, point x_f will never be actually constructed and convergence of the sequence $\{x_i\}$ cannot be observed. One says that the algorithm is not *implementable*, but merely *conceptual*. An implementable algorithm for solving (4.1) will be examined later.

4.3 Introduction to convergence analysis

In order to analyze Algorithm SD, we embed it in a class of algorithms characterized by the following algorithm model. Here, $a : V \to V$, V a normed vector space, represents an *iteration map*; and $\Delta \subset V$ is a set of "desirable" points.

Algorithm Model 1
Data. $x_0 \in V$
$i = 0$
while $x_i \notin \Delta$ do {
  $x_{i+1} = a(x_i)$
  $i = i + 1$
  }
stop

Theorem 4.2 *Suppose*

(i) $a(\cdot)$ *is continuous in* Δ^c *(*Δ^c *is the complement of* Δ*)*

(ii) There exists $v : \Delta^c \to V$ *such that* $v(a(x)) < v(x) \quad \forall x \in \Delta^c$

Then, if the sequence $\{x_i\}$ *constructed by Algorithm model 1 is infinite, every accumulation point of* $\{x_i\}$ *is desirable (i.e., belongs to* Δ*).*

Exercise 4.5 *Prove the theorem.*

Exercise 4.6 *Give an example (i.e., exhibit* $a(\cdot)$*,* $v(\cdot)$ *and* Δ*) showing that condition (ii), in Theorem 4.2, cannot be dropped.*

Remark 4.4

1. Note that Algorithm Model 1 does not imply any type of "optimization" idea. The result of Theorem 4.2 will hold if one can show the existence of a function $v(\cdot)$ that, together with $a(\cdot)$, would satisfy conditions (i) to (iii). This idea is related to that of a Lyapunov function for the "discrete-time system" $x_{i+1} = a(x_i)$ (but assumptions on v are weaker, and the resulting sequence may be unbounded).

2. The result of Theorem 4.2 is stronger than it may appear at first glance.

   (i) if $\{x_i\}$ is bounded (e.g., if all level sets of f are bounded) and V is finite-dimensional, accumulation points do exist.

   (ii) if accumulation point(s) exist(s) (which implies that Δ is nonempty), and if Δ is a finite set (which is often the case), there are simple techniques that can force the entire sequence to converge to one of these accumulation points. For example, if Δ is the set of stationary points of v, one might restrict the step $\|x_{i+1} - x_i\|$ to never be larger than some constant multiple of $\|\nabla v(x_i)\|$ (step-size limitation).

Exercise 4.7 *Consider Algorithm SD (steepest descent) with $H = \mathbf{R}^n$ and define, for any $x \in \mathbf{R}^n$*

$$t(x) = \arg\min_t \{f(x - t\nabla f(x)) \ : \ t \geq 0\}$$

where we assume that $t(x)$ is uniquely defined (unique global minimizer) for every $x \in \mathbf{R}^n$. Also suppose that $t(\cdot)$ is locally bounded, i.e., for any bounded set K, there exists $M > 0$ s.t. $|t(x)| < M$ for all $x \in K$. Show that the hypotheses of Theorem 4.2 are satisfied with $\Delta = \{x \in \mathbf{R}^n : \frac{\partial f}{\partial x}(x) = 0\}$ and $v = f$. [Hint: the key point is to show that $a(\cdot)$ is continuous, i.e., that $t(\cdot)$ is continuous. This does hold because, since f is continuously differentiable, the curve below (Figure **4.1***) does not change too much in a neighborhood of x.]*

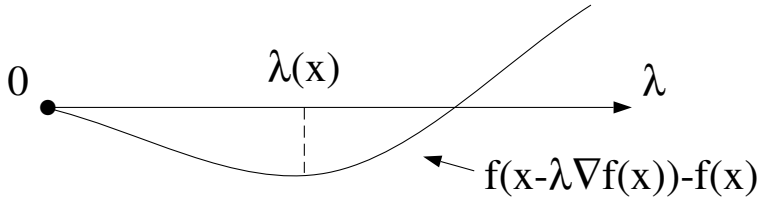


Figure **4.1**: Steepest descent with exact search

Remark 4.5 We just proved that Algorithm SD yields accumulation points $\hat{x}$ (if any) such that $\frac{\partial f}{\partial x}(\hat{x}) = 0$. There is no guarantee, however, that $\hat{x}$ is even a local minimizer (e.g., take the case where x_0 is a local *maximizer*). Nevertheless, this will very likely be the case, since the cost function decreases at each iteration (and thus, local minimizers are the only 'stable' points.)

In many cases $t(x)$ will not be uniquely defined for all x and, when it is, $a(\cdot)$ may not be continuous. (See, e.g., the "Armijo" line search discussed below.) Also, the iteration map may have memory, i.e., may not be a function of x alone. This issue is addressed in Algorithm Model 2 below, which is based on a point-to-set iteration map

$$A \ : \ V \to 2^V$$

(2^V is the set of subsets of V).

Algorithm Model 2
Data. $x_0 \in V$
$i = 0$
while $x_i \notin \Delta$ do {
  pick $x_{i+1} \in A(x_i)$
  $i = i + 1$
  }
stop

The advantages of using a point to set iteration map are that

(i) compound algorithms can be readily analyzed (two or more algorithms are intertwined)

(ii) this can include algorithms for which the iteration depends on some past information (i.e., conjugate gradient methods)

(iii) algorithms not satisfying the conditions of the previous theorem (e.g., a not continuous) may satisfy the conditions of the theorem below.

The algorithm above, with the convergence theorem below, will allow us to analyze an implementable algorithm. The following theorem is due to Polak [26].

Theorem 4.3 *Suppose that there exists a function $v : V \to \mathbf{R}$ such that*

(i) $v(\cdot)$ is continuous in Δ^c

(ii) $\forall x \in \Delta^c \; \exists \; \epsilon > 0, \; \delta > 0$ such that

$$v(y') - v(x') \leq -\delta \quad \forall x' \in B(x, \epsilon) \quad \forall y' \in A(x') \tag{4.8}$$

Then, if the sequence $\{x_i\}$ constructed by Algorithm model 2 is infinite, every accumulation point of $\{x_i\}$ is desirable (i.e., belongs to Δ).

Remark 4.6 (4.8) indicates a *uniform decrease* in the neighborhood of any non-desirable point. Note that a similar property was implied by (ii) and (iii) in Theorem 4.2.

Lemma 4.1 *Let $\{t_i\} \subset \mathbf{R}$ be a monotonically decreasing sequence such that $t_i \overset{K}{\to} t^*$ for some $K \subset \mathbb{N}, \; \mathrm{t}^* \in \mathbf{R}$. Then $t_i \searrow t^*$.*

Exercise 4.8 *Prove the lemma.*

Proof of Theorem 4.3
By contradiction. Suppose $x_i \overset{K}{\to} \hat{x} \notin \Delta$. Since $v(\cdot)$ is continuous, $v(x_i) \overset{K}{\to} v(\hat{x})$. Since, in view of (ii),

$$v(x_{i+1}) < v(x_i) \quad \forall i$$

it follows from the lemma above that

$$v(x_i) \to v(\hat{x}) \; . \tag{4.9}$$

Now let ϵ, δ correspond to $\hat{x}$ in assumption (ii). Since $x_i \overset{K}{\to} \hat{x}$, there exists i_0 such that $\forall i \geq i_0, \; i \in K$, x_i belongs to $B(\hat{x}, \epsilon)$. Hence, $\forall i \geq i_0, \; i \in K$,

$$v(y) - v(x_i) \leq -\delta \quad \forall y \in A(x_i) \tag{4.10}$$

and, in particular

$$v(x_{i+1}) \leq v(x_i) - \delta \quad \forall i \geq i_0, \; i \in K \tag{4.11}$$

But this contradicts (4.9) and the proof is complete.

■

Exercise 4.9 *Show that Algorithm SD (steepest descent) with $H = \mathbf{R}^n$ satisfies the assumptions of Theorem 4.3. Hence $x_i \overset{K}{\to} \hat{x}$ implies $\frac{\partial f}{\partial x}(\hat{x}) = \theta$ (assuming that* $\operatorname{argmin}_t\{f(x_i - t\nabla f(x_i))\}$ *is always nonempty).*

In the following algorithm, a line search due to Armijo (Larry Armijo, 20th century American mathematician) replaces the exact line search of Algorithm SD, making the algorithm implementable. This line search imposes a decrease of $f(x_i)$ at each iteration, which is common practice. Note however that such "monotone decrease" in itself is not sufficient for inducing convergence to stationary points. Two ingredients in the Armijo line search insure that "sufficient decrease" is achieved: (i) the back-tracking technique insures that, away from stationary points, the step will not be vanishingly small, and (ii) the "Armijo line" test insures that, whenever a reasonably large step is taken, a reasonably large decrease is achieved.
We present this algorithm in a more general form, with a search direction h_i ($h_i = -\text{grad}f(x_i)$ corresponds to *Armijo-gradient*).

Algorithm 2 (Armijo step-size rule)
Parameters $\alpha, \beta \in (0, 1)$
Data $x_0 \in V$
$i = 0$
while $\frac{\partial f}{\partial x}(x_i) \neq \theta$ do {
  compute an appropriate search direction h_i
  $t = 1$
  while $f(x_i + th_i) - f(x_i) > \alpha t f'(x_i; h_i)$ do $t := \beta t$
  $x_{i+1} = x_i + th_i$
  $i = i + 1$
  }
stop

To get an intuitive picture of this step-size rule, let us define a function $\phi_i : \mathbf{R} \to \mathbf{R}$ by

$$\phi_i(t) = f(x_i + th_i) - f(x_i)$$

Using the chain rule we have

$$\phi_i'(0) = \frac{\partial f}{\partial x}(x_i + 0h_i)h_i = f'(x_i; h_i)$$

so that the condition to be satisfied by $t = \beta^k$ can be written as

$$\phi_i(t) \leq \alpha t \phi_i'(0)$$

Hence the Armijo rule prescribes to choose the step-size t_i as the least power of β (hence the largest value, since $\beta < 1$) at which the curve $\phi(t)$ is below the straight line $\alpha t\phi_i'(0)$, as shown on the Figure **4.2**. In the case of the figure $k = 2$ will be chosen. We see that this step-size will be well defined as long as

$$f'(x_i; h_i) < 0$$

which insures that the straight lines on the picture are downward. Let us state and prove this precisely.

Proposition 4.1 *Suppose that* $\frac{\partial f}{\partial x}(x_i) \neq 0$ *and that* h_i *is such that*

$$f'(x_i; h_i) < 0$$

Then there exists an integer k *such that* $t = \beta^k$ *satisfies the line search criterion.*

Proof

$$\begin{aligned} \phi_i(t) \;=\; \phi_i(t) - \phi_i(0) \;&=\; t\phi_i'(0) + o(t) \\ &=\; t\alpha\phi_i'(0) + t(1-\alpha)\phi_i'(0) + o(t) \\ &=\; t\alpha\phi_i'(0) + t\left((1-\alpha)\phi_i'(0) + \frac{o(t)}{t}\right) \quad t \neq 0 \end{aligned}$$

Since $\phi_i'(0) = f'(x_i; h_i) < 0$, the expression within braces is negative for $t > 0$ small enough, thus $\phi_i(t) < t\alpha\phi_i'(0)$ for $t > 0$ small enough.

■

We will now apply Theorem 4.3 to prove convergence of Algorithm 2. We just have to show that condition (ii) holds (using $v \equiv f$, (i) holds by assumption). For simplicity, we let $V = \mathbf{R}^n$.

THE NEXT THEOREM TO BE SIMPLIFIED AND BE STATED AND PROVED FOR THE SPECIAL CASE OF h_i BEING THE NEGATIVE GRADIENT DIRECTION, WITH A REMARK MENTIONING THE GENERALIZATION.

Theorem 4.4 *Let* $H(x)$ *denote the set of search directions that could possibly be constructed by Algorithm 2 when* x *is the current iterate. (In the case of steepest descent,* $H(x) =$

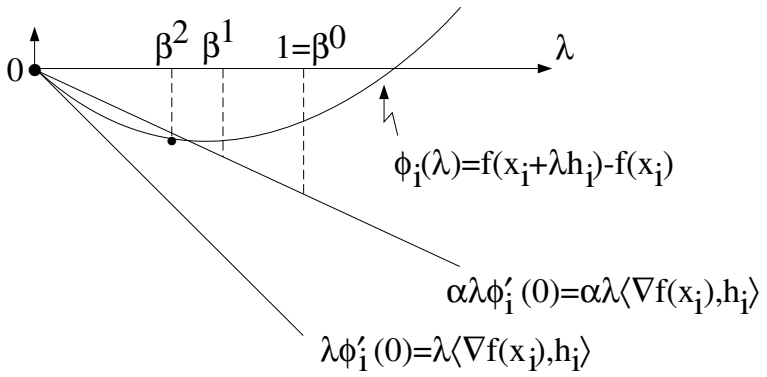


Figure **4.2**: Armijo rule

$\{-\text{grad}f(x)\}$.) *Suppose that $H(x)$ is bounded away from zero near non-stationary points, i.e., for any $\hat{x}$ such that $\frac{\partial f}{\partial x}(\hat{x}) \neq 0$, there exists $\epsilon > 0$ such that*

$$\inf\{\|h\| : h \in H(x), \|x - \hat{x}\| \leq \epsilon\} > 0. \tag{4.12}$$

Further suppose that for any $\hat{x}$ for which $\frac{\partial f}{\partial x}(\hat{x}) \neq 0$ there exist positive numbers ϵ and ρ such that, $\forall x \in B(\hat{x}, \epsilon)$, $\forall h \in H(x)$, it holds

$$\langle \nabla f(x), h\rangle \leq -\rho \|\nabla f(x)\| \; ||h||, \tag{4.13}$$

where $\|\cdot\|$ is the norm induced by the inner product. Then $x_i \overset{K}{\to} x^*$ implies $\frac{\partial f}{\partial x}(x^*) = 0$.

Proof (For simplicity we assume the standard Euclidean inner product.)

$$f(x + th) = f(x) + \int_0^1 \langle \nabla f(x + tth), th\rangle dt$$

Thus

$$\begin{aligned} f(x+th) - f(x) - \alpha t\langle \nabla f(x), h\rangle = t\int_0^1 \langle \nabla f(x &+ \sigma th) - \nabla f(x), h\rangle d\sigma \\ &+ t(1-\alpha)\langle \nabla f(x), h\rangle \\ \leq t(\sup_{\sigma\in[0,1]} |\langle \nabla f(x+\sigma th) - \nabla f(x), h\rangle| &+ (1-\alpha)\langle \nabla f(x), h\rangle) \quad \forall t \geq 0 \end{aligned} \tag{4.14}$$

Suppose now that $\hat{x}$ is such that $\nabla f(\hat{x}) \neq 0$ and let ϵ, ρ and C satisfy the hypotheses of the theorem. Substituting (4.13) into (4.14) yields (since $\alpha \in (0,1)$ and $t \geq 0$), using Schwartz inequality,

$$\begin{aligned} f(x+th) - f(x) - \alpha t\langle \nabla f(x), h\rangle &\leq t||h||(\sup_{\sigma\in[0,1]} ||\nabla f(x+\sigma th) - \nabla f(x)|| - (1-\alpha)\rho||\nabla f(x)||) \\ &\forall \quad x \in B(\hat{x}, \epsilon), \quad \forall h \in H(\hat{x}) \end{aligned} \tag{4.15}$$

Assume now that $h(x) = -\nabla f(x)$ [the proof for the general case is left as a (not entirely trivial) exercise.] First let us pick $\epsilon' \in (0, \epsilon]$ s.t., for some $\eta > 0$, (using continuity of ∇f)

$$(1-\alpha)\rho||\nabla f(x)|| > \eta > 0 \quad \forall x \in B(\hat{x}, \epsilon')$$

Also by continuity of ∇f, $\exists C$ s.t. $||\nabla f(x)|| \leq C \quad \forall x \in B(\hat{x}, \epsilon')$. Since $\bar{B}(\hat{x}, \epsilon')$ is compact, ∇f is *uniformly continuous* over $\bar{B}(\hat{x}, \epsilon')$. Thus, there exists $\bar{\epsilon} > 0$ such that

$$||\nabla f(x+v) - \nabla f(x)|| < \eta \quad \forall ||v|| < \bar{\epsilon} \quad \forall x \in \bar{B}(\hat{x}, \epsilon')$$

Thus,

$$||\nabla f(x - \sigma t \nabla f(x)) - \nabla f(x)|| < \eta \quad \forall \sigma \in [0,1], \; t \in [0, \frac{\bar{\epsilon}}{C}], \; x \in \bar{B}(\hat{x}, \epsilon')$$

which implies

$$\sup_{\sigma\in[0,1]} ||\nabla f(x-\sigma t\nabla f(x))-\nabla f(x)|| < \eta \quad \forall t\in[0,\bar{t}],\ x\in\bar{B}(\hat{x},\epsilon')$$

with $\bar{t}=\frac{\bar{\epsilon}}{C}>0$. Thus (4.15) yields

$$f(x-t\nabla f(x))-f(x)+\alpha t||\nabla f(x)||^2<0 \quad \forall t\in(0,\bar{t}],\ x\in\bar{B}(\hat{x},\epsilon') \tag{4.16}$$

Let us denote by $k(x)$ the value of k_i constructed by Algorithm 2 if $x_i=x$. Then, from (4.16) and the definition of $k(x)$

$$k(x)\le k^* \stackrel{\Delta}{=} \max(0,\tilde{k}) \quad \forall x\in B(\hat{x},\epsilon') \tag{4.17}$$

where $\tilde{k}$ is such that

$$\beta^{\tilde{k}}\le\bar{t}<\beta^{\tilde{k}-1} \tag{4.18}$$

(since $\beta^{\tilde{k}}$ will then always satisfy inequality (4.16). The iteration map $A(x)$ (singleton valued in this case) is

$$A(x)=\{x-\beta^{k(x)}\nabla f(x)\}.$$

and, from the line search criterion, using (4.17) and (4.18)

$$\begin{aligned} f(A(x))-f(x) &\le -\alpha\beta^{k(x)}||\nabla f(x)||^2 \\ &\le -\alpha\beta^{k^*}||\nabla f(x)||^2 \quad \forall x\in\bar{B}(\hat{x},\epsilon') \\ &\le -\alpha\beta^{k^*}\frac{\eta^2}{(1-\alpha)^2\rho^2} \quad \forall x\in\bar{B}(\hat{x},\epsilon') \end{aligned}$$

and condition (ii) of Theorem 4.3 is satisfied with

$$\delta=\alpha\beta^{k^*}\frac{\eta^2}{(1-\alpha)^2\rho^2}>0.$$

Hence $x_i\stackrel{K}{\to}x^*$ implies $x^*\in\Delta$, i.e., $\nabla f(x^*)=0$.

■

Remark 4.7 Condition (4.13) expresses that the angle between h and $(-\text{grad}f(x))$, (in the 2D plane spanned by these two vectors) is *uniformly* bounded by $\cos^{-1}(\rho)$ (note that $\rho>1$ cannot possibly satisfy (4.13) except if both sides are $=0$). In other words, this angle is *uniformly bounded away* from 90°. This angle just being less than 90° for all x, insuring that $h(x)$ is always a descent direction, is indeed not enough. Condition (4.12) prevents $h(x)$ from collapsing (resulting in a very small step $||x_{k+1}-x_k||$) except near a desirable point.

Exercise 4.10 *Prove that Theorem 4.4 is still true if the Armijo search is replaced by an exact search (as in Algorithm SD).*

Remark 4.8

1. A key condition in Theorem 4.3 is that of *uniform* descent in the neighborhood of a non-desirable point. Descent by itself may not be enough.

2. The best values for α and β in Armijo step-size are not obvious a priori. More about all this can be found in [5, 26].

3. Many other step-size rules can be used, such as golden section search, quadratic or cubic interpolation, Goldstein step-size. With some line searches, a stronger convergence result than that obtained above can be proved, under the additional assumption that f is bounded from below (note that, without such assumption, the optimization problem would not be well defined): Irrespective of whether or not $\{x_k\}$ has accumulation points, the sequence of gradients $\{\nabla f(x_k)\}$ always converges to zero. See, e.g., [24, section 3.2].

Note on the assumption $x_i \to x^*$. The convergence results given earlier assert that, under some assumptions, every accumulation point of the sequence generated by a suitable algorithm (e.g., Armijo gradient) satisfies the first order necessary condition of optimality. A much "nicer" result would be that the entire sequence converges. The following exercises address this question.

Exercise 4.11 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable. Suppose $\{x_i\}$ is constructed by the Armijo-gradient algorithm. Suppose that $\{x : \frac{\partial f}{\partial x}(x) = \theta\}$ is finite and suppose that $\{x_i\}$ has an accumulation point $\hat{x}$. Then $x_i \to \hat{x}$.*

Exercise 4.12 *Let $\hat{x}$ be an isolated stationary point of $f : \mathbf{R}^n \to \mathbf{R}$ and let $\{x_i\}$ be a sequence with the property that all its accumulation points are stationary, and that one of those is $\hat{x}$. (A stationary point $\hat{x}$ of f is said to be isolated if there exists $\epsilon > 0$ such that f has no stationary point in $B(\hat{x}, \epsilon) \setminus \{\hat{x}\}$.) Further suppose that there exists $\rho > 0$ such that the index set $K := \{i : x_i \in B(\hat{x}, \rho)\}$ is such that $\|x_{i+1} - x_i\| \to 0$ as $i \to \infty$, $i \in K$. Prove that under these assumptions $x_i \to \hat{x}$ as $i \to \infty$.*

The assumption in Exercise 4.12 is often satisfied. In particular, any local minimum satisfying the 2nd order *sufficient* condition of optimality is isolated (why?). Finally if $x_{i+1} = x_i - t\text{grad}f(x_i)$ and $|t| \leq 1$ (e.g., Armijo-gradient) then $\|x_{i+1} - x_i\| \leq \|\text{grad}f(x_i)\|$ which goes to 0 on sub-sequences converging to a stationary point; for other algorithms, suitable *step-size limitation* schemes will yield the same result.

Exercise 4.13 *Let $f : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable and let $\{x_i\}$ be a bounded sequence with the property that every accumulation point $\hat{x}$ satisfies $\frac{\partial f}{\partial x}(\hat{x}) = 0$. Then $\frac{\partial f}{\partial x}(x_i) \to 0$ as $i \to \infty$.*

4.4 Second order optimality conditions

Here we consider only $V = \mathbf{R}^n$ (the analysis in the general case is slightly more involved). Consider again the problem

$$\min\{f(x) \;\big|\; x \in \mathbf{R}^n\} \tag{4.19}$$

Theorem 4.5 *(2nd order necessary condition) Suppose that f is twice differentiable and let $\hat{x}$ be a local minimizer for (4.19). Then $\nabla^2 f(\hat{x})$ is positive semi-definite, i.e.*

$$d^T \nabla^2 f(\hat{x})(\hat{x}) d \geq 0 \quad \forall d \in \mathbf{R}^n$$

Proof. Let $d \in \mathbf{R}^n$, $\|d\| = 1$, and let $t > 0$. Since $\hat{x}$ is a local minimizer, $\nabla f(\hat{x}) = 0$. Second order expansion of f around $\hat{x}$ yields

$$0 \leq f(\hat{x} + td) - f(\hat{x}) = \frac{t^2}{2}\left[d^T \nabla^2 f(\hat{x}) d + \frac{o_2(td)}{t^2}\right] \tag{4.20}$$

with $\frac{o_2(h)}{\|h\|^2} \to 0$ as $h \to 0$. The claim then follows by letting $t \to 0$.

Theorem 4.6 *(2nd order sufficiency condition) Suppose that f is twice differentiable, that $\frac{\partial f}{\partial x}(\hat{x}) = 0$ and that $\nabla^2 f(\hat{x})$ is positive definite. Then $\hat{x}$ is a strict local minimizer for (4.19).*

Proof. Let $m > 0$ be the smallest eigenvalue of $\nabla^2 f(\hat{x})$. (It is positive indeed since V is assumed finite-dimensional.) Then, since $h^T \nabla^2 f(\hat{x}) h \geq m\|h\|^2$ for all h,

$$f(\hat{x} + h) - f(\hat{x}) \geq \|h\|^2 \left(\frac{m}{2} + \frac{o_2(h)}{\|h\|^2}\right) \quad \forall h \neq \theta \; .$$

Let $\epsilon > 0$ be such that $\frac{|o_2(h)|}{\|h\|^2} < \frac{m}{2}$ for all $\|h\| < \epsilon$, $h \neq \theta$. Then

$$f(\hat{x} + h) > f(\hat{x}) \quad \forall \|h\| \leq \epsilon \; , \; h \neq \theta,$$

proving the claim.

■

Alternatively, under the further assumption that the second derivative of f is continuous, Theorem 4.6 can be proved by making use of (B.15), and using the fact that, due to the assumed continuity of the second derivative, $\left(\frac{\partial^2 f}{\partial x^2}(\hat{x} + th)h\right) h \geq (m/2)\|h\|^2$ for all h small enough and $t \in (0, 1)$.

Exercise 4.14 *Show by a counterexample that the 2nd order sufficiency condition given above is not valid when the space is infinite-dimensional. [Hint: Consider the space of sequences in $\mathbf{R}$ with finitely nonzero entries (equivalently, the space of univariate polynomials), with an appropriate norm.] Show that the condition remains sufficient in the infinite-dimensional case if it is expressed as: there exists $m > 0$ such that, for all h,*

$$\left(\frac{\partial^2 f}{\partial x^2}(\hat{x})h\right) h \geq (m/2)\|h\|^2$$

Remark 4.9 Consider the following "proof" for Theorem 4.6. "Let $d \neq 0$. Then $\langle d, \frac{\partial^2 f}{\partial x^2}(\hat{x})d\rangle = \delta > 0$. Let $h = td$. Proceeding as in (1), (4.20) yields $f(\hat{x}+td) - f(\hat{x}) > 0 \quad \forall t \in (0, \bar{t}]$ for some $\bar{t} > 0$, which shows that $\hat{x}$ is a local minimizer." This argument is in error. Why? (Note that if this argument was correct, it would imply that the result also holds on infinite-dimensional spaces. However, on such spaces, it is not sufficient that $\langle d, \frac{\partial^2 f}{\partial x^2}(\hat{x})d\rangle$ be positive for every nonzero d: it must be bounded away from zero for $\|d\| = 1$.)

Remark 4.10 Note that, in the proof of Theorem 4.6, it is not enough that $o_2(h)/\|h\|^2$ goes to zero along straight lines. (Thus twice Gateaux differentiable is not enough.)

Exercise 4.15 *Exhibit an example where f has a strict minimum at some $\hat{x}$, with $\frac{\partial^2 f}{\partial x^2}(\hat{x}) \succeq 0$ (as required by the 2nd order necessary condition), but such that there is no neighborhood of $\hat{x}$ where $\frac{\partial^2 f}{\partial x^2}(x)$ is everywhere positive semi-definite. (Try $x \in \mathbf{R}^2$; while examples in $\mathbf{R}$ do exist, they are contrived.) Simple examples exist where there is a neighborhood of $\hat{x}$ where the Hessian is nowhere positive semi-definite (except at $\hat{x}$). [Hint: First ignore the strictness requirement.]*

4.5 Minimization of convex functions

Convex functions have very nice properties in relation with optimization as shown by the exercise and theorem below.

Exercise 4.16 *The set of global minimizers of a convex function is convex (without any differentiability assumption). Also, every local minimizer is global. If such minimizer exists and f is strictly convex, then it is the unique global minimizer.*

Theorem 4.7 *Suppose $f : V \to \mathbf{R}$, is differentiable and convex. Then $\frac{\partial f}{\partial x}(x^*) = 0$ implies that x^* is a global minimizer for f. If f is strictly convex, x^* is the unique minimizer (and hence is strict).*

Proof. If f is convex, then, $\forall x \in V$

$$f(x) \geq f(x^*) \; + \; \frac{\partial f}{\partial x}(x^*)(x - x^*)$$

and, since $\frac{\partial f}{\partial x}(x^*) = 0$

$$f(x) \geq f(x^*) \quad \forall x \in V$$

and x^* is a global minimizer. If f is strictly convex, then $\forall x \neq x^*$

$$f(x) > f(x^*) \; + \; \frac{\partial f}{\partial x}(x^*)(x - x^*)$$

hence

$$f(x) > f(x^*) \quad \forall x \neq x^*$$

and x^* is strict and is the unique global minimizer.

■

Remark 4.11 Let $V = \mathbf{R}^n$ and suppose f is strongly convex. Then there is a global minimizer (why?). By the previous theorem it is unique and $\frac{\partial f}{\partial x}$ vanishes at no other point.

Exercise 4.17 *Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is strictly convex and has a minimizer $\hat{x}$, and suppose that the sequence $\{x_i\}$ is such that*

(i) every accumulation point $\tilde{x}$ satisfies $\frac{\partial f}{\partial x}(\tilde{x}) = \theta$;

(ii) $f(x_{i+1}) \leq f(x_i)$ for all i.

Then $x_i \to \hat{x}$.

4.6 Conjugate direction methods

(see [22])
We restrict the discussion to $V = \mathbf{R}^n$.

Steepest descent type methods can be very slow.

Exercise 4.18 *Let $f(x, y) = \frac{1}{2}(x^2 + ay^2)$ where $a > 0$. Consider the steepest descent algorithm with exact minimization. Given $(x_i, y_i) \in \mathbf{R}^2$, obtain formulas for x_{i+1} and y_{i+1}. Using these formulas, give a qualitative discussion of the performance of the algorithm for $a = 1$, a very large and a very small. Verify numerically using, e.g., MATLAB.*

If the objective function is quadratic and $x \in \mathbf{R}^2$, two function evaluations and two gradient evaluations are enough to identify the function exactly (why?). Thus there ought to be a way to reach the solution in two iterations. As most functions look quadratic locally, such method should give good results in the general case. Clearly, such a method must have memory (to remember previous function and gradient values). It turns out that a very simple idea gives answers to these questions. The idea is that of *conjugate direction*.

Definition 4.1 *Given a symmetric matrix Q, two vectors d_1 and d_2 are said to be Q-orthogonal, or conjugate with respect to Q if $d_1^T Q d_2 = 0$*

Fact. If Q is positive definite and $d_0, \ldots, d_k$ are Q-orthogonal and are all nonzero, then these vectors are linearly independent (and thus there can be no more than n such vectors).

Proof. See [22].

Theorem 4.8 *(Expanding Subspace Theorem)*
Consider the quadratic function $f(x) = \frac{1}{2}x^T Q x + b^T x$ with $Q \succ 0$ and let $h_0, h_1, \ldots, h_{n-1}$ be a sequence of Q-orthogonal vectors in $\mathbf{R}^n$. Then given any $x_0 \in \mathbf{R}^n$ if the sequence $\{x_k\}$ is generated according to $x_{k+1} = x_k + t_k h_k$, where t_k minimizes $f(x_k + t h_k)$, then x_k minimizes f over the linear variety (or affine set $x_0 +$ span $\{h_0, \ldots, h_{k-1}\}$.

Proof. Since f is convex we can write

$$f(x_k + h) \geq f(x_k) + \nabla f(x_k)^T h$$

Thus it is enough to show that, for all $h \in \text{ sp } \{h_0, \ldots, h_{k-1}\}$,

$$\nabla f(x_k)^T h = 0$$

i.e., since $h \in \text{ span } \{h_0, \ldots, h_{k-1}\}$ if and only if $-h \in \text{ span } \{h_0, \ldots, h_{k-1}\}$, that

$$\nabla f(x_k)^T h = 0 \quad \forall h \in \text{ span } \{h_0, \ldots, h_{k-1}\} \ ,$$

which holds if and only if

$$\nabla f(x_k)^T h_i = 0 \qquad i = 0, \ldots, k-1.$$

We prove this by induction on k, for $i = 0, 1, 2, \ldots$ First, for any i, and $k = i + 1$,

$$\nabla f(x_{i+1})^T h_i = \frac{\partial}{\partial t} f(x_i + t h_i) = 0$$

Suppose it holds for some $k > i$. Then

$$\begin{aligned} \nabla f(x_{k+1})^T h_i &= (Q x_{k+1} + b)^T h_i \\ &= (Q x_k + t_k Q h_k + b)^T h_i \\ &= \nabla f(x_k)^T h_i + t_k Q h_k^T h_i = 0 \end{aligned}$$

This first term vanishes due to the induction hypothesis, the second due to Q-orthogonality of the h_i's.

■

Corollary 4.1 *x_n minimizes $f(x) = \frac{1}{2} x^T Q x + b^T x$ over $\mathbf{R}^n$, i.e., the given iteration yields the minimizer for any quadratic function in no more than n iterations.*

Remark 4.12 The minimizing step size t_k is given by

$$t_k = -\frac{(Q x_k + b)^T h_k}{h_k^T Q h_k}$$

Conjugate gradient method

There are many ways to choose a set of conjugate directions. The conjugate gradient method selects each direction as the negative gradient added to a linear combination of the previous directions. It turns out that, in order to achieve Q-orthogonality, one must use

$$h_{k+1} = -\nabla f(x_{k+1}) + \beta_k h_k \tag{4.21}$$

i.e., only the preceding direction h_k has a nonzero coefficient. This is because, if (4.21) was used to construct the previous iterations, then $\nabla f(x_{k+1})$ is already conjugate to $h_0, \ldots, h_{k-1}$. Indeed, first notice that h_i is always a descent direction (unless $\nabla f(x_i) = 0$), so that $t_i \neq 0$. Then, for $i < k$

$$\begin{aligned}
\nabla f(x_{k+1})^T Q h_i &= \nabla f(x_{k+1})^T \left(\frac{1}{t_i} Q(x_{i+1} - x_i) \right) \\
&= \frac{1}{t_i} \nabla f(x_{k+1})^T (\nabla f(x_{i+1}) - \nabla f(x_i)) \\
&= \frac{1}{t_i} \nabla f(x_{k+1})^T (\beta_i h_i - h_{i+1} + h_i - \beta_{i-1} h_{i-1}) = 0
\end{aligned}$$

where we have used (4.21) for $k = i + 1$ and $k = i$, and the Expanding Subspace Theorem. The coefficient β_k is chosen so that $h_{k+1}^T Q h_k = 0$. One gets

$$\beta_k = \frac{\nabla f(x_{k+1})^T Q h_k}{h_k^T Q h_k}. \tag{4.22}$$

Non-quadratic objective functions

If x^* is a minimizer for f, $\nabla^2 f(x^*)$ is positive semi-definite. Generically (i.e., in 'most' cases), it will be strictly positive definite, since matrices are generically non-singular. Thus (since $\nabla f(x^*) = 0$)

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T Q (x - x^*) + o_2(x - x^*)$$

with $Q = \nabla^2 f(x^*) \succ 0$. Thus, close to x^*, f looks like a quadratic function with positive definite Hessian matrix and the conjugate gradient algorithm should work very well in such a neighborhood of the solution. However, (4.22) cannot be used for β_k since Q is unknown and since we do not want to compute the second derivative. Yet, for a quadratic function, it can be shown that

$$\beta_k = \frac{||\nabla f(x_{k+1}||^2}{||\nabla f(x_k)||^2} = \frac{(\nabla f(x_{k+1}) - \nabla f(x_k))^T \nabla f(x_{k+1})}{||\nabla f(x_k)||^2} \tag{4.23}$$

The first expression yields the Fletcher-Reeves conjugate gradient method. The second one gives the Polak-Ribière conjugate gradient method.

Exercise 4.19 *Prove (4.23).*

Algorithm 3 (conjugate gradient, Polak-Ribière version)
Data $x_o \in \mathbf{R}^n$
$i = 0$
$h_0 = -\nabla f(x_0)$
while $\nabla f(x_i) \neq 0$ do {
  $t_i \in \arg\min_t f(x_i + t h_i)$ (exact search)
  $x_{i+1} = x_i + t_i h_i$
  $h_{i+1} = -\nabla f(x_{i+1}) + \beta_i h_i$

$i = i + 1$
}
stop

The Polak-Ribière formula uses

$$\beta_i = \frac{(\nabla f(x_{i+1}) - \nabla f(x_i))^T \nabla f(x_{i+1})}{||\nabla f(x_i)||^2} \ . \tag{4.24}$$

It has the following advantage over the Fletcher-Reeves formula: away from a solution there is a possibility that the search direction obtained be not very good, yielding a small step $||x_{k+1} - x_k||$. If such a difficulty occurs, $||\nabla f(x_{k+1}) - \nabla f(x_k)||$ will be small as well and P-R will yield $-\nabla f(x_{k+1})$ as next direction, thus "resetting" the method.
By inspection, one verifies that h_i is a descent direction for f at x_i, i.e., $\nabla f(x_i)^T h_i < 0$ whenever $\nabla f(x_i) \neq 0$. The following stronger statement can be proved.

Fact. If f is twice continuously differentiable and strongly convex then $\exists \rho > 0$ such that

$$\langle \nabla f(x_i), h_i \rangle \leq -\rho ||\nabla f(x_i)|| \ ||h_i|| \quad \forall i$$

where $\{h_i\}$ and $\{x_i\}$ are as constructed by Algorithm 3 (in particular, this assumes an exact line search).

Exercise 4.20 *Show that*

$$||h_i|| \geq ||\nabla f(x_i)|| \quad \forall i.$$

As pointed out earlier one can show that the convergence theorem of Algorithm 2 still holds in the case of an exact search (since an exact search results in a larger decrease).

Exercise 4.21 *Show how the theorem just mentioned can be applied to Algorithm 3, by specifying $H(x)$.*

Thus, all accumulation points are stationary and, since f is strongly convex, $x_i \to \hat{x}$ the unique global minimizer of f. An implementable version of Algorithm 3 can be found in [26].
If f is not convex, it is advisable to periodically reset the search direction (i.e., set $h_i = -\nabla f(x_i)$ whenever i is a multiple of some number k; e.g., $k = n$ to take advantage of the quadratic termination property).

4.7 Rates of convergence

(see [25])
Note: Our definitions are simpler and not exactly equivalent to the ones in [25].

Quotient convergence rates

Definition 4.2 *Suppose $x_i \to x^*$. One says that $\{x_i\}$ converges to x^* with a Q-order of $p(\geq 1)$ and a corresponding Q-factor = γ if there exists i_0 such that, for all $i \geq i_0$*

$$||x_{i+1} - x^*|| \leq \gamma ||x_i - x^*||^p \ .$$

Obviously, for a given initial point x_0, and supposing $i_0 = 0$, the larger the Q-order, the faster $\{x_i\}$ converges and, for a given Q-order, the smaller the Q-factor, the faster $\{x_i\}$ converges. Also, according to our definition, if $\{x_i\}$ converges with Q-order $= p$ it also converges with Q-order $= p'$ for any $p' \leq p$, and if it converges with Q-order $= p$ and Q-factor $= \gamma$ it also converges with Q-order $= p$ and Q-factor $= \gamma'$ for any $\gamma' \geq \gamma$. Thus it may be more appropriate to say that $\{x_i\}$ converges with *at least* Q-order $= p$ and *at most* Q-factor $= \gamma$. But what is more striking is the fact that a larger Q-order will overcome any initial conditions, as shown in the following exercise. (If $p = 1$, a smaller Q-factor also overcomes any initial conditions.)

Exercise 4.22 *Let $x_i \to x^*$ be such that $||x_{i+1} - x^*|| = \gamma \ ||x_i - x^*||^p$ for all i, with $\gamma > 0$, $p \geq 1$, and suppose that $y_i \to y^*$ is such that*

$$||y_{i+1} - y^*|| \leq \delta ||y_i - y^*||^q \ \forall i \ \text{ for some } \delta > 0$$

Show that, if $q > p$, for any x_0, y_0, $x_0 \neq x^*$ $\exists N$ such that

$$||y_i - y^*|| < ||x_i - x^*|| \quad \forall i \geq N$$

Q-linear convergence

If $\gamma \in (0, 1)$ and $p = 1$, convergence is called *Q-linear*. This terminology comes from the fact that, in that case, we have (assuming $i_0 = 0$),

$$||x_i - x^*|| \leq \gamma^i ||x_0 - x^*|| \quad \forall i$$

and, hence

$$\log ||x_i - x^*|| \leq i \log \gamma + \log ||x_0 - x^*|| \quad \forall i$$

so that $-\log \|x_i - x^*\|$ (which, when positive, is roughly proportional to the number of exact figures in x_i) is linear (more precisely, affine) in i.

If $x_i \to x^*$ Q-linearly with Q-factor$= \gamma$, clearly $\frac{||x_{i+1} - x^*||}{||x_i - x^*||} \leq \gamma$ for i larger enough. This motivates the following definition.

Definition 4.3 *$x_i \to x^*$ Q-superlinearly if*

$$\frac{||x_{i+1} - x^*||}{||x_i - x^*||} \to 0 \text{ as } i \to \infty$$

Exercise 4.23 *Show, by a counterexample, that a sequence $\{x_i\}$ can converge Q-superlinearly, without converging with any Q-order $p > 1$. (However, Q-order larger than 1 implies Q-superlinear convergence.)*

Exercise 4.24 *Show that, if $x_i \to x^*$ Q-superlinearly, $||x_{i+1} - x_i||$ is a good estimate of the "current error" $||x_i - x^*||$ in the sense that*

$$\lim_{i\to\infty} \frac{||x_i - x^*||}{||x_{i+1} - x_i||} = 1$$

Definition 4.4 *If $x_i \to x^*$ with a Q-order $p = 2$, $\{x_i\}$ is said to converge q-quadratically.*

q-quadratic convergence is very fast: for large i, the number of exact figures is *doubled at each iteration.*

Exercise 4.25 *[25]. The Q-factor is norm-dependent (unlike the Q-order). Let*

$$x_i \begin{cases} (.9)^k & \binom{1}{0} \quad \text{for } k \text{ even} \\ (.9)^k & \binom{1/\sqrt{2}}{1/\sqrt{2}} \quad \text{for } k \text{ odd} \end{cases}$$

and consider the norms $\sqrt{x_1^2 + x_2^2}$ and $\max(|x_1|, |x_2|)$. Show that in both cases the sequence converges with Q-order= 1, but that only in the 1st case convergence is Q-linear ($\gamma > 1$ in the second case).

Root convergence rates

Definition 4.5 *One says that $x_i \to x^*$ with an R-order equal to $p \geq 1$ and an R-factor equal to $\gamma \in (0, 1)$ if there exists i_0 such that, for all $i \geq i_0$*

$$||x_{i_0+i} - x^*|| \leq \gamma^i \delta \quad \text{for } p = 1 \tag{4.25}$$
$$||x_{i_0+i} - x^*|| \leq \gamma^{p^i} \delta \quad \text{for } p > 1 \tag{4.26}$$

for some $\delta > 0$. (Note: by increasing δ, i_0 can always be set to 0.)

Equivalently there exists $\delta' > 0$ and $\gamma' \in (0, 1)$ such that, for all i,

$$||x_i - x^*|| \leq \gamma^i \delta' \quad \text{for } p = 1 \tag{4.27}$$
$$||x_i - x^*|| \leq \gamma'^{p^i} \delta' \qquad p > 1 \tag{4.28}$$

(take $\gamma' = \gamma^{1/p^{i_0}}$).

Again with the definition as given, it would be appropriate to used the phrases 'at least' and 'at most'.

Exercise 4.26 *Show that, if $x_i \to x^*$ with Q-order= $p \geq 1$ and Q-factor= $\gamma \in (0, 1)$ then $x_i \to x^*$ with R-order= p and R-factor= γ. Show that the converse is not true. Finally, exhibit a sequence $\{x_i\}$ converging*

(i) with Q-order $p \geq 1$ which does not converge with any R-order= $p' > p$, with any R-factor

(ii) with Q-order $p \geq 1$ and Q-factor γ, not converging with R-order p with any R-factor $\gamma' < \gamma$.

Definition 4.6 *If $p = 1$ and $\gamma \in (0, 1)$, convergence is called R-linear.*
If $x_i \to x^*$ R-linearly we see from (4.27) that

$$\limsup_{i\to\infty} ||x_i - x^*||^{\frac{1}{i}} \leq \lim_{i\to\infty} \gamma\delta^{\frac{1}{i}} = \gamma$$

Definition 4.7 *$\{x_i\}$ is said to converge to x^* R-superlinearly if*

$$\lim_{i\to\infty} ||x_i - x^*||^{\frac{1}{i}} = 0$$

Exercise 4.27 *Show that, if $x_i \to x^*$ Q-superlinearly, then $x_i \to x^*$ R-superlinearly.*

Exercise 4.28 *Show, by a counterexample, that a sequence $\{x_i\}$ can converge R-superlinearly without converging with any R-order $p > 1$*

Definition 4.8 *If $x_i \to x^*$ with R-order $p = 2$, $\{x_i\}$ is said to converge R-quadratically.*

Remark 4.13 R-order, as well as R-factor, are norm independent.

Rate of convergence of first order algorithms (see [22, 26])
Consider the following algorithm

Algorithm
Data $x_0 \in \mathbf{R}^n$
$i = 0$
while $\frac{\partial f}{\partial x}(x_i) \neq 0$ do {
  obtain h_i
  $t_i \in \arg\min\{f(x_i + th_i) \mid t \geq 0\}$
  $x_{i+1} = x_i + t_i h_i$
  $i = i + 1$
}
stop

Suppose that there exists $\rho > 0$ such that

$$\nabla f(x_i)^T h_i \leq -\rho||\nabla f(x_i)||\ ||h_i|| \quad \forall i, \tag{4.29}$$

where $\|\cdot\|$ is the Euclidean norm.
and that $h_i \neq 0$ whenever $\frac{\partial f}{\partial x}(x_i) \neq 0$ (condition $||h_i|| \geq c||\frac{\partial f}{\partial x}(x_i)||$ is obviously superfluous if we use an exact search, since, then, x_{i+1} depends only on the *direction* of h_i). We know that this implies that any accumulation point is stationary. We now suppose that, actually, $x_i \to x^*$. This will happen for sure if f is strongly convex. It will in fact happen in most cases when $\{x_i\}$ has some accumulation point.
Then, we can study the rate of convergence of $\{x_i\}$. We give the following theorem without proof (see [26]).

Theorem 4.9 *Suppose $x_i \to x^*$, for some x^*, where $\{x_i\}$ is constructed by the algorithm above, and assume that (4.29) holds. Suppose that f is twice continuously differentiable and that the second order sufficiency condition holds at x^* (as discussed above, this is a mild assumption). Let m, M, ϵ be positive numbers such that $\forall x \in B(x^*, \epsilon), \forall y \in \mathbf{R}^n$*

$$m||y||^2 \leq y^T \nabla^2 f(x) y \leq M||y||^2$$

[Such numbers always exist. Why?]. Then $x_i \to x^*$ R-linearly (at least) with an R-factor of (at most)

$$\gamma = \sqrt{1 - (\frac{\rho m}{M})^2} \tag{4.30}$$

If $\rho = 1$ (steepest descent), convergence is Q-linear with same Q-factor (with Euclidean norm).

Exercise 4.29 *Show that if $\rho < 1$ convergence is not necessarily Q-linear (with Euclidean norm).*

Remark 4.14

1. f as above could be called 'locally strongly convex', why?
2. without knowing anything else on h_i, the fastest convergence is achieved with $\rho = 1$ (steepest descent), which yields $\gamma_s = \sqrt{1 - (\frac{m}{M})^2}$. m and M are related to the smallest and largest eigenvalues of $\frac{\partial^2 f}{\partial x^2}(x^*)$ (why? how?). If $m \ll M$, convergence may be very slow again (see Figure **4.3**) (also see [5]).

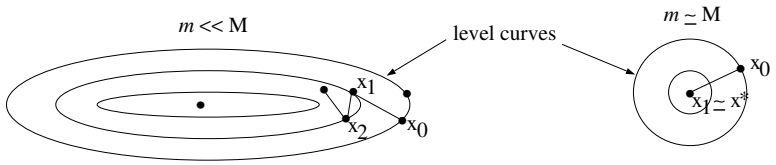


Figure **4.3**:

■

We will see below that if h_i is cleverly chosen (e.g., conjugate gradient method) the rate of convergence can be much faster.

If instead of using exact line search we use an Armijo line search, with parameters α, $\beta \in (0, 1)$, convergence is still R-linear and the R-factor is now given by

$$\gamma_a = \sqrt{1 - 4\beta\alpha(1-\alpha)(\frac{\rho m}{M})^2} \tag{4.31}$$

Remark 4.15

1. For $\alpha = 1/2$ and $\beta \simeq 1$, γ_a is close to γ_s, i.e., Armijo gradient converges as fast as steepest descent. Note, however that, the larger β is, the more computer time is going to be needed to perform the Armijo line search, since β^k will be very slowly decreasing when k increases.

2. Hence it appears that the rate of convergence does not by itself tell how fast the problem is going to be solved (even asymptotically). The time needed to complete one iteration has to be taken into account. In particular, for the rate of convergence to have any significance at all, the work per iteration must be bounded. See exercise below.

■

Exercise 4.30 *Suppose $x_i \to x^*$ with $x_i \neq x^*$ for all i. Show that for all p, γ there exists a sub-sequence $\{x_{i_k}\}$ such that $x_{i_k} \to x^*$ with Q-order $= p$ and Q-factor $= \gamma$.*

Exercise 4.31 *Consider two algorithms for the solution of some given problem. Algorithms SD and 2 construct sequences $\{x_k^1\}$ and $\{x_k^2\}$, respectively, both of which converge to x^*. Suppose $x_0^1 = x_0^2 \in B(x^*, 1)$ (open unit ball) and suppose*

$$\|x_{k+1}^i - x^*\| = \|x_k^i - x^*\|^{p_i}$$

with $p_1 > p_2 > 0$. Finally, suppose that, for both algorithms, the CPU time needed to generate x_{k+1} from x_k is bounded (as a function of k), as well as bounded away from 0. Show that there exists $\bar{\epsilon} > 0$ such that, for all $\epsilon \in (0, \bar{\epsilon})$, $\{x_k^1\}$ enters the ball $B(x^*, \epsilon)$ in less total CPU time than $\{x_k^2\}$ does. Thus, under bounded, and bounded away from zero, time per iteration, Q-orders can be meaningfully compared. (This is in contrast with the point made in Exercise 4.30.)

Conjugate direction methods

The rates of convergence (4.30) and (4.31) are conservative since they do not take into account the way the directions h_i are constructed, but only assume that they satisfy (4.29). The following modification of Algorithm 3 can be shown to converge superlinearly.

Algorithm 4 (conjugate gradient with periodic reinitialization)
Parameter $k > 0$ (integer)
Data $x_0 \in \mathbf{R}^n$
$i = 0$
$h_0 = -\nabla f(x_0)$
while $\nabla f(x_i) \neq 0$ do {
  pick $t_i \in \operatorname{argmin}\{f(x_i + t h_i) : t \geq 0\}$
  $x_{i+1} = x_i + t_i\, h_i$
  $h_{i+1} = -\nabla f(x_{i+1}) + \beta_i h_i$

with $\beta_i = \begin{cases} 0 & \text{if } i \text{ is a multiple of } k \\ \frac{(\nabla f(x_{i+1}) - \nabla f(x_i))^T \nabla f(x_{i+1})}{||\nabla f(x_i)||^2} & \text{otherwise} \end{cases}$

$i = i + 1$

}

stop

Exercise 4.32 *Show that any accumulation point of the sub-sequence $\{x_i\}_{i=\ell k}$ of the sequence $\{x_i\}$ constructed by Algorithm 4 is stationary.*

Exercise 4.33 *(Pacer step) Show that, if f is strongly convex in any bounded set, then the sequence $\{x_i\}$ constructed by Algorithm 4 converges to the minimum x^* and the rate of convergence is at least R-linear.*

Convergence is in fact n-step q-quadratic if $k = n$. If it is known that $x_i \to x^*$, clearly the strong convexity assumption can be replaced by strong convexity around x^*, i.e., 2nd order sufficiency condition.

Theorem 4.10 *Suppose that $k = n$ and suppose that the sequence $\{x_i\}$ constructed by Algorithm 4 is such that $x_i \to x^*$, at which point the 2nd order sufficiency condition is satisfied. Then $x_i \to x^*$ n-step q-quadratically, i.e. $\exists q, l_0$ such that*

$$||x_{i+n} - x^*|| \leq q||x_i - x^*||^2 \text{ for } i = ln,\ l \geq l_0.$$

This should be compared with the quadratic rate obtained below for Newton's method: Newton's method achieves the minimum of a quadratic convex function in 1 step (compared to n steps here).

Exercise 4.34 *Show that n-step q-quadratic convergence does not imply R-superlinear convergence. Show that the implication would hold under the further assumption that, for some $C > 0$,*

$$\|x_{k+1} - x^*\| \leq C\|x_k - x^*\| \quad \forall k.$$

4.8 Newton's method

Let us first consider Newton's method (Isaac Newton, English mathematician, 1642–1727) for solving a system of equations. Consider the system of equations

$$F(x) = \theta \tag{4.32}$$

with $F : V \to V$, V a normed vector space, and F is differentiable. The idea is to replace (4.32) by a linear approximation at the current estimate of the solution (see Figure **4.4**). Suppose x_i is the current estimate. Assuming Fréchet-differentiability, consider the equation

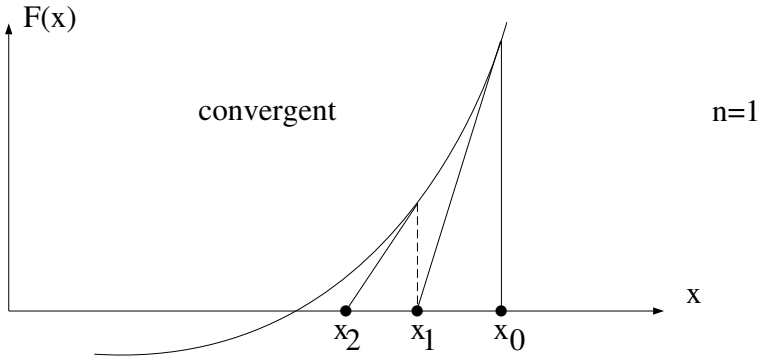


Figure **4.4**: Newton's iteration

$$\tilde{F}_i(x) := F(x_i) + \frac{\partial F}{\partial x}(x_i)(x - x_i) = \theta \tag{4.33}$$

and we denote by x_{i+1} the solution to (4.33) (assuming $\frac{\partial F}{\partial x}(x_i)$ is invertible).
Note that, from (4.33), x_{i+1} is given by

$$x_{i+1} = x_i - \frac{\partial F}{\partial x}(x_i)^{-1} F(x_i) \tag{4.34}$$

but it should not be computed that way but rather by solving the linear system (4.33) (much cheaper than computing an inverse). It turns out that, under suitable conditions, x_i converges very fast to a solution x^*.
Hence, in particular, if $F : \mathbf{R}^n \to \mathbf{R}^n$, Newton's method is invariant under scaling of the individual components of x.

Theorem 4.11 *Suppose V is a Banach space and $F : V \to V$ is twice continuously differentiable. Let x^* be such that $F(x^*) = \theta$ and suppose that $\frac{\partial F}{\partial x}(x^*)$ is invertible. Then there exists $\rho > 0$ and $C > 0$ such that, if $\|x_0 - x^*\| < \rho$, $x_i \to x^*$ and $\|x_{i+1} - x^*\| \leq C\|x_i - x^*\|^2$ for all i.*

Proof. (We use the following instance of the Inverse Function Theorem: Let $f : V \to V$ be continuously differentiable, V a Banach space and let $x_0 \in V$ be such that $\frac{\partial F}{\partial x}(x_0)$ is invertible. Then there exists $\rho > 0$ such that $\frac{\partial F}{\partial x}(x)^{-1}$ exists and is continuous on $B(x_0, \rho)$. See, e.g. [23, Section 9.7, Problem 1].) Let x_i be such that $\frac{\partial F}{\partial x}(x_i)$ is invertible. Then, from (4.34),

$$\|x_{i+1} - x^*\| = \left\| x_i - x^* - \frac{\partial F}{\partial x}(x_i)^{-1} F(x_i) \right\| \leq \left\| \frac{\partial F}{\partial x}(x_i)^{-1} \right\| \cdot \left\| F(x_i) + \frac{\partial F}{\partial x}(x_i)(x^* - x_i) \right\|,$$

where the induced norm is used for the (inverse) linear map. Now, there exist (i) $\rho_1 > 0$ and $\beta_1 > 0$ such that

$$\left\| \frac{\partial F}{\partial x}(x)^{-1} \right\| \leq \beta_1 \quad \forall x \in B(x^*, \rho_1)$$

and (ii) $\rho_2 > 0$ and $\beta_2 > 0$ such that

$$\left\| F(x) + \frac{\partial F}{\partial x}(x)(x^* - x) \right\| \leq \beta_2 \|x - x^*\|^2 \quad \forall x \in B(x^*, \rho_2)$$

(Existence of β_1 and ρ_1 follow from the Inverse Function Theorem; existence of β_2, for ρ_2 small enough follows from continuity of the second derivative of F and from Corollary B.2.) Further, let $\rho > 0$, with $\rho < \min\{\rho_1, \rho_2\}$ be such that

$$\beta_1 \beta_2 \rho < 1.$$

It follows that, whenever $x_i \in B(x^*, \rho)$,

$$\|x_{i+1} - x^*\| \leq \beta_1 \beta_2 \|x_i - x^*\|^2$$

$$\leq \beta_1 \beta_2 \rho \|x_i - x^*\| \leq \|x_i - x^*\|$$

so that, if $x_0 \in B(x^*, \rho)$, then $x_i \in B(x^*, \rho)$ for all i, and thus

$$\|x_{i+1} - x^*\| \leq \beta_1 \beta_2 \|x_i - x^*\|^2 \quad \forall i \tag{4.35}$$

and

$$\|x_{i+1} - x^*\| \leq \beta_1 \beta_2 \rho \|x_i - x^*\| \leq (\beta_1 \beta_2 \rho)^i \|x_0 - x^*\| \quad \forall i \ .$$

Since $\beta_1 \beta_2 \rho < 1$, it follows that $x_i \to x^*$ as $i \to \infty$. In view of (4.35), convergence is q-quadratic.

■

We now turn back to our optimization problem

$$\min\{f(x) | x \in V\}.$$

Since we are looking for points x^* such that $\frac{\partial f}{\partial x}(x^*) = \theta$, we want to solve a system of nonlinear equations and the theory just presented applies, provided f is twice differentiable. The Newton iteration now amounts to solving the linear system

$$\frac{\partial f}{\partial x}(x_i) + \frac{\partial^2 f}{\partial x^2}(x_i)(x - x_i) = 0$$

i.e., finding a stationary point for the *quadratic approximation to f*

$$f(x_i) + \frac{\partial f}{\partial x}(x_i)(x - x_i) + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}(x_i)(x - x_i) \right)(x - x_i).$$

In particular, if f is quadratic (i.e., $\frac{\partial f}{\partial x}$ is linear), one Newton iteration yields such point (which may or may not be the minimizer) exactly. Quadratic convergence is achieved, e.g., if f is three times continuously differentiable and the 2nd order sufficiency condition holds at x^*, so that $\frac{\partial^2 f}{\partial x^2}(x^*)$ is non-singular. We will now show that we can obtain stronger convergence

properties when the Newton iteration is applied to a minimization problem (than when it is applied to a general root-finding problem). In particular, we want to achieve global convergence (convergence for any initial guess x_0). We can hope to achieve this because the optimization problem has more structure than the general equation solving problem, as shown in the next exercise.

Exercise 4.35 *Exhibit a function $F : \mathbf{R}^2 \to \mathbf{R}^2$ which is not the gradient of any C^2 function $f : \mathbf{R}^2 \to \mathbf{R}$.*

Exercise 4.36 *Newton's method is invariant under affine transformations of the domain $z \mapsto Lz + b$, where $L : V \to V$ is an invertible, bounded linear map. [Note that, if L is a bounded linear map and F is continuously differentiable, then $G : z \mapsto F(Lz + b)$ also is continuously differentiable. Why?] Express this statement in a mathematically precise form, and prove it. Next, turning to the application to minimization, show (e.g., by exhibiting an example) that steepest descent (e.g., with exact line search) is not invariant under such transformations, a significant shortcoming in comparison to Newton's method, since selecting a good scaling is then the user's responsibility.*

Remark 4.16 Global strong convexity of f (over all of $\mathbf{R}^n$) does *not* imply global convergence of Newton's method to minimize f. (f = integral of the function plotted on Figure **4.5**. gives such an example).

Exercise 4.37 *Come up with a globally strongly convex function $f : \mathbf{R} \to \mathbf{R}$ and an initial point x_0 such that Newton's iteration started at x_0 does not converge to the unique minimizer of f. Experiment with Matlab, trying various initial points, and comment (in particular, you should observe local quadratic convergence).*

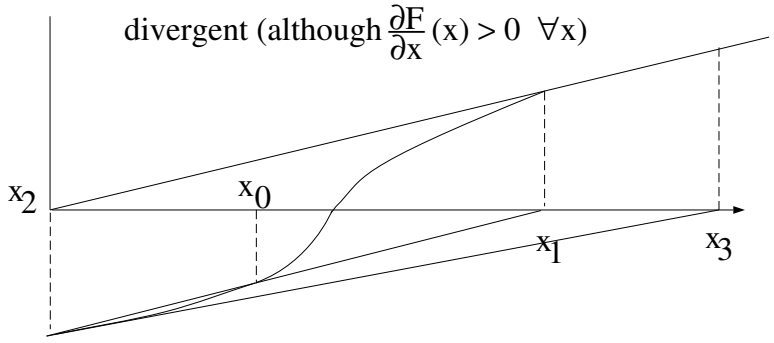


Figure **4.5**: Example of non-convergence of Newton's method for minimizing a strongly convex function f, equivalently for locating a point at which its derivative F (pictured, with derivative F' bounded away from zero) vanishes.

■

Global convergence will be obtained by making use of a suitable step-size rule, e.g., the Armijo rule.
Now suppose $V = \mathbf{R}^n$.[1] Note that the Newton direction $h^{\mathrm{N}}(x)$ at some x is given by

$$h^{\mathrm{N}}(x) = -\nabla^2 f(x)^{-1} \nabla f(x).$$

When $\nabla^2 f(x) \succ 0$, $h^{\mathrm{N}}(x)$ is minus the gradient associated with the (local) inner product

$$\langle u, v \rangle = u^T \nabla^2 f(x) v.$$

Hence, in such case, Newton's method is a special case of steepest descent, albeit with a norm that changes at each iteration.

Armijo–Newton

The idea is the following: replace the Newton iterate $x_{i+1} = x_i - \nabla^2 f(x_i)^{-1} \nabla f(x_i)$ by a suitable step in the Newton direction, i.e.,

$$x_{i+1} = x_i - t_i \nabla^2 f(x_i)^{-1} \nabla f(x_i) \tag{4.36}$$

with t_i suitably chosen. By controlling the length of each step, one hopes to prevent instability. Formula (4.36) fits into the framework of Algorithm 2, with $h(x) := h^{\mathrm{N}}(x)$. Following Algorithm 2, we define t_i in (4.36) via the Armijo step-size rule.

Note that if $\nabla^2 f(x)$ is positive definite, $h(x)$ is a descent direction. Global convergence with an Armijo step, however, requires more (see (4.12)–(4.13)).

Exercise 4.38 *Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is in C^2 and suppose $\nabla^2 f(x) \succ 0$ for all $x \in \mathbf{R}^n$. Then*

(i) For every $\hat{x} \in L$ there exists $C > 0$ and $\rho > 0$ such that, for all x close enough to $\hat{x}$,

$$||h^{\mathrm{N}}(x)|| \geq C ||\nabla f(x)||$$

$$h^{\mathrm{N}}(x)^T \nabla f(x) \ \leq -\rho ||h^{\mathrm{N}}(x)|| \ ||\nabla f(x)||$$

so that the assumptions of Theorem 4.4 hold.

(ii) The sequence $\{x_k\}$ generated by the Armijo-Newton algorithm converges to the unique global minimizer.

Armijo-Newton yields global convergence. However, the q-quadratic rate may be lost since nothing insures that the step-size t_i will be equal to 1, even very close to a solution x^*. However, as shown in the next theorem, this will be the case if the α parameter is less than $1/2$.

Theorem 4.12 *Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is in C^3 and let an initial point x_0 be given. Consider Algorithm 2 with $\alpha \in (0, 1/2)$ and $h_i := h^{\mathrm{N}}(x_i)$, and suppose that $\hat{x}$ is an accumulation point of $\{x_i\}$ with ($\nabla f(\hat{x}) = 0$ and) $\nabla^2 f(\hat{x}) \succ 0$ (second order sufficiency condition). Then there exists i_0 such that $t_i = 1$ for all $i \geq i_0$, and $x_i \to \hat{x}$ q-quadratically.*

[1] The same ideas apply for any Hilbert space, but some additional notation is needed.

Exercise 4.39 *Prove Theorem 4.12.*

We now have a globally convergent algorithm, with (locally) a q-quadratic rate (if f is thrice continuously differentiable). However, we needed an assumption of strong convexity on f on bounded sets (for the iteration to be well-defined).

Suppose now that f is not strongly convex (perhaps not even convex). We noticed earlier that, around a local minimizer, the strong convexity assumption is likely to hold. Hence, we need to steer the iterate x_i towards the neighborhood of such a local solution and then use Armijo-Newton for local convergence. Such a scheme is called a *stabilization* scheme. Armijo-Newton stabilized by a gradient method would look like the following.

Algorithm
Parameters. $\alpha \in (0, 1/2)$, $\beta \in (0, 1)$
Data. $x_0 \in \mathbf{R}^n$
$i = 0$
while $\nabla f(x_i) \neq 0$ do {
set

$$H_i := \nabla^2 f(x_i) + \delta_i I,$$

where $\delta_i \geq 0$ is appropriately selected;
set

$$h_i := -H_i^{-1} \nabla f(x_i);$$

compute Armijo step size t_i;
set $x_{i+1} := x_i + t_i h_i$
}
stop.

In defining H_i, the non-negative scalar δ_i should be selected large enough (so h_i is close enough to the being along the negative gradient direction) that global convergence ensures, while ideally being set to zero from some iteration onwards, so that local the Armijo-Newton iteration takes over, yielding q-quadratic convergence. The choice δ-selection rule is critical, to avoid the iteration entering into an infinite loop by repeated switching between $\delta_i = 0$ and a "large" value, possibly even hindering global convergence.

4.9 Variable metric methods

Two major drawbacks of Newton's method are as follows:

1. for $h_i = -\nabla^2 f(x_i)^{-1} \nabla f(x_i)$ to be a descent direction for f at x_i, the Hessian $\nabla^2 f(x_i)$ must be positive definite. (For this reason, we had to stabilize it with a gradient method.)

2. second derivatives must be computed ($\frac{n(n+1)}{2}$ of them!) and a linear system of equations has to be solved.

The variable metric methods avoid these 2 drawbacks. The price paid is that the rate of convergence is not quadratic anymore, but merely superlinear (as in the conjugate gradient method). The idea is to construct increasingly better estimates S_i of the inverse of the Hessian, making sure that those estimates remain positive definite. Since the Hessian $\nabla^2 f$ is generally positive definite around a local solution, the latter requirement presents no contradiction.

Algorithm
Data. $x_0 \in \mathbf{R}^n$, $S_0 \in \mathbf{R}^{n\times n}$, positive definite (e.g., $S_0 = I$)
while $\nabla f(x_i) \neq 0$ do {
$h_i = -S_i \nabla f(x_i)$
pick $t_i \in \arg\min_t \{f(x_i + t h_i) | t \geq 0\}$
$x_{i+1} = x_i + t_i h_i$
compute S_{i+1}, positive definite, using some update formula
}
stop

■

If the S_i are bounded and uniformly positive definite, all accumulation points of the sequence constructed by this algorithm are stationary (why?).

Exercise 4.40 *Consider the above algorithm with the step-size rule $t_i = 1$ for all i instead of an exact search and assume f is strongly convex. Show that, if $\|S_i - \nabla^2 f(x_i)^{-1}\| \to 0$, convergence is Q-superlinear (locally). [Follow the argument used for Newton's method. In fact, if $\|S_i - \nabla^2 f(x_i)^{-1}\| \to 0$ fast enough, convergence may be quadratic.]*

A number of possible update formulas have been suggested. The most popular one, due independently to Broyden, Fletcher, Goldfarb and Shanno (BFGS) is given by

$$S_{i+1} = S_i + \frac{\gamma_i \gamma_i^T}{\delta_i^T \gamma_i} - \frac{S_i \delta_i \delta_i^T S_i}{\delta_i^T S_i \delta_i}$$

where

$$\begin{aligned} \delta_i &\triangleq x_{i+1} - x_i \\ \gamma_i &\triangleq \nabla f(x_{i+1}) - \nabla f(x_i) \end{aligned}$$

Convergence

If f is three times continuously differentiable and *strongly convex*, the sequence $\{x_i\}$ generated by the BFGS algorithm converges superlinearly to the solution x^*.

Remark 4.17

1. BFGS has been observed to perform remarkably well on non convex cost functions

2. Variable metric methods, much like conjugate gradient methods, use past information in order to improve the rate of convergence. Variable metric methods require more storage than conjugate gradient methods (an $n \times n$ matrix) but generally exhibit much better convergence properties.

Exercise 4.41 *Justify the name "variable metric" as follows. Given a symmetric positive definite matrix M, define*

$$||x||_M = (x^T M x)^{1/2} \qquad (= \text{ new metric})$$

and show that

$$\inf_h \{f(x + \alpha h) : ||h||_{S_i^{-1}} = 1\} \qquad (P_\alpha)$$

is achieved for h close to $\hat{h} := \frac{-S_i \nabla f(x)}{||S_i \nabla f(x)||_{S_i^{-1}}}$ for α small. Specifically, show that, given any $\tilde{h} \neq \hat{h}$, with $\|\tilde{h}\|_{S_i^{-1}} = 1$, there exists $\bar{\alpha} > 0$ such that

$$f(x + \alpha\hat{h}) < f(x + \alpha\tilde{h}) \quad \forall \alpha \in (0, \bar{\alpha}].$$

In particular, if $\frac{\partial^2 f}{\partial x}(x) > 0$, the Newton direction is the direction of steepest descent in the corresponding norm.

Chapter 5

Constrained Optimization

Consider the problem

$$\min\{f(x) \ : \ x \in \Omega\} \qquad (P)$$

where $f : V \to \mathbf{R}$ is continuously Fréchet differentiable and Ω is a subset of V, a normed vector space. Although the case of interest here is when Ω is not open (see Exercise 4.1), no such assumption is made.

Remark 5.1 Note that this formulation is very general. It allows in particular for binary variables (e.g., $\Omega = \{x : x(x-1) = 0\}$) or integer variables (e.g., $\Omega = \{x : \sin(x\pi) = 0\}$). These two examples are in fact instances of smooth equality-constrained optimization: see section 5.2 below.

5.1 Abstract Constraint Set

In the unconstrained case, we obtained a first order condition of optimality we approximated f in the neighborhood of $\hat{x}$ with its first order expansion at $\hat{x}$. Specifically we had, for all small enough h and some little-o function,

$$0 \leq f(\hat{x}+h) - f(\hat{x}) = \frac{\partial f}{\partial x}(\hat{x}) + o(h),$$

and since the direction (and orientation) of h was unrestricted, we could conclude that $\frac{\partial f}{\partial x}(\hat{x}) = \theta$.
In the case of problem (P) however, the inequality $f(\hat{x}) \leq f(\hat{x}+h)$ is not known to hold for every small h. It is known to hold when $\hat{x}+h \in \Omega$ though. Since Ω is potentially very complicated to specify, some approximation of it near $\hat{x}$ is needed. A simple yet very powerful such class of approximations is the class of conic approximations. Perhaps the most obvious one is the "radial cone".

Definition 5.1 *The radial cone* $\mathrm{RC}(x,\Omega)$ *to* Ω *at* x *is defined by*

$$\mathrm{RC}(x,\Omega) = \{h \in V \ : \ \exists \bar{t} > 0 \text{ s.t. } x + th \in \Omega \quad \forall t \in (0,\bar{t}]\}.$$

The following result is readily proved (as an extension of the proof of Theorem 4.1 of unconstrained minimization, or with a simplification of the proof of the next theorem).

Proposition 5.1 *Suppose x^* is a local minimizer for (P). Then*

$$\frac{\partial f}{\partial x}(x^*)h \geq 0 \quad \forall h \in \mathrm{cl}(\mathrm{coRC}(x^*, \Omega))$$

While this result is useful, it has a major drawback: When nonlinear equality constraints are present, $\mathrm{RC}(x^*, \Omega)$ is usually empty, so that theorem is vacuous. This motivates the introduction of the "tangent cone". But first of all, let us define what is meant by "cone".

Definition 5.2 *A set $C \subseteq V$ is a cone if $x \in C$ implies $\alpha x \in C$ for all $\alpha > 0$.*

In other words given $x \in C$, the entire ray from the origin through x belongs to C (but possibly not the origin itself). A cone may be convex or not.

Example 5.1 (Figure **5.1**)

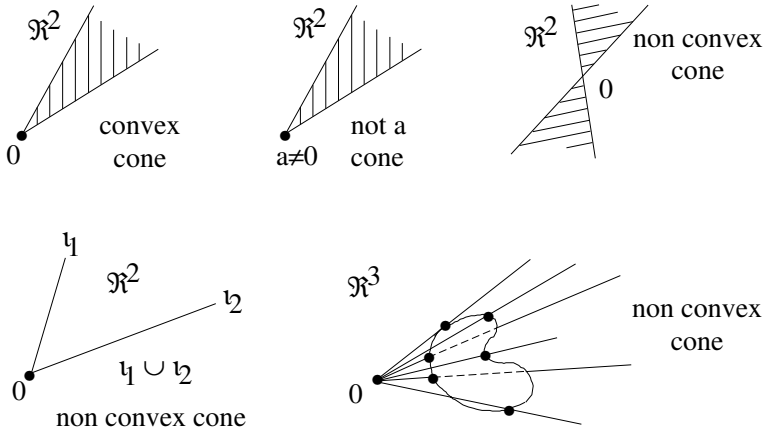


Figure **5.1**:

A cone may or may not include the origin of the space. If it does, it is pointed. A cone is salient if it does not include a nontrivial subspace. (A subspace S is nontrivial subspace if $S \neq \{\theta\}$.)

Exercise 5.1 *Show that a cone C is convex if and only if $\alpha x + \beta y \in C$ for all $\alpha, \beta \geq 0$, $\alpha + \beta > 0$, $x, y \in C$.*

Exercise 5.2 *Show that a cone C is convex if and only if $\frac{1}{2}(x + y) \in C$ for all $x, y \in C$. Show by exhibiting a counterexample that this statement is incorrect if "cone" is replaced with "set".*

Exercise 5.3 *Prove that radial cones are cones.*

To address the difficulty just encountered with radial cones, let us focus for a moment on the case $\Omega = \{(x, y) \ : \ y = f(x)\}$ (e.g., with x and y scalars), and with f nonlinear. As just observed, the radial cone is empty in this situation. From freshman calculus we do know how to approximate such set around $\hat{x}$ though: just replace f with its first-order Taylor expansion, yielding $\{(x, y) \ : \ y = f(\hat{x}) + \frac{\partial f}{\partial x}(\hat{x})(x - \hat{x})\}$, where we have replaced the curve with its "tangent" at $\hat{x}$. Hence we have replaced the radial-cone specification (that a ray belongs to the cone if short displacements along that ray yield points within Ω) with a requirement that short displacements along a candidate tangent direction to our curve yield points "little-o-close" to the curve. Merging the two ideas yields the "tangent cone", a super-set of the radial cone. In the case of Figure **5.2** below, the tangent cone is the closure of the radial cone: it includes the radial cone as well as the two "boundary line" which are the tangents to the "boundary curves" of Ω. (It is of course not always the case that the tangent cone is the closure of the radial cone: Just think of the case that motivated this discussion, where the radial cone was empty.)

Definition 5.3 *Given an normed space V, $\Omega \subseteq V$, and $x \in \Omega$, the tangent cone* $\mathrm{TC}(x, \Omega)$ *to* Ω *at x is defined by*

$$\mathrm{TC}(x, \Omega) = \{h \in V \ : \ \exists o(\cdot), \bar{t} > 0 \text{ s.t. } x + th + o(t) \in \Omega \ \ \forall t \in (0, \bar{t}], \ \frac{\|o(t)\|}{t} \to 0 \text{ as } t \to 0, \ t > 0\} \ .$$

The next exercise shows that this definition meets the geometric intuition discussed above.

Exercise 5.4 *Let*

$$\Omega = \mathrm{epi} f := \{(y, z) \in \mathbf{R}^{n+1} : z \geq f(y)\},$$

with $f : \mathbf{R}^n \to \mathbf{R}$ continuously differentiable. Verify that, given $(\hat{y}, \hat{z})$ with $\hat{z} = f(\hat{y})$,

$$\mathrm{TC}((\hat{y}, \hat{z}), \Omega) = \{(v, w) : w \geq \frac{\partial f}{\partial y}(\hat{y})v\},$$

i.e., the tangent cone is the half space that lies "above" the subspace parallel to the tangent hyperplane—when $n = 1$, the line through the origin parallel to the tangent line—to f at $(\hat{y}, \hat{z})$.

Exercise 5.5 *Show that tangent cones are cones.*

Exercise 5.6 *Given an normed space V, $\Omega \subseteq V$, and $x \in \Omega$, show that*

$$\mathrm{TC}(x, \Omega) = \{h \in V \ : \ \exists o(\cdot) \text{ s.t. } x + th + o(t) \in \Omega \ \ \forall t > 0, \ \frac{o(t)}{t} \to 0 \text{ as } t \to 0, t > 0\} \ .$$

Exercise 5.7 *Let $\varphi : \mathbf{R} \to \Omega$ be continuously differentiable. Then, for all t, $\varphi'(t)$ and $-\varphi'(t)$ both belong to* $\mathrm{TC}(\varphi(t), \Omega)$.

Some authors require that $o(\cdot)$ be continuous. This may yield a smaller set (and hence a weaker version of the necessary condition of optimality), as shown in the next exercise.

Exercise 5.8 *Let* $\Omega : \{0, 1, \frac{1}{2}, \frac{1}{3}, \ldots\}$. *Show that* $\mathrm{TC}(0, \Omega) = \{h \in \mathbf{R} : h \geq 0\}$ *but there exists no* continuous *little-o function such that* $th + o(t) \in \Omega$ *for all* $t > 0$ *small enough.*

TC need not be convex, even if Ω is defined by smooth equality and inequality constraints (although $\mathcal{N}\left(\frac{\partial g}{\partial x}(x)\right)$ and $S(x)$, introduced below, are convex). Example: $\Omega = \{x \in \mathbf{R}^2 : x_1 x_2 \leq 0\}$. Also note that $x^* + \mathrm{TC}(x^*, \Omega)$ is an approximation to Ω around x^*.

Theorem 5.1 *Suppose* x^* *is a local minimizer for (P). Then*

$$\frac{\partial f}{\partial x}(x^*)h \geq 0 \quad \forall h \in \mathrm{cl}(\mathrm{coTC}(x^*, \Omega)).$$

Proof. Let $h \in \mathrm{TC}(x^*, \Omega)$. Then

$$\exists o(\cdot) \ \ni \ x^* + th + o(t) \in \Omega \quad \forall t \geq 0 \text{ and } \lim_{t \to 0} \frac{o(t)}{t} = 0. \tag{5.1}$$

By definition of a local minimizer, $\exists\ \bar{t} > 0 \ \ni$

$$f(x^* + th + o(t)) \geq f(x^*) \quad \forall t \in [0, \bar{t}] \tag{5.2}$$

But, using the definition of derivative

$$\begin{aligned} f(x^* \ &+ \ th + o(t)) - f(x^*) \\ &= \ \frac{\partial f}{\partial x}(x^*)(th + o(t)) + \hat{o}(th + o(t)) \text{ with } \frac{\hat{o}(th)}{t} \to 0 \text{ as } t \to 0 \\ &= \ t\frac{\partial f}{\partial x}(x^*)h + \tilde{o}(t) \end{aligned}$$

with $\tilde{o}(t) = \frac{\partial f}{\partial x}(x^*)o(t) + \hat{o}(th + o(t))$.
Hence

$$f(x^* + th + o(t)) - f(x^*) = t\left(\frac{\partial f}{\partial x}(x^*)h + \frac{\tilde{o}(t)}{t}\right) \geq 0 \ \forall t \in (0, \bar{t}]$$

so that

$$\frac{\partial f}{\partial x}(x^*)h + \frac{\tilde{o}(t)}{t} \geq 0 \quad \forall t \in (0, \bar{t}]$$

It is readily verified that $\frac{\tilde{o}(t)}{t} \to 0$ as $t \to 0$. Thus, letting $t \searrow 0$, one obtains

$$\frac{\partial f}{\partial x}(x^*)h \geq 0.$$

The remainder of the proof is left as an exercise.

■

Dual cone. In a Hilbert space X, if $x^* = \theta$ solves (P), then $\langle \mathrm{grad} f(x^*), h\rangle \geq 0$ for all h in the tangent cone, i.e., $\mathrm{grad} f(x^*)$ belongs to the dual cone to the tangent cone.

Definition 5.4 *Given a cone* K, *its* dual cone K^* *is given by*

$$K^* := \{y \in X : \langle y, x\rangle \geq 0 \ \forall x \in K\}.$$

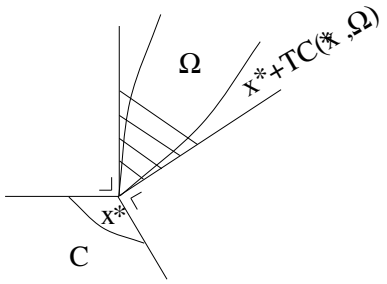


Figure **5.2**: In the situation pictured (where $x^* = \theta$ for clarity), the tangent cone to Ω at x^* is the closure of the radial cone to Ω at x^*.

In Figure **5.2**, when $x^* = \theta$, cone C is the negative of the dual cone, which is the polar cone, of $\mathrm{TC}(x, \Omega)$.

The necessary condition of optimality we just obtained is not easy to use. By considering specific types of constraint sets Ω, we hope to obtain a simple expression for TC, thus a convenient condition of optimality. We first consider the equality constrained problem.

5.2 Equality Constraints - First Order Conditions

For simplicity, we now restrict ourselves to $V = \mathbf{R}^n$. Consider the problem

$$\min\{f(x) \ : \ g(x) = \theta\} \tag{5.3}$$

i.e.

$$\min\{f(x) \ : \ x \in \underbrace{\{x \ : \ g(x) = \theta\}}_{\Omega}\}$$

where $f \ : \ \mathbf{R}^n \to \mathbf{R}$ and $g \ : \ \mathbf{R}^n \to \mathbf{R}^\ell$ are both continuously Fréchet differentiable. Let x^* be such that $g(x^*) = \theta$. Let $h \in \mathrm{TC}(x^*, \Omega)$, i.e., suppose there exists a $o(\cdot)$ function such that

$$g(x^* + th + o(t)) = \theta \ \forall t \geq 0.$$

Since $g(x^*) = \theta$, we readily conclude that $\frac{\partial g}{\partial x}(x^*)h = \theta$. Hence $\mathrm{TC}(x^*, \Omega) \subseteq \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$, i.e., if $h \in \mathrm{TC}(x^*, \Omega)$ then $\nabla g_i(x^*)^T h = 0$ for every i: tangent vectors are orthogonal to the gradients of the components of g.

Exercise 5.9 *Prove that, furthermore,*

$$\mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) \subseteq \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right).$$

Definition 5.5 *The pair (x^*, g) is said to be non-degenerate if*

$$\mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right). \tag{5.4}$$

Below, we investigate conditions that guarantee non-degeneracy. Note that, indeed, non-degeneracy may fail to hold. In particular, unlike $\mathrm{TC}(x^*, \Omega)$ and the closure of its convex hull, $\mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$ depends not only on Ω, but also on the way Ω is formulated. For example, with $x \in \mathbf{R}$, $\{x : x = 0\} \equiv \{x : x^2 = 0\}$ but $\mathcal{N}(g'(0))$ is $\{0\}$ for the left-hand side and $\mathbf{R}$ for the right-hand side.
We now derive a first order optimality conditions for the case when (5.4) does hold.

Theorem 5.2 *Suppose x^* is a local minimizer for (5.3) and suppose that (x^*, g) is non-degenerate, i.e., $\mathrm{cl}\,\mathrm{coTC}(x^*, \Omega) = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$. Then*

$$\frac{\partial f}{\partial x}(x^*)h = 0 \quad \forall h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$$

Proof. From Theorem 5.1,

$$\frac{\partial f}{\partial x}(x^*)h \geq 0 \quad \forall h \in \mathrm{cl}(\mathrm{coTC}(x^*, \Omega) \tag{5.5}$$

i.e.,

$$\frac{\partial f}{\partial x}(x^*)h \geq 0 \quad \forall h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right). \tag{5.6}$$

Now obviously $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$ implies $-h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$. Hence

$$\frac{\partial f}{\partial x}(x^*)h = 0 \quad \forall h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) \tag{5.7}$$

■

Remark 5.2 Theorem 5.2 can also be proved directly, without making use of Theorem 5.1, by invoking the "Inverse Function Theorem". See, e.g., [23, Section 9.3, Lemma 1].

Remark 5.3 Our non-degeneracy condition is in fact a type of "constraint qualification"; more on this later. In fact, as mentioned in Remark 5.22 it is the least restrictive constraint qualification for our equality-constrained problem.

Corollary 5.1 *(Lagrange multipliers. Joseph-Louis Lagrange, Italian-born mathematician, 1736–1813.) Under the same assumptions, $\exists \lambda^* \in \mathbf{R}^m$ such that*

$$\nabla f(x^*) + \frac{\partial g}{\partial x}(x^*)^T \lambda^* = \theta \tag{5.8}$$

i.e.

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda^{*,j} \nabla g^j(x^*) = \theta \tag{5.9}$$

Proof. From the theorem above

$$\nabla f(x^*) \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)^{\perp} = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)^{\perp} = \mathcal{R}\left(\frac{\partial g}{\partial x}(x^*)^T\right),$$

i.e.,

$$\nabla f(x^*) = \frac{\partial g}{\partial x}(x^*)^T \tilde{\lambda} \tag{5.10}$$

for some $\tilde{\lambda}$. The proof is complete if one sets $\lambda^* = -\tilde{\lambda}$.

■

Next, we seek intuition concerning (5.4) by means of examples.

Example 5.2 (Figure **5.3**)

(1) $m = 1$ $n = 2$. Claim: $\nabla g(x^*) \perp \mathrm{TC}(x^*, \Omega)$ (why ?). Thus, from picture, $\mathrm{TC}(x^*, \Omega) = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$ (assuming $\nabla g(x^*) \neq \theta$), so (5.4) holds.

(2) $m = 2$ $n = 3$. Again, $\mathrm{TC}(x^*, \Omega) = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$, so again (5.4) holds. (We assume here that $\nabla g_1(x^*) \neq \theta \neq \nabla g_2(x^*)$ and it is clear from the picture that these two gradients are not parallel to each other.)

(3) $m = 2$ $n = 3$. Here TC is a line but $\mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$ is a plane (assuming $\nabla g_1(x^*) \neq \theta \neq \nabla g_2(x^*)$). Thus $\mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) \neq \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$. Note that x^* could be a local minimizer with $\nabla f(x^*)$ as depicted, although $\nabla f(x^*)^T h < 0$ for some $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$.

Now let $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$, i.e., suppose $\frac{\partial g}{\partial x}(x^*)h = \theta$. We want to find $o(\cdot) : \mathbf{R} \to \mathbf{R}^n$ s.t.

$$s(t) \stackrel{\Delta}{=} x^* + th + o(t) \in \Omega \quad \forall t \geq 0$$

Geometric intuition (see Figure **5.4**) suggests that we could try to find $o(t)$ orthogonal to h. (This does not work for the third example in Figure **5.3**.). Since $h \in \mathcal{N}(\frac{\partial g}{\partial x}(x^*))$, we try with $o(t)$ in the range of $\frac{\partial g}{\partial x}(x^*)^T$, i.e.,

$$s(t) = x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi(t)$$

for some $\varphi(t) \in \mathbf{R}^\ell$. We want to find $\varphi(t)$ such that $s(t) \in \Omega \quad \forall t$, i.e., such that $g(x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi(t)) = \theta \ \forall t$, and to see under what condition φ exists and is a "little o" function. We will make use of the implicit function theorem.

Theorem 5.3 (Implicit Function Theorem (IFT)*; see, e.g., [25]. Let* $F : \mathbf{R}^m \times \mathbf{R}^{n-m} \to \mathbf{R}^m$*,* $m < n$ *and* $\hat{x}_1 \in \mathbf{R}^m$*,* $\hat{x}_2 \in \mathbf{R}^{n-m}$ *be such that*

(a) $F \in C^1$

(b) $F(\hat{x}_1, \hat{x}_2) = \theta$

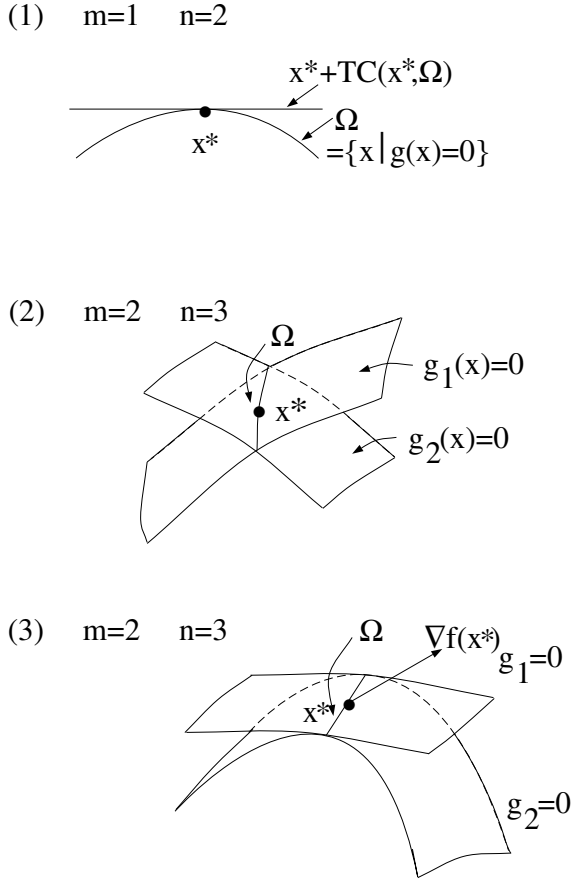


Figure **5.3**: Inclusion (5.4) holds in case (1) and (2), but not in case (3).

(c) $\frac{\partial F}{\partial x_1}(\hat{x}_1, \hat{x}_2)$ *is non-singular.*

Then $\exists \epsilon > 0$ *and a function* $\Phi : B(\hat{x}_2, \epsilon) \to \mathbf{R}^m$ *such that*

(i) $\hat{x}_1 = \Phi(\hat{x}_2)$

(ii) $F(\Phi(x_2), x_2) = \theta \quad \forall x_2 \in B(\hat{x}_2, \epsilon)$

and Φ *is the only function satisfying (i) and (ii);*

(iii) $\Phi \in C^1$ *in* $B(\hat{x}_2, \epsilon)$

(iv) $D\Phi(x_2) = -[D_1 F(\Phi(x_2), x_2)]^{-1} D_2 F(\Phi(x_2), x_2) \quad \forall x_2 \in B(\hat{x}_2, \epsilon)$*, where* D *denotes differentiation and* D_i *differentiation with respect to the* i*th argument.*

■

Interpretation (Figure **5.5**)
Let $n = 2$, $m = 1$ i.e. $x_1 \in \mathbf{R}, x_2 \in \mathbf{R}$. [The idea is to "solve" the system of equations for x_1 (locally around $(\hat{x}_1, \hat{x}_2)$.] Around x^*, ("likely" $\frac{\partial}{\partial x_1} F(x_1^*, x_2^*) \neq \theta$) x_1 is a well defined continuous function of x_2 : $x_1 = \Phi(x_2)$. Around $\tilde{x}$, ($\frac{\partial}{\partial x_1} F(\tilde{x}_1, \tilde{x}_2) = \theta$) and x_1 is not everywhere defined (specifically, it is not defined for $x_2 < \tilde{x}_2$) ((c) is violated). Note that

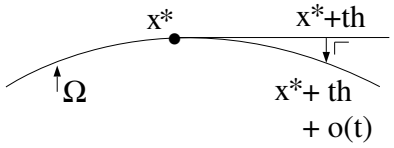


Figure **5.4**:

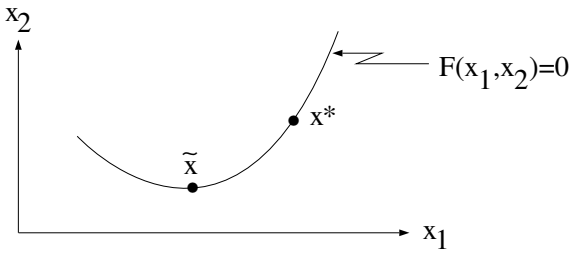


Figure **5.5**:

(iv) is obtained by differentiating (ii) using the chain rule. We are now ready to prove the following theorem.

Exercise 5.10 *Let A be a full rank $m \times n$ matrix with $m \leq n$. Then AA^T is non-singular. If $m > n$, AA^T is singular.*

Definition 5.6 *$x \in \Omega$ is a regular point for problem (5.3) if $\frac{\partial g}{\partial x}(x^*)$ is surjective, i.e., has full row rank.*

Theorem 5.4 *Suppose that $g(x^*) = \theta$ and x^* is a regular point for (5.3). Then (x^*, g) is non-degenerate, i.e.,* $\mathrm{TC}(x^*, \Omega) = \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$.

Proof. Let $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$. We want to find $\varphi : \mathbf{R} \to \mathbf{R}^\ell$ such that $s(t) := x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi(t)$ satisfies

$$g(s(t)) = \theta \quad \forall t > 0$$

i.e.,

$$g\left(x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi(t)\right) = \theta \quad \forall t > 0.$$

Consider the function $\tilde{g} : \mathbf{R}^\ell \times \mathbf{R} \to \mathbf{R}^\ell$

$$\tilde{g} : (\varphi, t) \mapsto g\left(x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi\right)$$

We now use the IFT on $\tilde{g}$ with $\hat{\varphi} = \theta$, $\hat{t} = 0$. We have

(i) $\tilde{g} \in C^1$

(ii) $\tilde{g}(\theta, 0) = \theta$

(iii) $\frac{\partial \tilde{g}}{\partial \varphi}(\theta, \theta) = \frac{\partial g}{\partial x}(x^*)\frac{\partial g}{\partial x}(x^*)^T$

Hence $\frac{\partial \tilde{g}}{\partial \varphi}(\theta, 0)$ is non-singular and IFT applies i.e. $\exists \varphi : \mathbf{R} \to \mathbf{R}^\ell$ and $\bar{t} > 0$ such that $\varphi(0) = \theta$ and

$$\tilde{g}(\varphi(t), t) = \theta \quad \forall t \in [-\bar{t}, \bar{t}]$$

i.e.

$$g\left(x^* + th + \frac{\partial g}{\partial x}(x^*)^T \varphi(t)\right) = \theta \quad \forall t \in [-\bar{t}, \bar{t}].$$

Now note that a differentiable function that vanishes at 0 is a"o" function if and only if its derivative vanishes at 0. To exploit this fact, note that

$$\frac{d\varphi}{dt}(t) = -\left(\frac{\partial}{\partial \varphi}\tilde{g}(\varphi(t), t)\right)^{-1} \frac{\partial}{\partial t}\tilde{g}(\varphi(t), t) \quad \forall t \in [-\bar{t}, \bar{t}]$$

But, from the definition of $\tilde{g}$ and since $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$

$$\frac{\partial}{\partial t}\tilde{g}(\theta, 0) = \frac{\partial g}{\partial x}(x^*)h = \theta$$

i.e.

$$\frac{d\varphi}{dt}(0) = \theta$$

and

$$\varphi(t) = \varphi(0) + \frac{d\varphi}{dt}(0)t + o(t) = o(t)$$

so that

$$g(x^* + th + o'(t)) = \theta \quad \forall t$$

with

$$o'(t) = \frac{\partial g}{\partial x}(x^*)^T o(t)$$

which implies that $h \in \mathrm{TC}(x^*, \Omega)$.

■

Remark 5.4 Note that, in order for $\frac{\partial g}{\partial x}(x^*)$ to be full row rank, it is necessary that $m \leq n$, a natural condition for problem (5.3).

Remark 5.5 Regularity of x^* is not necessary in order for (5.8) to hold. For example, consider cases where two components of g are identical (in which case there are no regular points), or cases where x^* happens to also be an unconstrained local minimizer (so (5.8) holds trivially) but is not a regular point, e.g., $\min x^2$ subject to $x = 0$. Also, it is a simple exercise to show that, if g is affine, then every x^* such that $g(x^*) = \theta$ is regular.

Remark 5.6 It follows from Theorem 5.4 that, when x^* is a regular point, $\mathrm{TC}(x^*, \Omega)$ is convex and closed, i.e.,

$$\mathrm{TC}(x^*, \Omega) = \mathrm{cl}(\mathrm{coTC}(x^*, \Omega)).$$

Indeed, is a closed subspace. This subspace is typically referred to as the tangent plane to Ω at x^*.

Remark 5.7 Suppose now g is scalar-valued. The set $L_\alpha(g) := \{x : g(x) = \alpha\}$ is often referred to as the α-level set of g. Let $x^* \in L_\alpha(g)$ for some α. Regardless of whether x^* is regular for g, we have

$$\mathrm{TC}(x^*, L_\alpha(g)) \subseteq \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) = \{\nabla g(x^*)\}^\perp,$$

i.e., $\nabla g(x^*)$ is orthogonal to $\mathrm{TC}(x^*, L_\alpha(g))$. It is said to be normal at x^* to $L_\alpha(g)$, the (unique) level set (or level surface, or level curve) of g that contains x^*.

Remark 5.8 Also of interest is the connection between (i) the gradient at some $\hat{x} \in \mathbf{R}$ of a scalar, continuously differentiable function $f : \mathbf{R}^n \to \mathbf{R}$ and (ii) the normal at $\hat{x}, f(\hat{x})$ to its graph

$$G := \{(x, z) \in \mathbf{R}^{n+1} \ : \ z = f(x)\}.$$

(Note that $\nabla f(x)$ lies in $\mathbf{R}^n$ while G belongs to R^{n+1}.) Thus let $\hat{x} \in \mathbf{R}^n$ and $\hat{z} = f(\hat{x})$. Then (i) $(\hat{x}, \hat{z})$ is regular for the function $z - f(x)$ and (ii) $(g, -1)$ is orthogonal to the tangent plane at $(\hat{x}, \hat{z})$ to G (i.e., is normal to G) if and only if $g = \nabla f(\hat{x})$. Thus, the gradient at $\hat{x}$ of f is the horizontal projection of the downward normal (makes an angle of more than $\pi/2$ with the positive z-axis) at $(\hat{x}, \hat{z})$ to the graph of f, when the normal is scaled so that its vertical projection (which is always nonzero since f is differentiable at $\hat{x}$) has unit magnitude.

Exercise 5.11 *Prove claims (i) and (ii) in Remark 5.8.*

Remark 5.9 Regularity of x^* is not necessary in order for (5.8) to hold. For example, consider cases where two components of g are identical (in which case there are no regular points), or cases where x^* happens to also be an unconstrained local minimizer (so (5.8) holds trivially) but is not a regular point, e.g., $\min x^2$ subject to $x = 0$.

Corollary 5.2 *Suppose that x^* is a local minimizer for (5.3) (without full rank assumption). Then $\exists \lambda^* \in \mathbf{R}^{\ell+1}$, $\lambda^* = (\lambda^{*,0}, \lambda^{*,1}, \ldots, \lambda^{*,\ell}) \neq \theta_{\ell+1}$, such that*

$$\lambda^{*,0} \nabla f(x^*) + \sum_{j=1}^{\ell} \lambda^{*,j} \nabla g^j(x^*) = \theta, \tag{5.11}$$

i.e., $\{\nabla f(x^*), \nabla g^j(x^*), j = 1, \ldots, \ell\}$ are linearly dependent.

Proof. If $\frac{\partial g}{\partial x}(x^*)$ is not full rank, then (5.11) holds with $\lambda^{*,0} = 0$ ($\nabla g^j(x^*)$ are linearly dependent). If $\frac{\partial g}{\partial x}(x^*)$ has full rank, (5.11) holds with $\lambda^{*,0} = 1$, from Corollary 5.1. ■

Remark 5.10

1. How does the optimality condition (5.9) help us solve the problem? Just remember that x^* must satisfy the constraints, i.e.,

$$g(x^*) = \theta \tag{5.12}$$

   Hence we have a system of $n + \ell$ equations with $n + \ell$ unknown (x^* and λ^*). Now, keep in mind that (5.9) is *only a necessary* condition. Hence, all we can say is that, if there exists a local minimizer satisfying the full rank assumption, it must be among the solutions of (5.9)+(5.12).

2. Solutions with $\lambda^{*,0} = 0$ are *degenerate* in the sense that the cost function does not enter the optimality condition. If $\lambda^{*,0} \neq 0$, dividing both sides of (5.11) by $\lambda^{*,0}$ yields (5.9).

Lagrangian function

If one defines $L : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ by

$$L(x, \lambda) = f(x) + \sum_{j=1}^{m} \lambda^j \; g^j(x)$$

then, the optimality conditions (5.9)+(5.12) can be written as: $\exists \lambda^* \in \mathbf{R}^m$ s.t.

$$\frac{\partial}{\partial x} L(x^*, \lambda^*) \;=\; \theta \;, \tag{5.13}$$

$$\frac{\partial}{\partial \lambda} L(x^*, \lambda^*) \;=\; \theta \;. \tag{5.14}$$

(In fact, $\frac{\partial L}{\partial \lambda}(x^*, \lambda) = \theta \; \forall \lambda$.) L is called the Lagrangian for problem (5.3).

5.3 Equality Constraints – Second Order Condition

Assume $V = \mathbf{R}^n$. In view of the second order condition for the unconstrained problems, one might expect a second order necessary condition of the type

$$h^T \nabla^2 f(x^*) h \geq 0 \quad \forall h \in \mathrm{cl}(\mathrm{co}(TC(x^*, \Omega)) \tag{5.15}$$

since it should be enough to consider directions h in the tangent plane to Ω at x^*. The following exercise show that this is not true in general. (In a sense, the tangent cone is a mere first-order approximation to Ω near x^*, and is not accurate enough for for (5.15) to hold.)

Exercise 5.12 *Consider the problem*

$$\min\{f(x, y) \equiv -x^2 + y \;:\; g(x, y) \equiv y - kx^2 = 0\} \;, \quad k > 1 \;.$$

Show that $(0, 0)$ *is the unique global minimizer and that it does not satisfy (5.15). ((0,0) does not minimize* f *over the tangent cone (=tangent plane).)*

The correct second order conditions will involve the Lagrangian. (The statement given here is more restrictive than strictly necessary. Specifically, while surjectivity of the Jacobian of the constraints is sufficient for the theorem to hold, it also holds under milder appropriate conditions.)

Theorem 5.5 *(2nd order necessary condition). Suppose x^* is a local minimizer for (5.3) and suppose that $\frac{\partial g}{\partial x}(x^*)$ is surjective. Also suppose that f and g are twice continuously differentiable. Then there exists $\lambda \in \mathbf{R}^m$ such that (5.13) holds and*

$$h^T \nabla^2_{xx} L(x^*, \lambda) h \ \geq 0 \quad \forall h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) \tag{5.16}$$

Proof. Let $h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) = \mathrm{TC}(x^*, \Omega)$ (since $\frac{\partial g}{\partial x}(x^*)$ has full row rank). Then $\exists o(\cdot)$ s.t. $x^* + th + o(t) \equiv s(t) \in \Omega \quad \forall t \geq 0$ i.e.

$$g(s(t)) = \theta \quad \forall t \geq 0 \tag{5.17}$$

and, since x^* is a local minimizer, for some $\bar{t} > 0$

$$f(s(t)) \geq f(x^*) \quad \forall t \in [0, \bar{t}] \tag{5.18}$$

We can write $\forall t \in [0, \bar{t}]$

$$0 \leq f(s(t)) - f(x^*) = \nabla f(x^*)^T (th + o(t)) + \frac{1}{2}(th + o(t))^T \left(\nabla^2 f(x^*)(th + o(t))\right) + o_2(t) \tag{5.19}$$

and, for $j = 1, 2, \ldots, m$,

$$0 = g^j(s(t)) - g^j(x^*) = \nabla g^j(x^*)^T (th + o(t)) + \frac{1}{2}(th + o(t))^T \left(\nabla^2 g^j(x^*)(th + o(t))\right) + o_2^j(t) \tag{5.20}$$

where $\frac{o_2(t)}{t^2} \to 0$, as $t \to 0$, $\frac{o_2^j(t)}{t^2} \to 0$ as $t \to 0$ for $j = 1, 2, \ldots, m$. (Note that the first term in the RHS of (5.19) is generally not 0, because of the "o" term, and is likely to even dominate the second term; this is why conjecture (5.15) is incorrect.) We have shown (1st order condition) that there exists $\lambda \in \mathbf{R}^m$ such that

$$\nabla f(x^*) + \sum_{j=1}^{m} \lambda^j \nabla g^j(x^*) = 0$$

Hence, multiplying the jth equation in (5.20) by λ^j and adding all of them, together with (5.19), we get

$$0 \leq L(s(t), \lambda) - L(x^*, \lambda) = \nabla_x L(x^*, \lambda)^T (th + o(t)) + \frac{1}{2}(th + o(t))^T \nabla^2_{xx} L(x^*, \lambda)(th + o(t)) + \tilde{o}_2(t)$$

yielding

$$0 \leq \frac{1}{2}(th + o(t))^T \nabla^2_{xx} L(x^*, \lambda)(th + o(t)) + \tilde{o}_2(t)$$

with

$$\tilde{o}_2(t) = o_2(t) + \sum_{j=1}^{m} \lambda^j o_2^j(t)$$

which can be rewritten as

$$0 \leq \frac{t^2}{2} h^T \nabla^2_{xx} L(x^*, \lambda) h + \bar{o}_2(t) \quad \text{with } \frac{\bar{o}_2(t)}{t^2} \to 0 \text{ as } t \to 0.$$

Dividing by t^2 and letting $t \searrow 0$ $(t > 0)$, we obtain the desired result.

■

Just like in the unconstrained case, the above proof cannot be used directly to obtain a sufficiency condition, by changing '$\leq$' to '$<$' and proceeding backwards. In fact, an additional difficulty here is that all we would get is

$$f(x^*) < f(s(t)) \quad \text{for any } h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) \text{ and } s(t) = x^* + th + o_h(t),$$

for some "little o" function o_h and for all $t \in (o, \bar{t}_h]$ for some $\bar{t}_h > 0$. It is not clear whether this would ensure that

$$f(x^*) < f(x) \quad \forall x \in \Omega \cap B(x^*, \epsilon) \setminus \{x^*\}, \text{ for some } \epsilon > 0.$$

It turns out that it does.

Theorem 5.6 *(2nd order sufficiency condition). Suppose that f and g are twice continuously differentiable. Suppose that $x^* \in \mathbf{R}^n$ is such that*

(i) $g(x^*) = \theta$

(ii) (5.13) is satisfied for some $\lambda^* \in \mathbf{R}^m$

(iii) $h^T \nabla^2_{xx} L(x^*, \lambda^*) h > 0 \quad \forall h \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right),\ h \neq \theta$

Then x^* is a strict local minimizer for (5.3)

Proof. By contradiction. Suppose x^* is not a strict local minimizer. Then $\forall \epsilon > 0$, $\exists x \in B(x^*, \epsilon)$, $x \neq x^*$ s.t. $g(x) = \theta$ and $f(x) \leq f(x^*)$ i.e., $\exists$ a sequence $\{h_k\} \subset \mathbf{R}^n$, $||h_k|| = 1$ and a sequence of positive scalars $t_k \to 0$ such that

$$g(x^* + t_k h_k) = \theta \tag{5.21}$$

$$f(x^* + t_k h_k) \leq f(x^*) \tag{5.22}$$

We can write

$$\begin{aligned} 0 &\geq f(x_k) - f(x^*) = t_k \nabla f(x^*)^T h_k + \frac{1}{2} t_k^2 h_k^T \nabla^2 f(x^*) h_k + o_2(t_k h_k) \\ 0 &= g^j(x_k) - g^j(x^*) = t_k \nabla g^j(x^*)^T h_k + \frac{1}{2} t_k^2 h_k^T \nabla^2 g^j(x^*) h_k + o_2^j(t_k h_k) \end{aligned}$$

Multiplying by the multipliers λ^{*^j} given by (ii) and adding, we get

$$0 \geq \frac{1}{2} t_k^2 h_k^T \nabla_{xx}^2 L(x^*, \lambda^*) h_k + \tilde{o}_2(t_k h_k)$$

i.e.

$$\frac{1}{2} h_k^T \nabla_{xx}^2 L(x^*, \lambda^*) h_k + \frac{\tilde{o}_2(t_k h_k)}{t_k^2} \leq 0 \quad \forall k \tag{5.23}$$

Since $||h_k|| = 1 \quad \forall k$, $\{h_k\}$ lies in a compact set and there exists h^* and K s.t. $h_k \stackrel{K}{\to} h^*$. Without loss of generality, assume $h_k \to h^*$. Taking $k \to \infty$ in (5.23), we get

$$(h^*)^T \nabla_{xx}^2 L(x^*, \lambda^*) h^* \leq 0.$$

Further, (5.21) yields

$$0 = t_k \left(\frac{\partial g}{\partial x}(x^*)\ h_k + \frac{o(t_k)}{t_k} \right) \tag{5.24}$$

Since $t_k > 0$ and $h_k \to h^*$, and since $\frac{o(t_k)}{t_k} \to 0$ as $k \to \infty$, this implies

$$\frac{\partial g}{\partial x}(x^*) h^* = \theta, \text{ i.e., } h^* \in \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*) \right) .$$

This is in contradiction with assumption (iii).

■

Exercise 5.13 *Give an alternate proof of Theorem 4.6, using the line of proof used above for Theorem 5.6. Where do you make use of the assumption that the domain is finite dimensional? (Recall Exercise 4.14.)*

5.4 Inequality Constraints – First Order Conditions

Consider the problem

$$\min\{f^0(x)\ :\ f(x) \leq \theta\} \tag{5.25}$$

where $f^0 : \mathbf{R}^n \to \mathbf{R}$ and $f : \mathbf{R}^n \to \mathbf{R}^m$ are continuously differentiable, and "$\leq$" is meant component-wise. Recall that, in the equality constrained case (see (5.3)), we obtained two first order conditions: a strong condition

$$\exists \lambda^* \text{ such that } \nabla f(x^*) + \frac{\partial g}{\partial x}(x^*)^T \lambda^* = \theta \tag{5.26}$$

subject to the assumption that $\frac{\partial g}{\partial x}(x^*)$ has rank m, and a weaker condition

$$\exists (\lambda_0^*, \lambda^*) \neq \theta \text{ such that } \lambda_0^* \nabla f(x^*) + \frac{\partial g}{\partial x}(x^*)^T \lambda^* = \theta, \tag{5.27}$$

not subject to such assumption. For the inequality constraint case, we shall first obtain a weak condition (F. John condition; Fritz John, German-born mathematician, 1910–1994) analogous to (5.27). Then we will investigate when a strong condition (Karush-Kuhn-Tucker condition) holds. (William Karush, American mathematician, 1917–1997; Harold W. Kuhn, American mathematician, 1925–2014; Albert W. Tucker, Canadian mathematician, 1905–1995. From wikipedia: "The KKT conditions were originally named after Harold W. Kuhn, and Albert W. Tucker, who first published the conditions in 1951. Later scholars discovered that the necessary conditions for this problem had been stated by William Karush in his master's thesis in 1939.") We use the following notation. For $x \in \mathbf{R}^n$

$$J(x) = \{j \in \{1,\dots,m\} \ : \ f^j(x) \geq 0\} \tag{5.28}$$

$$J_0(x) = \{0\} \cup J(x) \tag{5.29}$$

i.e., $J(x)$ is the set of indices of active or violated constraints at x whereas $J_0(x)$ includes the index of the cost function as well.

Exercise 5.14 *Let $s \in \mathbf{R}^m$ be a vector of "slack" variables. Consider the 2 problems*

$$(P_1) \ \min_x\{f^0(x) \ : \ f^j(x) \leq 0, \ j = 1,\dots,m\}$$

$$(P_2) \ \min_{x,s}\{f^0(x) \ : \ f^j(x) + (s^j)^2 = 0, \ j = 1,\dots,m\}$$

(i) Carefully prove that if $(\hat{x}, \hat{s})$ solves (P_2) then $\hat{x}$ solves (P_1).

(ii) Carefully prove that if $\hat{x}$ solves (P_1) then $(\hat{x}, \hat{s})$ solves (P_2), where $\hat{s}^j = \sqrt{-f^j(\hat{x})}$.

(iii) Use Lagrange multipliers rule and part (ii) to prove (carefully) the following weak result:

If $\hat{x}$ solves (P_1) and $\{\nabla f^j(\hat{x}) : j \in J(\hat{x})\}$ is a linearly-independent set, then there exists a vector $\mu \in \mathbf{R}^m$ such that

$$\nabla f^0(\hat{x}) + \frac{\partial f}{\partial x}(\hat{x})^T \mu = \theta \tag{5.30}$$

$$\mu^j = 0 \quad \text{if } f^j(\hat{x}) < 0 \ \ j = 1, 2, \dots, m \tag{5.31}$$

[This result is weak because: (i) there is no constraint on the sign of μ, (ii) the linear independence assumption is significantly stronger than necessary.]

To obtain stronger conditions, we now proceed, as in the case of equality constraints, to characterize $\mathrm{TC}(x^*, \Omega)$. For $x \in \Omega$ we first define the set of "first-order strictly feasible" directions

$$\tilde{S}_f(x) := \{h \ : \ \nabla f^j(x)^T h < 0 \quad \forall j \in J(x)\}.$$

It is readily checked that $\tilde{S}_f(x^*)$ is a convex cone. Further, as shown next, it is a subset of the radial cone, hence of the tangent cone.

Theorem 5.7 *If $x \in \Omega$ (i.e., $f(x) \leq \theta$),*

$$\tilde{S}_f(x) \subseteq \mathrm{RC}(x, \Omega) \ .$$

Proof. Let $h \in \tilde{S}_f(x)$. For $j \in J(x)$, we have, since $f^j(x) = 0$,

$$f^j(x+th) = f^j(x+th) - f^j(x) \ = \ t\ \nabla f^j(x^*)^T h + o^j(t) \tag{5.32}$$

$$= \ t\ \left(\nabla f^j(x)^T h + \frac{o^j(t)}{t} \right). \tag{5.33}$$

Since $\nabla f^j(x)^T h < 0,$ there exists $\bar{t}^j > 0$ such that

$$\nabla f^j(x)^T h + \frac{o^j(t)}{t} < 0 \quad \forall t \in (0, \bar{t}^j] \tag{5.34}$$

and, with $\bar{t} = \min\{\bar{t}^j \ : \ j \in J(x)\} > 0$,

$$f^j(x+th) < 0 \quad \forall j \in J(x), \quad \forall t \in (0, \bar{t}]. \tag{5.35}$$

For $j \in \{1, \ldots, \ m\} \backslash J(x),\ f^j(x) < 0$, thus, by continuity $\exists\ \tilde{t} > 0$ s.t.

$$f^j(x+th) < 0 \quad \forall j \in \{1, \ldots, m\} \backslash J(x), \quad \forall t \in (0, \tilde{t}]$$

Thus

$$x + th \in \Omega \quad \forall t \in (0, \min(\bar{t}, \tilde{t})]$$

■

Theorem 5.8 *Suppose $x^* \in \Omega$ is a local minimizer for (P). Then*

$$\nabla f^0(x^*)^T h \geq 0 \quad \text{for all } h \text{ s.t. } \nabla f^j(x^*)^T h \ < 0 \ \forall j \in J(x^*)$$

or, equivalently

$$\nexists h \text{ s.t. } \nabla f^j(x^*)^T h \ < 0 \quad \forall j \in J_0(x^*).$$

Proof. Follows directly from Theorems 5.7 and 5.1.

■

Remark 5.11 Since $\mathrm{RC}(x, \Omega) \subset \mathrm{TC}(x, \Omega))$, it follows from Theorem 5.7 that

$$\mathrm{cl}\, \tilde{S}_f(x^*) \subseteq \mathrm{cl}\, \mathrm{RC}(x, \Omega) \subset \mathrm{cl}\, \mathrm{co}\, \mathrm{TC}(x, \Omega)).$$

However, as further discussed below, it does not follow from Theorem 5.8 that $\nabla f^0(x^*)^T h \geq 0$ for all h such that $\nabla f^j(x^*)^T h \leq 0$ for all $j \in J(x^*)$.

Definition 5.7 *A set of vectors $a_1, \ldots, a_k \in \mathbf{R}^n$ is said to be positively linearly independent if*

$$\left.\begin{array}{l} \sum_{i=1}^k \mu^i a_i = 0 \\ \mu^i \geq 0, \quad i = 1, \ldots, k \end{array}\right\} \qquad \textit{implies } \mu_i = 0 \qquad i = 1, 2, \ldots, k$$

If the condition does not hold, they are positively linearly dependent.

Note that, if $a_1, \ldots,\ a_k$ are linearly independent, then they are positively linearly independent.

The following proposition gives a geometric interpretation to the necessary condition just obtained.

Theorem 5.9 *Given $\{a_1, \ldots, a_k\}$, a finite collection of vectors in $\mathbf{R}^n$, the following three statements are equivalent*

(i) $\nexists h \in \mathbf{R}^n$ *such that* $a_j^T h \ < 0 \quad j = 1, \ldots, k$

(ii) $\theta \in \text{co}\{a_1, \ldots, a_k\}$

(iii) the a_j's are positively linearly dependent.

Proof

(i)⇒(ii): By contradiction. If $\theta \notin \text{co}\{a_1, \ldots, a_k\}$ then by the separation theorem (see Appendix B) there exists h such that

$$v^T h \ < \ 0 \quad \forall v \in \text{co}\{a_1, \ldots, a_k\} \ .$$

In particular $a_j^T h < 0$ for $j = 1, \ldots, k$. This contradicts (i)

(ii)⇒(iii): If $\theta \in \text{co}\{a_1, \ldots, a_k\}$ then (see Exercise B.24 in Appendix B), for some α_j,

$$\theta = \sum_{j=1}^{k} \alpha_j a_j \quad \alpha_j \geq 0 \quad \sum_{1}^{k} \alpha_j = 1$$

Since positive numbers which sum to 1 cannot be all zero, this proves (iii).

(iii)⇒(i): By contradiction. Suppose (iii) holds, i.e., $\sum \alpha_i a_i = \theta$ for some $\alpha_i \geq 0$, not all zero, and there exists h such that $a_j^T h < 0$, $j = 1, \ldots, k$. Then

$$0 = \left(\sum_{1}^{k} \alpha_j a_j \right)^T h = \sum_{1}^{k} \alpha_j a_j^T h.$$

But since the α_j's are non negative, not all zero, this is a contradiction.

■

Corollary 5.3 *Suppose x^* is a local minimizer for (5.25). Then*

$$\theta \in \text{co}\{\nabla f^j(x^*) \ : \ j \in J_0(x^*)\}$$

Corollary 5.4 *(F. John conditions). Suppose x^* is a local minimizer for (5.25). Then there exist $\mu^{*0}, \mu^{*1}, \ldots, \mu^{*m}$, not all zero such that*

(i) $\mu^{*0}\nabla f^0(x^*) + \sum_{j=1}^{m} \mu^{*j}\nabla f^j(x^*) = \theta$

(ii) $f^j(x^*) \leq 0 \qquad j = 1, \ldots, m$

(iii) $\mu^{*j} \geq 0 \qquad j = 0, 1, \ldots, m$

(iv) $\mu^{*j} f^j(x^*) = 0 \qquad j = 1, \ldots, m$

Proof. From (iii) in Theorem 5.9, $\exists \mu^{*j}$, $\forall j \in J_0(x^*)$, $\mu^{*j} \geq 0$, not all zero such that

$$\sum_{j \in J_0(x^*)} \mu^{*j} \nabla f^j(x^*) = 0.$$

By defining $\mu^{*j} = 0$ for $j \notin J_0(x^*)$ we obtain (i). Finally, (ii) just states that x^* is feasible, (iii) directly follows and (iv) follows from $\mu^{*j} = 0 \quad \forall j \notin J_0(x^*)$.

■

Remark 5.12 Condition (iv) in Corollary 5.4 is called *complementary slackness*. In view of conditions (ii) and (iii) it can be equivalently stated as

$$(\mu^*)^T f(x^*) = \sum_{1}^{m} \mu^{*j} f^j(x^*) = 0$$

■

The similarity between the above F. John conditions and the weak conditions obtained for the equality constrained case is obvious. Again, if $\mu^{*0} = 0$, the cost f^0 does not enter the conditions at all. These conditions are degenerate. (For example, $\min x$ s.t. $x^3 \geq 0$ has $x^* = 0$ as global minimizer and the F. John condition only holds with $\mu^{*0} = 0$. See more meaningful examples below.)
We are now going to try and obtain conditions for the optimality conditions to be non-degenerate, i.e., for the first multiplier μ^{*0} to be nonzero (in that case, it can be set to 1 by dividing through by μ^{*0}). The resulting optimality conditions are called Karush-Kuhn-Tucker (or Kuhn-Tucker) optimality conditions. The general condition for $\mu^{*0} \neq 0$ is called Kuhn-Tucker constraint qualification (KTCQ). But before considering it in some detail, we give a simpler (but more restrictive) condition, which is closely related to the "regularity" condition of equality-constrained optimization.

Proposition 5.2 *Suppose x^* is a local minimizer for (5.25) and suppose $\{\nabla f^j(x^*) \ : \ j \in J(x^*)\}$ is a positively linearly independent set of vectors. Then the F. John conditions hold with $\mu^{*0} \neq 0$.*

Proof. By contradiction. Suppose $\mu^{*0} = 0$. Then, from Corollary 5.4 above, $\exists\ \mu^{*j} \geq 0,\ j = 1, \ldots, m$, *not all zero* such that

$$\sum_{j=1}^{m} \mu^{*j} \nabla f^j(x^*) = \theta$$

and, since $\mu^{*j} = 0 \quad \forall j \notin J_0(x^*)$,

$$\sum_{j \in J(x^*)} \mu^{*j} \nabla f^j(x^*) = \theta$$

which contradicts the positive linear independence assumption.

■

Note: Positive linear independence of $\{\nabla f^j(x^*) \ : \ j \in J(x^*)\}$ is not necessary in order for μ^{*0} to be nonzero. See, e.g, again, $\min x$ s.t. $x^3 \geq 0$ has $x^* = 0$.
We now investigate weaker conditions under which a strong result holds. For $x \in \Omega$, define

$$S_f(x) := \{h \ : \ \nabla f^j(x)^T h \ \leq 0 \ \ \forall j \in J(x)\}. \tag{5.36}$$

It turns out that the strong result holds whenever $S_f(x^*)$ is equal the convex hull of the closure of the tangent cone. (It is readily verified that $S_f(x^*)$ is a closed convex cone.)

Remark 5.13 In the case of equality constraints recast as inequality constraints ($g(x) \leq 0, -g(x) \leq 0$), $S_f(x)$ reduces to $\mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$!

Remark 5.14 The inclusion $\tilde{S}_f(x^*) \subseteq \mathrm{TC}(x^*, \Omega)$ (Theorem 5.7) does not imply that

$$S_f(x^*) \subseteq \mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) \tag{5.37}$$

since, in general

$$\mathrm{cl}(\tilde{S}_f(x^*)) \subseteq S_f(x^*) \tag{5.38}$$

but equality in (5.38) need not hold. (As noted below (see Proposition 5.6), it holds if and only if $\tilde{S}_f(x^*)$ is nonempty.) Condition (5.37) is known as the KTCQ condition (see below).

The next exercise shows that the inclusion opposite to (5.37) always holds. That inclusion is analogous to the inclusion $\mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) \subseteq \mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right)$ in the equality-constrained case.

Exercise 5.15 *If $x^* \in \Omega$, then*

$$\mathrm{cl}(\mathrm{coTC}(x^*, \Omega)) \subseteq S_f(x^*).$$

Definition 5.8 *The Kuhn-Tucker constraint qualification (KTCQ) is satisfied at $\hat{x} \in \Omega$ if*

$$\mathrm{cl}(\mathrm{coTC}(\hat{x}, \Omega)) = \{h \ : \ \nabla f^j(\hat{x})^T h \leq 0 \quad \forall j \in J(\hat{x})\} \ \ (= S_f(\hat{x})) \tag{5.39}$$

Thus KTCQ is satisfied iff (5.37) holds. In particular this holds whenever $\tilde{S}_f(x) \neq \emptyset$ (see Proposition 5.6 below).

Example 5.3 In (1), (2), and (3) below, the gradients of the active constraints are not positively linearly independent. In (1) and (2) though, KTCQ does hold.

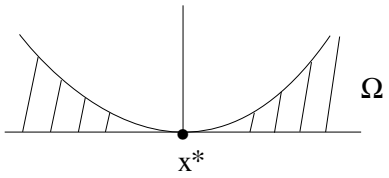


Figure **5.6**:

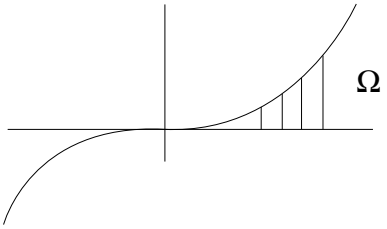


Figure **5.7**:

(1) (Figure **5.6**)

$$\left.\begin{array}{ll} f^1(x) & \equiv x_2 - x_1^2 \\ f^2(x) & \equiv -x_2 \\ x^* & = (0,0) \end{array}\right\} \quad \Rightarrow \begin{array}{ll} \tilde{S}_f(x^*) & = \emptyset \\ S_f(x^*) & = \{h \text{ s.t. } h_2 = 0\} = \mathrm{TC}(x^*, \Omega) \end{array}$$

but (5.37) holds anyway

(2) (Figure **5.7**)

$$\left.\begin{array}{ll} f^1(x) & \equiv x_2 - x_1^3 \\ f^2(x) & = -x_2 \\ x^* & = (0,0) \end{array}\right\} \quad \Rightarrow \begin{array}{ll} \tilde{S}_f(x^*) & = \emptyset \\ S_f(x^*) & = \{h \text{ s.t. } h_2 = 0\} \\ \mathrm{TC}(x^*, \Omega) & = \{h : h_2 = 0, h_1 \geq 0\} \\ & \text{and (5.37) does not hold} \end{array}$$

If f^2 is changed to $x_1^4 - x_2$, KTCQ still does not hold, but with the cost function $f^0(x) = x_2$, the KKT conditions do hold! (Hence KTCQ is sufficient but <u>not necessary</u> in order for KKT to hold. Also see Exercise 5.20 below.)

(3) Example similar to that given in the equality constraint case; $\mathrm{TC}(x^*, \Omega)$ and $S_f(x^*)$ are the same as in the equality case (Figure **5.8**). KTCQ does not hold.

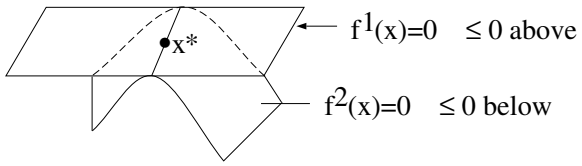


Figure **5.8**:

(4) $\{(x, y, z) : z - (x^2 + 2y^2) \geq 0, z - (2x^2 + y^2) \geq 0\}$. At $(0, 0, 0)$, the gradients are positively linearly independent, though they are not linearly independent. $\tilde{S}_f(x^*) \neq \emptyset$, so that KTCQ does hold.

Remark 5.15 The equality constraint $g(x) = \theta$ can be written instead as the pair of inequalities $g(x) \leq \theta$ and $-g(x) \leq \theta$. If there are no other constraints, then we get $S_f(\hat{x}) = \mathcal{N}\left(\frac{\partial g}{\partial x}(\hat{x})\right)$ so that KTCQ holds if and only if $(\hat{x}, g)$ is non-degenerate. (In particular, if $\hat{x}$ is regular, then KTCQ does hold.) However, none of the sufficient conditions we have discussed (for KTCQ to hold) are satisfied in such case (implying that none are necessary)! (Also, as noted in Remark 5.5, in the equality-constrained case, regularity is (sufficient but) not necessary for a "strong" condition—i.e., with $\lambda^{*,0} \neq 0$—to hold.)

Exercise 5.16 *(due to H.W. Kuhn). Obtain the sets $S_f(x^*), \tilde{S}_f(x^*)$ and $\mathrm{TC}(x^*, \Omega)$ at minimizer x^* for the following examples (both have the same Ω): (i) minimize $(-x_1)$ subject to $x_1 + x_2 - 1 \leq 0,\ x_1 + 2x_2 - 1 \geq 0,\ x_1 \geq 0, x_2 \geq 0$, (ii) minimize $(-x_1)$ subject to $(x_1 + x_2 - 1)(x_1 + 2x_2 - 1) \leq 0$; $x_1 \geq 0,\ x_2 \geq 0$. In each case, indicate whether KTCQ holds.*

Exercise 5.17 *In the following examples, exhibit the tangent cone at $(0, 0)$ and determine whether KTCQ holds*

(i) $\Omega = \{(x, y) \ : \ (x-1)^2 + y^2 - 1 \leq 0,\ x^2 + (y-1)^2 - 1 \leq 0\}$

(ii) $\Omega = \{(x, y) \ : \ -x^2 + y \leq 0,\ -x^2 - y \leq 0.$

(iii) $\Omega = \{(x, y) \ : \ -x^2 + y \leq 0,\ -x^2 - y \leq 0,\ -x \leq 0\}$

(iv) $\Omega = \{(x, y) \ : \ xy \leq 0\}$.

Remark 5.16 Convexity of f^j, $j = 0, 1, \ldots, m$ does not imply KTCQ. E.g., for $\Omega := \{x : x^2 \leq 0\} \subset \mathbf{R}$ (or, say, $\Omega := \{(x, y) : x^2 \leq 0\} \subset \mathbf{R}^2$), KTCQ does not hold at the origin.

The following necessary condition of optimality follows trivially.

Proposition 5.3 *Suppose x^* is a local minimizer for (P) and KTCQ is satisfied at x^*. Then $\nabla f^0(x^*)^T h \geq 0\ \forall h \in S_f(x^*)$, i.e.,*

$$\nabla f^0(x^*)^T h \geq 0\ \forall h \text{ s.t. } \nabla f^j(x^*)^T h \leq 0 \quad \forall j \in J(x^*) \tag{5.40}$$

Proof. Follows directly from (5.39) and Theorem 5.1.

■

Remark 5.17 As mentioned in Remark 5.22 below, KTCQ is in fact the least restrictive constraint qualification for our inequality-constrained problem.

Remark 5.18 The above result may appear to the reader to be only slightly stronger than Theorem 5.8. It turns out that this is enough to ensure $\mu^{*0} \neq 0$ though. This follows directly from Farkas's Lemma, named after Gyula Farkas (Hungarian mathematician, 1847–1930).

■

Proposition 5.4 *(Farkas's Lemma). Consider* $a_1, \ldots, a_k,\ b \in \mathbf{R}^n$. *Then* $b^T x \leq 0 \quad \forall x \in \{x\ :\ a_i^T x\ \leq 0,\ i = 1, \ldots, k\}$ *if and only if*

$$\exists \lambda^i \geq 0,\ i = 1, 2, \ldots, k \text{ s.t. } b = \sum_{i=1}^{k} \lambda^i a_i$$

Proof. ($\Leftarrow$): Obvious. ($\Rightarrow$): Consider the set

$$C = \{y\ :\ y = \sum_{i=1}^{k} \lambda_i a_i,\ \lambda_i \geq 0, i = 1, \ldots,\ k\}\ .$$

Exercise 5.18 *Prove that* C *is a closed convex cone.*

Our claim can now be simply expressed as $b \in C$. We proceed by contradiction. Thus suppose $b \notin C$. Then, by Exercise B.26, there exists $\tilde{x}$ such that $\tilde{x}^T b > 0$ and $\tilde{x}^T v \leq 0$ for all $v \in C$, in particular, $\tilde{x}^T a_i \leq 0$ for all i, contradicting the premise.

■

Theorem 5.10 *(Karush-Kuhn-Tucker). Suppose* x^* *is a local minimizer for (5.26) and KTCQ holds at* x^*. *Then there exists* $\mu^* \in \mathbf{R}^m$ *such that*

$$\begin{aligned} \nabla f^0(x^*) \ &+\ \sum_{j=1}^{m} \mu^{*j} \nabla f^j(x^*) = \theta \\ \mu^{*j} \ &\geq\ 0 \qquad j = 1, \ldots, m \\ f^j(x*) \ &\leq\ 0 \qquad j = 1, \ldots, m \\ \mu^{*j} f^j(x^*) \ &=\ 0 \qquad j = 1, \ldots, m \end{aligned}$$

Proof. From (5.40) and Farkas's Lemma $\exists \mu^{*j},\ j \in J(x^*)$ such that

$$\nabla f^0(x^*) + \sum_{j \in J(x^*)} \mu^{*j} \nabla f^j(x^*) = \theta$$

$$\text{with } \mu^{*j} \geq 0,\ j \in J(x^*)$$

Setting $\mu^{*j} = 0$ for $j \notin J(x^*)$ yields the desired results.

■

Remark 5.19 An interpretation of Farkas's Lemma is that the closed convex cone C and the closed convex cone

$$D := \{x\ :\ a_i^T x \geq 0,\ i = 1, \ldots, k\}$$

are dual of each other; see Definition 5.4. (Hint: $\langle b, x\rangle \leq 0$ for all $x \in \{x : \langle a_i, x\rangle \leq 0\}$ iff $\langle b, x\rangle \geq 0$ for all $x \in \{x : \langle a_i, x\rangle \geq 0\}$.) In the case of a subspace S (subspaces are cones),

the dual cone is simply the orthogonal complement, i.e., $S^* = S^\perp$ (check it). In that special case, the fundamental property of linear maps L

$$\mathcal{N}(L) = \mathcal{R}(L^*)^\perp$$

can be expressed in the notation of cone duality as

$$\mathcal{N}(L) = \mathcal{R}(L^*)^*,$$

which shows that our proofs of Corollary 5.1 (equality constraint case) and Theorem 5.10 (inequality constraint case), starting from Theorem 5.2 and Proposition 5.3 respectively, are analogous.

We pointed out earlier a sufficient condition for $\mu^{*0} \neq 0$ in the F. John conditions. We now consider two other important sufficient conditions and we show that they imply KTCQ.

Definition 5.9 *$h : \mathbf{R}^n \to \mathbf{R}$ is affine if $h(x) = a^T x + b$.*

Proposition 5.5 *Suppose the f^j's are affine. Then $\Omega = \{x : f^j(x) \leq 0,\ j = 1, 2, \ldots, m\}$ satisfies KTCQ at any $x \in \Omega$.*

Exercise 5.19 *Prove Proposition 5.5.*

The next exercise shows that KTCQ (a property of the description of the constraints) is not necessary in order for the KKT conditions to hold for some objective function.

Exercise 5.20 *Consider again the optimization problems in Exercise 5.16. In both cases (i) and (ii) check whether the Kuhn-Tucker optimality conditions hold. Then repeat with the cost function x_2 instead of $-x_1$, which does not move the minimizer x^*. (This substitution clearly does not affect whether KTCQ holds!)*

Remark 5.20 Note that, as a consequence of this, the strong optimality condition holds for any equality constrained problem whose constraints are all affine (with no "regular point" assumption).

Now, we know that it always holds that

$$\mathrm{cl}(\mathrm{TC}(x^*, \Omega)) \subseteq S_f(x^*) \text{ and } \mathrm{cl}(\tilde{S}_f(x^*)) \subseteq \mathrm{cl}(\mathrm{TC}(x^*, \Omega)) \ .$$

Thus, a sufficient condition for KTCQ is

$$S_f(x^*) \subset \mathrm{cl}(\tilde{S}_f(x^*)).$$

The following proposition establishes when this holds. (Clearly, the only instance except for the trivial case when both $S_f(x^*)$ and $\tilde{S}_f(x^*)$ are empty.)

Proposition 5.6 *(Mangasarian-Fromovitz) Suppose that there exists* $\hat{h} \in \mathbf{R}^n$ *such that* $\nabla f^j(\hat{x})^T \hat{h} < 0 \quad \forall j \in J(\hat{x})$, *i.e., suppose* $\tilde{S}_f(\hat{x}) \neq \emptyset$. *Then* $S_f(\hat{x}) \subset \mathrm{cl}(\tilde{S}_f(\hat{x}))$ *(and thus, KTCQ holds at* $\hat{x}$*).*

Proof. Let $h \in S_f(\hat{x})$ and let $h_i = \frac{1}{i}\hat{h} + h \quad i = 1, 2, \ldots$ Then

$$\nabla f^j(\hat{x})^T h_i = \frac{1}{i} \underbrace{\nabla f^j(\hat{x})^T \hat{h}}_{<0} + \underbrace{\nabla f^j(\hat{x})^T h}_{\leq 0} \quad < 0 \quad \forall j \in J(\hat{x}) \tag{5.41}$$

so that $h_i \in \tilde{S}_f(\hat{x}) \quad \forall i$. Since $h_i \to h$ as $i \to \infty$, $h \in \mathrm{cl}(\tilde{S}_f(\hat{x}))$.

■

Exercise 5.21 *Suppose the gradients of active constraints are positively linearly independent. Then KTCQ holds. (Hint:* $\exists h$ *s.t.* $\nabla f^j(x)^T h\rangle < 0 \; \forall j \in J(x)$*.)*

Exercise 5.22 *Suppose the gradients of the active constraints are linearly independent. Then the KKT multipliers are unique.*

Exercise 5.23 *(First order sufficient condition of optimality.) Consider the problem*

$$\text{minimize } f_0(x) \;\; \text{s.t.} \;\; f_j(x) \leq 0 \;\; j = 1, \ldots, m,$$

where $f_j : \mathbf{R}^n \to \mathbf{R}$, $j = 0, 1, \ldots, m$, *are continuously differentiable. Suppose the F. John conditions hold at* x^* *with multipliers* μ_j^*, $j = 0, 1, \ldots, m$, *not all zero. In particular* x^* *satisfies the constraints,* $\mu_j^* \geq 0, j = 0, 1, \ldots, m$, $\mu_j^* f_j(x^*) = 0, j = 1, \ldots, m$, *and*

$$\mu_0^* \nabla f_0(x^*) + \sum_{j=1}^{m} \mu_j^* \nabla f_j(x^*) = \theta.$$

(There is no assumption that these multipliers are unique in any way.) Further suppose that there exists a subset J *of* $J_0(x^*)$ *of cardinality* n *with the following properties: (i)* $\mu_j^* > 0$ *for all* $j \in J$, *(ii)* $\{\nabla f_j(x^*) \;:\; j \in J\}$ *is a linearly independent set of vectors. Prove that* x^* *is a strict local minimizer.*

5.5 Mixed Constraints – First Order Conditions

We consider the problem

$$\min\{f^0(x) \;:\; f(x) \leq \theta, \; g(x) = \theta\} \tag{5.42}$$

with

$$f^0 \;:\; \mathbf{R}^n \to \mathbf{R}, \; f \;:\; \mathbf{R}^n \to \mathbf{R}^m, \; g \;:\; \mathbf{R}^n \to \mathbf{R}^\ell, \text{ all } C_1$$

We first obtain an extended Karush-Kuhn-Tucker condition. We define

$$S_{f,g}(x^*) = \{h \ : \ \nabla f^j(x^*)^T h \ \leq 0 \quad \forall j \in J(x^*), \ \frac{\partial g}{\partial x}(x^*)h = 0\}$$

As earlier, KTCQ is said to hold at x^* if

$$S_{f,g}(x^*) = \mathrm{cl}(\mathrm{coTC}(x^*, \Omega)). \tag{5.43}$$

Remark 5.21 In the pure equality case ($m = 0$) this reduces to $\mathcal{N}\left(\frac{\partial g}{\partial x}(x^*)\right) = \mathrm{cl}(\mathrm{coTC}(x^*, \Omega))$, which is less restrictive than regularity of x^*.

Exercise 5.24 *Again $S_{f,g}(x^*) \supseteq \mathrm{cl}(\mathrm{coTC}(x^*, \Omega))$ always holds.*

Fact. If $\{\nabla g^j(x), j = 1, \ldots, \ell\} \cup \{\nabla f^j(x), j \in J(x)\}$ is a linearly independent set of vectors, then KTCQ holds at x.

Exercise 5.25 *Prove the Fact.*

Theorem 5.11 *(extended KKT conditions). If x^* is a local minimizer for (5.42) and KTCQ holds at x^* then $\exists \mu^* \in \mathbf{R}^m, \ \lambda^* \in \mathbf{R}^\ell$ such that*

$$\begin{aligned} \mu^* &\geq \theta \\ f(x^*) &\leq \theta, \ g(x^*) = \theta \\ \nabla f^0(x^*) + \sum_{j=1}^{m} \mu^{*j} \nabla f^j(x^*) + \sum_{j=1}^{\ell} \lambda^{*j} \nabla g^j(x^*) &= \theta \\ \mu^{*j} f^j(x^*) &= 0 \quad j = 1, \ldots, m \end{aligned}$$

Proof (sketch)
Express $S_{f,g}(x^*)$ by means of pairs of inequalities of the form $\nabla g^j(x^*)^T h \leq 0$, $g^j(x^*)^T h \geq 0$, and use Farkas's lemma.

■

Remark 5.22 As shown in the next exercise, constraint qualification (5.43) is the least restrictive valid constraint qualification: it is necessary in some appropriate sense. The non-degeneracy condition we considered in the context of equality-constrained optimization and the ("restricted") KTCQ we considered obtained in the pure inequality-constraint case are special cases of the above and hence are necessary in the same sense.

Exercise. Prove the following: Constraint qualification (5.43) is necessary in the sense that, if it does not hold at x^* then there exists a continuously differentiable objective function f^0 that attains a (constrained) local minimum at x^* at which the extended KKT conditions do not hold. [Hint: Invoke the main result in paper [15] (which involves the concept of polar cone) and show that the CQ condition used there is equivalent to the above.]

Without the KTCQ assumption, the following (weak) result holds (see, e.g., [6], section 3.3.5).

Theorem 5.12 *(extended F. John conditions). If x^* is a local minimizer for (5.42) then $\exists \mu^{*j},\ j = 0, 1, 2, \ldots, m$ and $\lambda^{*j}\ j = 1, \ldots, \ell$, not all zero, such that*

$$\begin{aligned} \mu^{*j} &\geq 0 \quad j = 0, 1, \ldots, m \\ f(x^*) &\leq \theta, \quad g(x^*) = \theta \end{aligned}$$

$$\sum_{j=0}^{m} \mu^{*j} \nabla f^j(x^*) + \sum_{j=1}^{\ell} \lambda^{*j} \nabla g^j(x^*) = \theta$$

$$\mu^{*j} f^j(x^*) = 0 \quad j = 1, \ldots, m \text{ (complementary slackness)}$$

■

Note the constraint on the sign of the μ^{*j}'s (but not of the λ^{*j}'s) and the complementary slackness condition for the inequality constraints.

Exercise 5.26 *The argument that consists in splitting again the equalities into sets of 2 inequalities and expressing the corresponding F. John conditions is inappropriate. Why?*

Theorem 5.13 *(Convex problems, sufficient condition). Consider problem (5.42). Suppose that $f^j,\ j = 0, 1, 2, \ldots, m$ are convex and that $g^j,\ j = 1, 2, \ldots, \ell$ are affine. Under those conditions, if x^* is such that $\exists \mu^* \in \mathbf{R}^m,\ \lambda^* \in \mathbf{R}^\ell$ which, together with x^*, satisfy the KKT conditions, then x^* is a global minimizer for (5.42).*

Proof. Define $\ell\ :\ \mathbf{R}^n \to \mathbf{R}$ as

$$\ell(x) := L(x, \mu^*, \lambda^*) = f^0(x) + \sum_{j=1}^{m} \mu^{*^j} f^j(x) + \sum_{j=1}^{\ell} \lambda^{*^j} g^j(x)$$

with μ^* and λ^* as given in the theorem.

(i) $\ell(\cdot)$ is convex (prove it)

(ii) $\nabla \ell(x^*) = \nabla f^0(x^*) + \sum\limits_{j=1}^{m} \mu^{*^j} \nabla f^j(x^*) + \sum\limits_{j=1}^{\ell} \lambda^{*^j} \nabla g^j(x^*) = \theta$

since (x^*, μ^*, λ^*) is a KKT triple.

(i) and (ii) imply that x^* is a global minimizer for ℓ, i.e.,

$$\ell(x^*) \leq \ell(x) \quad \forall x \in \mathbf{R}^n$$

in particular,

$$\ell(x^*) \leq \ell(x) \quad \forall x \in \Omega$$

i.e.

$$f^0(x^*) + \sum_{j=1}^{m} \mu^{*^j} f^j(x^*) + \sum_{j=1}^{\ell} \lambda^{*^j} g^j(x^*) \leq f^0(x) + \sum_{j=1}^{m} \mu^{*^j} f^j(x) + \sum_{j=1}^{\ell} \lambda^{*^j} g^j(x) \quad \forall x \in \Omega\ .$$

Since (x^*, μ^*, λ^*) is a KKT triple this simplifies to

$$f^0(x^*) \leq f^0(x) + \sum_{j=1}^{m} \mu^{*j} f^j(x) + \sum_{j=1}^{\ell} \lambda^{*j} g^j(x) \quad \forall x \in \Omega$$

and, since for all $x \in \Omega$, $g(x) = \theta$, $f(x) \leq \theta$ and $\mu^* \geq \theta$,

$$f^0(x^*) \leq f^0(x) \quad \forall x \in \Omega \ .$$

■

Exercise 5.27 *Under the assumptions of the previous theorem, if f^0 is strictly convex, then x^* is the unique global minimizer for (P).*

Remark 5.23 Our assumptions require that g be affine (not just convex). In fact, what we really need is that $\ell(x)$ be convex and this might not hold if g is merely convex. For example $\{(x, y) \ : \ x^2 - y = 0\}$ is obviously not convex.

5.6 Mixed Constraints – Second order Conditions

Theorem 5.14 *(necessary condition). Suppose that x^* is a local minimizer for (5.42) and suppose that $\{\nabla f^j(x^*), j \in J(x^*)\} \cup \{\nabla g^j(x^*), \ j = 1, \ldots, \ell\}$ is a linearly independent set of vectors. Then there exist $\mu^* \in \mathbf{R}^n$ and $\lambda^* \in \mathbf{R}^\ell$ such that the KKT conditions hold and*

$$h^T \nabla^2_{xx} L(x^*, \mu^*, \lambda^*) h \geq 0 \quad \forall h \in \{h \ : \ \frac{\partial g}{\partial x}(x^*) h = \theta, \ \nabla f^j(x^*)^T h = 0 \quad \forall j \in J(x^*)\} \quad (5.44)$$

with $L(x, \mu, \lambda) \stackrel{\Delta}{=} f^0(x) + \sum_{j=1}^{m} \mu^j f^j(x) + \sum_{j=1}^{\ell} \lambda^j g^j(x)$.

Proof. It is clear that x^* is also a local minimizer for the problem

$$\text{minimize } f^0(x) \quad \text{s.t.} \qquad \begin{array}{l} g(x) = \theta \\ f^j(x) = 0 \quad \forall j \in J(x^*) \ . \end{array}$$

The claims is then a direct consequence of the second order necessary condition of optimality for equality constrained problems. ■

Remark 5.24

1. The linear independence assumption is more restrictive than KTCQ (and is more restrictive than necessary; see related comment in connection with Theorem 5.5). It insures *uniqueness* of the KKT multipliers (prove it).

2. Intuition may lead to believe that (5.44) should hold for all h in the larger set

$$S_{f,g}(x^*) := \{h \ : \ \frac{\partial g}{\partial x}(x^*)h = \theta, \nabla f^j(x^*)^T h \leq 0 \quad \forall j \in J(x^*)\} \ .$$

The following scalar example shows that this is not true:

$$\min\{\log(x) \ : \ x \geq 1\} \qquad \text{(with } x \in \mathbf{R}\text{)}$$

(Note that a *first order* sufficiency condition holds for this problem: see Exercises 5.23 and 5.28)

■

There are a number of (non-equivalent) second-order sufficient conditions (SOSCs) for problems with mixed (or just inequality) constraints. The one stated below strikes a good trade-off between power and simplicity.

Theorem 5.15 *(SOSC with strict complementarity). Suppose that $x^* \in \mathbf{R}^n$ is such that*

(i) KKT conditions (see Theorem 5.11) hold with μ^*, λ^* as multipliers, and $\mu^{*j} > 0 \quad \forall j \in J(x^*)$.

(ii) $h^T \nabla^2_{xx} L(x^*, \mu^*, \lambda^*)h > 0 \quad \forall h \in \{h \neq \theta \ : \ \frac{\partial g}{\partial x}(x^*)h = \theta, \nabla f^j(x^*)^T h = 0 \quad \forall j \in J(x^*)\}$

Then x^* is a strict local minimizer.

Proof. See [22].

Remark 5.25 Without strict complementarity, the result is not valid. (*Example*: $\min\{x^2 - y^2 \ : \ y \geq 0\}$ with $(x^*, y^*) = (0, 0)$ which is a KKT point but not a local minimizer.) An alternative (stronger) condition is obtained by dropping the strict complementarity assumption but replacing in (ii) $J(x^*)$ with its subset $I(x^*) := \{j \ : \ \mu^{*^j} > 0\}$, the set of indices of binding constraints. Notice that if $\mu^{*^j} = 0$, the corresponding constraint does not enter the KKT conditions. Hence, if such "non-binding" constraints is removed, x^* will still be a KKT point.

Exercise 5.28 *Show that the second order sufficiency condition still holds if condition (ii) is replaced with*

$$h^T \nabla^2_{xx} L(x^*, \mu^*, \lambda^*)h > 0 \quad \forall h \in S_{f,g}(x^*) \backslash \{0\} \ .$$

Find an example where this condition holds while (ii) does not.

This condition is overly strong: see example in Remark 5.24.

Remark 5.26 If $\text{sp}\{\nabla f^j(x^*) \ : \ j \in I(x^*)\} = \mathbf{R}^n$, then condition (iii) in the sufficient condition holds trivially, yielding a *first order sufficiency condition.* (Relate this to Exercise 5.23.)

5.7 Glance at Numerical Methods for Constrained Problems

Penalty functions methods

$$(P) \qquad \min\{f^0(x) \ : \ f(x) \leq 0, \ g(x) = \theta\}$$

The idea is to replace (P) by a sequence of unconstrained problems

$$(P_i) \qquad \min \phi^i(x) \equiv f^0(x) + c_i P(x)$$

with $P(x) = ||g(x)||^2 + ||f(x)_+||^2$, $(f(x)_+)^j = \max(0, f^j(x))$, and where c_i grows to infinity. Note that $P(x) = 0$ if and only if $x \in \Omega$ (feasible set). The rationale behind the method is that, if c_i is large, the *penalty term* $P(x)$ will tend to push the solution x_i (P_i) towards the feasible set. The norm used in defining $P(x)$ is arbitrary, although the Euclidean norm has clear computational advantages ($P(x)$ continuously differentiable: this is the reason for squaring the norms).

Exercise 5.29 *Show that if $||\cdot||$ is any norm in $\mathbf{R}^n$, $||\cdot||$ is not continuously differentiable at θ.*

First, let us suppose that each P_i can be solved for a global minimizer x_i (conceptual version).

Theorem 5.16 *Suppose $x_i \overset{K}{\to} \hat{x}$, then $\hat{x}$ solves (P).*

Exercise 5.30 *Prove Theorem 5.16*

As pointed out earlier, the above algorithm is purely conceptual since it requires *exact* computation of a *global* minimizer for each i. However, using one of the algorithms previously studied, given $\epsilon_i > 0$, one can construct a point x_i satisfying

$$||\nabla \phi^i(x_i)|| \leq \epsilon_i, \tag{5.45}$$

by constructing a sequence $\{x^j\}$ such that $\nabla \phi^i(x^j) \to \theta$ as $j \to \infty$ and stopping computation when (5.45) holds. We choose $\{\epsilon_i\}$ such that $\epsilon_i \to 0$ as $i \to \infty$.
For simplicity, we consider now problems with *equality* constraints *only*.

Exercise 5.31 *Show that $\nabla \phi^i(x) = \nabla f(x) + c_i \frac{\partial g}{\partial x}(x)^T g(x)$*

Theorem 5.17 *Suppose that $\forall x \in \mathbf{R}^n$, $\frac{\partial g}{\partial x}(x)$ has full row rank. Suppose that $x_i \overset{K}{\to} x^*$. Then $x^* \in \Omega$ and $\exists \lambda^* \ni$*

$$\nabla f(x^*) + \frac{\partial g}{\partial x}(x^*)^T \lambda^* = \theta \tag{5.46}$$

i.e., the first order necessary condition of optimality holds at x^*. Moreover, $c_i g(x_i) \overset{K}{\to} \lambda^*$.

Exercise 5.32 *Prove Theorem 5.17*

Remark 5.27 The main drawback of this penalty function algorithm is the need to drive c_i to ∞ in order to achieve convergence to a solution. When c_i is large (P_i) is very difficult to solve (slow convergence and numerical difficulties due to ill-conditioning). In practice, one should compute x_i for a few values of c_i, set $\gamma_i = 1/c_i$ and define $x_i = x(\gamma_i)$, and then extrapolate for $\gamma_i = 0$ (i.e., $c_i = \infty$). Another approach is to modify (P_i) as follows, yielding the method of multipliers, due to Hestenes and Powell.

$$(P_i) \qquad \min f(x) + \frac{1}{2}c_i||g(x)||^2 + \lambda_i^T g(x) \tag{5.47}$$

where

$$\lambda_{i+1} = \lambda_i + c_i g(x_i) \tag{5.48}$$

with x_i the solution to (P_i).

It can be shown that convergence to the solution x^* can now be achieved without having to drive c_i to ∞, but merely by driving it above a certain threshold $\hat{c}$ (see [3] for details.)

■

Methods of feasible directions (inequality constraints)

For unconstrained problems, the value of the function ∇f at a point x indicates whether x is stationary and, if not, in some sense how far it is from a stationary point. In constrained optimization, a similar role is played by optimality functions. Our first optimality function is defined as follows. For $x \in \mathbf{R}^n$

$$\theta_1(x) = \min_{h \in S} \ \max\left\{\nabla f^j(x)^T h, \ j \in J_0(x)\right\} \tag{5.49}$$

with $S \stackrel{\Delta}{=} \{h \ : \ ||h|| \leq 1\}$. Any norm could be used but we will focus essentially on the Euclidean norm, which does not favor any direction. Since the "max" is taken over a finite set (thus a compact set) it is continuous in h. Since S is compact, the minimum is achieved.

Proposition 5.7 *For all $x \in \mathbf{R}^n$, $\theta_1(x) \leq 0$. Moreover, if $x \in \Omega$, $\theta_1(x) = 0$ if and only if x is a F. John point.*

Exercise 5.33 *Prove Proposition 5.7*

We thus have an optimality function through which we can identify F. John points. Now suppose that $\theta_1(x) < 0$ (hence x is not a F. John point) and let $\hat{h}$ be a minimizer in (5.49). Then $\nabla f^j(x)^T \hat{h} < 0$ for all $j \in J_0(x)$, i.e., $\hat{h}$ is a descent direction for the cost function and all the active constraints, i.e., a *feasible descent* direction. A major drawback of $\theta_1(x)$ is its lack of continuity, due to jump in the set $J_0(x)$ when x hits a constraint boundary. Hence $|\theta_1(x)|$ may be large even if x is very close to a F. John point. This drawback is avoided by the following optimality function.

$$\theta_2(x) = \min_{h \in S} \max\left\{\nabla f^0(x)^T h; \ f^j(x) + \nabla f^j(x),^T h, \ j = 1, \ldots, m\right\} \tag{5.50}$$

Exercise 5.34 *Show that $\theta_2(x)$ is continuous.*

Proposition 5.8 *Suppose $x \in \Omega$. Then $\theta_2(x) \leq 0$ and, moreover, $\theta_2(x) = 0$ if and only if x is a F. John point.*

Exercise 5.35 *Prove Proposition 5.8*

Hence θ_2 has the same properties as θ_1 but, furthermore, it is continuous. A drawback of θ_2, however, is that its computation requires evaluation of the *gradients* of *all* the constraints, as opposed to just those of the active constraints for θ_1.

We will see later that computing θ_1 or θ_2, as well as the minimizing h, amounts to solving a quadratic program, and this can be done quite efficiently. We now use $\theta_2(x)$ in the following optimization algorithm, which belongs to the class of methods of feasible directions.

Algorithm (method of feasible directions)
Parameters $\alpha, \beta \in (0,1)$
Data $x_0 \in \Omega$
$i = 0$
while $\theta_2(x_i) \neq 0$ do {
  obtain $h_i = h(x_i)$, minimizer in (5.50)
  $k = 0$
  repeat {
    if $(f^0(x_i + \beta^k h_i) - f(x_i) \leq \alpha\beta^k \nabla f(x_i)^T h$ & $f^j(x_i + \beta^k h_i) \leq 0$ for $j = 1, 2, \ldots, m)$
  then break
    $k = k + 1$
    }
  forever
  $x_{i+1} = x_i + \beta^k h_i$
  $i = i + 1$
  }
stop

We state without proof a corresponding convergence theorem (the proof is essentially patterned after the corresponding proof in the unconstrained case).

Theorem 5.18 *Suppose that $\{x_i\}$ is constructed by the above algorithm. Then $x_i \in \Omega \quad \forall i$ and $x_i \xrightarrow{K} \hat{x}$ implies $\theta_2(\hat{x}) = 0$ (i.e. $\hat{x}$ is a F. John point).*

Note: The proof of this theorem relies crucially on the continuity of $\theta_2(x)$.

Newton's method for constrained problems
Consider again the problem

$$\min\{f(x) \ : \ g(x) = 0\} \tag{5.51}$$

We know that, if x^* is a local minimizer for (5.51) and $\frac{\partial g}{\partial x}(x^*)$ has full rank, then $\exists \lambda \in \mathbf{R}^m$ such that

$$\begin{cases} \nabla_x L(x^*, \lambda^*) & = 0 \\ g(x^*) & = 0 \end{cases}$$

which is a system of $n+m$ equations with $n+m$ unknowns. We can try to solve this system using Newton's method. Define $z = (x, \lambda)$ and

$$F(z) = \begin{pmatrix} \nabla_x L(x, \lambda) \\ g(x) \end{pmatrix} .$$

We have seen that, in order for Newton's method to converge locally (and quadratically), it is sufficient that $\frac{\partial F}{\partial z}(z^*)$ be non singular, with $z^* = (x^*, \lambda^*)$, and that $\frac{\partial F}{\partial z}$ be Lipschitz continuous around z^*.

Exercise 5.36 *Suppose that 2nd order sufficiency conditions of optimality are satisfied at x^* and, furthermore, suppose that $\frac{\partial g}{\partial x}(x^*)$ has full rank. Then $\frac{\partial F}{\partial z}(z^*)$ is non singular.*

As in the unconstrained case, it will be necessary to "stabilize" the algorithm in order to achieve convergence from any initial guess. Again such stabilization can be achieved by

- using a suitable step-size rule
- using a composite algorithm.

Sequential quadratic programming
This is an extension of Newton's method to the problem

$$\min\{f^0(x) \ : \ g(x) = 0, f(x) \leq 0\} . \qquad (P)$$

Starting from an estimate (x_i, μ_i, λ_i) of a KKT triple, solve the following minimization problem:

$$\min_v \{\nabla_x L(x_i, \mu_i, \lambda_i)^T v + \frac{1}{2} v^T \nabla^2_{xx} L(x_i, \mu_i, \lambda_i) v \quad : \\ g(x_i) \quad + \quad \frac{\partial g}{\partial x}(x_i) v = 0, \ f(x_i) + \frac{\partial f}{\partial x}(x_i) v \leq 0\} \ (P_i)$$

i.e., the constraints are linearized and the cost is quadratic (but is an approximation to L rather than to f^0). (P_i) is a quadratic program (we will discuss those later) since the cost is quadratic (in v) and the constraints linear (in v). It can be solved exactly and efficiently. Let us denote by v_i its solution and ξ_i, η_i the associated multipliers. The next iterate is

$$\begin{aligned} x_{i+1} &= x_i + v_i \\ \mu_{i+1} &= \xi_i \\ \lambda_{i+1} &= \eta_i \end{aligned}$$

Then (P_{i+1}) is solved as so on. Under suitable conditions (including second order sufficiency condition at x^* with multipliers μ^*, λ^*), the algorithm converges locally quadratically if (x_i, μ_i, λ_i) is close enough to (x^*, μ^*, λ^*). As previously, $\frac{\partial^2 L}{\partial x^2}$ can be replaced by an estimate, e.g., using an update formula. It is advantageous to keep those estimates positive definite. To stabilize the methods (i.e., to obtain global convergence) suitable step-size rules are available.

Exercise 5.37 *Show that, if $m = 0$ (no inequality) the iteration above is identical to the Newton iteration considered earlier.*

5.8 Sensitivity

An important question for practical applications is to know what would be the effect of slightly modifying the constraints, i.e., of solving the problem

$$\min\{f^0(x) \ : \ g(x) = b_1, \ f(x) \leq b_2\} \tag{5.52}$$

with the components of b_1 and b_2 being small. Specifically, given a (local) minimizer x^* for the problem when $b := (b_1, b_2) = \theta$, (i) does there exist $\epsilon > 0$ such that, whenever $\|b\| \leq \epsilon$, the problem has a local minimizer $x(b)$ which is "close" to x^*? and (ii) if such ϵ exists, what can be said about the "value function" $V : B(\theta, \epsilon) \to \mathbf{R}$ given by

$$V(b) := f_0(x(b));$$

in particular, how about $\nabla V(\theta)$?
For simplicity, we first consider the case with equalities only:

$$\min\{f^0(x) \ : \ g(x) = b\}. \tag{5.53}$$

When $b = \theta$, the first-order conditions of optimality are

$$\left.\begin{array}{rl} \nabla f^0(x) + \frac{\partial g}{\partial x}(x)^T \lambda & = \theta_n \\ g(x) - b & = \theta_\ell \end{array}\right\}. \tag{5.54}$$

By hypothesis there is a solution x^*, λ^* to (5.54) when $b = \theta_\ell$. Consider now the left-hand side of (5.54) as a function $F(x, \lambda, b)$ of x, λ and b and try to solve (5.54) locally for x and λ using IFT (assuming f^0 and g are twice continuously Fréchet-differentiable)

$$\frac{\partial F}{\partial \begin{pmatrix} x \\ \lambda \end{pmatrix}}(x^*, \lambda^*, \theta_\ell) = \begin{bmatrix} \nabla^2_{xx} L(x^*, \lambda^*) & \frac{\partial g}{\partial x}(x^*)^T \\ \frac{\partial g}{\partial x}(x^*) & 0 \end{bmatrix} \tag{5.55}$$

which was shown to be non-singular, in Exercise 5.36, under the same hypotheses. Hence $x(b)$ and $\lambda(b)$ are well defined and continuously differentiable for $||b||$ small enough and they form a KKT pair for (5.53) since they satisfy (5.54). Furthermore, by continuity, they still satisfy the 2nd order sufficiency conditions (check this) and hence, $x(b)$ is a strict local minimizer for (5.53). Finally, by the chain rule, we have

$$\begin{aligned} \frac{\partial V}{\partial b}(\theta_\ell) &= \frac{\partial f^0}{\partial x}(x^*)\frac{\partial x}{\partial b}(\theta_\ell) \\ &= -\lambda^{*^T}\frac{\partial g}{\partial x}(x^*)\frac{\partial x}{\partial b}(\theta_\ell). \end{aligned}$$

where we have invoked (5.54). But by the second equality in (5.54)

$$\frac{\partial g}{\partial x}(x(b))\frac{\partial x(b)}{\partial b} = \frac{\partial g(x(b))}{\partial b} = \frac{\partial}{\partial b} b = I,$$

in particular,

$$\frac{\partial g}{\partial x}(x^*)\frac{\partial x}{\partial b}(\theta_\ell) = I.$$

Hence $\nabla V(\theta_\ell) = -\lambda^*$. ■

We now state the corresponding theorem for the general problem (5.52). Its proof follows the above. Strict complementarity is needed in order for $\mu(b)$ to still satisfy $\mu(b) \geq 0$.

Theorem 5.19 *Consider the family of problems (5.52) where f^0, f and g are twice continuously differentiable. Suppose that for $b = (b_1, b_2) = \theta_{\ell+m}$, there is a local solution x^* such that $S \stackrel{\Delta}{=} \{\nabla g^j(x^*),\ j = 1, 2, \ldots, \ell\} \cup \{\nabla f^j(x^*),\ j \in J(x^*)\}$ is a linearly independent set of vectors. Suppose that, together with the multipliers $\mu^* \in \mathbf{R}^m$ and $\lambda^* \in \mathbf{R}^\ell$, x^* satisfies SOSC with strict complementarity. Then there exists $\epsilon > 0$ such that for all $b \in B(\theta_\ell, \epsilon)$ there exists $x(b)$, $x(\cdot)$ continuously differentiable, such that $x(\theta_\ell) = x^*$ and $x(b)$ is a strict local minimizer for (5.52). Furthermore*

$$\nabla V(\theta_{\ell+m}) = [-\lambda^*; -\mu^*].$$

■

The idea of the proof is similar to the equality case. Strict complementarity is needed in order for $\mu(b)$ to still satisfy $\mu(b) \geq \theta$.

Exercise 5.38 *In the theorem above, instead of considering (5.55), one has to consider the matrix*

$$\frac{\partial F}{\partial \begin{pmatrix} x \\ \lambda \\ \mu \end{pmatrix}}(x^*, \lambda^*, \mu^*, \theta_{\ell+m}) = \begin{bmatrix} \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) & \frac{\partial g}{\partial x}(x^*)^T & \nabla f^1(x^*) & \cdots & \nabla f^m(x^*) \\ \frac{\partial g}{\partial x}(x^*) & & & & \\ \mu_1^* \nabla f^1(x^*)^T & & f_1(x^*) & & \\ \vdots & & & \ddots & \\ \mu_m^* \nabla f^m(x^*)^T & & & & f_m(x^*) \end{bmatrix}.$$

Show that, under the assumption of the theorem above, this matrix is non-singular. (Hence, again, one can solve locally for $x(b)$, $\lambda(b)$ and $\mu^j(b)$.)

Remark 5.28 Equilibrium price interpretation. For simplicity, consider a problem with a single constraint,

$$\min\{f^0(x) \ : \ f(x) \leq \theta\} \text{ with } f \ : \ \mathbf{R}^n \to \mathbf{R}.$$

(Extension to multiple constraints is straightforward.) Suppose that, at the expense of paying a price of p per unit of b, the producer can, by acquiring some amount of b, replace the inequality constraint used above by the less stringent

$$f(x) \leq b \quad , \quad b > 0$$

(conversely, he/she will save p per unit if $b < 0$), for a total additional cost to the producer of pb. From the theorem above, the resulting savings will be, to the first order

$$f^0(x(0)) - f^0(x(b)) \simeq -\frac{d}{db} f(x(0))b = \mu^* b$$

Hence, if $p < \mu^*$, it is to the producers' advantage to relax the constraint by relaxing the right-hand side, i.e., by acquiring some additional amount of b. If $p > \mu^*$, to the contrary, s/he can save by tightening the constraint. If $p = \mu^*$, neither relaxation nor tightening yields any gain (to first order). Hence μ^* is called the *equilibrium price*.

■

Note. As seen earlier, the linear independence condition in the theorem above (linear independence of the gradients of the active constraints) insures *uniqueness* of the KKT multipliers λ^* and μ^*. (Uniqueness is obviously required for the interpretation of μ^* as "sensitivity".)

Exercise 5.39 *Discuss the more general case*

$$\min\{f^0(x) \ : \ g(x, b_1) = \theta_\ell, f(x, b_2) \leq \theta_m\}.$$

5.9 Duality

See [32]. Most results given in this section do not require any differentiability of the objective and constraint functions. Also, some functions will take values on the *extended real line* (including $\pm\infty$). The crucial assumption will be that of *convexity*. The results are global.

We consider the inequality constrained problem

$$\min\{f^0(x) \ : \ f(x) \leq 0, \ x \in X\} \quad (P)$$

with $f^0 \ : \ \mathbf{R}^n \to \mathbf{R}$, $f \ : \ \mathbf{R}^n \to \mathbf{R}^m$ and X is a given subset of $\mathbf{R}^n$ (e.g., $X = \mathbf{R}^n$). As before, we define the Lagrangian function by

$$L(x, \mu) = f^0(x) + \sum_{j=1}^{m} \mu^j f^j(x)$$

Exercise 5.40 *(sufficient condition of optimality; no differentiability or convexity assumed). Suppose that $(x^*, \mu^*) \in X \times \mathbf{R}^m$ is such that $\mu^* \geq 0$ and*

$$L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*) \quad \forall \mu \geq 0, \ x \in X, \tag{5.56}$$

i.e., (x^*, μ^*) is a saddle point for L. Then x^* is a global minimizer for (P). (In particular, it is feasible for (P).)

■

We will see that under assumptions of convexity and a certain constraint qualification, the following converse holds: if x^* solves (P) then $\exists \mu^* \geq 0$ such that (5.56) holds. As we show below (Proposition 5.9), the latter is equivalent to the statement that

$$\min_{x \in X} \left\{ \sup_{\mu \geq 0} L(x, \mu) \right\} = \max_{\mu \geq 0} \left\{ \inf_{x \in X} L(x, \mu) \right\}$$

and that the left-hand side achieves its minimum at x^* and the right-hand side achieves its maximum at μ^*. If this holds, strong duality is said to hold. In such case, one could compute any μ^* which globally maximizes $\psi(\mu) = \inf_x L(x, \mu)$, subject to the simple constraint $\mu \geq 0$. Once μ^* is known, (5.56) shows that x^* is a minimizer of $L(x, \mu^*)$, unconstrained if $X = \mathbf{R}^n$. (Note: $L(x, \mu^*)$ may have other "spurious" minimizers, i.e., minimizers for which (5.56) does not hold; see below.)
Instead of $L(x, \mu)$, we now consider a more general function $F : \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$. First of all, "weak duality" always holds.

Lemma 5.1 *(Weak duality). Given two sets X and Y and a function $F : X \times Y \to \mathbf{R}$,*

$$\sup_{y \in Y} \inf_{x \in X} F(x, y) \leq \inf_{x \in X} \sup_{y \in Y} F(x, y) \tag{5.57}$$

Proof. We have successively

$$\inf_{x \in X} F(x, y) \leq F(x, y) \quad \forall x \in X, \ \forall y \in Y. \tag{5.58}$$

Hence, taking $\sup_{y \in Y}$ on both sides,

$$\sup_{y \in Y} \inf_{x \in X} F(x, y) \leq \sup_{y \in Y} F(x, y) \quad \forall x \in X \tag{5.59}$$

where now only x is free. Taking $\inf_{x \in X}$ on both sides (i.e., in the right-hand side) yields

$$\sup_{y \in Y} \inf_{x \in X} F(x, y) \leq \inf_{x \in X} \sup_{y \in Y} F(x, y) \tag{5.60}$$

■

In the sequel, we will make use of the following result, of independent interest.

Proposition 5.9 *Given two sets X and Y and a function $F : X \times Y \to \mathbf{R}$, the following statements are equivalent (under no regularity or convexity assumption)*

(i) $x^* \in X, \ y^* \in Y$ *and*

$$F(x^*, y) \leq F(x^*, y^*) \leq F(x, y^*) \quad \forall x \in X, \ \forall y \in Y$$

(ii)

$$\min_{x \in X} \{\sup_{y \in Y} F(x, y)\} = \max_{y \in Y} \{\inf_{x \in X} F(x, y)\}$$

where the common value is finite and the left-hand side is achieved at x^* and the right-hand side is achieved at y^*.

Proof. ((ii)⇒(i)) Let $\alpha = \min_x \left\{ \sup_y F(x,y) \right\} = \max_y \left\{ \inf_x F(x,y) \right\}$ and let x^* and y^* achieve the 'min' in the first expression and the 'max' in the second expression, respectively. Then

$$F(x^*,y) \leq \sup_y F(x^*,y) = \alpha = \inf_x F(x,y^*) \leq F(x,y^*) \quad \forall x,y.$$

Thus

$$F(x^*,y^*) \leq \alpha \leq F(x^*,y^*)$$

and the proof is complete.
((i)⇒(ii))

$$\inf_x \sup_y F(x,y) \leq \sup_y F(x^*,y) = F(x^*,y^*) = \inf_x F(x,y^*) \leq \sup_y \inf_x F(x,y).$$

By weak duality (see below) it follows that

$$\inf_x \sup_y F(x,y) = \sup_y F(x^*,y) = F(x^*,y^*) = \inf_x F(x,y^*) = \sup_y \inf_x F(x,y).$$

Further, the first and fourth equalities show that the "inf" and "sup" are achieved at x^* and y^*, respectively.

■

A pair (x^*,y^*) satisfying the conditions of Proposition 5.9 is referred to as a saddle point.

Exercise 5.41 *The set of saddle points of a function F is a Cartesian product, that is, if (x_1,y_1) and (x_2,y_2) are saddle points, then (x_1,y_2) and (x_2,y_1) also are. Further, F takes the same value at all its saddle points.*

Let us apply the above to $L(x,\mu)$. Thus, consider problem (P) and let $Y = \{\mu \in \mathbf{R}^m : \mu \geq 0\}$. Let $p : \mathbf{R}^n \to \mathbf{R} \cup \{+\infty\}$, and $\psi : \mathbf{R}^m \to \mathbf{R} \cup \{-\infty\}$ be given by

$$\begin{aligned} p(x) &= \sup_{\mu \geq 0} L(x,\mu) = \begin{cases} f^0(x) & \text{if } f(x) \leq 0 \\ +\infty & \text{otherwise} \end{cases} \\ \psi(\mu) &= \inf_{x \in X} L(x,\mu). \end{aligned}$$

Then (P) can be written

$$\text{minimize } p(x) \quad \text{s.t.} \quad x \in X \tag{5.61}$$

and weak duality implies that

$$\inf_{x \in X} p(x) \geq \sup_{\mu \geq 0} \psi(\mu). \tag{5.62}$$

Definition 5.10 *It is said that duality holds (or strong duality holds) if equality holds in (5.62).*

Remark 5.29 Some authors use the phrases "strong duality" or "duality holds" to means that not only there is no duality gap but furthermore the primal infimum and dual supremum are attained at some $x^* \in X$ and $\mu^* \in Y$.

If duality holds and $x^* \in X$ minimizes $p(x)$ over X and μ^* solves the dual problem

$$\text{maximize}\, \psi(\mu) \text{ s.t.} \mu \geq 0, \tag{5.63}$$

it follows that

$$L(x^*, \mu) \leq L(x^*, \mu^*) \leq L(x, \mu^*) \quad \forall x \in X,\ \forall \mu \geq 0 \tag{5.64}$$

and, in particular, x^* is a global minimizer for

$$\text{minimize } L(x, \mu^*) \quad \text{s.t.} \quad x \in X.$$

Remark 5.30 That some $\hat{x} \in X$ minimizes $L(x, \mu^*)$ over X is not sufficient for $\hat{x}$ to solve (P) (even if (P) does have a global minimizer). [Similarly, it is not sufficient that $\hat{\mu} \geq 0$ maximize $L(x^*, \mu)$ in order for $\hat{\mu}$ to solve (5.63); in fact it is immediately clear that $\hat{\mu}$ maximizing $L(x^*, \mu)$ implies nothing about $\hat{\mu}^j$ for $j \in J(x^*)$.] However, as we saw earlier, $(\hat{x}, \hat{\mu})$ satisfying both inequalities in (5.64) *is* enough for $\hat{x}$ to solve (P) and $\hat{\mu}$ to solve (5.63).

Suppose now that duality holds, i.e.,

$$\min_{x \in X} \{\sup_{\mu \geq 0} L(x, \mu)\} = \max_{\mu \geq 0} \{\inf_{x \in X} L(x, \mu)\} \tag{5.65}$$

(with the min and the max being achieved). Suppose we can easily compute a maximizer μ^* for the right-hand side. Then, by Proposition 5.9 and Exercise 5.40, there exists $x^* \in X$ such that (5.56) holds, and such x^* is a global minimizer for (P). From (5.56) such x^* is among the minimizers of $L(x, \mu^*)$. The key is thus whether (5.65) holds. This is known as (strong) duality. We first state without proof a more general result, about the existence of a saddle point for convex-concave functions. (This and other related results can be found in [6]. The present result is a minor restatement of Proposition 2.6.4 in that book.)

Theorem 5.20 *Let X and Y be convex sets and let $F : X \times Y \to \mathbf{R}$ be convex in its first argument and concave (i.e., $-F$ is convex) in its second argument, and further suppose that, for each $x \in X$ and $y \in Y$, the epigraphs of $F(\cdot, y)$ and $-F(x, \cdot)$ are closed. Further suppose that the set of minimizers of $\sup_{y \in Y} F(x, y)$ is nonempty and compact. Then*

$$\min_{x \in X} \sup_{y \in Y} F(x, y) = \max_{y \in Y} \inf_{x \in X} F(x, y).$$

We now prove this result in the specific case of our Lagrangian function L.

Proposition 5.10 *$\psi(\mu) \stackrel{\Delta}{=} \inf_x L(x, \mu)$ is concave (i.e., $-\psi$ is convex) (without convexity assumption)*

Proof. $L(x,\mu)$ is affine, hence concave, and the pointwise infimum of a set of concave functions is concave (prove it!).

■

We will now see that convexity of f^0 and f and a certain "stability" assumption (related to KTCQ) are sufficient for duality to hold. This result, as well as the results derived so far in this section, holds *without differentiability assumption*. Nevertheless, we will first prove the differentiable case with the additional assumption that $X = \mathbf{R}^n$. Indeed, in that case, the proof is immediate.

Theorem 5.21 *Suppose x^* solves (P) with $X = \mathbf{R}^n$, suppose that $f^{)}$ and f are differentiable and that KTCQ holds, and let μ^* be a corresponding KKT multiplier vector. Furthermore, suppose f^0 and f are convex functions. Then μ^* solves the dual problem, duality holds and x^* minimizes $L(x,\mu^*)$.*

Proof. Since (x^*,μ^*) is a KKT pair one has

$$\nabla f^0(x^*) + \sum_{j=1}^{m} \mu^{*^j} \nabla f^j(x^*) = 0 \tag{5.66}$$

and by complementary slackness

$$L(x^*,\mu^*) = f^0(x^*) = p(x^*).$$

Now, under our convexity assumption,

$$\ell(x) := f^0(x) + \sum_{j=1}^{m} \mu^{*^j} f^j(x) = L(x,\mu^*) \tag{5.67}$$

is convex, and it follows that x^* is a global minimum for $L(x,\mu^*)$. Hence $\psi(\mu^*) = L(x^*,\mu^*)$. It follows that μ^* solves the dual, duality holds, and, from (5.66), x^* solves

$$\nabla_x L(x^*,\mu^*) = 0$$

■

Remark 5.31 The condition above is also necessary for μ^* to solve the dual: indeed if $\hat{\mu}$ is not a KKT multiplier vector at x^* then $\nabla_x L(x^*,\hat{\mu}) \neq 0$ so that x^* does not minimize $L(\cdot,\hat{\mu})$, and $(x^*,\hat{\mu})$ is not a saddle point.

Remark 5.32

1. The only convexity assumption we used is convexity of $\ell(x) = L(x,\mu^*)$, which is weaker than convexity of f^0 and f (for one thing, the inactive constraints are irrelevant).

2. A weaker result, called local duality is as follows: If the min's in (5.61)-(5.63) are taken in a local sense, around a KKT point x^*, then duality holds with only *local* convexity of $L(x, \mu^*)$. Now, if 2nd order sufficiency conditions hold at x^*, then we know that $\nabla^2_{xx} L(x^*, \mu^*)$ is positive definite over the tangent space to the active constraints. If positive definiteness can be extended over the whole space, then local (strong) convexity would hold. This is in fact one of the ideas which led to the method of multipliers mentioned above (also called *augmented* Lagrangian).

Exercise 5.42 *Define again (see Method of Feasible Directions in section 5.7)*

$$\Theta(x) = \min_{||h||_2 \leq 1} \max\{\nabla f^j(x)^T h \ : \ j \in J_0(x)\}. \tag{5.68}$$

Using duality, show that

$$\Theta(x) = -\min\{||\sum_{j \in J_0(x)} \mu^j \nabla f^j(x)||_2 \ \Big| \ \sum_{j \in J_0(x)} \mu^j = 1, \mu^j \geq 0 \quad \forall j \in J_0(x)\} \tag{5.69}$$

and if h^* solves (5.68) and μ^* solves (5.69) then, if $\Theta(x) \neq 0$, we have.

$$h^* = \frac{-\sum_{j \in J_0(x)} \mu^{*^j} \nabla f^j(x)}{||\sum_{j \in J_0(x)} \mu^{*^j} \nabla f^j(x)||_2} \ . \tag{5.70}$$

Hint: first show that the given problem is equivalent to the minimization over $\left[\begin{smallmatrix} h \\ \tilde{h} \end{smallmatrix}\right] \in \mathbf{R}^{n+1}$

$$\min_{\substack{h \\ \tilde{h}}} \{\tilde{h} | \nabla f^j(x)^T h \leq \tilde{h} \quad \forall j \in J_0(x), \ h^T h \leq 1\} \ .$$

Remark 5.33 This shows that the search direction corresponding to Θ is the direction opposite to the nearest point to the origin in the convex hull of the gradients of the active constraints. Notice that applying duality has resulted in a problem (5.69) with fewer variables and simpler constraints than the original problem (5.68). Also, (5.69) is a *quadratic program* (see below).

We now drop the differentiability assumption on f^0 and f and merely assume that they are convex. We substitute for (P) the family of problems

$$\min\{f^0(x) \ : \ f(x) \leq b, \ x \in X\}$$

with $b \in \mathbf{R}^m$ and X a convex set, and we will be mainly interested in b in a neighborhood of the origin. We know that, when f^0 and f are continuously differentiable, the KKT multipliers can be interpreted as sensitivities of the optimal cost to variation of components of f. We will see here how this can be generalized to the non-differentiable case. When the generalization holds, duality will hold.

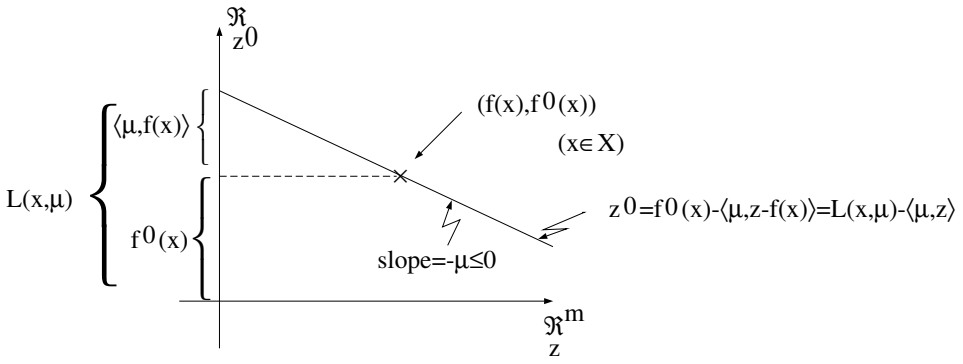


Figure **5.9**

The remainder of our analysis will take place in $\mathbf{R}^{m+1}$ where points of the form $(f(x), f^0(x))$ lie (see Figure **5.9**). We will denote vectors in $\mathbf{R}^{m+1}$ by (z, z_0) where $z^0 \in \mathbf{R}$, $z \in \mathbf{R}^m$. We also define $\bar{f} \ : \ X \to \mathbf{R}^{m+1}$ by

$$\bar{f}(x) = \begin{bmatrix} f(x) \\ f^0(x) \end{bmatrix}$$

and $\bar{f}(X)$ by

$$\bar{f}(X) = \{\bar{f}(x) \ : \ x \in X\}$$

On Figure **5.9**, the cross indicates the position of $\bar{f}(x)$ for some $x \in X$; and, for some $\mu \geq 0$, the oblique line represents the hyperplane $H_{x,\mu}$ orthogonal to $(\mu, 1) \in \mathbf{R}^{m+1}$, i.e.,

$$H_{x,\mu} = \{(z, z^0) \ : \ (\mu, 1)^T (z, z^0) = \alpha\},$$

where α is such that $\bar{f}(x) \in H_{x,\mu}$, i.e.,

$$\mu^T f(x) + f^0(x) = \alpha,$$

i.e.,

$$z^0 = L(x, \mu) - \mu^T z.$$

In particular, the oblique line intersects the vertical axis at $z^0 = L(x, \mu)$.
Next, we define the following objects:

$$\Omega(b) = \{x \in X \ : \ f(x) \leq b\},$$

$$B = \{b \in \mathbf{R}^m \ : \ \Omega(b) \neq \emptyset\},$$

$$V : \mathbf{R}^m \to \mathbf{R} \cup \{\pm\infty\}, \quad \text{with } V(z) := \inf_{x \in X} \{f^0(x) \ : \ f(x) \leq z\}.$$

V is the value function. It can take the values $+\infty$ (when $\{x \in X : f(x) \leq b\}$ is empty) and $-\infty$ (when V is unbounded from below on $\{x \in X : f(x) \leq b\}$).

Exercise 5.43 *If f^0, f are convex and X is a convex set, then $\Omega(b)$ and B and the epigraph of V are convex sets (so that V is a convex function), and V is monotonic non-increasing.*

(Note that, on the other hand, $\min\{f^0(x) : f(x) = b\}$ need not be convex in b: e.g., $f(x) = e^x$ and $f^0(x) = x$. Also, $\bar{f}(X)$ need not be convex (e.g., it could be just a curve as, for instance, if $n = m = 1$).)
The following exercise points out a simple geometric relationship between epi$V$ and $\bar{f}(X)$, which yields simple intuition for the central result of this section.

Exercise 5.44 *Prove that*

$$\text{cl}(\text{epi}V) = \text{cl}(\bar{f}(X) + \mathbf{R}^{m+1}_+),$$

where $\mathbf{R}^{m+1}_+$ *is the set of points in* $\mathbf{R}^{m+1}$ *with non-negative components. Further, if for all* z *such that* $V(z)$ *is finite, the "min" in the definition of* V *is attained, then*

$$\text{epi}V = \bar{f}(X) + \mathbf{R}^{m+1}_+.$$

This relationship is portrayed in Figure 4.10, which immediately suggests that the "lower tangent plane" to epi$V$ with "slope" $-\mu$ (i.e., orthogonal to $(\mu, 1)$, with $\mu \geq 0$), intersects the vertical axis at ordinate $\psi(\mu)$, i.e.,

$$\inf\{z^0 + \mu^T z : (z, z^0) \in \text{epi}V\} = \inf_{x \in X} L(x, \mu) = \psi(\mu) \ \forall \mu \geq 0. \tag{5.71}$$

We now provide a simple, rigorous derivation of this result.
First, the following identity is easily derived.

Exercise 5.45

$$\inf\{z^0 + \mu^T z : z^0 \geq V(z)\} = \inf\{z^0 + \mu^T z : z^0 > V(z)\}.$$

The claim now follows:

$$\begin{aligned}
\inf\{z^0 + \mu^T z : (z, z^0) \in \text{epi}V\} &= \inf_{(z,z^0) \in \mathbf{R}^{m+1}} \{z^0 + \mu^T z : z^0 \geq V(z)\} \\
&= \inf_{(z,z^0) \in \mathbf{R}^{m+1}} \{z^0 + \mu^T z : z^0 > V(z)\} \\
&= \inf_{(z,z^0) \in \mathbf{R}^{m+1}, x \in X} \{z^0 + \mu^T z : z^0 > f^0(x), f(x) \leq z\} \\
&= \inf_{(z,z^0) \in \mathbf{R}^{m+1}, x \in X} \{z^0 + \mu^T f(x)\} \\
&= \inf_{x \in X} \{f^0(x) + \mu^T f(x)\} \\
&= \inf_{x \in X} L(x, \mu) \\
&= \psi(\mu),
\end{aligned}$$

where we have used the fact that $\mu \geq 0$.
In the case of the picture, $V(0) = f^0(x^*) = \psi(\mu^*)$ i.e., duality holds and inf and sup are achieved and finite. This works because

(i) $V(\cdot)$ is convex

(ii) there exists a non-vertical supporting line (supporting hyperplane) to epi V at $(0, V(0))$. (A sufficient condition for this is that $0 \in$ int B.)

Note that if $V(\cdot)$ is continuously differentiable at 0, then $-\mu^*$ is its slope where μ^* is the KKT multiplier vector. This is exactly the sensitivity result we proved in section 5.8!
In the convex (not necessarily differentiable) case, condition (ii) acts as a substitute for KTCQ. It is known as the Slater constraint qualification (after Morton L. Slater, 20th century American mathematician). It says that there exists some feasible x at which none of the constraints is active, i.e., $f(x) < 0$. Condition (i), which implies convexity of epi V, implies the existence of a supporting hyperplane, i.e., $\exists (\mu, \mu^0) \neq 0$, s.t.

$$\mu^T 0 + \mu^0 V(0) \leq \quad \mu^T z + \mu^0 z^0 \quad \forall (z, z^0) \in \text{ epi } V$$

i.e.

$$\mu^0(V(0)) \quad \leq \quad \mu^0 z^0 + \mu^T z \quad \forall (z, z^0) \in \text{ epi } V \tag{5.72}$$

and the epigraph property of epi V ($(0, \beta) \in$ epi $V \quad \forall \beta > V(0)$) implies that $\mu^0 \geq 0$.

Exercise 5.46 *Let S be convex and closed and let $x \in \partial S$, where ∂S denotes the boundary of S. Then there exists a hyperplane H separating x and S. H is called supporting hyperplane to S at x. (Note: The result still holds without the assumption that S is closed, but the proof is harder. Hint for this harder result: If S is convex, then $\partial S = \partial \text{cl} S$.)*

Exercise 5.47 *Suppose f^0 and f^j are continuously differentiable and convex, and suppose Slater's constraint qualification holds. Then MFCQ holds at every feasible x^*.*

Under condition (ii), $\mu^0 > 0$, i.e., the supporting hyperplane is non-vertical. In particular suppose that $0 \in$ int B and proceed by contradiction: if $\mu^0 = 0$ (5.72) reduces to $\mu^T z \geq 0$ for all $(z, z^0) \in$ epi V, in particular for all z with $\|z\|$ small enough; since $(\mu^0, \mu) \neq 0$, this is impossible. Under this condition, dividing through by μ^0 we obtain:

$$\exists\ \mu^* \in \mathbf{R}^m \text{ s.t. } V(0) \leq z^0 + (\mu^*)^T z \quad \forall (z, z^0) \in \text{ epi } V\ .$$

Next, the fact that V is monotonic non-increasing implies that $\mu^* \geq 0$. (Indeed we can keep z^0 fixed and let any component of z go to $+\infty$.) Formalizing the argument made above, we now obtain, since, for all $x \in X$, $(f^0(x),\ f(x)) \in$ epi V,

$$f^0(x^*) = V(0) \leq f^0(x) + (\mu^*)^T f(x) = L(x, \mu^*) \quad \forall x \in X\ ,$$

implying that

$$f^0(x^*) \leq \inf_x\ L(x, \mu^*) = \psi(\mu^*)\ .$$

In view of weak duality, duality holds and $\sup_{\mu \geq 0} \psi(\mu)$ is achieved at $\mu^* \geq 0$.

Remark 5.34 Figure **5.10** may be misleading, as it focuses on the special case where $m = 1$ and the constraint is active at the solution (since $\mu^* > 0$). Figure 4.11 illustrates other cases. It shows $V(z)$ for $z = \gamma e_j$, with γ a scalar and e_j the jth coordinate vector, both in the case when the constraint is active and in the case it is not.

To summarize: if x^* solves (P) (in particular, $\inf_{x \in X} p(x)$ is achieved and finite) and conditions (i) and (ii) hold, then $\sup_{\mu \geq 0} \psi(\mu)$ is achieved and equal to $f^0(x^*)$ and x^* solves

$$\text{minimize } L(x, \mu^*) \quad \text{s.t.} \quad x \in X \ .$$

Note. Given a solution μ^* to the dual, there may exist x minimizing $L(x, \mu^*)$ s.t. x does not solve the primal. A simple example is given by the problem

$$\min \quad (-x) \quad \text{s.t.} \quad x \leq 0$$

where $L(x, \mu^*)$ is constant ($V(\cdot)$ is a straight line). For, e.g., $\hat{x} = 1$,

$$L(\hat{x}, \mu^*) \leq L(x, \mu^*) \quad \forall x$$

but it is not true that

$$L(\hat{x}, \mu) \leq L(\hat{x}, \mu^*) \quad \forall \mu \geq 0$$

5.10 Linear and Quadratic Programming

(See [22])
Consider the problem

$$\min\{c^T x \ : \ Gx = b_1, Fx \leq b_2\} \tag{5.73}$$

where

$$\begin{aligned} c &\in \mathbf{R}^n \\ G &\in \mathbf{R}^{m \times n} \\ F &\in \mathbf{R}^{k \times n} \end{aligned}$$

For simplicity, assume that the feasible set is nonempty and bounded. Note that if $n = 2$ and $m = 0$, we have a picture such as that on Figure **5.12**. Based on this figure, we make the following guesses

1. If a solution exists (there is one, in view of our assumptions), then there is a solution on a vertex (there may be a continuum of solutions, along an edge)

2. If all vertices "adjacent" to $\hat{x}$ have a larger cost, then $\hat{x}$ is optimal

Hence a simple algorithm would be

1. Find $x_0 \in \Omega$, a vertex. Set $i = 0$

2. If all vertices adjacent to x_i have higher cost, stop. Else, select an edge along which the directional derivative of $c^T x$ is the most negative. Let x_{i+1} be the corresponding adjacent vertex and iterate.

This is the basic idea of simplex algorithm.
In the sequel, we restrict ourselves to the following canonical form

$$\min\{c^T x \ : \ Ax = b, x \geq 0\} \tag{5.74}$$

where $c \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$

Proposition 5.11 *Consider the problem (5.73). Also consider*

$$\min\{c^T(v-w) \ : \ G(v-w) = b_1, \ F(v-w) + y = b_2, \ v \geq 0, \ w \geq 0, \ y \geq 0\}. \tag{5.75}$$

If $(\hat{v}, \hat{w}, \hat{y})$ solves (5.75) then $\hat{x} = \hat{v} - \hat{w}$ solves (5.73).

Exercise 5.48 *Prove Proposition 5.11.*

Problem (5.75) is actually of the form (5.74) since it can be rewritten as

$$\min \left\{ \begin{pmatrix} c \\ -c \\ 0 \end{pmatrix}^T \begin{pmatrix} v \\ w \\ y \end{pmatrix} \left| \begin{bmatrix} G & -G & 0 \\ F & -F & I \end{bmatrix} \begin{bmatrix} v \\ w \\ y \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{pmatrix} v \\ w \\ y \end{pmatrix} \geq 0 \right. \right\}$$

Hence, we do not lose any generality by considering (5.74) only. To give meaning to our first guess, we need to introduce a suitable notion of vertex or extreme point.

Definition 5.11 *Let $\Omega \in \mathbf{R}^n$ be a polyhedron (intersection of half spaces). Then $x \in \Omega$ is an extreme point of Ω if*

$$\left. \begin{array}{c} x = \lambda x_1 + (1-\lambda) x_2 \\ \lambda \in (0,1); \ x_1, \ x_2 \in \Omega \end{array} \right\} \Rightarrow x = x_1 = x_2$$

Proposition 5.12 *(See [22])*
Suppose (5.74) has a solution (it does under our assumptions). Then there exists x^*, an extreme point, such that x^* solves (5.74).

■

Until the recent introduction of new ideas in linear programming (Karmarkar and others), the only practical method for the solution of general linear programs was the *simplex* method. The idea is as follows:

1. obtain an extreme point of Ω, x_0, set $i = 0$.

2. if x_i is not a solution, pick among the components of x_i which are zero, a component such that its increase (x_i remaining on $Ax = b$) causes a decrease in the cost function. Increase this component, x_i remaining on $Ax = b$, until the boundary of Ω is reached (i.e., another component of x is about to become negative). The new point, x_{i+1}, is an extreme point *adjacent* to x_i with lower cost.

Obtaining an initial extreme point of Ω can be done by solving another linear program

$$\min\{\Sigma\epsilon^j \ : \ Ax - b - \epsilon = 0, \ x \geq 0, \ \epsilon \geq 0\} \tag{5.76}$$

Exercise 5.49 *Let $(\hat{x}, \hat{\epsilon})$ solve (5.76) and suppose that*

$$\min\{c^T x \ : \ Ax = b, \ x \geq 0\}$$

has a feasible point. Then $\hat{\epsilon} = 0$ and $\hat{x}$ is feasible for the given problem.

More about the simplex method can be found in [22].

Quadratic programming (See [9])
Problems with quadratic cost function and linear constraints frequently appear in optimization (although not as frequently as linear programs). We have twice met such problems when studying algorithms for solving general nonlinear problems: in some optimality functions and in the sequential quadratic programming method. As for linear programs, any quadratic program can be put into the following canonical form

$$\min\{\frac{1}{2}x^T Q x + c^T x \ : \ Ax = b, \ x \geq 0\} \tag{5.77}$$

We assume that Q is positive definite so that (5.77) (with strongly convex cost function and convex feasible set) admits at most one KKT point, which is then the global minimizer. K.T. conditions can be expressed as stated in the following theorem.

Theorem 5.22 *$\hat{x}$ solves (5.77) if and only if $\exists \psi \in \mathbf{R}^n$, $\epsilon \in \mathbf{R}^n \ni$*
$A\hat{x} = b$, $\hat{x} \geq 0$
$Q\hat{x} + c + A^T\psi - \epsilon = 0$
$\epsilon^T \hat{x} = 0$
$\epsilon \geq 0$

Exercise 5.50 *Prove Theorem 5.22.*

In this set of conditions, all relations except for the next to last one (complementary slackness) are linear. They can be solved using techniques analog to the simplex method (Wolfe algorithm; see [9]).

Figure **5.10**

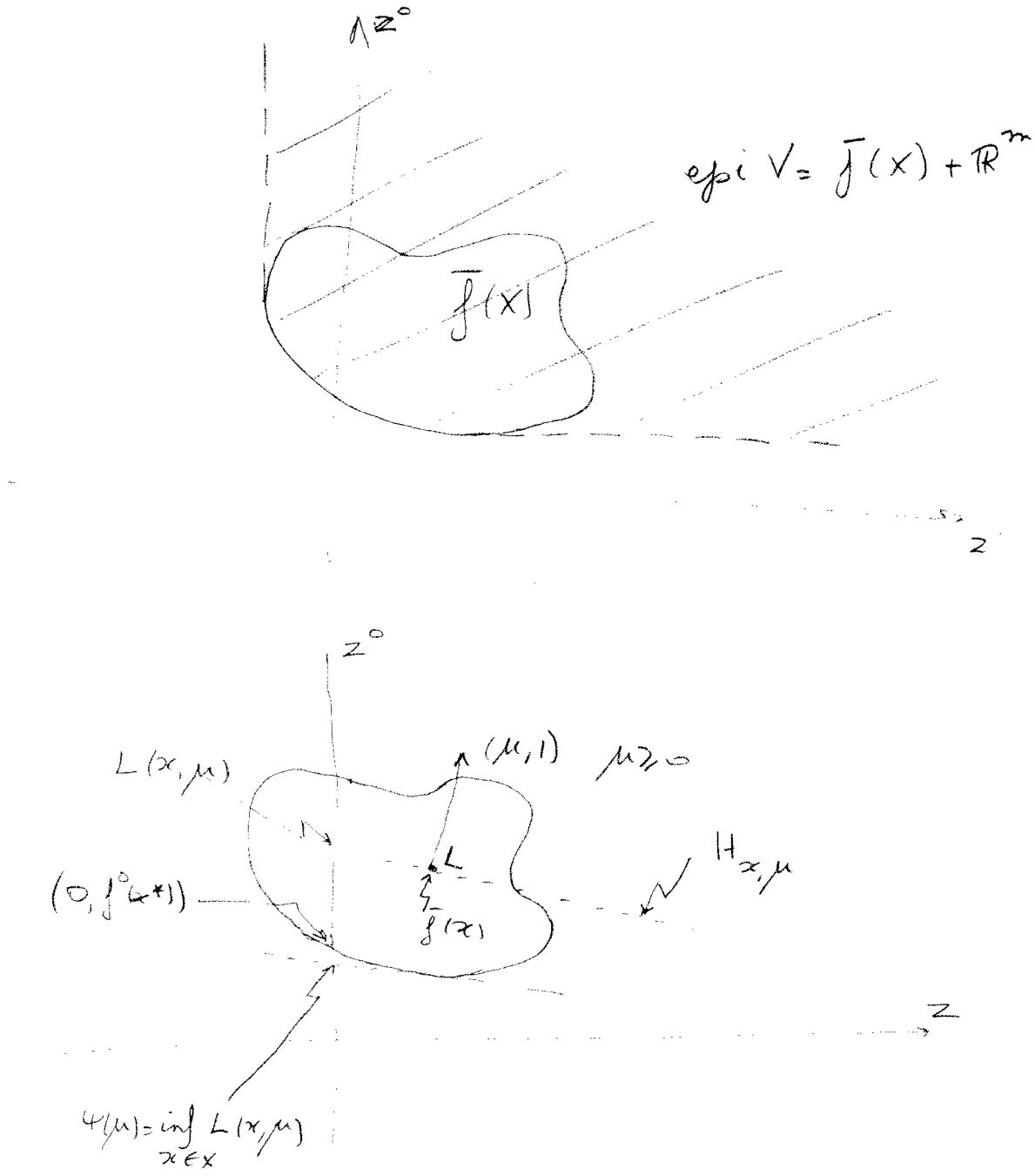


Fig. 4.10

Figure **5.11**

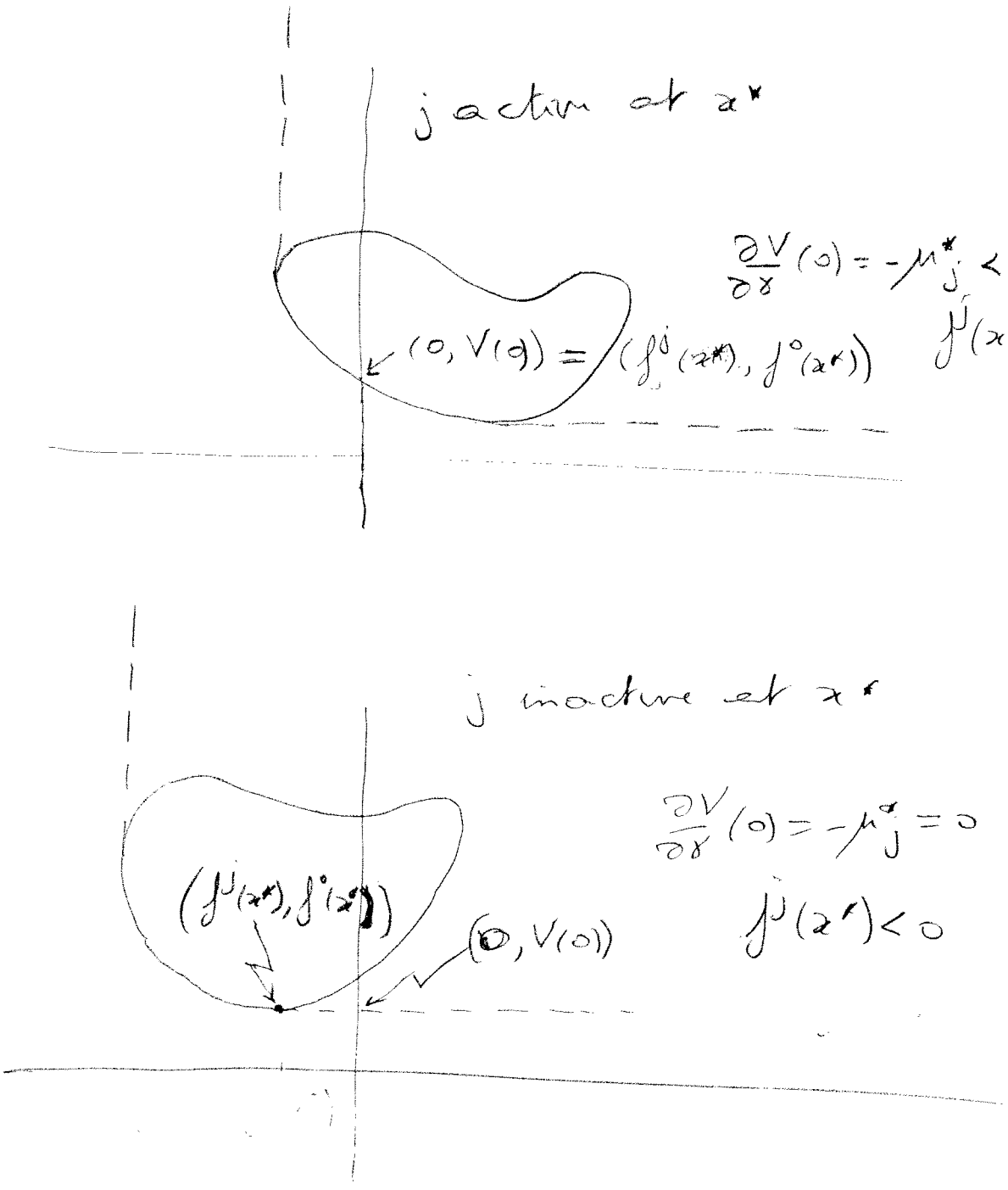


Fig. 4.11 (×××).

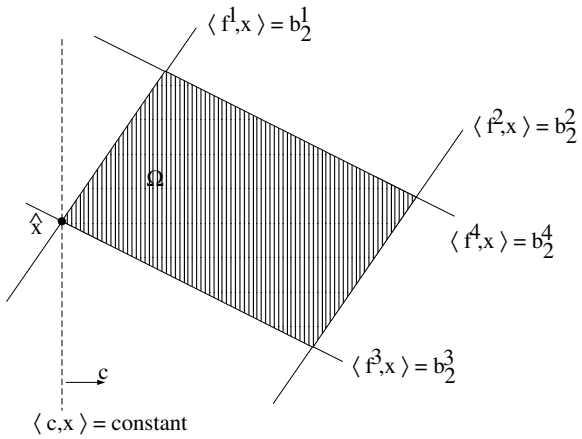


Figure **5.12**:

Chapter 6

Calculus of Variations and Pontryagin's Principle

This chapter deals with a subclass of optimization problems of prime interest to controls theorists and practitioners, that of optimal control problems, and with the classical field of calculus of variations (whose formalization dates back to the late 18th century), a precursor to continuous-time optimal control. Optimal control problem can be optimization problems in finite-dimensional spaces (like in the case of discrete-time optimal control with finite horizon), or in infinite-dimensional cases (like in the case of a broad class of continuous-time optimal control problems). In the latter case, finite-dimensional optimization ideas often still play a key role, at least when the underlying state space is finite-dimensional.

6.1 Introduction to the calculus of variations

In this section, we give a brief introduction to the classical calculus of variations, and following the discussion in [31], we show connections with optimal control and Pontryagin's Principle. (Also see [21].)
Let $X := [C^1([a,b])]^n$, a, $b \in \mathbf{R}$, let $A, B \in \mathbf{R}^n$ be given, let

$$\Omega = \{x \in X \ : \ x(a) = A, \ x(b) = B\}, \tag{6.1}$$

and let

$$J(x) = \int_a^b \mathcal{L}(t, x(t), \dot{x}(t))dt \tag{6.2}$$

where $\mathcal{L} : \mathbf{R}^n \times \mathbf{R}^n \times [a,b] \to \mathbf{R}$ is a given smooth function. $\mathcal{L}$ is typically referred to as "Lagrangian" in the calculus-of-variations literature; note that it is unrelated to the Lagrangian encountered in Chapter 5. The basic problem in the classical calculus of variations is

$$\text{minimize } J(x) \text{ s.t. } x \in \Omega. \tag{6.3}$$

Remark 6.1 Note at this point that problem (6.3) can be thought of as the "optimal control" problem

$$\text{minimize } \int_a^b \mathcal{L}(t, x(t), u(t))dt \quad \text{s.t. } \dot{x}(t) = u(t)\ \forall t,\ x(a) = A, x(b) = B,\ u \text{ continuous, } (6.4)$$

where minimization is to be carried out over the pair (x, u). The more general $\dot{x} = f(x, u)$, with moreover the values of $u(t)$ restricted to lie in some U (which could be the entire $\mathbf{R}^m$) amounts to generalizing Ω by imposing constraints on the values of $\dot{x}$ in (6.4), viz.,

$$\dot{x}(t) \in f(x(t), U)\ \forall t$$

(a "differential inclusion"), which can also be thought of as

$$\dot{x}(t) = v(t), \quad v(t) \in \tilde{U} := f(x(t), U) \quad \forall t,$$

i.e., as state-dependent constraint on the control values. E.g., $f(x, u) = \sin(u)$ with $U = \mathbf{R}^m$ yields the constraint $|\dot{x}(t)| \leq 1$ for all t. In general, such constraint are not allowed within the framework of the calculus of variations. If $f(x, U) = \mathbf{R}^n$ though (more likely to be the case when $U = \mathbf{R}^n$) then the problem does fit within (6.3), via (6.4), but requires the solution of an equation of the form $v = f(x, u)$ for u.

While the tangent cone is norm dependent (see Remark 6.2 below), the radial cone is not, so as a first approach we base our analysis on the latter. Indeed, it turns out that much can be said about this problem[1] without need to specify a norm on X.

Exercise 6.1 *Show that, for any $x \in \Omega$,*

$$\mathrm{RC}(x, \Omega) = \{h \in X \ :\ h(a) = h(b) = \theta\}.$$

Furthermore, $\mathrm{RC}(x^*, \Omega)$ *is a subspace. Finally, for all $x \in \Omega$, $x + \mathrm{RC}(x, \Omega) = \Omega$.*

Remark 6.2 The tangent cone $\mathrm{TC}(x, \Omega)$ (which unlike the radial cone, is norm-dependent) may be larger that the radial cone, depending on the norm. In particular, with the L_p norm, $p \in [1, \infty)$, $\mathrm{TC}(x, \Omega)$ is the entire space X, as we show next, assuming $a = 0$ and disregarding the constraint $x(b) = B$, for simplicity. Indeed, let $h \in X$ and w.l.o.g (since $\mathrm{TC}(x, \Omega)$ is a cone) assume $h(0) = 1$, and let $\varphi_p(\alpha, t)$ be defined by

$$\varphi_p(\alpha, t) = -\alpha \exp\left(-\frac{t}{\alpha^{2p+1}}\right).$$

Then $x + \alpha h + \varphi_p(\alpha, \cdot) \in \Omega$, and $\varphi_p(\alpha, \cdot)$ is a little-o function, since

$$\|\varphi_p(\alpha, \cdot)\|_p = \frac{\alpha^{\frac{p+1}{p}}}{p},$$

[1] Even concerning "weak" local minimizers (see Remark 4.2) whose definition, in contrast to that of local minimizers, does not rely on an underlying norm. In the sequel though, we focus on global minimizers.

In the sequel, prompted by the connection pointed out above with optimal control, we denote by $\frac{\partial \mathcal{L}}{\partial u}$ the derivative of $\mathcal{L}$ with respect to its second argument. It is readily established that J is G-differentiable. (Recall that, unlike F-differentiability, G-differentiability is independent of the norm on the domain of the function.)

Exercise 6.2 *Show that J is G-differentiable, with derivative $\frac{\partial J}{\partial x}$ given by*

$$\frac{\partial J}{\partial x}(x(\cdot))h = \int_a^b \left(\frac{\partial \mathcal{L}}{\partial x}(t, x(t), \dot{x}(t))h(t) + \frac{\partial \mathcal{L}}{\partial u}(t, x(t), \dot{x}(t))\dot{h}(t) \right) dt \quad \forall x, h \in X.$$

[Hint: First investigate directional differentiability, recalling that if J is differentiable then $\frac{\partial J}{\partial x}(x)h$ is the directional derivative of J at x in direction h.]

Since $\mathrm{RC}(x^*, \Omega)$ is a subspace, we know that, if x^* is a minimizer for J, then we must have $\frac{\partial}{\partial x} J(x^*)h = \theta$ for all $h \in \mathrm{RC}(x^*, \Omega)$. The following key result follows.

Proposition 6.1 *If $x^* \in \Omega$ is optimal for problem (6.3), then*

$$\int_a^b \left(\frac{\partial \mathcal{L}}{\partial x}(t, x^*(t), \dot{x}^*(t))h(t) + \frac{\partial \mathcal{L}}{\partial u}(t, x^*(t), \dot{x}^*(t),)\dot{h}(t) \right) dt = 0 \tag{6.5}$$

$$\forall h \in X \text{ such that } h(a) = h(b) = \theta.$$

Remark 6.3 Clearly, this results also holds for "stationary" points. E.g., see below the discussion of the Principle of Least Action.

We now proceed to transform (6.5) to obtain a more manageable condition: the Euler–Lagrange equation (Leonhard Euler, Swiss mathematician, 1707–1787). The following derivation is simple but *assumes that x^** is twice continuously differentiable. (An alternative, which does not require such assumption, is to use the DuBois-Reymond Lemma: see, e.g., [20, 21].) Integrating (6.5) by parts, one gets

$$\int_a^b \frac{\partial \mathcal{L}}{\partial x}(t, x^*(t), \dot{x}^*(t))h(t)dt + \left[\frac{\partial \mathcal{L}}{\partial u}(t, x^*(t), \dot{x}^*(t))h(t) \right]_a^b - \int_a^b \frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial u}(t, x^*(t), \dot{x}^*(t)) \right) h(t)dt = 0$$

$$\forall h \in X \quad \text{such that } h(a) = h(b) = \theta \quad \text{i.e,}$$

$$\int_a^b \left(\frac{\partial \mathcal{L}}{\partial x}(t, x^*(t), \dot{x}^*(t)) - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial u}(t, x^*(t), \dot{x}^*(t)) \right) h(t)dt = 0 \tag{6.6}$$

$$\forall h \in X \quad \text{such that } h(a) = h(b) = \theta.$$

Since the integrand is continuous, the Euler–Lagrange equation follows

Theorem 6.1 *(Euler–Lagrange Equation.) If $x^* \in \Omega$ is optimal for problem (6.3) then*

$$\frac{\partial \mathcal{L}}{\partial x}(t, x^*(t), \dot{x}^*(t)) - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial u}(t, x^*(t), \dot{x}^*(t)) = \theta \qquad \forall \quad t \in [a, b] \tag{6.7}$$

Indeed, it is a direct consequence of the following result.

Exercise 6.3 *If $f : \mathbf{R} \to \mathbf{R}^n$ is continuous and if*

$$\int_a^b f(t)h(t)dt = \theta \quad \forall h \in X \text{ s.t. } h(a) = h(b) = \theta$$

then

$$f(t) = \theta \quad \forall t \in [a, b].$$

■

Remark 6.4

1. Euler–Lagrange equation is a *necessary* condition of optimality. If x^* is a local minimizer for (6.3) in some norm (in particular, $x^* \in C^1$), then it satisfies E.-L. Problem (6.3) may also have solutions which are not in C^1, not satisfying E.-L.

2. E.-L. amounts to a second order ordinary differential equation in x with 2 point boundary conditions ($x^*(a) = A, x^*(b) = B$). Existence and uniqueness of a solution are not guaranteed in general.

Example 6.1 (see [14] for details)

Among all the curves joining 2 given points (x_1, t_1) and (x_2, t_2) in $\mathbf{R}^2$, find the one which generates the surface of minimum area when rotated about the t-axis.
The area of the surface of revolution generated by rotating the curve x around the t-axis is

$$J(x(\cdot)) = 2\pi \int_{t_1}^{t_2} x(t)\sqrt{1 + \dot{x}(t)^2}dt, \tag{6.8}$$

so that

$$\frac{\partial \mathcal{L}}{\partial x}(x, \dot{x}) = \sqrt{1 - \dot{x}^2}, \quad \frac{\partial \mathcal{L}}{\partial u}(x, \dot{x}) = \frac{x\dot{x}}{\sqrt{1 - \dot{x}^2}},$$

and the Euler–Lagrange equation can be integrated to give

$$x^*(t) = C \cosh \frac{t + C_1}{C} \tag{6.9}$$

where C and C_1 are constants to be determined using the boundary conditions.

■

Exercise 6.4 *Check (6.9).*

It can be shown that 3 cases are possible, depending on the positions of (x_1, t_1) and (x_2, t_2)

1. There are 2 curves of the form (6.9) passing through (x_1, t_1) and (x_2, t_2) (in limit cases, these 2 curves are identical). One of them solves the problem (see Figure **6.1**).

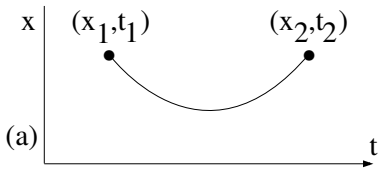


Figure **6.1**:

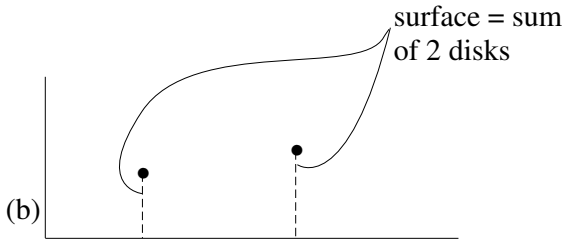


Figure **6.2**:

2. There are 2 curves of the form (6.9) passing through (x_1, t_1) and (x_2, t_2). One of them is a local minimizer. The global minimizer is as in (3) below (non smooth).

3. There is no curve of the form (6.9) passing through (x_1, t_1) and (x_2, t_2). Then there is no *smooth* curve that achieves the minimum. The solution is not C^1 and is shown in Figure **6.2** below (it is even not continuous).

Various extensions
The following classes of problems can be handled in a similar way as above.

(1) Variable end points

– $x(a)$ and $x(b)$ may be unspecified, as well as either a or b (free time problems)

– some of the above may be constrained without being fixed, e.g.,

$$g(x(a)) \leq 0$$

(2) Isoperimetric problems: One can have constraints of the form

$$K(x) \stackrel{\Delta}{=} \int_a^b G(x(t), \dot{x}(t), t)dt = \text{given constant}$$

For more detail, see [20, 21].

Theorem 6.2 *(Legendre second-order condition. Adrien-Marie Legendre, French mathematician, 1752–1833.) If $x^* \in \Omega$ is optimal for problem (6.3), then*

$$\nabla_u^2 \mathcal{L}(t, x^*(t), \dot{x}^*(t)) \succeq 0 \quad \forall t \in [a, b].$$

Again, see [20, 21] for details.

Toward Pontryagin's Principle (see [31])

Along the lines of like sub-section in section 2.1.1, but adapted to problem (6.4) (more general nonlinear objective but simpler dynamics), define the pre-Hamiltonian $H : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}$ by

$$H(t, x, p, u) := -\mathcal{L}(t, x, u) + p^T u \tag{6.10}$$

and, given a minimizer u^* and corresponding state trajectory x^* for problem (6.4) (with $\dot{x}^* = u^*$), let

$$p^*(t) := \nabla_u \mathcal{L}(t, x^*(t), u^*(t)) \quad \forall t, \tag{6.11}$$

Then the Euler–Lagrange equation yields

$$\dot{p}^*(t) = \nabla_x \mathcal{L}(t, x^*(t), u^*(t)) = -\nabla_x H(t, x^*(t), p^*(t), u^*(t)) \quad \forall t. \tag{6.12}$$

Next, since $\nabla_p H(t, x, p, u) = u$, and since $u^* = \dot{x}^*$, we have

$$\dot{x}^*(t) = \nabla_p H(t, x^*(t), p^*(t), \dot{x}^*(t)) \quad \forall t. \tag{6.13}$$

Further,

$$\nabla_u H(t, x^*(t), p^*(t), u^*(t)) = -p^*(t) + p^*(t) = 0 \quad \forall t. \tag{6.14}$$

Finally, Legendre's second-order condition yields

$$\nabla_u^2 H(t, x^*(t), p^*(t), \dot{x}^*(t)) \preceq 0 \quad \forall t. \tag{6.15}$$

Equations (6.14)-(6.13)-(6.12)-(6.15), taken together, are very close to Pontryagin's Principle applied to problem (6.4) (fixed initial and terminal states). The only missing piece is that, instead of (recall that $\dot{x}^*(t) = u^*(t)$)

$$H(t, x^*(t), p^*(t), u^*(t)) = \max_{v \in \mathbf{R}^n} H(t, x^*(t), p^*(t), v), \tag{6.16}$$

we merely have (6.14) and (6.15), which are necessary conditions for $u^*(t)$ to be such maximizer.

Exercise. Verify that this (in particular equation (6.11)) is consistent with the exercise immediately following Exercise 2.17 (Pontryagin's Principle for the case of linear dynamics and fixed terminal state).

Exercise. Reconcile definition (6.11) of $p^*(t)$ (in the case when the dynamics is $\dot{x} = u$) with definition (3.13) used in the context of dynamic programming. [Hint: Invoke (HJB).]

Remark 6.5 Condition (2.3) can be written as

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)),$$

where $\mathcal{H}$ is the Hamiltonian, which in the present case is also the Legendre-Fenchel transform of the Lagrangian $\mathcal{L}(t, x, \cdot)$, viz.

$$\mathcal{H}(t, x, p) = \sup_v \{p^T v - \mathcal{L}(t, x, v)\}.$$

also known as convex conjugate of $\mathcal{L}(t, x, \cdot)$. (Indeed, as the supremum of linear functions, the Legendre-Fenchel transform is convex—in our context, in p.)

Remark 6.6 In view of Remark 6.1 and with equations (6.14)-(6.13)-(6.12)-(6.15) in hand, it is tempting to conjecture that, subject to a simple modification, such "maximum principle" still holds in much more general cases, when $\dot{x} = u$ is replaced by $\dot{x} = f(x, u)$ and $u(t)$ is constrained to lie in a certain set U for all t: Merely replace $p^T u$ by $p^T f(x, u)$ in the definition (6.10) of H and, in (6.16), replace the unconstrained maximization by one over U. This intuition turns out to be essentially correct indeed, as we will see below (and as we have already seen, in a limited context, in section 2.3).

Connection with classical mechanics

Classical mechanics extensively refers to a "Hamiltonian" which is the total energy in the system. This quantity can be linked to the above as follows. (See [11, Section 1.4] for additional insight.)

Consider an isolated mechanical system and denote by $x(t)$ the vector of its position (and angle) variables. According to Hamilton's Principle of Least Action (which should be more appropriately called Principle of Stationary Action), the state of such system evolves so as to annihilate the derivative of the "action" $\mathcal{S}(x)$, where $\mathcal{S} : C^1[a, b] \to \mathbf{R}$ is given by

$$\mathcal{S}(x) = \int_a^b \mathcal{L}(x(t), \dot{x}(t), t)dt,$$

where

$$\mathcal{L}(t, x(t), \dot{x}(t)) = T - V,$$

T and V being the kinetic and potential energies. Again, define

$$p^*(t) := \nabla_u \mathcal{L}(t, x^*(t), \dot{x}^*(t)). \tag{6.17}$$

In classical mechanics, the potential energy does not depend on $\dot{x}(t)$, while the kinetic energy is of the form $T = \frac{1}{2}\dot{x}(t)^T M \dot{x}(t)$, where M is a symmetric, positive definite matrix. Substituting into (6.17) yields $p^*(t) = M\dot{x}^*(t)$. Hence, from (6.10),

$$H(t, x^*(t), p^*(t), \dot{x}^*(t)) = p^*(t)^T \dot{x}(t) - \mathcal{L}(t, x(t), \dot{x}(t)) = \dot{x}^*(t) M \dot{x}^*(t) - T + V = T + V,$$

i.e., the pre-Hamiltonian evaluated along the trajectory that makes the action stationary is indeed the total energy in the system. In this context, p^* is known as the momentum. For more details, see e.g., [21].

6.2 Discrete-Time Optimal Control

(see [32])

Consider the problem (time–varying system)

$$\min J(u) := \sum_{i=0}^{N-1} \mathcal{L}(i, x_i, u_i) + \psi(x_N) \quad \text{s.t.} \tag{6.18}$$

$$x_{i+1} = x_i + f(i, x_i, u_i),\ i = 0, 1, \ldots, N-1 \text{ (dynamics)}^2$$

[2] or equivalently, $\Delta x_i \triangleq x_{i+1} - x_i = f(i, x_i, u_i)$

$$g_0(x_0) = \theta,\ h_0(x_0) \leq \theta \text{ (initial state constraints)}$$
$$g_N(x_N) = \theta,\ h_N(x_N) \leq \theta \text{ (final state constraints)}$$
$$q_i(u_i) \leq \theta,\ i = 0, \ldots,\ N-1 \text{ (control constraints)}$$

where all functions are continuously differentiable in the x's and u's, with $u_i \in \mathbf{R}^m$, $x_i \in \mathbf{R}^n$, and where $\psi : \mathbf{R}^n \to \mathbf{R}$ and all other functions are real vector-valued. To keep things simple, we will not consider trajectory constraints of the type

$$r(i, x_i, u_i) \quad \leq \theta\ .$$

The given problem can be formulated as

$$\min_z \{f^0(z) \ :\ \bar{f}(z) \leq \theta,\ \bar{g}(z) = \theta\} \tag{6.19}$$

where

$$z = \begin{bmatrix} x_0 \\ \vdots \\ x_N \\ u_0 \\ \vdots \\ u_{N-1} \end{bmatrix} \in \mathbf{R}^{(N+1)n+Nm}$$

is an augmented vector on which to optimize and

$$f^0(z) = \sum_{i=0}^{N-1} \mathcal{L}(i, x_i, u_i) + \psi(x_N)$$

$$\bar{f}(z) = \begin{bmatrix} q_0(u_0) \\ \vdots \\ q_{N-1}(u_{N-1}) \\ h_0(x_0) \\ h_N(x_N) \end{bmatrix} \qquad \bar{g}(z) = \begin{bmatrix} x_1 - x_0 - f(0, x_0, u_0) \\ \vdots \\ x_N - x_{N-1} - f(N-1, x_{N-1}, u_{N-1}) \\ g_0(x_0) \\ g_N(x_N) \end{bmatrix}$$

(the dynamics are now handled as constraints). If $\hat{z}$ is optimal, the F. John conditions hold for (6.19), i.e.,

$$\exists(\underbrace{p^0}_{\psi}, \underbrace{\lambda_0, \ldots, \lambda_{N-1}}_{q}, \underbrace{\nu_0, \nu_N}_{h}) \geq 0,\ \underbrace{p_1, \ldots, p_N}_{\text{dynamics}},\ \underbrace{\eta_0, \eta_N}_{g}, \text{ not all zero s.t.}$$

$$\frac{\partial}{\partial x_0} \Rightarrow p^0 \nabla_x \mathcal{L}(0, \hat{x}_0, \hat{u}_0) - p_1 + \frac{\partial f}{\partial x}(0, \hat{x}_0, \hat{u}_0)^T p_1 + p_0 = \theta, \tag{6.20}$$

where (note that $p_0 \in \mathbf{R}^n$ is to be distinguished from $p^0 \in \mathbf{R}$)

$$p_0 := \frac{\partial g_0}{\partial x}(\hat{x}_0)^T \eta_0 + \frac{\partial h_0}{\partial x}(\hat{x}_0)^T \nu_0, \tag{6.21}$$

$$\frac{\partial}{\partial x_i} \Rightarrow p^0\nabla_x\mathcal{L}(i,\hat{x}_i,\hat{u}_i) + p_i - p_{i+1} - \frac{\partial f}{\partial x}(i,\hat{x}_i,\hat{u}_i)^T p_{i+1} = \theta \quad i = 1,\ldots,N-1 \tag{6.22}$$

$$\frac{\partial}{\partial x_N} \Rightarrow p^0\nabla\psi(\hat{x}_N) + p_N + \frac{\partial g_N}{\partial x}(\hat{x}_N)^T\eta_N + \frac{\partial h_N}{\partial x}(\hat{x}_N)^T\nu_N = \theta \tag{6.23}$$

$$\frac{\partial}{\partial u_i} \Rightarrow p^0\nabla_u\mathcal{L}(i,\hat{x}_i,\hat{u}_i) - \frac{\partial}{\partial u}f(i,\hat{x}_i,\hat{u}_i)^T p_{i+1} + \frac{\partial}{\partial u}q_i(\hat{u}_i)^T\lambda_i = \theta \quad i = 0,\ldots,N- \tag{6.24}$$

+ complementarity slackness

$$\lambda_i^T q_i(\hat{u}_i) = 0 \qquad i = \theta,\ldots,N-1 \tag{6.25}$$

$$\nu_0^T h_0(\hat{x}_0) = 0 \qquad \nu_N^T h_N(\hat{x}_N) = \theta \tag{6.26}$$

To simplify these conditions, let us define the pre-Hamiltonian function $H : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^{1+n} \times \{0,1,\ldots,N-1\} \to \mathbf{R}$

$$H(i,x,[p^0;p],u) = -p^0\mathcal{L}(i,x,u) + p^T f(i,x,u).$$

Then we obtain, with $\tilde{p}_i := [p_0; p_i]$,

$$(6.20) + (6.22) \Rightarrow p_i = p_{i+1} + \nabla_x H(i,\hat{x}_i,\tilde{p}_{i+1},\hat{u}_i) \qquad i = 0,1,\ldots,N-1 \tag{6.27}$$

$$(6.24) + (6.25) \Rightarrow \begin{bmatrix} -\frac{\partial}{\partial u}H(i,\hat{x}_i,\tilde{p}_{i+1},\hat{u}_i) + \frac{\partial}{\partial u}q_i(\hat{u}_i)^T\lambda_i = \theta \\ \lambda_i^T q_i(\hat{u}_i) = 0 \end{bmatrix} \quad i = 0,\ldots,N-1 \tag{6.28}$$

and (6.28) is a *necessary* condition of optimality (assuming KTCQ holds) for the problem

$$\max_u H(i,\hat{x}_i,\tilde{p}_{i+1},u) \text{ s.t. } q_i(u) \le \theta,\ i = 0,\ldots,N-1$$

to have $\hat{u}_i$ as a solution. (So this is a weak version of Pontryagin's Principle.) Also, (6.21)+(6.23) +(6.26) imply the transversality conditions

$$p_0 \ \perp \ \mathcal{N}\left[\begin{array}{c} \frac{\partial g_0}{\partial x}(\hat{x}_0) \\ \frac{\partial h_0^j}{\partial x}(\hat{x}_0),\ j \text{ s.t. } h_0^j(\hat{x}_0) = 0 \end{array}\right], \tag{6.29}$$

$$p_N + p^0\nabla\psi(x_N) \ \perp \ \mathcal{N}\left[\begin{array}{c} \frac{\partial g_N}{\partial x}(\hat{x}_N) \\ \frac{\partial h_N^j}{\partial x}(\hat{x}_N),\ j \text{ s.t. } h_N^j(\hat{x}_N) = 0 \end{array}\right]. \tag{6.30}$$

Remark 6.7 Simple examples show that it is indeed the case that a strong version of Pontryagin's Principle does not hold in general, in the discrete-time case. One such example can be found, e.g., in [10, Section 4.1, Example 37]. A less restrictive condition than convexity, that still guarantees that a strong Pontryagin's principle holds, is "directional convexity"; see [10, Section 4.2].

Remark 6.8 1. If KTCQ holds for the original problem, one can set $p^0 = 1$. Then if there are no constraints on x_N (no g_N's or h_N's) we obtain that $p_N = -\nabla\psi(x_N)$.

2. If $p^0 \neq 0$, then without loss of generality we can assume $p_0 = 1$ (this amounts to scaling all multipliers).

3. If x_0 is fixed, p_0 is free (i.e, no additional information is known about p_0). Every degree of freedom given to x_0 results in one constraint on p_0. The same is true for x_N and $p_N + p^0 \nabla \psi(x_N)$.

4. The vector p_i is known as the *co-state* or *adjoint variable* (or *dual variable*) at time i. Equation (6.27) is the adjoint difference equation. Note that problems (6.31) are decoupled in the sense that the kth problem can be solved for $\hat{u}_k$ as a function of $\hat{x}_k$ and $\hat{p}_{k+1}$ only. We will discuss this further in the context of continuous-time optimal control.

Remark 6.9 If all equality constraints (including dynamics) are *affine* and the objective function and inequality constraints are *convex*, then (6.28) is a necessary and sufficient condition for the problem

$$\max_u H(i, \hat{x}_i, \tilde{p}_{i+1}, u) \text{ s.t. } q_i(u) \leq 0,\ i = 1, \ldots, N-1 \tag{6.31}$$

to have a local minimum at $\hat{u}_i$, in particular, a *true* Pontryagin Principle holds: there exist vectors $p_0, p_1, \ldots, p_N$ satisfying (6.27) (dynamics) such that $\hat{u}_i$ solves the constraint optimization problem (6.31) for $i = 1, \ldots, N$, and such that the transversality conditions hold.

Remark 6.10 More general conditions under which a true maximum principle holds are discussed in [10, Section 6.2].

Remark 6.11 (Sensitivity interpretation of p_i.) From section 5.8, we know that, under appropriate assumptions (which imply, in particular, that we can choose $p^0 = 1$), if we modify problem (6.18) by changing the ith dynamic equation (and only the ith one) from

$$-x_{i-1} - f(i-1, x_{i-1}, u_{i-1}) + x_i = \theta$$

to

$$-x_{i-1} - f(i-1, x_{i-1}, u_{i-1}) + x_i = b,$$

then, if we denote by $\hat{u}(b)$ the new optimal control, we have

$$\nabla_b J(\hat{u}(b)) \Big|_{b=\theta} = p_i.$$

Now note that changing θ to b is equivalent to introducing at time i a perturbation that replaces x_i with $x_i - b$, so that varying b is equivalent to varying x_i in the opposite direction. Consider now the problem $(P_{i,x})$, where we start at time i from state x, with value function $V(k, x)$. It follows from the above that

$$\nabla_x V(i, x_i) = -p_i,$$

the discrete-time analog to (3.13). This can be verified using the dynamic-programming approach of section 3.1. Equation (3.4)

$$V(i,x) = \mathcal{L}(i,x,\hat{u}_i) + V(i+1, x + f(i,x,\hat{u}_i)),$$

yields

$$\nabla_x V(i,x) = \nabla_x \mathcal{L}(i,x,\hat{u}_i) + \left(I + \frac{\partial f}{\partial x}(i,x,\hat{u}_i)\right) \nabla_x V(i+1, x + f(i,x,\hat{u}_i)).$$

With $x := x_i$ and p_i substituted for $\nabla_x V(i,x_i)$, we get (6.27) indeed!

6.3 Continuous-Time Optimal Control

More on optimal control of linear systems

Definition 6.1 *Let $K \subset \mathbf{R}^n$, $x^* \in K$. We say that d is the inward (resp. outward) normal to a hyperplane supporting K at x^* if $d \neq 0$ and*

$$d^T x^* \leq d^T x \quad \forall x \in K \qquad (\text{resp. } d^T x^* \geq d^T x \quad \forall x \in K)$$

Remark 6.12 Equivalently

$$d^T(x - x^*) \geq 0 \quad \forall x \in K \qquad (\text{resp. } d^T(x - x^*) \leq 0 \quad \forall x \in K)$$

i.e., from x^*, all the directions towards a point in K make with d an angle of less (resp. more) than 90°.

Proposition 6.2 *Let $u^* \in \mathcal{U}$ and let $x^*(t) = \phi(t, t_0, x_0, u^*)$. Then u^* is optimal if and only if c is the inward normal to a hyperplane supporting $K(t_\mathrm{f}, t_0, x_0)$ at $x^*(t_\mathrm{f})$ (which implies that $x^*(t_\mathrm{f})$ is on the boundary of $K(t_\mathrm{f}, t_0, x_0)$).*

Exercise 6.5 *Prove Proposition 6.2.*

Further, as seen in Corollary 2.2, at every $t \in [t_0, t_\mathrm{f}]$, $p^*(t)$ is the outward normal at $x^*(t)$ to $K(t, t_0, x_0)$.

Exercise 6.6 *(i) Assuming that U is convex, show that $\mathcal{U}$ is convex. (ii) Assuming that $\mathcal{U}$ is convex, show that $K(t_\mathrm{f}, t_0, z)$ is convex.*

Remark 6.13 It can be shown that $K(t_\mathrm{f}, t_0, z)$ is convex even if U is not, provided we enlarge $\mathcal{U}$ to include all bounded measurable functions $u : [t_0, \infty) \to U$: see Theorem 1A, page 164 of [19].

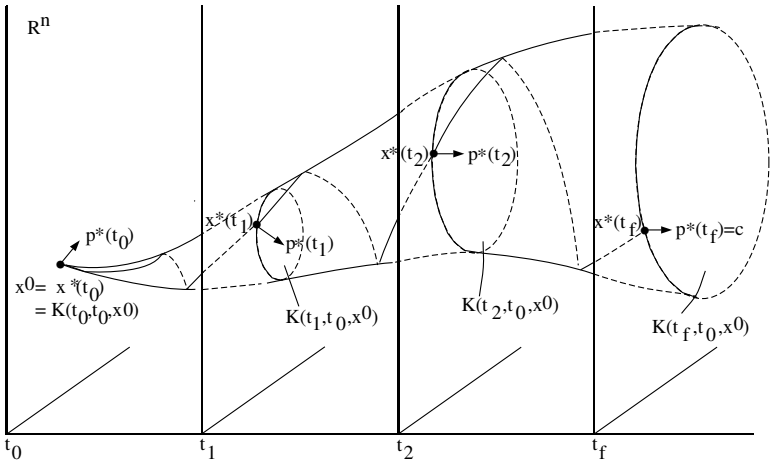


Figure **6.3**:

From (2.45), we see that, if u^* is optimal, i.e., if $c = -p^*(t_\mathrm{f})$ is the inward normal to a hyperplane supporting $K(t_\mathrm{f}, t_0, x_0)$ at $x^*(t_\mathrm{f})$ then, for $t_0 \leq t \leq t_\mathrm{f}$, $x^*(t)$ is on the boundary of $K(t, t_0, x_0)$ and $p^*(t)$ is the outward normal to a hyperplane supporting $K(t, t_0, x_0)$ at $x^*(t)$. This normal is obtained by transporting backwards in time, via the adjoint equation, the outward normal $p^*(t_\mathrm{f})$ at time t_f.

Suppose now that the objective function is of the form $\psi(x(t_\mathrm{f}))$, ψ continuously differentiable, instead of $c^T x(t_\mathrm{f})$ and suppose $K(t_\mathrm{f}, t_0, x_0)$ is convex (see remark above on convexity of $K(t, t_0, z)$).
We want to

$$\text{minimize} \quad \psi(x) \quad \text{s.t.} \quad x \in K(t_\mathrm{f}, t_0, x_0)$$

we know that if $x^*(t_\mathrm{f})$ is optimal, then

$$\nabla\psi(x^*(t_\mathrm{f}))^T h \geq 0 \quad \forall h \in \mathrm{cl}\,(\mathrm{coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, x_0)))$$

Claim: from convexity of $K(t_\mathrm{f}, t_0, x_0)$, this implies

$$\nabla\psi(x^*(t_\mathrm{f}))^T (x - x^*(t_\mathrm{f})) \geq 0 \quad \forall x \in K(t_\mathrm{f}, t_0, x_0). \tag{6.32}$$

Exercise 6.7 *Prove the claim.*

Note: Again, $\nabla\psi(x^*(t_\mathrm{f}))$ is an inward normal to $K(t_\mathrm{f}, t_0, x_0)$ at $x^*(t_\mathrm{f})$.
An argument identical to that used in the case of a linear objective function shows that a version of Pontryagin's Principle still holds in this case (but only as a *necessary* condition, since (6.32) is merely necessary), with the terminal condition on the adjoint equation being now

$$p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f})).$$

Note that the adjoint equation cannot anymore be integrated independently of the state equation, since $x^*(t_\mathrm{f})$ is needed to integrate p^* (backward in time from t_f to t_0).

Exercise 6.8 *By following the argument in these notes, show that a version of Pontryagin's Principle also holds for a discrete-time linear optimal control problem. (We saw earlier that it may not hold for a discrete nonlinear optimal control problem. Also see discussion in the next section of these notes.)*

Exercise 6.9 *If ψ is convex and U is convex, Pontryagin's Principle is again a necessary and sufficient condition of optimality.*

Optimal control of nonlinear systems

See [21, 32]. We consider the problem

$$\text{minimize } \psi(x(t_{\rm f})) \quad \text{s.t. } \dot{x}(t) = f(t, x(t), u(t)), \quad \text{a.e. } t \in [t_0, t_{\rm f}], \quad u \in \mathcal{U}, \tag{6.33}$$

where $x(t_0) := x_0$ is prescribed and x is absolutely continuous. Unlike the linear case, we do not have an explicit expression for $\phi(t_{\rm f}, t_0, x_0, u)$. We shall settle for a comparison between the trajectory x^* and trajectories x obtained by *perturbing* the control u^*. By considering *strong* perturbations of u^* we will be able to still obtain a *global Pontryagin Principle*, involving a *global* minimization of the pre-Hamiltonian. Some proofs will be omitted.

We assume that ψ is continuously differentiable, and impose the following regularity conditions on (6.33) (same assumptions as in Chapter 3)

(i) for each $t \in [t_0, t_{\rm f}]$, $f(t, \cdot, \cdot) \ : \ \mathbf{R}^n \times \mathbf{R}^m \to \mathbf{R}^n$ is continuously differentiable;

(ii) the functions $f, \frac{\partial f}{\partial x}, \frac{\partial f}{\partial u}$ are continuous on $[t_0, t_{\rm f}] \times \mathbf{R}^n \times \mathbf{R}^m$;

(iii) for every finite α, $\exists \beta, \gamma$ s.t.

$$||f(t, x, u)|| \leq \beta + \gamma ||x|| \quad \forall t \in [t_0, t_{\rm f}], x \in \mathbf{R}^n, \ u \in \mathbf{R}^m, \ ||u|| \leq \alpha.$$

Under these conditions, for any $\hat{t} \in [t_0, t_{\rm f}]$, any $z \in \mathbf{R}^n$, and $u \in$ PC, (6.33) has a unique continuous solution

$$x(t) = \phi(t, \hat{t}, z, u) \quad \hat{t} \leq t \leq t_{\rm f}$$

such that $x(\hat{t}\,) = z$. Let $x(t_0) = x_0$ be given. Again, let $U \subseteq \mathbf{R}^m$ and let

$$\mathcal{U} = \{u \ : \ [t_0, t_{\rm f}] \to U \ , \ u \in \text{PC}\}$$

be set of admissible controls. Let

$$K(t, t_0, z) = \{\phi(t, t_0, z, u) : u \in \mathcal{U}\}.$$

Problem (6.33) can be seen to be closely related to

$$\text{minimize } \psi(x) \quad \text{s.t. } x \in K(t_{\rm f}, t_0, x_0), \tag{6.34}$$

where x here belongs in $\mathbf{R}^n$, a finite-dimension space! Indeed, it is clear that $x^*(\cdot)$ is an optimal state-trajectory for (6.33) if and only if $x^*(t_{\rm f})$ solves (6.34). Characterization of a related optimal control will be obtained as a by-product of solving this problem.

Now, let $u^* \in \mathcal{U}$ be optimal and let x^* be the corresponding state trajectory. As in the linear case we must have (necessary condition)

$$\nabla\psi(x^*(t_\mathrm{f}))^T h \geq 0 \quad \forall h \in \mathrm{cl\,coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, x_0)), \tag{6.35}$$

where ψ is the (continuously differentiable) terminal cost. However, unlike in the linear case (convex reachable set), there is no explicit expression for this tangent cone. We will obtain a characterization for a subset of interest of this cone. This subset will correspond to a particular type of perturbations of $x^*(t_\mathrm{f})$. The specific type of perturbation to be used is motivated by the fact that we are seeking a 'global' Pontryagin Principle, involving a 'min' over the entire U set.

Let D be the set of discontinuity points of u^*. Let $\tau \in (t_0, t_\mathrm{f}),\ \tau \notin D$, and let $v \in U$. For $\epsilon > 0$, we consider the strongly perturbed control $u_{\tau,\epsilon}$ defined by

$$u_{\tau,v,\epsilon}(t) = \begin{cases} v & \text{for } t \in [\tau - \epsilon, \tau) \\ u^*(t) & \text{elsewhere} \end{cases}$$

This is often referred to as a "needle" perturbation. A key fact is that, as shown by the following proposition, even though v may be very remote from $u^*(\tau)$, the effect of this perturbed control on x is small. Such "strong" perturbations lead to a global Prontryagin's Principle even though local "tools" are used. Let $x_{\tau,v,\epsilon}$ be the trajectory corresponding to $u_{\tau,v,\epsilon}$. For small ϵ, this trajectory will be claose to x^* and, because $u_{\tau,v,\epsilon} \in \mathcal{U}$, $x_{\tau,v,\epsilon}(t_\mathrm{f})$ will be in $K(t_\mathrm{f}, t_0, x_0)$.

Proposition 6.3

$$x_{\tau,v,\epsilon}(t_\mathrm{f}) = x^*(t_\mathrm{f}) + \epsilon h_{\tau,v} + o(\epsilon)$$

with o satisfying $\frac{o(\epsilon)}{\epsilon} \to 0$ as $\epsilon \to 0$, and with

$$h_{\tau,v} = \Phi(t_\mathrm{f}, \tau)[f(\tau, x^*(\tau), v) - f(\tau, x^*(\tau), u^*(\tau))].$$

where Φ is the state transition matrix for the linear (time-varying) system $\dot{\xi}(t) = \frac{\partial f}{\partial x}(t, x^*(t), u^*(t))\xi(t)$.

See, e.g., [19, p. 246-250] for a detailed proof of this result. The gist of the argument is that (i)

$$x_{\tau,v,\epsilon}(\tau) = x^*(\tau - \epsilon) + \int_{\tau-\epsilon}^{\tau} f(t, x_{\tau,v,\epsilon}(t), v)dt = x^*(\tau - \epsilon) + \epsilon f(\tau, x^*(\tau), v) + o_\tau(\epsilon).$$

and

$$x^*(\tau) = x^*(\tau - \epsilon) + \int_{\tau-\epsilon}^{\tau} f(t, x^*(t), u^*(t))d\tau = x^*(\tau - \epsilon) + \epsilon f(\tau, x^*(\tau), u^*(t)) + o_\tau(\epsilon)$$

so that

$$x_{\tau,v,\epsilon}(\tau) - x^*(\tau) = \epsilon[f(\tau, x^*(\tau), v) - f(\tau, x^*(\tau), u^*(\tau))] + o(\epsilon).$$

and (ii) for $t > \tau$, with $\xi(t) := x_{\tau,v,\epsilon}(t) - x^*(t)$,

$$\dot{\xi}(t) := \dot{x}_{\tau,\epsilon}(t) - \dot{x}^*(t) = f(t, x_{\tau,\epsilon}(t), u^*(t)) - f(t, x^*(t), u^*(t)) = \frac{\partial f}{\partial x}(t, x^*(t), u^*(t))\xi(t) + o(\epsilon).$$

Exercise 6.10 $\forall \tau \in (t_0, t_\mathrm{f}), \tau \notin D,\ \forall v \in U,$

$$h_{\tau,v} \in \mathrm{TC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, x_0)) \tag{6.36}$$

This leads to the following Pontryagin Principle (*necessary* condition).

Theorem 6.3 *Let $\psi : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable and consider the problem*

$$\begin{array}{ll} \text{minimize} & \psi(x(t_\mathrm{f})) \quad \text{s.t.} \\ & \dot{x}(t) = f(t, x(t), u(t)) \text{ a.e. } t \in [t_0, t_\mathrm{f}] \\ & x(t_0) = x_0,\ u \in \mathcal{U},\ x \text{ continuous} \end{array}$$

Suppose $u^* \in \mathcal{U}$ is optimal and let $x^*(t) = \phi(t, t_0, x_0, u^*)$. Let $p^*(t),\ t_0 \le t \le t_\mathrm{f}$, continuous, satisfy the (linear) adjoint equation

$$\dot{p}^*(t) = -\frac{\partial f}{\partial x}(t, x^*(t), u^*(t))^T p^*(t) = -\nabla_x H(t, x^*(t), p^*(t), u^*(t)) \text{ a.e. } t \in [t_0, t_\mathrm{f}] \tag{6.37}$$

$$p^*(t_\mathrm{f}) = -\nabla \psi(x^*(t_\mathrm{f})) \tag{6.38}$$

with

$$H(t, x, p, u) = p^T f(t, x, u) \qquad \forall t, x, u, p.$$

Then u^* satisfies the Pontryagin Principle

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \ (= \max_{v \in U} H(t, x^*(t), p^*(t), v)\ ) \tag{6.39}$$

for all $t \in [t_0, t_\mathrm{f})$.

Proof (compare with linear case). In view of (6.36), we must have

$$\nabla \psi(x^*(t_\mathrm{f}))^T h_{t,v} \ge 0 \quad \forall t \in [t_0, t_\mathrm{f}], v \in U \tag{6.40}$$

Using the expression we obtained above for $h_{t,v}$, we get

$$\nabla \psi(x^*(t_\mathrm{f}))^T \Phi(t_\mathrm{f}, t)[f(t, x^*(t), v) - f(t, x^*(t), u^*(t))] \ge 0 \ \ \forall t \in (t_0, t_\mathrm{f}), t \notin D,\ \forall v \in U.$$

Since, in view of (6.38)–(6.37), $p^*(t) = -\Phi(t_\mathrm{f}, t) \nabla \psi(x^*(t_\mathrm{f}))$, we get

$$p^*(t)^T f(t, x^*(t), u^*(t)) \ge p^*(t)^T f(t, x^*(t), v)\ \forall t \in (t_0, t_\mathrm{f}), t \notin D,\ \forall v \in U.$$

Finally, for $t \in D$, the result follows from right-continuity of the two sides.

■

Remark 6.14 If $u^* \in \mathcal{U}$ is locally optimal, in the sense that $x^*(t_\mathrm{f})$ is a local minimizer for ψ in $K(t_\mathrm{f}, t_0, x_0)$, Pontryagin's Principle still holds (with a global minimization over U). Why?

The following exercise generalizes Exercise 2.24.

Exercise 6.11 *Prove that, if f does not explicitly depend on t, then $m(t) = \mathcal{H}(t, x^*(t), p^*(t))$ is constant. Assume that u^* is continuously differentiable on all the intervals over which it is continuous.*

Integral objective functions (Lagrange problems)

Suppose the objective function, instead of being $\psi(x(t_\mathrm{f}))$, is of the form

$$\int_{t_0}^{t_\mathrm{f}} \mathcal{L}(t, x(t), u(t))dt \ .$$

Such problems are known as a Lagrange problems, while terminal-state-cost problems are known as Mayer problems.

Lagrange problems can be converted to the Mayer form. To this end, we consider the *augmented system* with state variable $\tilde{x} = [x^0; x] \in \mathbf{R}^{1+n}$ as follows

$$\dot{\tilde{x}}(t) = \begin{bmatrix} \mathcal{L}(t, x(t), u(t)) \\ f(t, x(t), u(t)) \end{bmatrix} := \tilde{f}(t, x(t), u(t)) \text{ a.e. } t \in [t_0, t_\mathrm{f}],$$

$x^0(t_0) = 0,\ x^0(t_\mathrm{f})$ free. Now the problem is equivalent to minimizing

$$\psi(\tilde{x}(t_\mathrm{f})) := x^0(t_\mathrm{f})$$

with dynamics and constraints of the same form as before. After some simplifications, we get the following result.

Theorem 6.4 *Let $u^* \in \mathcal{U}$ be optimal, and let x^* be the associated state trajectory, and let*

$$\begin{aligned} H(t, x, p, u) &= -\mathcal{L}(t, x, u) + p^T f(t, x, u) \\ \mathcal{H}(t, x, p) &= \sup_{v \in U} H(t, x, p, v) \end{aligned}$$

Then there exists a function $p^* : [t_0, t_\mathrm{f}] \to \mathbf{R}^n$, continuous, satisfying

$$\dot{p}^*(t) = -\frac{\partial H}{\partial x}(t, x^*(t), p^*(t), u^*(t))^T \text{ a.e. } t \in [t_0, t_\mathrm{f}],$$

with $p^*(t_\mathrm{f}) = 0$. Furthermore

$$H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \quad \forall t.$$

Finally, if f_0 and f do not depend explicitly on t,

$$m(t) = \mathcal{H}(t, x^*(t), p^*(t)) = \text{constant}.$$

Exercise 6.12 *Prove Theorem 6.4.*

Remark 6.15 Note that the expression for H is formally quite similar to the negative of that for the "Lagrangian" used in constrained optimization. Indeed f_0 is the (integrand in the) cost function, and f specifies the "constraints" (dynamics in the present case). (But beware!! In the calculus of variations literature, the term "Lagrangian" refers to the integrand $\mathcal{L}$ in problem (6.2).)

Exercise 6.13 *Conversely, express terminal cost $\psi(x(t_\mathrm{f}))$ as an integral cost.*

Lagrange multiplier interpretation of $p^*(t)$

For each t, $p^*(t)$ can be thought of as a (vector of) Lagrange multiplier(s) for the constraint

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

or rather for

$$dx(t) = f(t, x(t), u(t))dt. \tag{6.41}$$

To see this, finely discretize time, let $\Delta := t_{i+1} - t_i$, and let $x_i := x(t_i)$, $u_i := u(t_i)$, and $f_i(x_i, u_i) := \Delta \cdot f(t_i, x_i, u_i)$. Then (6.41) is appropriate approximated by

$$x_{i+1} = x_i + f_i(x_i, u_i), \tag{6.42}$$

which is the dynamics used in section 6.2 (Discrete-time optimal control). Similarly, if we let $p_i = p^*(t_i)$, we can appropriately approximate the adjoint equation (6.37) with

$$p_{i+1} = p_i - \frac{\partial f_i}{\partial x}(x_i, u_i)^T p_i,$$

which is identical to (6.22), i.e., p_i is the Lagrange multiplier associated with (6.42). Also, recall that in Chapter 3), we noted in (3.13) that, when value function V is smooth enough,

$$p^*(t) := -\nabla_x V(t, x^*(t)),$$

which, again (see section 6.2) can be viewed in terms of the sensitivity interpretation of Lagrange multipliers.

Geometric approach to discrete-time case

We investigate to what extent the approach just used for the continuous-time case can also be used in the discrete-time case, the payoff being the geometric intuition.

A hurdle is immediately encountered: strong perturbations as described above cannot work in the discrete-time case. The reason is that, in order to build an approximation to the reachable set we must construct small perturbations of $x^*(t_\mathrm{f})$. In the discrete-time case, the time interval during which the control is varied cannot be made arbitrarily small (it is at least one time step), and thus the "smallness" of the perturbation must come from the perturbed value v. Consequently, at every t, $u^*(t)$ can only be compared to nearby values v and a true Pontryagin Principle cannot be obtained. We investigate what can be obtained by considering appropriate "weak" variations.

Thus consider the discrete-time system

$$x_{i+1} = x_i + f(i, x_i, u_i),\ i = 0, \ldots, N-1,$$

and the problem of minimizing $\psi(x_N)$, given a fixed x_0 and the constraint that $u_i \in U$ for all i. Suppose u_i^*, $i = 0, \ldots, N-1$, is optimal, and x_i^*, $i = 1, \ldots, N$ is the corresponding

optimal state trajectory. Given $k \in \{0, \ldots, N-1\}$, $\epsilon > 0$, and $w \in \mathrm{TC}(u_k^*, U)$, consider the weak variation

$$(u_{k,\epsilon})_i = \begin{cases} u_k^* + \epsilon w + o(\epsilon) & \text{i=k} \\ u_i^* & \text{otherwise} \end{cases}$$

where $o(\cdot)$ is selected in such a way that $u_{\epsilon,i} \in U$ for all i, which is achievable due to the choice of w. The next state value is then given by

$$\begin{aligned} (x_{k,\epsilon})_{k+1} &= x_{k+1}^* + f(k, x_k^*, u_k^* + \epsilon w + o(\epsilon)) - f(k, x_k^*, u_k^*) && (6.43) \\ &= x_{k+1}^* + \epsilon \frac{\partial f}{\partial u}(k, x_k^*, u_k^*) w + \tilde{o}(\epsilon). && (6.44) \end{aligned}$$

The final state is then given by

$$(x_{k,\epsilon})_N = x_N^* + \epsilon \Phi(N, k+1) \frac{\partial f}{\partial u}(k, x_k^*, u_k^*) w + \hat{o}(\epsilon).$$

which shows that $h_{k,w} := \Phi(N, k+1) \frac{\partial f}{\partial u}(k, x_k^*, u_k^*) w$ belongs to $\mathrm{TC}(x_N^*, K(N, 0, x_0))$. We now can proceed as we did in the continuous-time case. Thus

$$\nabla \psi(x_N^*)^T \Phi(N, k+1) \frac{\partial f}{\partial u}(k, x_k^*, u_k^*) w \geq 0 \quad \forall k, \forall w \in \mathrm{TC}(u_k^*, U).$$

Letting p_k^* solve the adjoint equation,

$$p_i = p_{i+1} + \frac{\partial f}{\partial x}(i, x_i, u_i) p_{i+1} \quad \text{with } p_N^* = -\nabla \psi(x_N^*),$$

i.e., $p_{k+1}^* = -\Phi(N, k+1)^T \nabla \psi(x_N^*)$, we get

$$(p_{k+1}^*)^T \frac{\partial f}{\partial u}(k, x_k^*, u_k^*) w \leq 0 \quad \forall k, \forall w \in \mathrm{TC}(u_k^*, U).$$

And defining $H(j, \xi, \eta, v) = \eta^T f(j, \xi, v)$ we get

$$\frac{\partial H}{\partial u}(k, x_k^*, p_{k+1}^*, u_k^*) w \leq 0 \quad \forall k, \forall w \in \mathrm{TC}(u_k^*, U),$$

which is a mere necessary condition of optimality for the minimization of $H(k, x_k^*, p_{k+1}^*, u)$ with respect to $u \in U$. It is a special case of the result we obtained earlier in this chapter for the discrete-time case.

Partially free initial state

Suppose now that $x(t_0)$ is not necessarily fixed but, more generally, it is merely constrained to satisfy $g^0(x(t_0)) = \theta$ where g^0 is a given continuously differentiable function. Let $T_0 = \{x : g^0(x) = \theta\}$. ($T_0 = \mathbf{R}^n$ is a special case of this, where the image space of g^0 has dimension 0. $T_0 = \{x_0\}$, i.e., $g^0(x) = x - x_0$, is the other extreme: fixed initial point.) Problem (6.34) becomes

$$\text{minimize } \psi(x) \quad \text{s.t. } x \in K(t_\mathrm{f}, t_0, T_0), \tag{6.45}$$

Note that, if $x^*(\cdot)$ is an optimal trajectory, then $K(t, t_0, T_0)$ contains $K(t, t_0, x^*(t_0))$ for all t, so that

$$\mathrm{TC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, x^*(t_0))) \subseteq \mathrm{TC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)).$$

Since $x^*(t_\mathrm{f}) \in K(t_\mathrm{f}, t_0, x^*(t_0))$, $x^*(t_\mathrm{f})$ is also optimal for (6.33) with fixed initial state $x_0 := x^*(t_0)$, and the necessary conditions we obtained for the fixed initial point problem apply to the present problem, with $x_0 := x^*(t_0)$—but, of course, $x^*(t_0)$ isn't known. We now obtain an additional condition (which will be of much interest, since we now have one more degree of freedom). Let $x^*(t_0)$ be the optimal initial point. From now on, the following additional assumption will be in force:

Assumption. $(x^*(t_0), g^0)$ is non-degenerate.
Let

$$d \in \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right) = \mathrm{TC}(x^*(t_0), T_0).$$

Then, given $\epsilon > 0$ there exists some little-o function o such that $x^*(t_0) + \epsilon d + o(\epsilon) \in T_0$. For $\epsilon > 0$ small enough, let $x_\epsilon(t_0) = x^*(t_0) + \epsilon d + o(\epsilon)$, and consider applying our optimal control u^* (for initial point $x^*(t_0)$), but starting from $x_\epsilon(t_0)$ as initial point. We now invoke the following result, given as an exercise.

Exercise 6.14 *Show that, if $d \in \mathrm{TC}(x^*(t_0), T_0)$, then*

$$\Phi(t_\mathrm{f}, t_0)d \in \mathrm{TC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)),$$

where Φ is as in Proposition 6.3. (Hint: Use the fact that $\frac{\partial x_\epsilon(t)}{\partial \epsilon}$ follows linearized dynamics. See, e.g., Theorem 10.1 in [17].)

It follows that optimality of $x^*(t_\mathrm{f})$ for (6.45) yields

$$\nabla\psi(x^*(t_\mathrm{f}))^T\Phi(t_\mathrm{f}, t_0)d \geq 0 \ \forall d \in \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right),$$

i.e., since $\mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right)$ is a subspace,

$$(\Phi(t_\mathrm{f}, t_0)^T\nabla\psi(x^*(t_\mathrm{f})))^T d = 0 \quad \forall d \in \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right).$$

Thus

$$p^*(t_0) \perp \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right) \tag{6.46}$$

which is known as a transversality condition. Note that $p^*(t_0)$ is thus no longer free. Indeed, for each degree of freedom "gained" on $x^*(t_0)$, we "lose" one on $p^*(t_0)$.

Remark 6.16 An alternative derivation of this result follows by observing (excuse the abuse of notation) that

$$\frac{\partial \psi}{\partial x_0}(x^*(t_\mathrm{f})d = 0 \quad \forall d \in \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right).$$

and

$$\frac{\partial \psi}{\partial x_0}(x^*(t_\mathrm{f})) = \frac{\partial \psi}{\partial x}(x^*(t_\mathrm{f}))\frac{\partial x(t_\mathrm{f})}{\partial x_0} = \frac{\partial \psi}{\partial x}(x^*(t_\mathrm{f}))\Phi(t_\mathrm{f}, t_0).$$

Constrained terminal state

Now suppose that the final state $x^*(t_\mathrm{f})$, instead of being entirely free, is possibly constrained, specifically, suppose $x^*(t_\mathrm{f})$ is required to satisfy $g^\mathrm{f}(x^*(t_\mathrm{f})) = 0$, where g^f is a given continuously differentiable function. Let $T_\mathrm{f} = \{x : g^\mathrm{f}(x) = 0\}$. ($T_\mathrm{f} = \{x_f\}$, with x_f given, is a special case of this, where g^f is given by $g^\mathrm{f}(x) \equiv x - x_f$.)
An important special case is that of completely free initial state. Indeed this is the "mirror image" of the case with free terminal state and (possibly) constrained initial state, which was considered in the previous subsection. It is the object of the next exercise.

Exercise. Obtain a Pontryagin Principle for problem (6.33) but with free initial state and possibly constrained (or even fixed) terminal state. Use an initial cost instead of a terminal cost, or consider the case of an integral objective function. [Hint: Reverse time.]

Turning to the general case, note that ψ should no longer be minimized over the entire reachable set, but only on its intersection with T_f. The Pontryagin Principle as previously stated no longer holds. Rather we can write

$$\nabla\psi(x^*(t_\mathrm{f}))^T h \geq 0 \ \forall h \in \text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0) \cap T_\mathrm{f}), \tag{6.47}$$

initial which involves the tangent cone to a smaller set than when the terminal state is unconstrained. We now make simplifying assumptions.

Assumptions Terminal-State (TS):

1. $(x^*(t_\mathrm{f}), g^\mathrm{f})$ is non-degenerate.
2. $$\text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0) \cap T_\mathrm{f}) = \text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)) \ \cap \ \text{cl coTC}(x^*(t_\mathrm{f}), T_\mathrm{f}). \tag{6.48}$$
3. the convex cone
$$C = \left\{ \begin{bmatrix} \nabla\psi(x^*(t_\mathrm{f}))^T h \\ \frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))h \end{bmatrix} : h \in \text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)) \right\}$$
is closed.

The first two assumptions amount to a type of constraint qualification for the constraint $x^*(t_\mathrm{f}) \in T_\mathrm{f}$. In contrast, the third one also involves the objective function.

Exercise 6.15 *Show that, if $x \in \Omega_1 \cap \Omega_2$ then* cl coTC$(x, \Omega_1 \cap \Omega_2) \subseteq$ cl coTC$(x, \Omega_1) \cap$ cl coTC(x, Ω_2). *Provide an example showing that equality does not always hold.*

Exercise 6.16 *Consider minimizing $f(x)$ subject to $x \in \Omega_1 \cap \Omega_2$, with $\Omega_i := \{x : g_i(x) \geq 0\}$, $i = 1, 2$, where $f, g_1, g_2 : \mathbf{R}^n \to \mathbf{R}$ are smooth. Let $\hat{x}$ be a local minimizer for this problem. Further assume that (i) $\nabla g_i(\hat{x}) \neq \theta$, $i = 1, 2$, and (ii)* TC$(x, \Omega_1 \cap \Omega_2) =$ TC$(x, \Omega_1) \cap$ TC(x, Ω_2). *Show that, under such assumptions, KKT holds at $\hat{x}$; i.e., without further assumptions on f, there exists $\hat{\lambda} \in \mathbf{R}^2$ such that*

$$\nabla f(\hat{x}) + \hat{\lambda}_1 \nabla g_1(\hat{x}) + \hat{\lambda}_2 \nabla g_1(\hat{x}) = \theta.$$

This shows that (i)+(ii) forms a constraint qualification indeed.

Under these three assumptionts, the Pontrygain's Principle can be readily proved, as follows. First, $\mathrm{TC}(x^*(t_\mathrm{f}), T_\mathrm{f}) = \mathcal{N}\left(\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))\right)$, and from (6.47) and (6.48), we obtain

$$\nabla\psi(x^*(t_\mathrm{f}))^T h \geq 0 \quad \forall h \in \text{ cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)) \cap \mathcal{N}\left(\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))\right). \tag{6.49}$$

Now let m_f be the dimension of the image-space of g^f (number of terminal equality constraints) and define

$$R = (-1, 0, \ldots, 0)^T \in \mathbf{R}^{m_\mathrm{f}+1}$$

Then, in view of (6.49), $R \notin C$. Since C is a closed convex cone, it follows from Exercise B.26 that there exists $\mu := [p_0^*; \pi] \in \mathbf{R}^{m_\mathrm{f}+1}$ such that $\mu^T R < 0$ (i.e., $p_0^* > 0$) and $\mu^T v \geq 0$ for all $v \in C$, i.e.,

$$\left(p_0^* \nabla\psi(x^*(t_\mathrm{f})) + \frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))^T \pi\right)^T h \geq 0 \quad \forall h \in \text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)),$$

or equivalently (since $p_0^* > 0$), by redefining π,

$$\left(\nabla\psi(x^*(t_\mathrm{f})) + \frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))^T \pi\right)^T h \geq 0 \quad \forall h \in \text{cl coTC}(x^*(t_\mathrm{f}), K(t_\mathrm{f}, t_0, T_0)),$$

(to be compared to (6.35)). If, instead of imposing $p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f}))$, we impose the condition

$$p^*(t_\mathrm{f}) = -\nabla\psi(x^*(t_\mathrm{f})) - \frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))^T \pi \quad \text{for some } \pi,$$

we obtain, formally, the same Pontryagin Principle as above. While we do not know π, the above guarantees that

$$p^*(t_\mathrm{f}) + \nabla\psi(x^*(t_\mathrm{f})) \perp \mathcal{N}\left(\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))\right).$$

This is again a transversality condition.
It can be shown that, without our assumptions except for the requirement that $(x^*(t_\mathrm{f}), g^{t_\mathrm{f}})$ is non-degenerate, the same result still holds except that the transversality condition becomes: there exists $p_0^* \geq 0$, with $(p^*(t_\mathrm{f}), p_0^*)$ not identically zero, such that

$$p^*(t_\mathrm{f}) + p_0^* \nabla\psi(x^*(t_\mathrm{f})) \perp \mathcal{N}\left(\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))\right). \tag{6.50}$$

This result is significantly harder to prove though. It is the central difficulty in the proof of Pontryagin's principle. Proofs are found in [19, 27, 33]. Also see [21].

General case (Bolza problems)

Finally, consider the case of function that includes both an integral term and a terminal-state term (such problem is known as a Bolza problem), and both initial and terminal states are possibly fixed or constrained. (Such problems are readily converted to the Mayer form, using the same transformation that we used earlier to transform Lagrange problems to the Mayer form.) The following theorem can be proved.

Theorem 6.5 *Consider the problem (assume $\mathcal{L}$ is smooth enough)*

$$\begin{aligned} \text{minimize} \quad & \int_{t_0}^{t_\mathrm{f}} \mathcal{L}(t,x(t),u(t))dt + \psi(x(t_\mathrm{f})) \quad \text{subject to} \\ & \dot{x}(t) = f(t,x(t),u(t)) \text{ a.e. } t \in [t_0,t_\mathrm{f}], \\ & g^0(x(t_0)) = \theta,\ g^\mathrm{f}(x(t_\mathrm{f})) = \theta, \\ & u \in \mathcal{U},\ x \text{ continuous}, \end{aligned}$$

and the associated pre-Hamiltonian $H : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^{1+n} \times \mathbf{R}^m \to \mathbf{R}$ and Hamiltonian $\mathcal{H} : \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^{1+n} \to \mathbf{R}$ given by

$$H(t,\xi,\tilde{\eta},u) := -\pi\mathcal{L}(t,\xi,u) + \eta^T f(t,\xi,u), \quad \mathcal{H}(t,\xi,\tilde{\eta}) := \sup_{u\in U} H(t,\xi,\tilde{\eta},u),$$

where $\tilde{\eta} := [\pi;\eta]$. Suppose $u^* \in \mathcal{U}$ is an optimal control, and x_0^* (satisfying $g^0(x_0^*) = \theta$) an optimal initial state, and let $x^*(\cdot)$ be the associated optimal trajectory (with $x^*(t_0) = x_0^*$). Then there exists an absolutely continuous function $p^* : [t_0,t_\mathrm{f}] \to \mathbf{R}^n$ and a scalar constant $p_0^* \geq 0$, with $\tilde{p}^* := [p_0^*;p^*]$ not identically zero, that satisfy

$$\begin{aligned} \dot{p}^*(t) &= -\left[\frac{\partial H}{\partial x}(t,x^*(t),\tilde{p}^*(t),u^*(t))\right]^T \quad \text{a.e. } t \in [t_0,t_\mathrm{f}] \\ H(t,x^*(t),\tilde{p}^*(t),u^*(t)) &= \mathcal{H}(t,x^*(t),\tilde{p}^*(t)) \quad \forall t \in [t_0,t_\mathrm{f}). \end{aligned}$$

Further, if $\frac{\partial g^0}{\partial x}(x^*(t_0))$ has full row rank, then

$$p^*(t_0) \perp \mathcal{N}\left(\frac{\partial g^0}{\partial x}(x^*(t_0))\right),$$

and if $\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))$ has full row rank, then

$$p^*(t_\mathrm{f}) + p_0^*\nabla\psi(x^*(t_\mathrm{f})) \perp \mathcal{N}\left(\frac{\partial g^\mathrm{f}}{\partial x}(x^*(t_\mathrm{f}))\right).$$

Also, if f does not depend explicitly on t then $\mathcal{H}(t,x^*(t),p^*(t))$ is constant. Finally, if Assumptions TS hold for the associated Mayer problem, then such $[p_0^*;p^*]$ exists for which $p_0^* = 1$.

Remark 6.17 When constraint qualification (6.48) does not hold for the associated Mayer problem, then p_0^* may have to be equal to zero (with p^* not identically zero). This makes the result "weak", just like the F. John condition of constrained optimization, because it involves neither $\mathcal{L}$ nor ψ, i.e., does not involve the function being minimized.

Remark 6.18 It can be shown that Assumptions TS, (which imply that p_0^* can be chosen strictly positive) also hold if a certain controllability condition is satisfied provided the control values are unconstrained, i.e., $U = \mathbf{R}^n$. See, e.g., [23].

Free final time

Suppose the final time is itself a decision variable (example: minimum-time problem).

Consider the problem

$$\begin{array}{ll}\text{minimize} & \displaystyle\int_{t_0}^{t_\mathrm{f}} \mathcal{L}(t,x(t),u(t))dt \text{ subject to} \\ & \dot{x}(t) = f(t,x(t),u(t)) \text{ a.e. } t\in[t_0,t_\mathrm{f}],\ x \text{ absolutely continuous,} \quad \text{(dynamics)} \\ & g^0(x(t_0)) = \theta,\ g^\mathrm{f}(x(t_\mathrm{f})) = \theta \qquad \text{(initial and final conditions)} \\ & u\in\mathcal{U} \qquad \text{(control constraints)} \\ & t_\mathrm{f}\geq t_0 \qquad \text{(final time constraint)}\end{array}$$

We analyze this problem by converting the variable-length time interval $[t_0,t_\mathrm{f}]$ into a fixed-length time interval $[0,1]$. Define $t(\cdot)$, absolutely continuous, to satisfy

$$\frac{dt(s)}{ds} = \alpha(s) \text{ a.e. } s\in[0,1]. \tag{6.51}$$

To fall back into a known formalism, we will consider s as the new time, $t(s)$ as a new state variables, and $\alpha(s)$ as a new control. Note that, clearly, given any optimal $\alpha^*(\cdot)$, there is an equivalent constant optimal α^*, equal to $t^*(1)-t_0$; accordingly, among the controls (u,α) that satisfy Pontryagin's principle, we can choose to focus only on those for which α is constant. The initial and final condition on $t(s)$ are

$$\begin{array}{rcl} t(0) & = & t_0 \\ t(1) & & \text{free}\end{array}$$

Denoting $z(s)=x(t(s)),\ v(s)=u(t(s))$ we obtain the state equation

$$\begin{array}{rcl} \displaystyle\frac{d}{ds}z(s) & = & \alpha f(t(s),z(s),v(s)) \text{ a.e. } s\in[0,1] \\ g^0(z(0)) & = & \theta,\ g^\mathrm{f}(z(1))=\theta.\end{array}$$

Now suppose that $(u^*,{t_\mathrm{f}}^*,x_0^*)$ is optimal for the original problem. Then the corresponding (v^*,α^*,x_0^*), with $\alpha^*={t_\mathrm{f}}^*-t_0$ is optimal for the transformed problem. Expressing the known conditions for this problem and performing some simplifications, we obtain the following result.

Theorem 6.6 *Same as above, with the pre-Hamiltonian*

$$H(t,x,\tilde{p},u) = -p_0\mathcal{L}(t,x,u) + p^T f(t,x,u),$$

and the additional necessary condition

$$\mathcal{H}({t_\mathrm{f}}^*, x^*({t_\mathrm{f}}^*), \tilde{p}^*({t_\mathrm{f}}^*)) = 0$$

where $\tilde{p}^*(t) = (p_0^*, p^*(t))$ and ${t_\mathrm{f}}^*$ is the optimal final time. Again, if $\mathcal{L}$ and f do not explicitly depend on t, then

$$\mathcal{H}(t,x^*(t),\tilde{p}^*(t)) = \text{ constant } = 0 \quad \forall t$$

Exercise 6.17 *Prove Theorem 6.6 by applying the previous results.*

Minimum time problem

Consider the following special case of the previous problem (with fixed initial and final states).

$$\begin{array}{rcl} \text{minimize} & t_{\mathrm{f}} & \text{subject to} \\ \dot{x}(t) & = & f(t, x(t), u(t)) \text{ a.e. } t \in [t_0, t_{\mathrm{f}}],\ x \text{ absolutely continuous} \\ x(t_0) & = & x_0, \quad x(t_{\mathrm{f}}) = x_f \\ u \in \mathcal{U}, & & t_{\mathrm{f}} \geq t_0 \quad (t_{\mathrm{f}} \text{ free}) \end{array}$$

The previous theorem can be simplified to give the following.

Theorem 6.7 *Let ${t_{\mathrm{f}}}^* \geq t_0$ and $u^* \in \mathcal{U}$ be optimal. Let x^* be the corresponding trajectory. Then there exists an absolutely continuous function $p^* : [t_0, {t_{\mathrm{f}}}^*] \to \mathbf{R}^n$, not identically zero such that*

$$\begin{array}{l} \dot{p}^*(t) = -\left[\dfrac{\partial f}{\partial x}(t, x^*(t), u^*(t))\right]^T p^*(t) \quad \text{a.e. } t \in [t_0, {t_{\mathrm{f}}}^*] \\ [p^*(t_0), p^*(t_{\mathrm{f}}) \quad \text{free}] \\ H(t, x^*(t), p^*(t), u^*(t)) = \mathcal{H}(t, x^*(t), p^*(t)) \qquad t \in [t_0, t_{\mathrm{f}}) \\ \mathcal{H}({t_{\mathrm{f}}}^*, x^*({t_{\mathrm{f}}}^*), p^*({t_{\mathrm{f}}}^*)) \geq 0 \end{array}$$

with

$$\begin{array}{rcl} H(t, x, p, u) & = & p^T f(t, x, u) \qquad (\mathcal{L} = 0) \\ \mathcal{H}(t, x, p) & = & \sup\limits_{v \in U} H(t, x, p, u) \end{array}$$

Also, if f does not depend explicitly on t then $\mathcal{H}(t, x^*(t), p^*(t))$ is constant.

Exercise 6.18 *Prove Theorem 6.7. Hint: The cost function can be transformed to an integral, with free final time t_{f}. Indeed,*

$$t_{\mathrm{f}} = t_0 + \int_{t_0}^{t_{\mathrm{f}}} (1) dt.$$

Remark 6.19 Note that p is determined only up to a constant scalar factor.

6.4 Applying Pontryagin's Principle

Given that the finite-dimensional maximization problem (maximization of the pre-Hamiltonian) for obtaining $u^*(t)$ at each time t involves $x^*(t)$ and $p^*(t)$, both of which are yet to be determined, it may first appear that Pontryagin's may be of little help for solving a "real" problem (e.g., numerically). The following approach comes to mind though:

1. Minimize the pre-Hamiltonian with respect to the control, yielding $u^*(t)$ in terms of $x^*(t)$ and $p^*(t)$ at every time t;
2. Plug the expression for u^* into the differential equations for x^* and p^*.
3. Solve the resulting differential equations in (x^*, p^*). This is in general a two-point boundary value problem.
4. Plug $x^*(t)$ and $p^*(t)$ into the expression obtained for $u^*(t)$ (for each t) in step 1 above.

One difficulty with the scheme outlined above is that, in most cases of practical interest, no "closed-form" solution can be obtained at step 1 for $u^*(t)$ in terms of $(x^*(t), p^*(t))$: the maximization is to be carried out for fixed values of t, e.g., on a fine time grid, and this has to be done concurrently with carrying out steps 3 and 4, since $x^*(t)$ and $p^*(t)$ must be known at the "current" time in order to be able to proceed with the (numerical) minimization.
Another difficulty is that the differential equation to be solved at step 3, in $2n$ variables (assuming p_0^* is strictly positive and hence can be set to 1), has n auxiliary conditions at time t_0 and n auxiliary conditions at time t_f. I.e., it is a "two-point boundary-value problem". Such problems are notoriously hard to analyze (let alone solve), as compared to initial value problems. (E.g., the question of existence and uniqueness of solutions for such problems is largely open.) If instead $x^*(t_0)$ and $p^*(t_0)$ were fully known, than a solution process would be as follows: Starting from $(x^*(t_0), p^*(t_0))$, proceed with a single step of numerical integration of the set of differential equations, with $u^*(t_0)$ a minimizer of the pre-Hamiltonian at time t_0, yielding values of x^*, p^*, and then (via minimization of the pre-Hamiltonian) u^*, at the next time point, and proceed. (Note that, in a real-time application, $x^*(t)$ could possibly be measured rather than computed—but this is not so for $p^*(t)$.)
Given that $(x^*(t_0), p^*(t_0))$ is not fully known though, the above cannot be implemented. A standard way to proceed is then to use a "shooting method", by which the n "missing" initial conditions are first "guessed", and the scheme of the previous paragraph is carried out based on this guess. If extreme luck strikes and, at the end of the process, it so happens that the conditions to be satisfied at t_f are met, then the problem is solved! If not, the same process is restarted with a revised (possibly educated) guess of the missing initial conditions, i.e., another "shot" is taken at the "target", until a satisfactory result is achieved. The choice of the next guess could be driven, e.g., by an optimization process that would work at minimizing an expression of the error in the values obtained at time t_f.
Next we consider in details a (linear) example that can be solved explicitly.

Example 6.2 (see [32]). Consider the motion of a point mass

$$m\ddot{x} + \sigma\dot{x} = u \ , \ x(t), u(t) \in \mathbf{R} \qquad \sigma, m > 0$$

Suppose that $u(t)$ is constrained by

$$|u(t)| \leq 1 \qquad \forall t$$

Starting from $x_0, \dot{x}_0$ we want to reach $x = 0$, $\dot{x} = 0$ in minimum time. Set $x_1 = x$, $x_2 = \dot{x}$, $\alpha = \frac{\sigma}{m} > 0$, $\beta = \frac{1}{m} > 0$. Let $U = [-1, 1]$. The state equation is

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & -\alpha \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \beta \end{bmatrix} \tag{6.52}$$

Since the system is linear, and the objective function is linear in (x, t_f), the Pontryagin Principle is a necessary and sufficient condition of optimality. We seek an optimal control u^* with corresponding state trajectory x^*. The pre-Hamiltonian H is

$$H(t, x, p, u) = p^T(Ax + Bu) = p^T[x_2; -\alpha x_2 + bu] = (p_1 - \alpha p_2)x_2 + bp_2 u.$$

(i) Maximize $H(t, x^*(t), p^*(t), v)$ with respect to v. Pontryagin's principle the yields $u^*(t)$ in terms of t, $x^*(t)$, and $p^*(t)$. Specifically, since $b > 0$,

$$\begin{array}{rl} u^*(t) = +1 & \text{when } p_2^*(t) > 0 \\ -1 & \text{when } p_2^*(t) < 0 \\ \text{anything} & \text{when } p_2^*(t) = 0 \end{array}$$

Unfortunately, this is not a *bona fide* feedback control law, because it involves $p^*(\cdot)$.

(ii) Use the results of (i) to integrate the adjoint equation:

$$\begin{bmatrix} \dot{p}_1^*(t) \\ \dot{p}_2^*(t) \end{bmatrix} = - \begin{bmatrix} 0 & 0 \\ 1 & -\alpha \end{bmatrix} \begin{bmatrix} p_1^*(t) \\ p_2^*(t) \end{bmatrix}, \tag{6.53}$$

yielding

$$\begin{aligned} p_1^*(t) &= p_1^*(0) \\ p_2^*(t) &= \frac{1}{\alpha} p_1^*(0) + e^{\alpha t}(-\frac{1}{\alpha} p_1^*(0) + p_2^*(0)). \end{aligned}$$

(The fact that, for the problem at hand, the adjoint equation does not involves $x^*(t)$ or $u^*(t)$, simplifies matters.) Note that we initially have no information on $p^*(0)$ or $p^*(t_f)$, since $x^*(0)$ and $x^*(t_f)$ are fixed. We need to determine $p^*(0)$ from the knowledge of $x^*(0)$ and $x^*(t_f)$, i.e., we have to determine $p^*(0)$ such that the corresponding u^* steers x^* from x_0 to the target $(0, 0)$. For this, would could just "guess" $p^*(0)$ and use trial-and-error ("shooting"), but the specific structure of the problem (in particular, we only need the sign of $p_2^*(t)$) allows us to do better.

(iii) Plug the solution of the adjoint equation into the expression we obtained for u^*. Observe that, because p_2^* is monotonic in t, $u^*(t)$ can change its value at most once, except if $p_2^*(t) = 0 \quad \forall t$. Clearly, the latter cannot occur since (check it) it would imply that p^* is identically zero, which the theorem rules out. The following cases can arise:

case 1. $-p_1^*(0) + \alpha p_2^*(0) > 0$ $\quad p_2^*$ strictly monotonic increasing. Then either

$$\begin{aligned} u^*(t) &= +1 \quad \forall\, t \\ &\text{or} \\ u^*(t) &= \begin{cases} +1 & t < \hat{t} \\ -1 & t > \hat{t} \end{cases} \text{ for some } \hat{t} \\ &\text{or} \\ u^*(t) &= -1 \quad \forall\, t \end{aligned}$$

case 2. $-p_1^*(0) + \alpha p_2^*(0) < 0$ $\quad p_2^*$ strictly monotonic decreasing. Then either

$$\begin{aligned} u^*(t) &= -1 \quad \forall\, t \\ &\text{or} \\ u^*(t) &= \begin{cases} -1 & t < \hat{t} \\ +1 & t > \hat{t} \end{cases} \text{ for some } \hat{t} \\ &\text{or} \\ u^*(t) &= +1 \quad \forall t \end{aligned}$$

case 3. $-p_1^*(0) + \alpha p_2^*(0) = 0$ $\quad p_2^*$ constant, $p_2^*(t) = \frac{1}{\alpha} p_1^*(0)$. Then either

$$\begin{aligned} &u^*(t) = -1 \quad \forall t \\ &\text{or} \\ &u^*(t) = 1 \quad \forall t \end{aligned}$$

Thus we have narrowed down the possible optimal controls to the controls having the following property.

$$\begin{aligned} &|u^*(t)| = 1 \quad \forall t \\ &u^*(t) \text{ changes sign at most once.} \end{aligned}$$

(iv) Integrate the state equation. The only piece of information we have not used yet is the knowledge of $x(0)$ and $x(t_f)$. We now investigate the question of which among the controls just obtained steers the given initial point to the origin. It turns out that exactly one such control will do the job, hence will be optimal.

We proceed as follows. Starting from $x = (0, 0)$, we apply all possible controls backward in time and check which yield the desired initial condition. Let $\xi(t) = x(t^* - t)$.

1. $u^*(t) = 1 \quad \forall t$

   We obtain the system

$$\begin{aligned} \dot{\xi}_1(t) &= -\xi_2(t) \\ \dot{\xi}_2(t) &= \alpha \xi_2(t) - b \end{aligned}$$

   with $\xi_1(0) = \xi_2(0) = 0$. This gives

$$\xi_1(t) = \frac{b}{\alpha}\left(-t + \frac{e^{\alpha t} - 1}{\alpha}\right), \; \xi_2(t) = \frac{b}{\alpha}(1 - e^{\alpha t}) < 0 \; \forall t > 0.$$

   Also, eliminating t yields

$$\xi_1 = \frac{1}{\alpha}\left(\frac{b}{\alpha}\log(1 - \frac{\alpha}{b}\xi_2) - \xi_2\right).$$

   Thus ξ_1 is increasing, ξ_2 is decreasing (see curve OA in Figure **6.4**).

2. $u^*(t) = -1 \quad \forall t$

$$\xi_1(t) = -\frac{b}{\alpha}(-t + \frac{e^{\alpha t} - 1}{\alpha}), \xi_2(t) = -\frac{b}{\alpha}(1 - e^{\alpha t})$$

Also, eliminating t yields

$$\xi_1 = \frac{1}{\alpha}\left(-\frac{b}{\alpha}\log(1 + \frac{\alpha}{b}\xi_2) + \xi_2\right).$$

Thus ξ_1 is decreasing, ξ_2 is increasing initially (see curve OB in Figure **6.4**)

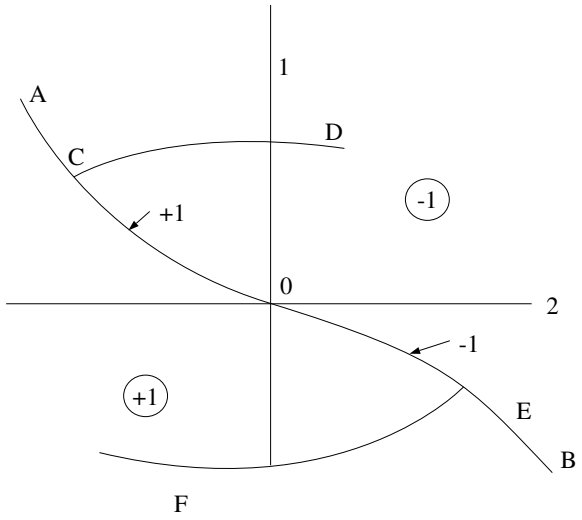


Figure **6.4**: ($\alpha < 1$ is assumed)

Suppose now that $v^*(t) := u^*(t^* - t) = +1$ until some time $\hat{t}$, and -1 afterward. Then the trajectory for ξ is of the type OCD (ξ_1 must keep increasing while $\xi_2 < 0$). If $u^*(t) = -1$ first, then $+1$, the trajectory is of the type OEF.
The reader should convince himself/herself that one and only one trajectory passes through any point in the plane. Thus the given control, inverted in time, must be *the optimal control* for initial conditions at the given point (assuming that an optimal control exists).
We see then that the optimal control $u^*(t)$ has the following properties, *at each time t*

$$\begin{aligned} &\text{if } x^*(t) \text{ is above BOA or on OB } u^*(t) = -1 \\ &\text{if } x^*(t) \text{ is below BOA or on OA } u^*(t) = 1 \end{aligned}$$

Thus we can synthesize the optimal control in feedback form: $u^*(t) = \psi(x^*(t))$ where the function ψ is given by

$$\psi(x_1, x_2) = \begin{cases} 1 & \text{if } (x_1, x_2) \text{ is below BOA or on OA} \\ -1 & \text{above BOA or on OB} \end{cases}$$

BOA is called the switching curve.

■

Example 6.3 Linear quadratic regulator

$$\begin{aligned}&\dot{x}(t) = A(t)x(t) + B(t)u(t) \qquad \text{(dynamics)} \quad A(\cdot), B(\cdot) \text{ continuous}\\ &x(0) = x_0 \text{ given,}\end{aligned}$$

we want to

$$\text{minimize} \quad J = \frac{1}{2}\int_0^1 \left(x(t)^T L(t)x(t) + u(t)^T R(t)u(t)\right) dt$$

where $L(t) = L(t)^T \succeq 0 \quad \forall t, \quad R(t) = R(t)^T \succ 0 \quad \forall t$, say, both continuous, and where the final state is free. Since the final state is free, we can select $p_0^* = 1$, so the pre-Hamiltonian H is given by

$$H(t, x, p, u) = -\frac{1}{2}(x^T L(t)x + u^T R(t)u) + p^T(A(t)x + B(t)u).$$

We first minimize $H(t, x^*(t), p^*(t), v)$ with respect to v in order to find u^* in terms of t, $x^*(t)$, and $p^*(t)$. Since $U = \mathbf{R}^n$, the Pontryagin Principle yields, as $R(t) > 0 \quad \forall t$,

$$R(t)u^*(t) - B^T(t)p^*(t) = 0\ ,$$

which we can explicitly solve for $u^*(t)$ in terms of $p^*(t)$. Thus

$$u^*(t) = R(t)^{-1}B(t)^T p^*(t).$$

Next, we plug this expression into the adjoint equation and state equation, yielding

$$\left.\begin{aligned}\dot{p}^*(t) &= -A^T(t)p^*(t) + L(t)x^*(t)\\ \dot{x}^*(t) &= A(t)x^*(t) + B(t)R(t)^{-1}B(t)^T p^*(t)\end{aligned}\right\} \quad (S)$$

with $p^*(1) = 0$, $x^*(0) = x_0$. Integrating, we obtain

$$\begin{bmatrix}p^*(1)\\ x^*(1)\end{bmatrix} = \begin{bmatrix}\Phi_{11}(1,t) & \Phi_{12}(1,t)\\ \Phi_{21}(1,t) & \Phi_{22}(1,t)\end{bmatrix}\begin{bmatrix}p^*(t)\\ x^*(t)\end{bmatrix}$$

Since $p^*(1) = 0$, the first row yields

$$p^*(t) = \Phi_{11}(1,t)^{-1}\Phi_{12}(1,t)x^*(t) \tag{6.54}$$

provided $\Phi_{11}(1,t)$ is non singular $\forall t$, which was proven to be the case (since $L(t)$ is positive semi-definite; see Theorem 2.1). Now let

$$K(t) = \Phi_{11}(1,t)^{-1}\Phi_{12}(1,t)$$

so that $p^*(t) = K(t)x^*(t)$. We now show that $K(t)$ satisfies a fairly simple equation. Note that $K(t)$ does not depend on the initial state x_0. From (6.54), $p^*(t) = K(t)x^*(t)$. Differentiating, we obtain

$$\begin{aligned}\dot{p}^*(t) &= \dot{K}(t)x^*(t) + K(t)[A(t)x^*(t) + B(t)R(t)^{-1}B(t)^T p^*(t)]\\ &= (\dot{K}(t) + K(t)A(t) + K(t)B(t)R(t)^{-1}B(t)^T K(t))x^*(t) \qquad (6.55)\end{aligned}$$

On the other hand, the first equation in (S) gives

$$\dot{p}^*(t) = -A(t)^T K(t) x^*(t) + L(t) x^*(t) \tag{6.56}$$

Since $K(t)$ does not depend on the initial state x_0, this implies

$$\dot{K}(t) = -K(t)A(t) - A(t)^T K(t) - K(t)B(t)R(t)^{-1}B(t)^T K(t) + L(t) \tag{6.57}$$

For the same reason, $p(1) = 0$ implies $K(1) = 0$. Equation (6.57) is a Riccati equation. It has a unique solution, which is a symmetric matrix. Note that we have

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

which is an optimal *feedback* control law. This was obtained in Chapter 2 using elementary arguments.

■

Exercise 6.19 *(Singular control. From [21].) Obtain the minimum-time control to bring the state from* $[1; 0]$ *to the origin under the dynamics*

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

where $u \in \mathcal{U}$ *and* $u(t) \in U := [-1.1]$ *for all* t*. Show by inspection that the sole optimal control* u^* *is identically zero. Show that, as asserted by Pontryagin's Principle, there exists a* p^* *satisfying the conditions of the principle, but that such* p^* *is such that, at every time* t*, the pre-Hamiltonian along* (x^*, p^*) *is minimized by every* $v \in [-1, 1]$*, so that Pontryagin's Principle is of no help for solving the problem. When such situation arises (which is not that rare in real-life problems), the term singular control (or singular arc, which refers to the time portion of the curve* $u^*(\cdot)$ *on which Pontryagin's Principle is of no help) is used.*

Appendix A

Generalities on Vector Spaces

Note. Some of the material contained in this appendix (and to a lesser extent in the second appendix) is beyond what is strictly needed for this course. We hope it will be helpful to many students in their research and in more advanced courses.

References: [23], [8, Appendix A]

Definition A.1 *Let* $\mathbf{F} = \mathbf{R}$ *or* $\mathbf{C}$ *and let* V *be a set.* V *is a vector space (linear space) over* $\mathbf{F}$ *if two operations, addition and scalar multiplication, are defined, with the following properties*

(a) $\forall x, y \in V,\ x + y \in V$ *and* V *is an Abelian (aka commutative) group for the addition operation (i.e., "+" is associative and commutative, there exists an additive identity* θ *and every* $x \in V$ *has an additive inverse* $-x$*).*

(b) $\forall \alpha \in \mathbf{F}, x \in V,\ \exists \alpha x \in V$ *and*

(i) $\forall x \in V,\ \forall \alpha, \beta \in \mathbf{F}$
$1x = x,\ \alpha(\beta x) = (\alpha\beta)x,\ 0x = \theta, \alpha\theta = \theta$

(ii) $x, y \in V,\ \forall \alpha, \beta \in \mathbf{F}$
$\alpha(x + y) = \alpha x + \alpha y$
$(\alpha + \beta)x = \alpha x + \beta x$

If $\mathbf{F} = \mathbf{R}$, V is said to be a real vector space. If $\mathbf{F} = \mathbf{C}$, it is said to be a complex vector space. Elements of V are often referred to as "vectors" or as "points".

Exercise A.1 *Let* $x \in V$*, a vector space. Prove that* $x + x = 2x$*.*

In the context of optimization and optimal control, the primary emphasis is on *real* vector spaces (i.e., $\mathbf{F} = \mathbf{R}$). In the sequel, unless explicitly indicated otherwise, we will assume this is the case.

Example A.1 $\mathbf{R}$, $\mathbf{R}^n$, $\mathbf{R}^{n\times m}$ (set of $n \times n$ real matrices); the set of all univariate polynomials of degree less than n; the set of all continuous functions $f : \mathbf{R}^n \to \mathbf{R}^k$; the set $C[a, b]$ of all continuous functions over an interval $[a, b] \subset \mathbf{R}$. (All of these with the usual $+$ and $\cdot$ operations.) The 2D plane (or 3D space), with an origin (there is no need for coordinate axes!), with the usual vector addition (parallelogram rule) and multiplication by a scalar.

Exercise A.2 *Show that the set of functions $f : \mathbf{R} \to \mathbf{R}$ such that $f(0) = 1$ is not a vector space.*

Definition A.2 *A set $S \subset V$ is said to be a subspace of V if it is a vector space in its own right with the same "+" and "·" operations as in V.*

Definition A.3 *Let V be a linear space. The family of vectors $\{x_1, \ldots, x_n\} \subset V$ is said to be linearly independent if any relation of the form*

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n = 0$$

implies

$$\alpha_1 = \alpha_2 = \ldots = \alpha_n = 0.$$

Given a finite collection of vectors, its span is given by

$$\text{sp}\,(\{b_1, \ldots, b_n\}) := \left\{ \sum_{i=1}^{n} \alpha_i b_i : \quad \alpha_i \in \mathbf{R},\ i = 1, \ldots, n \right\}.$$

Definition A.4 *Let V be a linear space. The family of vectors $\{b_1, \ldots, b_n\} \subset V$ is said to be a basis for V if (i) $\{b_1, \ldots, b_n\}$ is a linearly independent family, and (ii) $V = \text{sp}(\{b_1, \ldots, b_n\})$.*

Definition A.5 *For $i = 1, \ldots, n$, let $e_i \in \mathbf{R}^n$ be the n-tuple consisting of all zeros, except for a one in position i. Then $\{e_1, \ldots, e_n\}$ is the canonical basis for $\mathbf{R}^n$.*

Exercise A.3 *The canonical basis for $\mathbf{R}^n$ is a basis for $\mathbf{R}^n$.*

Exercise A.4 *Let V be a vector space and suppose $\{b_1, \ldots, b_n\}$ is a basis for V . Prove that, given any $x \in V$, there exists a unique n-tuple of scalars, $\{\alpha_1, \ldots, \alpha_n\}$ such that $x = \sum_{i=1}^{n} \alpha_i b_i$. (Such n-tuple referred to as the coordinate vector of x in basis $\{b_1, \ldots, b_n\}$.)*

Exercise A.5 *Suppose $\{b_1, \ldots, b_n\}$ and $\{b'_1, \ldots, b'_m\}$ both form bases for V. Then $m = n$.*

Definition A.6 *If a linear space V has a basis consisting of n elements then V is said to be finite-dimensional or of dimension n. Otherwise, it is said to be infinite-dimensional.*

Example A.2 $\mathbf{R}^n$ is n-dimenstional. The set $\mathbf{R}^{n\times m}$ of $n \times m$ real matrices form an nm-dimensional vector space. Univariate polynomials of degree $< n$ form an n-dimensional vector space. Points in the plane, once one such point is selected to be the origin (but no coordinate system has been selected yet), form a 2-dimensional vector space. (Come up with a basis for each of the preceding examples!) $C[a, b]$ is infinite-dimensional. The set of all univariate polynomials form an infinite-dimensional vector space, and so does that set of all scalar sequences with no more than finitely many nonzero entries. These last two examples are "isomorphic" to each other.

Exercise A.6 *Prove that the vector space of all univariate polynomials is infinite-dimensional.*

Suppose V is finite-dimensional and let $\{b_1, \ldots, b_n\} \subset V$ be a basis. Then, given any $x \in V$ there exists a unique n-tuple $(\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$ (the coordinates, or components of x) such that $x = \sum_{i=1}^{n} \alpha_i b_i$ (prove it). Conversely, to every $(\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$ corresponds a unique $x = \sum_{i=1}^{n} \alpha_i b_i \in V$. Moreover the coordinates of a sum (in V) are the sums (in $\mathbf{R}$) of the corresponding coordinates, and similarly for scalar multiples. Thus, once a basis has been selected, any n-dimensional vector space (over $\mathbf{R}$) can be thought of as $\mathbf{R}^n$ itself. (Every n-dimensional vectors space is said to be isomorphic to $\mathbf{R}^n$.)

Normed vector spaces

Definition A.7 *A norm on a vector space V is a function $\|\cdot\|: V \to \mathbf{R}$ with the properties*

(i) $\forall \alpha \in \mathbf{R}, \quad \forall x \in V, \quad \|\alpha x\| = |\alpha| \|x\|$ *(positive homegeneous)*

(ii) $\|x\| > 0 \quad \forall x \in V \setminus \{\theta\}$

(iii) $\forall x, y \in V, \quad \|x + y\| \leq \|x\| + \|y\|$ *(triangle inequality).*

A normed vector space is a pair $(V, \|\cdot\|)$ where V is a vector space and $\|\cdot\|$ is a norm on V. Often, when the specific norm is irrelevant or clear from the context, we simply refer to "normed vector space V".

Example A.3 In $\mathbf{R}^n$, $||x||_1 = \sum_{i=1}^{n} |x_i|$, $||x||_2 = \left(\sum_{i=1}^{n} (x_i)^2 \right)^{1/2}$, $||x||_p = \left(\sum_{i=1}^{n} (x_i)^p \right)^{1/p}$, $p \in [1, \infty)$, $||x||_\infty = \max_i |x_i|$; in the space of bounded continuous functions from $\mathbf{R} \to \mathbf{R}$, $||f||_\infty = \sup_t |f(t)|$; in $C[0,1]$, $\|f\|_p = \left(\int_0^1 |f(t)|^p dt \right)^{1/p}$, $p \in [1, \infty)$.

Note that the p-norm requires that $p \geq 1$. Indeed, when $p < 1$, the triangle inequality does not hold. E.g., take $p = 1/2$, $x = (1, 0)$, $y = (0, 1)$.
Once the concept of norm has been introduced, one can talk about balls and convergent sequences.

Definition A.8 *Given a normed vector space $(V, \|\cdot\|)$, a sequence $\{x_n\} \subset V$ is said to be convergent (equivalently, to converge) to $x^* \in V$ if $\|x_n - x^*\| \to 0$ as $n \to \infty$.*

Remark A.1 Examples of sequences that converge in one norm and not in another are well known. For example, it is readily checked that, in the space P of univariate polynomials, equivalently, of scalar sequences with only finitely many nonzero terms, the sequence $\{z^k\}$ (i.e., the kth term in the sequence is the monomial given by the kth power of the variable) does not converge in norm $\|\cdot\|_1$ (sum of absolute values of coefficients) but converges to zero in the norm $||p|| = \sum \frac{1}{i} |p_i|$ where p_i is the coefficient of the ith power term. As another example consider the sequence of piecewise continuous function x_k from $[0, 1]$ to $\mathbf{R}$ with $x_k(t) = k$ for $t \in [0, (1/k^2)]$ and 0 otherwise. Check that this sequence converges to θ in

norm $\|\cdot\|_2$ but does not converge in norm $\|\cdot\|_\infty$. Examples where a sequence converges to two different limits in two different norms are more of a curiosity. The following one is due to Tzvetan Ivanov from Catholic University of Louvain (UCL) and Dmitry Yarotskiy from Ludwig Maximilian Universität München. Consider the space P defined above, and for a polynomial $p \in P$, write $p(z) = \sum_i p_i z^i$. Consider the following two norms on P:

$$\|p\|_a = \max\left\{|p_0|, \max_{i\geq 1}\left\{\frac{|p_i|}{i}\right\}\right\},$$

$$\|p\|_b = \max\left\{|p_0 + \sum_{i\geq 1} p_i|, \max_{i\geq 1}\left\{\frac{|p_i - p_0|}{i}\right\}\right\}.$$

(Note that the coefficients p_i of p in the basis $\{1, z^1, z^2, ...\}$, used in norm a, are replaced in norm b by the coefficients of p in the basis $\{1, z^1 - 1, z^2 - 1, ...\}$.) As above, consider the sequence $\{x_k\} = \{z^k\}$ of monomials of increasing power. It is readily checked that x_k tends to zero in norm a, but that it tends to the "constant" polynomial $z^0 = 1$ in norm b, since $\|x_k - 1\|_b = \frac{2}{k}$ tends to zero.

Inner-product spaces

Definition A.9 *Let V be a (possibly complex) vector space. The function $\langle\cdot,\cdot\rangle : \ V\times V \to \mathbf{F}$ is called an inner product (scalar product) if*

$$\begin{aligned} \langle y, x\rangle &= \overline{\langle x, y\rangle} \quad \forall x, y \in V \\ \langle x, \alpha y + \beta z\rangle &= \alpha\langle x, y\rangle + \beta\langle x, z\rangle \quad \forall x, y, z, \in V,\ \alpha, \beta \in \mathbf{F} \\ \langle x, x\rangle &> 0 \quad \forall x \neq \theta \end{aligned}$$

A vector space endowed with an inner-product is termed inner-product space, or pre-Hilbert space. It is readily checked that, if $\langle\cdot,\cdot\rangle$ is an inner product on V, then the function $\|\cdot\|$ given by

$$\|x\| = \langle x, x\rangle^{1/2}$$

is a norm on V (the norm *derived from* the inner product). (Check it.) Hence every inner-product space is a normed vector space. Unless otherwise noted, the notation $\|x\|$, when x belongs to an inner-product space refers to $\langle x, x\rangle^{1/2}$.

Remark A.2 Some authors use a sightly different definition for the inner product, with the second condition replaced by $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$, or equivalently (given the fourth condition) $\langle x, \alpha y\rangle = \bar{\alpha}\langle x, y\rangle$. Note that the difference is merely notational, since $(x, y) := \langle y, x\rangle$ satisfies such definition. The definition given here has the advantage that it is satisfied by the standard dot product in $\mathbf{C}^n$, $x \cdot y := x^* y = \bar{x}^T y$, rather than by the slightly less "friendly" $x^T \bar{y}$. (In $\mathbf{R}^n$, these two definitions are equivalent, given the symmetry property $\langle y, x\rangle = \langle x, y\rangle$.)

Example A.4 Let V be (real and) finite-dimensional, and let $\{b^i\}_{i=1}^n$ be a basis. Then $\langle x, y\rangle = \sum_{i=1}^n \xi_i \eta_i$, where $\xi \in \mathbf{R}^n$ and $\eta \in \mathbf{R}^n$ are the vectors of coordinates of x and y in basis $\{b^i\}_{i=1}^n$, is an inner product. It is known as the Euclidean inner product associated to basis $\{b^i\}_{i=1}^n$.

The following exercise characterizes all inner products on finite-dimensional vector spaces.

Exercise A.7 *Let V be an n-dimensional inner-product space over $\mathbf{R}$, with a basis $\{b_i\}$. Prove the following statement. A mapping $\langle \cdot, \cdot \rangle : V \times V \to \mathbf{R}$ is an inner product on V if and only if there exists a symmetric positive definite matrix M such that*

$$\langle x, y \rangle = \xi^T M \eta \quad \forall x, y \in V,$$

where ξ (resp., η) is the column-vector of components of x (resp., y) in basis $\{b_i\}$.

Note that, if $M = M^T \succ 0$, then $M = A^T A$ for some square non-singular matrix A, so that $x^T M u = (Ax)^T (Ay) = x'^T y'$, where x' and y' are the coordinates of x and y in a new basis. Hence, up to a change of basis, an inner product over $\mathbf{R}^n$ always takes the form $x^T y$.

Example A.5 $V = C[t_0, t_\mathrm{f}]$, the space of all continuous functions from $[t_0, t_\mathrm{f}]$ to $\mathbf{R}$, with

$$\langle x, y \rangle = \int_{t_0}^{t_\mathrm{f}} x(t) y(t) \, dt.$$

This inner product is known as the L_2 inner product.

Example A.6 $V = C[t_0, t_\mathrm{f}]^m$, the space of continuous functions from $[t_0, t_\mathrm{f}]$ to $\mathbf{R}^m$. For $x(\cdot) = (\xi_1(\cdot), \ldots, \xi_m(\cdot))$, $y(\cdot) = (\eta_1(\cdot), \ldots, \eta_m(\cdot))$

$$\langle x, y \rangle = \int_{t_0}^{t_\mathrm{f}} \sum_{i=1}^{m} \xi_i(t) \eta_i(t) \, dt = \int_{t_0}^{t_\mathrm{f}} x(t)^T y(t) \, dt.$$

This inner product is again known as the L_2 inner product. The same inner product is valid for the space of piecewise-continuous functions $\mathcal{U}$ considered in Chapter 2.

Exercise A.8 *(Gram matrix). Let V be an inner-product space, let $v_1, \ldots, v_k \in V$ and let G be a $k \times k$ matrix with (i, j) entry given by $G_{ij} := \langle v_i, v_j \rangle$. Then $G \succeq \theta$, and $G \succ \theta$ if and only if the v_is are linearly independent.*

Exercise A.9 *Let V be the vector space of univariate quadratic polynomials (more precisely, polynomials of degree no higher than two) over $[0, 1]$, endowed with the inner product*

$$\langle p, q \rangle := \int_0^1 p(t) q(t) dt \quad \forall p, q \in V.$$

For $p, q \in V$, let $P, Q \in R^3$ be the associated vectors of coefficients (say, $P = [p_0; p_1; p_2]$). Obtain a symmetric matrix S such that

$$\langle p, q \rangle = P^T S Q, \quad \forall p, q \in V.$$

Theorem A.1 *(Cauchy-Bunyakovskii-Schwarz (CBS) inequality, after Baron Augustin-Louis Cauchy, French mathematician, 1789–1857; Viktor Ya. Bunyakovsky, Russian mathematician, 1804–1889; K. Hermann A. Schwarz, German mathematician, 1843–1921.) If V is a vector space with inner product $\langle \cdot, \cdot \rangle$,*

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\| \quad \forall x, y \in V.$$

Moreover both sides are equal if and only if $x = \theta$ or $y = \lambda x$ for some $\lambda \in \mathbf{R}$.

Exercise A.10 *Prove the CBS inequality. [Hint: $\langle x + \alpha y,\ x + \alpha y \rangle \geq 0 \quad \forall \alpha \in \mathbf{R}$.]*

Theorem A.2 *(Parallelogram law.) In an inner-product space V, the sum of the squares of the norms of the two diagonals of a parallelogram is equal to the sum of the squares of the norms of its four sides, i.e., for every $x, y \in V$,*

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2), \tag{A.1}$$

where $\| \cdot \|$ is the norm induced by the inner product.

Exercise A.11 *Prove Theorem A.2.*

Fact. Given a normed space V whose norm satisfies (A.1),

$$\langle x, y \rangle := \frac{1}{4} \left(\|x + y\|^2 - \|x - y\|^2 \right) \quad \forall x, y \in V$$

is an inner product; further, $\langle x, x \rangle = \|x\|^2$ for all $x \in V$.

Definition A.10 *Given $x, y \in V$, x and y are said to be orthogonal if $\langle x, y \rangle = 0$. Given a set $S \subset V$, the set*

$$S^\perp = \{ x \in V \ : \ \langle x, s \rangle = 0 \quad \forall s \in S \}.$$

is called orthogonal complement of S.

Exercise A.12 *Prove that $S^\perp$ is a subspace.*

Exercise A.13 *Prove that, if S is a subset of an inner-product space, then $S \subseteq (S^\perp)^\perp$.*

Example A.7 Equality may not hold even when S is a subspace. E.g., let V be the space of continuous functions with the L_2 inner product, and $S \subset V$ the set of all polynomials. Then $S \neq (S^\perp)^\perp = V$.

Exercise A.14 *Consider a plane (e.g., a blackboard or sheet of paper) together with a point in that plane declared to be the origin. With an origin in hand, we can add vectors (points) in the plane using the parallelogram rule, and multiply vectors by scalars, and it is readily checked that all vector space axioms are satisfied; hence we have a vector space V. Two non-collinear vectors e_1 and e_2 of V form a basis for V. Any vector $x \in V$ is now uniquely specified by its components in this basis; let us denote by x^E the column vector of its components. Now, let us say that two vectors $x, y \in V$ are perpendicular if the angle $\theta(x, y)$ between them (e.g., measured with a protractor on your sheet of paper) is $\pi/2$, i.e., if $\cos\theta(x, y) = 0$. Clearly, in general, $(x^E)^T y^E = 0$ is not equivalent to x and y being perpendicular. (In particular, of course, $(e_1^E)^T e_2^E = 0$ (since $e_1^E = [1, 0]^T$ and $e_2^E = [0, 1]^T$), while e_1 and e_2 may not be perpendicular to each other.)* **Question:** *Determine a symmetric positive-definite matrix S such that $\langle x, y\rangle_S := (x^E)^T S y^E = 0$ if and only if x and y are perpendicular.*

Gram-Schmidt ortho-normalization (Erhard Schmidt, Balto-German mathematician, 1876–1959)

Let V be a finite-dimensional inner product space and let $\{b_1, \ldots, b_n\}$ be a basis for V. Let

$$\begin{aligned} u_1 &= b_1,\ e_1 = \frac{u_1}{\|u_1\|_2} \\ u_k &= b_k - \sum_{i=1}^{k-1} \langle b_k, e_i\rangle e_i, \quad e_k = \frac{u_k}{\|u_k\|_2},\ k = 2; \ldots, n. \end{aligned}$$

Then $\{e_1, \ldots, e_n\}$ is an orthonormal basis for V, i.e., $\|e_i\|_2 = 1$ for all i, and $\langle e_i, e_j\rangle = 0$ for all $i \neq j$ (Check it).

Closed sets, open sets

Definition A.11 *Let V be a normed vector space. A subset $S \subset V$ is closed (in V) if every $x \in V$ for which there is a sequence $\{x_n\} \subset S$ that converges to x, belongs to S. A subset $S \subset V$ is open if its complement is closed. The closure $\mathrm{cl}S$ of a set S is the smallest closed set that contains S, i.e., the intersection of all closed sets that contain S (see the next exercise). The interior $\mathrm{int}S$ of a set S is the largest open set that is contained in S, i.e., the union of all open sets that contain S.*

Exercise A.15 *Prove the following, which shows that Definition A.11 is valid. The intersection $\cap_\alpha S_\alpha$ of an arbitrary (possibly uncountable) family of closed sets is closed. The union $\cup_\alpha S_\alpha$ of an arbitrary (possibly uncountable) family of open sets is open.*

Exercise A.16 *Prove that $\mathrm{int}S = (\mathrm{cl}S^{\mathrm{c}})^{\mathrm{c}}$ and $\mathrm{cl}S = (\mathrm{int}S^{\mathrm{c}})^{\mathrm{c}}$. .*

Exercise A.17 *Let V be a normed vector space and let $S \subset V$. Show that the closure of S is the set of all limit points of sequences of S that converge in V.*

Exercise A.18 *Show that a subset S of a normed vector space is open if and only if given any $\hat{x} \in S$ there exists $\epsilon > 0$ such that $\{x : \|x - \hat{x}\| < \epsilon\} \subset S$.*

Exercise A.19 *Show that, in a normed linear space, every finite-dimensional subspace is closed. In particular, all subspaces of $\mathbf{R}^n$ are closed.*

Example A.8 Given a positive integer n, the set of polynomials of degree $\leq n$ in one variable over [0,1] is a finite-dimensional subspace of $C[0,1]$. The set of all univariate polynomials over [0,1] is an infinite-dimensional subspace of $C[0,1]$; it is not closed in either of the norms of Example A.3 (prove it).

Exercise A.20 *Given any set E in an inner product space, $E^\perp$ is a closed subspace (in the norm derived from the inner product).*

Exercise A.21 *If S is a subspace of an inner product space, $(S^\perp)^\perp = \mathrm{cl}S$. In particular, if S is a finite-dimensional subspace of an inner product space, then $S^{\perp\perp} = S$. [Hint: choose an orthogonal basis for S.]*

Definition A.12 *A set S in a normed vector space is bounded if there exists $\rho > 0$ s.t. $S \subset \{x : \|x\| \leq \rho\}$*

Definition A.13 *A subset S of a normed vector space is said to be (sequentially) compact if, given any sequence $\{x_k\}_{k=0}^{\infty} \subset S$, there exists a sub-sequence that converges to a point of S, i.e., there exists an infinite index set $K \subseteq \{0, 1, 2, \ldots\}$ and $x^* \in S$ such that $x_k \longrightarrow x^*$ as $k \longrightarrow \infty$, $k \in K$.*

Remark A.3 The concept of compact set is also used in more general "topological spaces" than normed vector spaces, but with a different definition ("Every open cover includes a finite sub-cover"). In such general context, the concept introduced in Definition A.13 is referred to as "sequential compactness" and is weaker than compactness. In the case of normed vector spaces (or, indeed, of general "metric spaces"), compactness and sequential compactness are equivalent.

Fact. (Bolzano-Weierstrass, Heine-Borel). Let $S \subset \mathbf{R}^n$. Then S is compact if and only if it is closed and bounded. (For a proof see, e.g., wikipedia.)

Example A.9 The "simplest" infinite-dimensional vector space may be the space P of univariate polynomials (of arbitrary degrees), or equivalently the space of finite-length sequences (infinite sequences with finitely many nonzero entries). Consider P together with the ℓ_∞ norm (maximum absolute value among the (finitely many) non-zero entries). The closed unit ball in P is not compact. For example, the sequence x_k, where x_k is the monomial z^k, z being the unknown, which clearly belongs to the unit ball, does not have a converging sub-sequence. Similarly, the close unit ball B in ℓ_1 (absolutely summable real sequences) is not compact. For example, the following continuous function is unbounded over B (example due to Nuno Martins):

$$f(x) = \max\{x_n\} \text{ if } x_n \leq \frac{1}{2}\forall n, \text{ and } \frac{1}{2} + \max\{n(x_n - \frac{1}{2})\} \text{ otherwise.} \qquad \blacksquare$$

In fact, it is a important result due to Riesz that the closed unit ball of a normed vector space is compact if and only if the space if finite-dimensional. (See, e.g., [18, Theorem 6, Ch. 5].)

Definition A.14 *Supremum, infimum. Given a set $S \subseteq \mathbf{R}$, the supremum $\sup S$ (resp. infimum $\inf S$) of S is the lowest/leftmost (resp., highest/rightmost) x such that $s \leq x$ (resp. $s \geq x$) for all $s \in S$. If there is no such x, then $\sup S := +\infty$ (resp. $\inf S := -\infty$), and if S is empty, then $\sup S := -\infty$ and $\inf S := +\infty$. Finally, if $\sup S$ (resp. $\inf S$) belongs to S, the supremum (resp. infimum) is said to be attained, and is known as the maximum $\max S$ (resp. minimum $\min S$) of S.*

It is an axiom of $\mathbf{R}$ (i.e., part of the definition of $\mathbf{R}$) that every upper-bounded subset of $\mathbf{R}$ has a finite supremum and every lower-bounded subset of $\mathbf{R}$ has a finite infimum.

Definition A.15 *Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be normed spaces, and let $f : V \to W$. Then f is continuous at $\hat{x} \in v$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that $\|f(x) - f(\hat{x})\|_W < \epsilon$ for all x such that $\|x - \hat{x}\|_V < \delta$. If f is continuous at $\hat{x}$ for all $\hat{x} \in V$, it is said to be continuous.*

Exercise A.22 *Prove that, in any normed vector space, the norm is continuous with respect to itself.*

Exercise A.23 *Let V be a normed space and let S be a compact set in V. Let $f : V \to \mathbf{R}$ be continuous. Then there exists $\underline{x}, \bar{x} \in S$ such that*

$$f(\underline{x}) \leq f(x) \leq f(\bar{x}) \quad \forall x \in S$$

i.e., the supremum and infimum of $\{f(x) \ : \ x \in S\}$ are attained.

Exercise A.24 *Let $f : V \to \mathbf{R}$ be continuous, V a normed vector space. Then, for all $\alpha \in \mathbf{R}$, the sub-level set $\{x : f(x) \leq \alpha\}$ is closed.*

Definition A.16 *Let $f : \mathbf{R}^n \longrightarrow \mathbf{R}^m$ and let $S \subset \mathbf{R}^n$. Then f is uniformly continuous over S if for all $\epsilon > 0$ there exists $\delta > 0$ such that*

$$\left.\begin{array}{r} x, y \in S \\ \|x - y\| < \delta \end{array}\right\} \Longrightarrow \|f(x) - f(y)\| < \epsilon$$

Exercise A.25 *Let $S \subset \mathbf{R}^n$ be compact and let $f : \mathbf{R}^n \longrightarrow \mathbf{R}^m$ be continuous over S. Then f is uniformly continuous over S.*

For a given vector space, it is generally possible to define many different norms.

Definition A.17 *Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on a same vector space V are equivalent if there exist $N, M > 0$ such that $\forall x \in V, \quad N||x||_a \leq ||x||_b \leq M||x||_a$.*

Exercise A.26 *Verify that the above is a bona fide equivalence relationship, i.e., that it is reflexive, symmetric and transitive.*

We will see that equivalent norms can often be used interchangeably. The following result is thus of great importance.

Exercise A.27 *Prove that, if V is a finite-dimensional vector space, all norms on V are equivalent. [Hint. First select an arbitrary basis $\{b_i\}_{i=1}^n$ for V. Then show that $\|\cdot\|_\infty : V \to \mathbf{R}$, defined by $\|x\|_\infty := \max |x_i|$, where the x_i's are the coordinates of x in basis $\{b_i\}$, is a norm. Next, show that, if $\|\cdot\|$ is an arbitrary norm, it is a continuous function from $(V, \|\cdot\|_\infty)$ to $(\mathbf{R}, |\cdot|)$. Finally, conclude by using Exercise A.23.]*

Exercise A.28 *Does the result of Exercise A.27 extend to some infinite-dimensional context, say, to the space of univariate polynomials of arbitrary degrees? If not, where does your proof break down?*

Exercise A.29 *Suppose that $\|\cdot\|_a$ and $\|\cdot\|_b$ are two equivalent norms on a vector space V and let the sequence $\{x_k\} \subset V$ be such that the sequence $\{\|x_k\|_a^{1/k}\}$ converges. Then the sequence $\{\|x_k\|_b^{1/k}\}$ also converges and both limits are equal. Moreover, if V is a space of matrices and x_k is the kth power of a given matrix A, then the limit exists and is the spectral radius $\rho(A)$, i.e., the radius of the smallest disk centered at the origin containing all eigenvalues of A.*

Definition A.18 *Given a normed vector space $(V, \|\cdot\|)$, a sequence $\{x_n\} \subset V$ is said to be a Cauchy sequence if $\|x_n - x_m\| \to 0$ as $n, m \to \infty$, i.e., if for every $\epsilon > 0$ there exists N such that $n, m \geq N$ implies $\|x_n - x_m\| < \epsilon$.*

Exercise A.30 *Every convergent sequence is a Cauchy sequence.*

Exercise A.31 *Let $\{x_i\} \subset V$ and $\|.\|_a$ and $\|.\|_b$ be 2 equivalent norms on V. Then a set S us open (resp. closed) w.r.t. norm $\|.\|_a$ if and only if it is open (resp. closed) w.r.t. norm $\|.\|_b$. Furthermore*

(i) $\{x_i\}$ converges to x^* with respect to norm a if and only if it converges to x^* with respect to norm b.

(ii) $\{x_i\}$ is Cauchy w.r.t. norm a if and only if it is Cauchy w.r.t. norm b

Hence, in $\mathbf{R}^n$, we can talk about converging sequences and Cauchy sequences without specifying the norm.

Exercise A.32 *Suppose $\{x^k\} \subset \mathbf{R}^n$ is such that, for some $b \in \mathbf{R}^n$ and $\alpha > 0$, $\|x^{k+1} - x^k\| \leq \alpha b^T(x^{k+1} - x^k)$ for all k. Prove that, if $\{b^T x^k\}$ is bounded, then $\{x^k\}$ converges. Hint: Show that the sum $\sum_{k=0}^{N} \|x^{k+1} - x^k\|$ remains bounded as $N \to \infty$, and that this implies that $\{x^k\}$ is Cauchy. Such situation may arise when attempting to construct a maximizing sequence for the maximization of $b^T x$ over a certain set.*

Definition A.19 *A normed linear space V is said to be complete if every Cauchy sequence in V converges to a point of V.*

Exercise A.33 *Prove that $\mathbf{R}^n$ is complete*

Hence every finite-dimensional vector space is complete. Complete normed spaces are known as Banach spaces (Stefan Banach, Polish mathematician, 1892–1945). Complete inner product spaces (with norm derived from the inner product) are known as Hilbert spaces (David Hilbert, German mathematician, 1862–1943). $\mathbf{R}^n$ is a Hilbert space.

Example A.10 $C[0,1]$ with the sup norm is Banach. The vector space of polynomials over [0,1], with the sup norm, is not Banach, nor is $C[0,1]$ with an L_p norm, p finite. The space of square-summable real sequences, with norm derived from the inner product

$$\langle x, y \rangle = \sqrt{\sum_{i=1}^{\infty} x_i y_i}$$

is a Hilbert space.

Exercise A.34 *Exhibit an example showing that the inner product space of Example A.5 is not a Hilbert space. (Hence the set $\mathcal{U}$ of admissible controls is not complete. Even when enlarged to include piecewise-continuous functions it is still not complete.)*

While the concepts of completeness and closedness are somewhat similar in spirit, they are clearly distinct. In particular, completeness applies to vector spaces and closedness applies to subsets of vectors spaces. Yet, for example, the vector space P of univariate polynomials is not complete under the sup norm, but it is closed (as a subset of itself), and its closure (in itself) is itself; indeed, all vector spaces are closed subsets of themselves. Note however that P can also be thought of as a subspace of the (complete under the sup norm) space $C([0,1])$ of continuous functions over $[0,1]$. Under the sup norm, P is not a closed (in $C([0,1])$) subspace. (Prove it.) More generally, closedness and completeness are related by the following result.

Theorem A.3 *Let V be a normed vector space, and let W be a Banach space that contains a subspace V' with the properly that V and V' are isometric normed vector spaces. (Two normed vector spaces are isometric if they are isomorphic as vector spaces and the isomorphism leaves the norm invariant.) Then V' is closed (in W) if and only if V is complete. Furthermore, given any such V there exists such W and V' such that the closure of V' (in W) is W itself. (Such W is known as the completion of V. It is isomorphic to a certain normed space of equivalence classes of Cauchy sequences in V.) In particular, a subspace S of a Banach space V is closed if and only if it is complete as a vector space (i.e., if and only if it is a Banach space).*

You may think of incomplete vector spaces as being "porous", with pores being elements of the completion of the space. You may also think of non-closed subspaces as being porous; pores are elements of the "mother" space whenever that space is complete.

Direct sums and orthogonal projections

Definition A.20 *Let S and T be two subspaces of a linear space V. The sum $S+T := \{s+t: \quad s\in S, t\in T\}$ of S and T is called a direct sum if $S\cap T=\{\theta\}$. The direct sum is denoted $S\oplus T$.*

Exercise A.35 *Given two subspaces S and T of V, $V=S\oplus T$ if and only if for every $v\in V$ there is a unique decomposition $v=s+t$ such that $s\in S$ and $t\in T$.*

It can be shown that, if S is a *closed* subspace of a Hilbert space H, then

$$H = S\oplus S^{\perp}.$$

Equivalently, $\forall x\in H$ there is a unique $y\in S$ such that $x-y\in S^{\perp}$. The (linear) map $P: x\mapsto y$ is called orthogonal projection of x *onto the subspace* S.

Exercise A.36 *Prove that if P is the orthogonal projection onto S, then*

$$\|x-Px\| = \inf\{\|x-s\|,\ s\in S\},$$

where $\|\cdot\|$ is the norm derived from inner product.

Linear maps

Definition A.21 *Let V, W be vector spaces. A map $L: V\to W$ is said to be linear if*

$$L(\alpha x_1+\beta x_2) = \alpha L(x_1)+\beta L(x_2) \qquad \begin{array}{l}\forall \alpha,\beta\in\mathbf{R}\\ \forall x_1x_2\in V\end{array}$$

Exercise A.37 *Let $A: V\to W$ be linear, where V and W have dimension n and m, respectively, and let $\{b_i^V\}$ and $\{b_i^W\}$ be bases for V and W. Show that A can be represented by a matrix, i.e., there exists a matrix M_A such that, for any $x\in V$, $y\in W$ such that $y=A(x)$, it holds that $v_y = M_A\cdot v_x$, where v_x and v_y are $n\times 1$ and $m\times 1$ matrices (column vectors) with entries given by the components of x and y, and "$\cdot$" is the usual matrix product. The entries in the ith column of M_A are the components in $\{b_i^W\}$ of Ab_i^V.*

Linear maps from V to W themselves form a vector space $\mathcal{L}(V,W)$.
Given a linear map $L\in\mathcal{L}(V,W)$, its *range* is given by

$$\mathcal{R}(L) = \{Lx: \quad x\in V\}\subseteq W$$

and its *nullspace* (or *kernel*) by

$$\mathcal{N}(L) = \{x\in V: \quad Lx=\theta_W\}\subseteq V.$$

Exercise A.38 *$\mathcal{R}(L)$ and $\mathcal{N}(L)$ are subspaces of W and V respectively.*

Exercise A.39 *Let V and W be vector spaces, with V finite-dimensional, and let $L : V \to W$ be linear. Then $\mathcal{R}(L)$ is also finite-dimensional, of dimension no larger than that of V.*

Definition A.22 *A linear map $L : V \to W$ is said to be surjective if $\mathcal{R}(L) = W$; it is said to be injective if $\mathcal{N}(L) = \{\theta_V\}$.*

Exercise A.40 *Prove that a linear map $A : \mathbf{R}^n \to \mathbf{R}^m$ is surjective if and only if the matrix that represents it has full row rank, and that it is injective if and only if the matrix that represents is has full column rank.*

Bounded linear maps

Let V be $C[0,1]$ with the L_1 norm, let $W = \mathbf{R}$, and let $Lx = x(0)$; L is a linear map. Then, $|Lx|$ reaches arbitrarily large values without $\|x\|$ being large, for instance, on the unit ball in L_1: with $x_k(t) := k\exp(-kt)$, $\|x_k\|_1 = 1 - \exp(-k) < 1$ for all $k > 0$, but $x(0) = k$ becomes arbitrarily large as k increases. Such linear map is said to be "unbounded". For another example, let V be the vector space of continuously differentiable functions on $[0,1]$ with $\|x\| = \max_{t\in[0,1]} |x(t)|$. Let $W = \mathbf{R}$ and let L be defined to $Lx = x'(0)$. Then L is an unbounded linear map. (Think of the sequence $x_k(t) = \sin(kt)$.)

Definition A.23 *Let V, W be normed vector spaces. A linear map $L \in \mathcal{L}(V,W)$ is said to be bounded if there exists $c > 0$ such that $\|Lx\|_W \leq c\|x\|_V$ for all $x \in V$. If L is bounded, the operator norm (induced norm) of L is defined by*

$$\|L\| = \inf\{c : \quad \|Lx\|_W \leq c\|x\|_V \quad \forall x \in V\}$$

The set of bounded linear maps from V to W is a vector space. It is denoted by $\mathcal{B}(V,W)$.

Exercise A.41 *Show that $\|\cdot\|$ as defined above is a norm on $\mathcal{B}(V,W)$*

Exercise A.42 *Let V be a finite-dimensional normed vector space, and let L be a linear map over V. Prove that L is bounded.*

Unbounded linear maps exist whenever that space is infinite dimensional, as shown with the following example with the "smallest" infinite-dimensional space.

Example A.11 [23, Example 4, p.105] On the space of finitely nonzero infinite sequences with norm equal to the maximum of the absolute values of the entries, define, for $x : \{\xi_1, \ldots, \xi_n, 0, 0, \ldots\}$,

$$f(x) = \sum_{k=1}^{\infty} k\xi_k.$$

The functional f is clearly linear but unbounded.

Now for an example from linear system theory.

Example A.12 Let L be a linear time-invariant dynamical input-output system, say with scalar input and scalar output. If the input space and output space both are endowed with the ∞-norm, then the induced norm of L is

$$\|L\| = \int |h(t)|dt$$

where h is the system's unit impulse response. L is bounded if and only if h is absolutely integrable, which is the case if and only if the system is bounded-input/bounded-output stable—i.e., if the ∞-norm of the output is finite whenever that of the input is.

An important characterization of bounded linear maps is as follows.

Exercise A.43 *Let V, W be normed vector spaces and let $L \in \mathcal{L}(V, W)$. The following are equivalent: (i) L is bounded; (ii) L is continuous over V; (iii) L is continuous at θ_V. Moreover, if L is bounded then $\mathcal{N}(L)$ is closed.*

Example A.13 Let V be the vector space of continuously differentiable functions on $[0, 1]$ with $\|x\| = \max_{t\in[0,1]} |x(t)|$. Let $W = \mathbf{R}$ and again let L be defined to $Lx = x'(0)$, an unbounded linear map. It can be verified that $\mathcal{N}(L)$ is not closed. For example, let $x_k(t) = \frac{kt^3}{1+kt^2}$. Then $x_k \in \mathcal{N}(L)$ for all k and $x_k \to \hat{x}$ with $\hat{x}(t) = t$, but $L\hat{x} = 1$.

Exercise A.44 *Show that*

$$\|L\| = \sup_{\|x\|_V \le 1} \|Lx\|_W = \sup_{x\neq\theta} \frac{\|Lx\|_W}{\|x\|_V} = \sup_{\|x\|_V=1} \|Lx\|_W \ .$$

Also, $\|Lx\|_W \le \|L\|\|x\|_V$ for all $x \in V$.

Exercise A.45 *Prove that*

$$\|AB\| \le \|A\| \cdot \|B\| \quad \forall A \in \mathcal{B}(W, Z), \quad B \in \mathcal{B}(V, W)$$

with AB defined by

$$AB(x) = A(B(x)) \quad \forall x \in V.$$

Theorem A.4 (Riesz–Fréchet Theorem) *(e.g., [28, Theorem 4-12]). (Frigyes Riesz, Hungarian mathematician, 1880–1956; Maurice R. Fréchet, French mathematician, 1898–1973.) Let H be a Hilbert space and let $L \in \mathcal{B}(H, \mathbf{R})$ (i.e., L is a bounded linear functional on H). Then there exists $\ell \in H$ such that*

$$L(x) = \langle \ell, x \rangle \quad \forall x \in H.$$

Adjoint of a linear map

Definition A.24 *Let V, W be two spaces endowed with inner products $\langle\cdot,\cdot\rangle_V$ and $\langle\cdot,\cdot\rangle_W$ respectively and let $L \in \mathcal{L}(V,W)$. An adjoint map to L is a map $L^* : W \to V$ satisfying*

$$\langle L^* y, x\rangle_V = \langle y, Lx\rangle_W \quad \forall x \in V, y \in W.$$

Fact. If V is a Hilbert space, then every $L \in \mathcal{B}(V,W)$ has an adjoint.

Exercise A.46 *Suppose L has an ajoint map L^*. Show that (i) L has no other adjoint map, i.e., the adjoint map (when it exists) is unique; (ii) L^* linear; (iii) L^* has an adjoint, with $(L^*)^* = L$; (iv) if L is bounded, then L^* also is, and $\|L^*\| = \|L\|$.*

Exercise A.47 $(A+B)^* = A^* + B^*$; $(\alpha A)^* = \alpha A^*$; $(AB)^* = B^* A^*$; $(A^*)^* = A$.

When $L^* = L$ (hence $V = W$), L is said to be self-adjoint.

Exercise A.48 *Let L be a linear map with adjoint L^*. Show that $\langle x, Lx\rangle = \frac{1}{2}\langle x, (L+L^*)x\rangle$ for all x.*

Exercise A.49 *Let L be a linear map from V to W, where V and W are finite-dimensional vector spaces, and let M_L be its matrix representation in certain bases $\{b_V^i\}$ and $\{b_W^i\}$. Let S_n and S_m be $n \times n$ and $m \times m$ symmetric positive definite matrices. Obtain the matrix representation of L^* under the inner products $\langle x_1, x_2\rangle_V := \xi_1^T S_n \xi_2$ and $\langle y_1, y_2\rangle_W := \eta_1^T S_m \eta_2$, where ξ_k and η_k, $k = 1, 2$, are corresponding vectors of coordinates in bases $\{b_V^i\}$ and $\{b_W^i\}$. In particular show that if the Euclidean inner product is used for both spaces (i.e., S_n and S_m are both the identity), then $M_{L^*} = M_L^T$, so that L is self-adjoint if and only if M_L is symmetric.*

Exercise A.50 *Let $\mathcal{U}$ be given by Example A.6. Consider the map $L : \mathcal{U} \to \mathbf{R}^n$ given by*

$$L(u) = \int_{t_0}^{t_\mathrm{f}} G(\sigma) u(\sigma) d\sigma,$$

with $G : [t_0, t_\mathrm{f}] \to \mathbf{R}^{n\times m}$ continuous. Assume the inner product on $\mathcal{U}$ given by $\langle u, v\rangle := \int_{t_0}^{t_\mathrm{f}} u(t)^T v(t) dt$ and, in $\mathbf{R}^n$, $\langle x, y\rangle := x^T y$. Then L is linear and bounded. Verify that L has an adjoint $L^* : \mathbf{R}^n \to \mathcal{U}$ given by

$$(L^* x)(t) = G(t)^T x.$$

[Proving boundedness of L takes some effort. Hint: If $\varphi : [0,1] \to \mathbf{R}$ is continuous (or piecewise continuous), then $\|\varphi\|_1 \le \|\varphi\|_2$, which can be proved by craftily invoking the CBS inequality.]

Exercise A.51 *Prove that the linear map L in Exercise A.50 is bounded. [Hint: $(\int_0^1 x(t)dt)^2 \leq \int_0^1 x(t)^2 dt$, which can be proved from Jensen's inequality.]*

Exercise A.52 *Consider the linear time-invariant state-space model (A, B, and C are real)*

$$\dot{x} = Ax + Bu \tag{A.2}$$
$$y = Cx \tag{A.3}$$

and the associated transfer-function matrix $G(s) = C(sI - A)^{-1}B$ (for $s \in \mathbf{C}$). The time-invariant state-space model, with input y and output v,

$$\dot{p} = -A^T p + C^T y \tag{A.4}$$
$$v = -B^T p \tag{A.5}$$

is said to be adjoint to (A.2)-(A.3). (Note the connection with system (2.19), when $L = C^T C$.) Show that the transfer-function matrix $G^\sim$ associated with the adjoint system (A.4)-(A.5) is given by $G^\sim(s) = G(-s)^T$; in particular, for every $\omega \in \mathbf{R}$, $G^\sim(j\omega) = G(j\omega)^H$, where $j := \sqrt{-1}$ and a H superscript denotes the complex conjugate transpose of a matrix (which is the adjoint with respect to the complex Euclidean inner product $\langle u, v\rangle = \xi^H \eta$, where ξ and η are the vectors of coordinates of u and v, and ξ^H is the complex conjugate transpose of ξ). This justifies referring to p as the adjoint variable, or co-state. The triple $(-A^T, C^T, -B^T)$ is also said to be dual to (A, B, C).

Exercise A.53 *Let $\mathcal{G}$ be a linear map from $C[0,1]^m$ to $C[0,1]^p$, m and p positive integers, defined by*

$$(\mathcal{G}u)(t) = \int_0^t G(t,\tau)u(\tau)d\tau, \quad t \in [0,1], \tag{A.6}$$

where the matrix $G(t,\tau)$ depends continuously on t and τ. Let $C[0,1]^m$ and $C[0,1]^p$ be endowed with the L_2 inner product, i.e., $\langle r, s\rangle = \int_0^1 r(t)^T s(t)dt$. Then $\mathcal{G}$ is linear and bounded. Prove that $\mathcal{G}$ has an adjoint $\mathcal{G}^*$ given by

$$(\mathcal{G}^* y)(t) = \int_t^1 G(\tau,t)^T y(\tau)d\tau, \quad t \in [0,1]. \tag{A.7}$$

Further, suppose that G is given by

$$G(t,\tau) = C(t)\Phi(t,\tau)B(\tau)\mathbf{1}(t-\tau),$$

where $\mathbf{1}(t)$ is the unit step and $\Phi(t,\tau)$ is the state-transition matrix associated with a certain matrix $A(t)$, or equivalently, since $G(t,\tau)$ only affects (A.6) for $\tau < t$, $G(t,\tau) = C(t)\Phi(t,\tau)B(\tau)$. Then $\mathcal{G}$ is the mapping from u to y generated by

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = 0, \tag{A.8}$$
$$y(t) = C(t)x(t) \tag{A.9}$$

and $\mathcal{G}^*$ is the mapping from z to v generated by

$$\dot{p}(t) = -A(t)^T p(t) + C(t)^T z(t), \quad p(1) = 0, \tag{A.10}$$
$$v(t) = -B(t)^T p(t). \tag{A.11}$$

Observe from (A.7) that $\mathcal{G}^*$ is anticausal, i.e., the value of the output at any given time in $[0,1]$ depends only on present and future values of the input. Also note that, when A, B, and C are constant, the transfer function matrix associated with $\mathcal{G}$ is $\mathbf{G}(s) = C(sI - A)^{-1}B$ and that associated with $\mathcal{G}^*$ is $\mathbf{G}^{\sim}(s)$, discussed in Exercise A.52.

Remark A.4 In connection with Exercise A.53, note that, with $z := y$, $u := -v$ is the optimal control for problem (P) of section 2.1.1, with $t_0 = 0$, $t_f = 1$, $L = C^T C$, and $Q = 0$. Hence the optimal control can be automatically generated by the simple feedback loop.

$$\dot{x}(t) = Ax(t) + BB^T p(t),$$
$$\dot{p}(t) = -A^T p(t) + C^T C x(t),$$

Unfortunately, the initial costate $p(0)$ is unknown, so the adjoint system cannot be integrated. Of course, because $p(1)$ is known, $p(0)$ can be pre-calculated BUT it will have to be assumed that no disturbance or model errors affects that computation. I.e., the feedback-looking implementation would not enjoy the benefits provided by true feedback. (When $p(1)$ is known, this "feedback" loop effectively integrates the Riccati equation!

Theorem A.5 *(Fundamental Theorem of linear algebra) Let V, W be inner product spaces, let $L \in \mathcal{L}(V,W)$, with adjoint L^*. Then $\mathcal{N}(L)$ and $\mathcal{N}(L^*)$ are closed and*

(a) $\mathcal{N}(L^*) = \mathcal{R}(L)^\perp$; $\mathcal{N}(L) = \mathcal{R}(L^*)^\perp$;

(b) $\mathcal{N}(L^*) = \mathcal{N}(LL^*)$; $\mathcal{N}(L) = \mathcal{N}(L^*L)$;

(c) $\mathrm{cl}(\mathcal{R}(L)) = \mathrm{cl}(\mathcal{R}(LL^*))$, *and if* $\mathcal{R}(LL^*)$ *is closed, then* $\mathcal{R}(L) = \mathcal{R}(LL^*)$.

Proof. Closedness of $\mathcal{N}(L)$ and $\mathcal{N}(L^*)$ follows from (a).

(a)

$$\begin{aligned} y \in \mathcal{N}(L^*) &\Leftrightarrow L^* y = \theta_V \\ &\Leftrightarrow \langle L^* y, x\rangle = 0 \quad \forall x \in V \\ &\Leftrightarrow \langle y, Lx\rangle = 0 \quad \forall x \in V \\ &\Leftrightarrow y \in \mathcal{R}(L)^\perp. \end{aligned}$$

(We have used the fact that, if $\langle L^* y, x\rangle = 0 \quad \forall x \in V$, then, in particular, $\langle L^* y, L^* y\rangle = 0$, so that $L^* y = \theta_V$.)

(b)

$$y \in \mathcal{N}(L^*) \;\Leftrightarrow\; L^* y = \theta_V \Rightarrow LL^* y = \theta_W \Leftrightarrow y \in \mathcal{N}(LL^*)$$

and

$$y \in \mathcal{N}(LL^*) \Leftrightarrow LL^* y = \theta_W \Rightarrow \langle y, LL^* y\rangle = 0 \Leftrightarrow \langle L^* y, L^* y\rangle = 0 \Leftrightarrow L^* y = \theta_V \Leftrightarrow y \in \mathcal{N}(L^*)$$

(c) $\mathcal{R}(L)^\perp = \mathcal{N}(L^*) = \mathcal{N}(LL^*) = \mathcal{R}(LL^*)^\perp$. Thus $\mathcal{R}(L)^{\perp\perp} = \mathcal{R}(LL^*)^{\perp\perp}$. The result follows. Finally, if $\mathcal{R}(LL^*)$ is closed, then

$$\mathcal{R}(LL^*) \subseteq \mathcal{R}(L) \subseteq \mathrm{cl}(\mathcal{R}(L)) = \mathrm{cl}(\mathcal{R}(LL^*)) = \mathcal{R}(LL^*),$$

which implies that $\mathcal{R}(LL^*) = \mathcal{R}(L) = \mathrm{cl}\,\mathcal{R}(L) = \mathrm{cl}\,\mathcal{R}(LL^*)$.

■

Now let $L \in \mathcal{L}(V,W)$, with adjoint L^*, and suppose that $\mathcal{R}(L) = \mathcal{R}(LL^*)$. Let $\hat{w} \in \mathcal{R}(L)\,(= \mathcal{R}(LL^*))$. Then there exists $\hat{v} \in \mathcal{R}(L^*)$ such that $L\hat{v} = \hat{w}$. Further, $v \in V$ satisfies the equation $Lv = \hat{w}$ if and only if $v - \hat{v} \in \mathcal{N}(L)\,(= \mathcal{R}(L^*)^\perp)$. Hence, for any such v, $\hat{v}$ is the orthogonal projection of v (i.e., of the solution set of $Lv = \hat{w}$) on the subspace $\mathcal{R}(L^*)$, which implies that, unless $v = \hat{v}$, $\langle \hat{v}, \hat{v}\rangle < \langle v, v\rangle$. (Indeed, since $\langle v - \hat{v}, \hat{v}\rangle = 0$,

$$\langle v, v\rangle = \langle \hat{v} - (\hat{v} - v), \hat{v} - (\hat{v} - v)\rangle = \langle \hat{v}, \hat{v}\rangle + \langle \hat{v} - v, \hat{v} - v\rangle > \langle \hat{v}, \hat{v}\rangle$$

.) This leads to the solution of the linear least squares problem, stated next.

Theorem A.6 *Let V, W be inner-product spaces and let $L \in \mathcal{L}(v,w)$, with adjoint L^*. Suppose that $\mathcal{R}(L) = \mathcal{R}(LL^*)$. Let $w \in \mathcal{R}(L)$. Then the problem*

$$\text{minimize } \langle v, v\rangle \text{ s.t. } Lv = w \tag{A.12}$$

has a unique minimizer v_0. Further $v_0 \in \mathcal{R}(L^*)$.

Corollary A.1 *Let V, W be inner-product spaces and let $L : V \to W$ be linear, with adjoint L^*. Suppose that $\mathcal{R}(L) = \mathcal{R}(LL^*)$. Then*

$$V = \mathcal{R}(L^*) \oplus \mathcal{R}(L^*)^\perp = \mathcal{R}(L^*) \oplus \mathcal{N}(L).$$

See Figure A.1.

Proof. Let $v \in V$. Let $v_0 \in \mathcal{R}(L^*)$ be such that $Lv_0 = Lv$. (Such v_0 exists and is unique (Theorem A.6).) Then $v - v_0 \in \mathcal{N}(L) = \mathcal{R}(L^*)^\perp$ and $v = v_0 + (v - v_0)$.

■

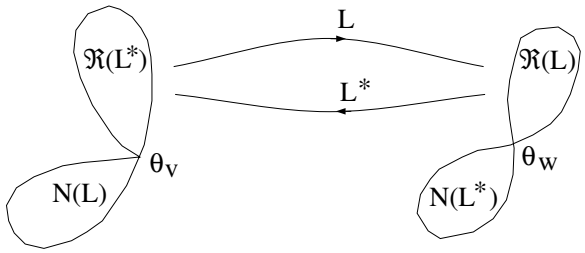


Figure A.1: Structure of a linear map

Moore-Penrose pseudo-inverse

Let $L \in \mathcal{L}(V,W)$, V and W inner-product spaces, with adjoint L^*. Suppose that $\mathcal{R}(L) = \mathcal{R}(LL^*)$. It follows from the above that $L|_{\mathcal{R}(L^*)} : \mathcal{R}(L^*) \to \mathcal{R}(L)$ is a bijection. Let $L^\dagger|_{\mathcal{R}(L)} : \mathcal{R}(L) \to \mathcal{R}(L^*)$ denote its inverse. Further, define $L^\dagger$ on $\mathcal{N}(L^*)$ by

$$L^\dagger w = \theta_V \quad \forall w \in \mathcal{N}(L^*).$$

Exercise A.54 *Suppose $W = \mathcal{R}(L) \oplus \mathcal{N}(L^*)$. (For instance, W is Hilbert and $R(L)$ is closed.) Prove that $L^\dagger$ has a unique linear extension to W.*

This extension is the Moore-Penrose pseudo-inverse (after Eliakim H. Moore, American mathematician, 1862–1932; Roger Penrose, English mathematician, born 1931). $L^\dagger$ is linear and $LL^\dagger$ restricted to $\mathcal{R}(L)$ is the identity in W, and $L^\dagger L$ restricted to $\mathcal{R}(L^*)$ is the identity in V; i.e.,

$$LL^\dagger L = L \quad \text{and } L^\dagger L L^\dagger = L^\dagger. \tag{A.13}$$

Exercise A.55 *Let $L : R^n \to R^m$ be a linear map with matrix representation M_L (an $m \times n$ matrix). We know that the restriction of L to $\mathcal{R}(L^*)$ is one-to-one, onto $\mathcal{R}(L)$. Thus there is an inverse map from $\mathcal{R}(L)$ to $\mathcal{R}(L^*)$. The Moore-Penrose pseudo-inverse $L^\dagger$ of L is a linear map from R^m to R^n that agrees with the just mentioned inverse on $\mathcal{R}(L)$ and maps to θ every point in $\mathcal{N}(L^*)$. Prove the following.*

- *Such $L^\dagger$ is uniquely defined, i.e., for an arbitrary linear map L, there exists a linear map, unique among linear maps, that satisfies all the listed condition.*
- *Let $M_L = U\Sigma V^T$ be the singular value decomposition of M_L, and let k be such that the nonzero entries of Σ are its (i,i) entries, $i = 1, \ldots, k$. Then the matrix representation of $L^\dagger$ is given by $M_L^\dagger := V\Sigma^\dagger U^T$ where $\Sigma^\dagger$ (an $n \times m$ matrix) has its (i,i) entry, $i = 1, \ldots, k$ equal to the inverse of that of Σ and all other entries equal to zero.*

In particular,

- *If L is one-to-one, then L^*L is invertible and $L^\dagger = (L^*L)^{-1}L^*$.*
- *If L is onto, then LL^* is invertible and $L^\dagger = L^*(LL^*)^{-1}$.*
- *If L is one-to-one and onto, then L is invertible and $L^\dagger = L^{-1}$.*

Appendix B

On Differentiability and Convexity

B.1 Differentiability

[2, 13, 25]

First, let $f: \mathbf{R} \to \mathbf{R}$. We know that f is *differentiable* at x^* if

$$\lim_{t \to 0} \frac{f(x^* + t) - f(x^*)}{t}$$

exists, i.e. if there exists $a \in \mathbf{R}$ such that

$$\frac{f(x^* + t) - f(x^*)}{t} = a + \varphi(t)$$

with $\varphi(t) \to 0$ as $\to 0$, i.e., if there exists $a \in \mathbf{R}$ such that

$$f(x^* + t) \;=\; f(x^*) + at + o(t) \quad \forall t \in \mathbf{R} \tag{B.1}$$

where $o: \mathbf{R} \to \mathbf{R}$ satisfies $\frac{o(t)}{t} \to 0$ as $t \to 0$. The number a is called the *derivative* at x^* and is noted $f'(x^*)$. Obviously, we have

$$f'(x^*) \;=\; \lim_{t \to 0} \frac{f(x^* + t) \;-\; f(x^*)}{t}.$$

Equation (B.1) shows that $f'(x^*)t$ is a linear (in t) approximation to $f(x^* + t) - f(x^*)$. Our first goal in this appendix is to generalize this to $f: V \to W$, with V and W more general (than $\mathbf{R}$) vector spaces. In such context, a formula such as (B.1) (with $a := f'(x^*)$) imposes that, for $t \in V$, $f'(x^*)t \in W$. Hence is it natural to stipulate that $f'(x^*)$ should be a linear map from V to W. This leads to Definitions B.3 and B.4 below. But first we consider the case of $f: \mathbf{R}^n \to \mathbf{R}^m$.

Definition B.1 *The function $f: \mathbf{R}^n \to \mathbf{R}^m$ has partial derivatives at x^* if $\exists\, a_{ij} \in \mathbf{R}$,* for $i = 1, \ldots, m$ $j = 1, \ldots, n$ *such that*

$$f_i(x^* + te_j) \;=\; f_i(x^*) + a_{ij}t + o_{ij}(t) \quad \forall t \in \mathbf{R}$$

with $\frac{o_{ij}(t)}{t} \to 0$ as $t \to 0$, for $i = 1, \ldots, m$, $j = 1, \ldots, n$. *The numbers a_{ij} are typically denoted $\frac{\partial f_i}{\partial x_j}(x^*)$.*

Example B.1 Let $f:\ \mathbf{R}^2 \to \mathbf{R}$ be defined by

$$f(x,y) = \begin{cases} 0 & \text{if } xy = 0 \\ 1 & \text{elsewhere} \end{cases}$$

Then f has partial derivatives at $(0,0)$, yet it is not continuous at (0,0). ■

Hence, existence of partial derivatives at x^* does not imply continuity at x^*. Also, the notion of partial derivative does not readily extend to functions whose domain is infinite-dimensional. For both of these reasons, we next consider notions of differentiability which, while being more restrictive than mere existence of partial derivatives, readily generalize to more general domains (and codomains) for which partial derivatives can't even be defined. Before doing so, we note the following fact, which applies when f has finite dimensional domain and co-domain.

Fact. (e.g., [13, Theorem 13.20]) Let $f:\ \mathbf{R}^n \to \mathbf{R}^m$, and let $\Omega \subseteq \mathbf{R}^n$ open set. If the partial derivatives of (the components of) f exist and are continuous throughout Ω, then f is continuous on Ω.

We now consider $f:\ V \to W$, where V and W are vector spaces, and W is equipped with a norm.

Definition B.2 *f is 1-sided (2-sided) directionally differentiable at $x^* \in V$ if for all $h \in V$ there exists $a_h \in W$ such that*

$$f(x^* + th) \ = \ f(x^*) + ta_h + o_h(t) \quad \forall t \in \mathbf{R} \tag{B.2}$$

with

$$\frac{1}{t}\ \|o_h(t)\| \ \to \ \theta \quad \text{as} \quad t \to 0 \ \ \text{(for any given } h) \quad (2-\text{sided})$$
$$\frac{1}{t}\ \|o_h(t)\| \ \to \ \theta \quad \text{as} \quad t \downarrow 0 \ \ \text{(for any given } h) \quad (1-\text{sided})$$

a_h is the directional derivative of f at x^* in direction h, often denoted $f'(x^*;h)$.

Definition B.3 *f is Gâteaux– (or G–) differentiable at x^* if there exists a linear map $A: V \to W$ such that*

$$f(x^* + th) = f(x^*) + tAh + o_h(t) \quad \forall h \in V \quad \forall t \in \mathbf{R} \tag{B.3}$$

with, for fixed $h \in V$,

$$\frac{1}{t}\ o_h(t) \to \theta \ \ \text{as} \ \ t \to 0.$$

A is termed G-derivative of f at x^* and is usually denoted $\frac{\partial f}{\partial x}(x^*)$.

(René E. Gâteaux, French mathematician, 1889-1914.)

Note: The term G-differentiability is also used in the literature to refer to other concepts of differentiability. Here we follow the terminology used in [25].

If f is G-differentiable at x^* then it is 2-sided directionally differentiable at x^* in all directions, and the directional derivative in direction h is the image of h under the mapping defined by the G-derivative, i.e.,

$$f'(x^*;h) = \frac{\partial f}{\partial x}(x^*)h \quad \forall h \in V.$$

G-differentiability at x^* still does not imply continuity at x^* though, even when V and W are finite-dimensional! See Exercise B.5 below.
Suppose now that V is also equipped with a norm.

Definition B.4 *f is Fréchet– (or F–) differentiable at x^* if there exists a continuous linear map $A : V \to W$ such that*

$$f(x^* + h) = f(x^*) + Ah + o(h) \quad \forall h \in V \tag{B.4}$$

with

$$\frac{\|o(h)\|_W}{\|h\|_V} \to \theta \quad \text{as} \quad h \to \theta.$$

A is termed F-derivative of f at x^* and is usually denoted $\frac{\partial f}{\partial x}(x^*)$.

In other words, f is F-differentiable at x^* if there exists a continuous linear map $A : V \to W$ such that

$$\frac{1}{\|h\|}(f(x^* + h) - f(x^*) - Ah) \to \theta \text{ as } h \to \theta \quad .$$

Using the $\frac{\partial f}{\partial x}$ notation, we can write (B.4) as

$$f(x^* + h) = f(x^*) + \frac{\partial f}{\partial x}(x^*)h + o(h)$$

whenever $f : V \to W$ is F-differentiable at x^*.
If $f : \mathbf{R}^n \to \mathbf{R}^m$ is F-differentiable at x^* then, given bases for $\mathbf{R}^n$ and $\mathbf{R}^m$, $\frac{\partial f}{\partial x}(x^*)$ can be represented by a matrix (as any linear map from $\mathbf{R}^n$ to $\mathbf{R}^m$).

Exercise B.1 *If $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at x^* then, given any bases for $\mathbf{R}^n$ and $\mathbf{R}^m$, the entries of the matrix representation of $\frac{\partial f}{\partial x}(x^*)$ are the partial derivatives $\frac{\partial f_i}{\partial x_j}(x^*)$.*

It should be clear that, if f is F-differentiable at x with F-derivative $\frac{\partial f}{\partial x}(x)$, then (i) it is G-differentiable at x with G-derivative $\frac{\partial f}{\partial x}(x)$, (ii) it is 2-sided directionally differentiable at x in all directions, with directional derivative in direction h given by $\frac{\partial f}{\partial x}(x)h$, and (iii) if $f : \mathbf{R}^n \to \mathbf{R}^m$, then f has partial derivatives $\frac{\partial f_i}{\partial x_j}(x)$ at x equal to $\left(\frac{\partial f}{\partial x}(x)\right)_{ij}$.

The difference between Gâteaux and Fréchet is that, for the latter, $\frac{\|o(h)\|_W}{\|h\|_V}$ must tend to θ no matter how h goes to θ whereas, for the former, convergence is along straight lines th, with fixed h. Further, F-differentiability at x requires that $\frac{\partial f}{\partial x}(x^*) \in \mathcal{B}(V,W)$ (bounded linear map). Clearly, every Fréchet-differentiable function at x is continuous at x (why?).
The following exercises taken from [25] show that each definition is "strictly stronger" than the previous one.

Exercise B.2 *(proven in [25])*

1. *If f is Gâteaux-differentiable, its Gâteaux derivative is unique.*
2. *If f is Fréchet differentiable, it is also Gâteaux differentiable and its Fréchet derivative is given by its (unique) Gâteaux-derivative.*

Exercise B.3 *Define $f : \mathbf{R}^2 \to \mathbf{R}$ by $f(x) = x_1$ if $x_2 = 0$, $f(x) = x_2$, if $x_1 = 0$, and $f(x) = 1$ otherwise. Show that the partial derivatives $\frac{\partial f}{\partial x_1}(0)$ and $\frac{\partial f}{\partial x_2}(0)$ exist, but that f is not directionally differentiable at $(0,0)$.*

Exercise B.4 *Define $f : \mathbf{R}^2 \to \mathbf{R}$ by*

$$f(x) \;=\; \mathrm{sgn}(x_2)\min(|x_1|,|x_2|).$$

Show that, for any $h \in \mathbf{R}^2$,

$$\lim_{t\to 0}(1/t)[f(th)-f(0,0)] \;=\; f(h),$$

and thus that f is 2-sided directionally differentiable at $0,0$, but that f is not G-differentiable at $0,0$.

Exercise B.5 *Define $f : \mathbf{R}^2 \to \mathbf{R}$ by*

$$f(x) = \begin{cases} 0 & \text{if } x_1 = 0 \\ \dfrac{2x_2 e^{-\frac{1}{x_1^2}}}{x_2^2+\left(e^{-\frac{1}{x_1^2}}\right)^2} & \text{if } x_1 \neq 0. \end{cases}$$

Show that f is G-differentiable at $(0,0)$, but that f is not continuous (and thus not F-differentiable) at $(0,0)$.

Remark B.1 For $f : \mathbf{R}^m \to \mathbf{R}^n$, again from the equivalence of norms, F-differentiability does not depend on the particular norm.

Gradient of a differentiable functional over a Hilbert space

Let $f : H \to \mathbf{R}$ where H is a Hilbert space and suppose f is Fréchet-differentiable at x. Then, in view of the Riesz-Fréchet theorem (see, e.g., [28, Theorem 4-12]), there exists a unique $g \in H$ such that

$$\langle g, h\rangle = \frac{\partial f}{\partial x}(x)h \quad \forall h \in H.$$

Such g is called the gradient of f at x, and denoted $\mathrm{grad} f(x)$, i.e.,

$$\langle \mathrm{grad} f(x), h\rangle = \frac{\partial f}{\partial x}(x)h \quad \forall h \in H.$$

When $H = \mathbf{R}^n$ and $\langle\cdot,\cdot\rangle$ is the Euclidean inner product $\langle x, y\rangle = x^T y$, we will often denote the gradient of f at x by $\nabla f(x^*)$.

Exercise B.6 *Note that the gradient depends on the inner product to which it is associated. For example, suppose $f : \mathbf{R}^n \to \mathbf{R}$. Let S be a symmetric positive definite $n \times n$ matrix, and define $\langle x, y\rangle = x^T S y$. Prove that* $\mathrm{grad} f(x) = S^{-1} \frac{\partial f}{\partial x}(x)^T$. *In particular, under the Euclidean inner product, $\nabla f(x) = \frac{\partial f}{\partial x}(x)^T$.*

In the sequel, unless otherwise specified, "differentiable" will mean "Fréchet-differentiable". An important property, which may not hold if f is merely Gâteaux differentiable, is given by the following fact.

Fact [25]. If $\phi : X \to Y$ and $\theta : Y \to Z$ are F-differentiable, respectively at $x^* \in X$ and at $\phi(x^*) \in Y$, then $h : X \to Z$ defined by $h(x) = \theta(\phi(x))$ is F-differentiable and the following chain rule applies

$$\begin{aligned} \frac{\partial}{\partial x} h(x^*) &= \frac{\partial}{\partial y}\theta(y') \frac{\partial}{\partial x}\phi(x^*)|_{y' = \phi(x^*)} \\ (&= \frac{\partial}{\partial y}\theta(\phi(x^*)) \frac{\partial}{\partial x}\phi(x^*)) \end{aligned}$$

Exercise B.7 *Prove this fact using the notation used in these notes.*

Exercise B.8 *Compute the "total" derivative of $\theta(\phi(x), \psi(x))$ with respect to x, with θ, ϕ, and ψ F-differentiable maps between appropriate spaces.*

Exercise B.9 *Let Q be an $n \times n$ (not necessarily symmetric) matrix and let $b \in \mathbf{R}^n$. Let $f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle b, x\rangle$, where $\langle x, y\rangle = x^T S y$, with $S = S^T > 0$. Show that f is F-differentiable and obtain its gradient with respect to the same inner product.*

Remark B.2

1. We will say that a function is *differentiable* (in any of the previous senses) if it is differentiable everywhere.
2. When x is allowed to move, $\frac{\partial f}{\partial x}$ can be viewed as a function of x, whose values are bounded linear maps:

$$\frac{\partial f}{\partial x} : V \to \mathcal{B}(V, W), \quad x \mapsto \frac{\partial f}{\partial x}(x).$$

First order "exact" expansions

Mean Value Theorem (e.g., [25]) Let $\phi : \mathbf{R} \to \mathbf{R}$ and suppose ϕ is continuous on $[a, b] \subset \mathbf{R}$ and differentiable on (a, b). Then, we know that there exists $\xi \in (a, b)$ such that

$$\phi(b) - \phi(a) = \phi'(\xi)(b - a) \tag{B.5}$$

i.e.,

$$\phi(a+h) - \phi(a) = \phi'(a+th)h, \quad \text{for some } t \in (0,1) \tag{B.6}$$

We have the following immediate consequence for functionals.

Fact. Suppose $f:\ V \to \mathbf{R}$ is differentiable. Then for any $x,\ h \in V$ there exists $t \in (0,1)$ such that

$$f(x+h)\ -\ f(x)\ =\ \frac{\partial f}{\partial x}(x+th)h \tag{B.7}$$

Proof. Consider $\phi:\ \mathbf{R} \to \mathbf{R}$ defined by $\phi(s):\ f(x+sh)$. By the result above there exists $t \in (0,1)$ such that

$$f(x+h)\ -\ f(x)\ =\ \phi(1)\ -\ \phi(0)\ =\ \phi'(t)\ =\ \frac{\partial f}{\partial x}(x+th)h \tag{B.8}$$

where we have applied the chain rule for F-derivatives.

■

It is important to note that this result is generally not valid for $f:\ V \to \mathbf{R}^m,\ m > 1$, because to different components of f will correspond different values of t. For this reason, we will often make use, as a substitute, of the **fundamental theorem of integral calculus**, which requires continuous differentiability (though the weaker condition of "absolute continuity", which implies existence of the derivative almost everywhere, is sufficient):

Definition B.5 *$f:\ V \to W$ is said to be continuously Fréchet-differentiable if it is Fréchet-differentiable and its Fréchet derivative $\frac{\partial f}{\partial x}$ is a continuous function from V to $\mathcal{B}(V,W)$.*

The following fact strengthens the result we quoted earlier that, when f maps $\mathbf{R}^n$ to $\mathbf{R}^m$, continuity of its partial derivatives implies its own continuity.
Fact. (E.g., [2], Theorem 2.5.) Let $f:\mathbf{R}^n \to \mathbf{R}^m$, and let Ω be an open subset of $\mathbf{R}^n$. If the partial derivatives of (the components of) f exist on Ω and are continuous at $\hat{x} \in \Omega$, then f is continuously (Fréchet) differentiable at $\hat{x}$.

Now, if $\phi\ :\ \mathbf{R} \to \mathbf{R}$ is continuously differentiable on $[a,b] \subset \mathbf{R}$, then the fundamental theorem of integral calculus asserts that

$$\phi(b) - \phi(a)\ =\ \int_a^b \phi'(\xi)d\xi. \tag{B.9}$$

Now For a continuously differentiable $g:\ \mathbf{R}^1 \to \mathbf{R}^m$, we define the integral of g by

$$\int_a^b g(t)dt\ =\ \begin{bmatrix} \int_a^b g^1(t)dt \\ \vdots \\ \int_a^b g^m(t)dt \end{bmatrix} \tag{B.10}$$

For $f:\ V \to \mathbf{R}^m$, we then obtain

Theorem B.1 *(Fundamental Theorem of integral calculus) If $f : V \to \mathbf{R}^m$ is continuously (Fréchet) differentiable, then for all $x, h \in V$*

$$f(x+h) - f(x) = \int_0^1 \frac{\partial f}{\partial x}(x+th)h \; dt$$

Proof. Define $\phi : \mathbf{R}^1 \to \mathbf{R}^m$ by $\phi(s) = f_i(x+sh)$. Apply (B.9), (B.10) and the chain rule. ■

Note. For $f : V \to W$, W a Banach space, the same result holds, with a suitable definition of the integral (of "integral-regulated" functions, see [2]).

Corollary B.1 *. Let V be a normed vector space. If $f : V \to \mathbf{R}^n$ is continuously differentiable, then for all $x \in V$*

$$f(x+h) = f(x) + O(\|h\|),$$

in the sense that f is locally Lipschitz; i.e., given any $x^* \in V$ there exist $\rho > 0$, $\delta > 0$, and $\beta > 0$ such that for all $h \in V$ with $\|h\| \leq \delta$, and all $x \in B(x^*, \rho)$,

$$\|f(x+h) - f(x)\| \leq \beta \|h\| \; . \tag{B.11}$$

Further, if V is finite-dimensional, then given any $\rho > 0$, and any $r > 0$, there exists $\beta > 0$ such that (B.11) holds for all $x \in B(x^*, \rho)$ and all $h \in V$ with $\|h\| \leq r$.

Exercise B.10 *Prove Corollary B.1.*

A stronger and more general version of the second statement in Corollary B.1 is as follows. (See, e.g, [2], Theorem 2.3.)
Fact. Let V and W be normed vector spaces, and let B be an open ball in V. Let $f : V \to W$ be differentiable on B, and suppose $\frac{\partial f}{\partial x}(x)$ is bounded on B. Then there exists $\beta > 0$ such that, for all $x \in B$, $h \in V$ such that $x + h \in B$,

$$\|f(x+h) - f(x)\| \leq \beta \|h\| \; .$$

(In this version, B need not be "small".)

Second derivatives

Definition B.6 *Suppose $f : V \to W$ is differentiable on V and use the induced norm for $\mathcal{B}(V, W)$. If*

$$\frac{\partial f}{\partial x} : \; V \to \mathcal{B}(V, W), \quad x \mapsto \frac{\partial f}{\partial x}(x)$$

is itself differentiable, then f is twice differentiable and the derivative of $\frac{\partial f}{\partial x}$ at $x \in V$ is noted $\frac{\partial^2 f}{\partial x^2}(x)$ and is called second derivative of f at x. Thus

$$\frac{\partial^2 f}{\partial x^2}(x) : V \to \mathcal{B}(V, W). \tag{B.12}$$

and

$$\frac{\partial^2 f}{\partial x^2} : \; V \to \mathcal{B}(V, \mathcal{B}(V, W)), \quad x \mapsto \frac{\partial^2 f}{\partial x^2}(x).$$

Fact. If $f: V \to W$ is twice continuously Fréchet-differentiable then its second derivative is symmetric in the sense that for all $u,\ v \in V$, $x \in V$

$$\left(\frac{\partial^2 f}{\partial x^2}(x)u\right)v \ = \ \left(\frac{\partial^2 f}{\partial x^2}(x)v\right)u \quad (\in W) \tag{B.13}$$

[Note: the reader may want to resist the temptation of viewing $\frac{\partial^2 f}{\partial x^2}(x)$ as a "cube" matrix. It is simpler to think about it as an abstract linear map.]
Now let $f: H \to \mathbf{R}$, where H is a Hilbert space, be twice differentiable. Then, in view of the Riess-Fréchet Theorem, $\mathcal{B}(H, \mathbf{R})$ is isomorphic to H, and in view of (B.12) $\frac{\partial^2 f}{\partial x^2}(x)$ can be thought of as a bounded linear map $\mathrm{Hess} f(x): H \to H$. This can be made precise as follows. For any $x \in H$, $\frac{\partial f}{\partial x}(x): H \to \mathbf{R}$ is a bounded linear functional and, for any $u \in H$, $\frac{\partial^2 f}{\partial x^2}(x)u: H \to \mathbf{R}$ is also a bounded linear functional. Thus in view of the Riesz-Fréchet representation Theorem, there exists a unique $\psi_u \in H$ such that

$$\left(\frac{\partial^2 f}{\partial x^2}(x)u\right)v = \langle \psi_u, v\rangle \quad \forall v \in H.$$

For given $x \in H$, The map from $u \in H$ to $\psi_u \in H$ is linear and bounded (why?). Let us denote this map by $\mathrm{Hess} f(x) \in \mathcal{B}(H, H)$ (after L. Otto Hesse, German mathematician, 1811–1874), i.e., $\psi_u = \mathrm{Hess} f(x)u$. We get

$$\left(\frac{\partial^2 f}{\partial x^2}(x)u\right)v = \langle \mathrm{Hess} f(x)u, v\rangle \quad \forall x, u, v \in H.$$

In view of (B.13), if f is twice continuously differentiable, $\mathrm{Hess} f(x)$ is self-adjoint. If $H = \mathbf{R}^n$ and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, then $\mathrm{Hess} f(x)$ is represented by an $n \times n$ symmetric matrix, which we will denote by $\nabla^2 f(x)$. In the sequel though, following standard usage, we will often abuse notation and use $\nabla^2 f$ and $\frac{\partial^2 f}{\partial x^2}$ interchangeably.

Fact. If $f: V \to W$ is twice F-differentiable at $x \in V$ then

$$f(x+h) \ = \ f(x) + \frac{\partial f}{\partial x}(x)h + \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}(x)h\right)h + o_2(h) \tag{B.14}$$

with

$$\frac{||o_2(h)||}{||h||^2} \to 0 \text{ as } h \to \theta.$$

Exercise B.11 *Prove the above in the case $W = \mathbf{R}^n$. Hint: first use Theorem B.1.*

Finally, a second order integral expansion.

Theorem B.2 *If $f: V \to \mathbf{R}^m$ is twice continuously differentiable, then for all $x, h \in V$*

$$f(x+h) = f(x) + \frac{\partial f}{\partial x}(x)h \ + \ \int_0^1 (1-t)\left(\frac{\partial^2 f}{\partial x^2}(x+th)h\right)h \ dt \tag{B.15}$$

(This generalizes the relation, with $\phi : \mathbf{R} \to \mathbf{R}$ twice continuously differentiable,

$$\phi(1) - \phi(0) - \phi'(0) = \int_0^1 (1-t)\phi''(t)dt$$

Check it by integrating by parts. Let $\phi(s) = f_i(x + s(y-x))$ to prove the theorem.)

Corollary B.2 . *Let V be a normed vector space. Suppose $f : V \to \mathbf{R}^n$ is twice continuously differentiable, Then given any $x^* \in V$ there exists $\rho > 0$, $\delta > 0$, and $\beta > 0$ such that for all $h \in V$ with $\|h\| \leq \delta$, and all $x \in B(x^*, \rho)$,*

$$\|f(x+h) - f(x) - \frac{\partial f}{\partial x}(x)h\| \leq \beta\|h\|^2. \tag{B.16}$$

Further, if V is finite-dimensional, then given any $\rho > 0$, and any $r > 0$, there exists $\beta > 0$ such that (B.16) holds for all $x \in B(x^*, \rho)$ and all $h \in V$ with $\|h\| \leq r$.

Again, a more general version of this result holds. (See, e.g, [2], Theorem 4.8.)
Fact. Let V and W be normed vector spaces, and let B be an open ball in W. Let $f : W \to W$ be twice differentiable on B, and suppose $\frac{\partial^2 f}{\partial x^2}(x)$ is bounded on B. Then there exists $\beta > 0$ such that, for all $x \in B$, $h \in V$ such that $x + h \in B$,

$$\|f(x+h) - f(x) - \frac{\partial f}{\partial x}(x)h\| \leq \beta\|h\|^2 \ .$$

[Again, in this version, B need not be "small".]

B.2 Some elements of convex analysis

[25]

Let V be a vector space.

Definition B.7 *A set $S \in V$ is said to be convex if, $\forall x, y \in S,\ \forall \lambda \in (0,1)$,*

$$\lambda x \ + \ (1-\lambda)y \in S. \tag{B.17}$$

Further, a function $f \ : \ V \to \mathbf{R}$ is convex on the convex set $S \subseteq X$ if $\forall x, y \in S,\ \forall \lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \tag{B.18}$$

i.e., if the arc lies below the chord (see Figure B.1).

However, this definition may fail (in the sense of the right-hand side of (B.18) not being well defined) when f is allowed to take on values of both $-\infty$ and $+\infty$, which is typical in convex analysis. (The definition is valid if f is "proper": see below.) A more general definition is as follows.

Definition B.8 *A function $f : S \to \mathbf{R} \cup \{\pm\infty\}$ is convex on convex set $S \subset V$ if its epigraph*

$$\text{epi } f := \{(x, z) \in S \times \mathbf{R} : z \geq f(x)\}$$

is a convex set.

The (effective) domain $\text{dom} f$ of a convex function $f : \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$ is the set of points $x \in \mathbf{R}^n$ such that $f(x) < \infty$; f is proper if its domain is non-empty and, for all $x \in \mathbf{R}^n$, $f(x) \neq -\infty$.

Fact. If $f : \mathbf{R}^n \to \mathbf{R}$ is convex, then it is continuous. More generally, if $f : \mathbf{R}^n \to \mathbf{R} \cup \{\pm\infty\}$ is convex and proper, then it is continuous over the relative interior of its domain (in particular, over the interior of its domain). See, e.g., [6], section 1.4 and [25].

Definition B.9 *A set $S \subseteq V$ is strictly convex if, $\forall x, y \in S,\ x \neq y,\ \forall \lambda \in (0, 1)$,*

$$\lambda x + (1 - \lambda)y \in \text{int}(S),$$

where $\text{int}(S)$ denotes the interior of S. A function $f : V \to \mathbf{R}$ is said to be strictly convex on the convex set $S \subseteq V$ if, $\forall x, y \in S,\ x \neq y,\ \forall \lambda \in (0, 1)$,

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda) f(y). \tag{B.19}$$

Definition B.10 *A convex combination of k points $x_1, \ldots, x_k \in V$ is any point x expressible as*

$$x \ = \ \sum_{i=1}^{k} \lambda_i x_i \tag{B.20}$$

with $\lambda_i \geq 0$, for $i \ = \ 1, \ldots, k$ and $\sum_{i=1}^{k} \lambda_i = 1$.

Exercise B.12 *Show that a set S is convex if and only if it contains the convex combinations of all its finite subsets. Hint: use induction.*

Fact 1. [25]. (Jensen's inequality.) A function $f : V \to \mathbf{R}$ is convex on the convex set $S \subset V$, if and only if for any finite set of points $x_1, ..., x_k \in S$ and any $\lambda_i \geq 0,\ i = 1, ..., k$ with $\sum_{i=1}^{k} \lambda_i = 1$, one has

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i).$$

Exercise B.13 *Prove the above. (Hint: Use Exercise B.12 and mathematical induction.)*

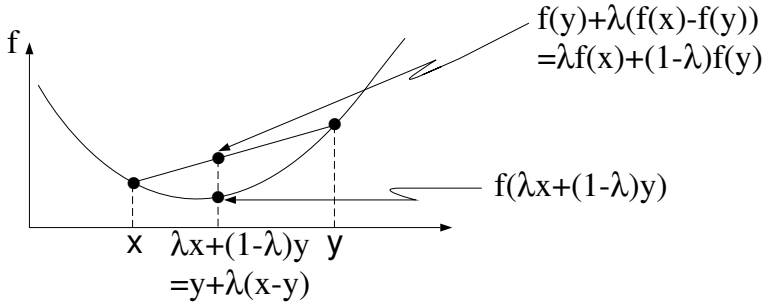


Figure B.1:

If the domain of f is $\mathbf{R}^n$, convexity implies continuity.

Convex Hull

Definition B.11 *The convex hull of a set X, denoted co$X$, is the smallest convex set containing X. The following exercise shows that it makes sense to talk about the "smallest" such set.*

Exercise B.14 *Show that $\cap\{Y : Y$ convex, $X \subseteq Y\}$ is convex and contains X. Since it is contained in any convex set containing X, it is the 'smallest' such set.*

Exercise B.15 *Prove that*

$$\mathrm{co}X = \bigcup_{k=1}^{\infty} \left\{ \sum_{i=1}^{k} \lambda^i x_i : x_i \in X, \lambda^i \geq 0 \ \ \forall i, \quad \sum_{i=1}^{k} \lambda^i = 1 \right\}$$

In other words, co$X$ is the set of all convex combinations of finite subsets of X. In particular, if X is finite, $X = \{x_1, \ldots, x_\ell\}$

$$\mathrm{co}X = \left\{ \sum_{i=1}^{\ell} \lambda^i x_i : \lambda^i \geq 0 \ \ \forall i, \quad \sum_{i=1}^{\ell} \lambda^i = 1 \right\}$$

(Hint: To prove $\subseteq$, show that $X \subseteq RHS$ (right-hand side) and that RHS is convex. To prove $\supseteq$, use mathematical induction.)

Exercise B.16 *(due to Constantin Carathéodory, Greek-born mathematician, 1873–1950; e.g. [12]). Show that, in the above exercise, for $X \subset \mathbf{R}^n$, it is enough to consider convex combinations of $n+1$ points, i.e.,*

$$\mathrm{co}X = \left\{ \sum_{i=1}^{n+1} \lambda^i x_i : \lambda^i \geq 0, \quad x_i \in X, \quad \sum_{i=1}^{n+1} \lambda^i = 1 \right\}$$

and show by example (say, in $\mathbf{R}^2$) that n points is generally not enough.

Exercise B.17 *Prove that the convex hull of a compact subset of $\mathbf{R}^n$ is compact and that the closure of a convex set is convex. Show by example that the convex hull of a closed subset of $\mathbf{R}^n$ need not be closed.*

Suppose now V is a normed space.

Proposition B.1 *Suppose that $f : V \to \mathbf{R}$ is differentiable on a convex subset S of V. Then $f : V \to \mathbf{R}$ is convex on S if and only if, for all $x, y \in S$*

$$f(y) \geq f(x) + \frac{\partial f}{\partial x}(x)(y - x). \tag{B.21}$$

Proof. (only if) (see Figure B.2). Suppose f is convex. Then, $\forall x, y \in V,\ \lambda \in [0, 1]$,

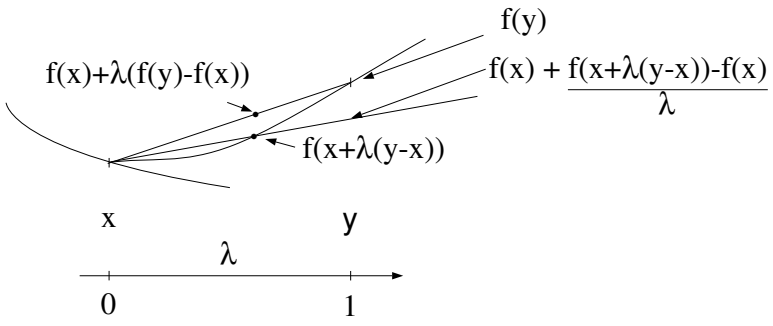


Figure B.2:

$$f(x + \lambda(y - x)) = f(\lambda y + (1 - \lambda)x) \leq \lambda f(y) + (1 - \lambda) f(x) = f(x) + \lambda(f(y) - f(x)) \tag{B.22}$$

i.e.

$$f(y) - f(x) \ \geq \ \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \quad \forall \lambda \in (0, 1],\ \forall x, y \in V \tag{B.23}$$

and, when $\lambda \searrow 0$, since f is differentiable,

$$f(y) - f(x) \geq \frac{\partial f}{\partial x}(x)(y - x) \quad \forall x, y \in V \tag{B.24}$$

(if) (see Figure B.3). Suppose (B.21) holds $\forall x, y \in V$. Then, for given $x, y \in V$ and $z = \lambda x + (1-\lambda)y$

$$f(x) \geq f(z) + \frac{\partial f}{\partial x}(z)(x - z) \tag{B.25}$$

$$f(y) \geq f(z) + \frac{\partial f}{\partial x}(z)(y - z), \tag{B.26}$$

and $\lambda\times$(B.25) $+(1-\lambda)\times$(B.26) yields

$$\begin{aligned} \lambda f(x) + (1-\lambda) f(y) &\geq f(z) + \frac{\partial f}{\partial x}(z)(\lambda x + (1-\lambda)y - z) \\ &= f(\lambda x + (1-\lambda)y) \end{aligned}$$

■

Fact. Moreover, f is strictly convex if and only if inequality (B.21) is strict whenever $x \neq y$ (see [25]).

Definition B.12 *$f : V \to \mathbf{R}$, V a normed space, is said to be strongly convex over a convex set S if f is continuously differentiable on S and there exists $m > 0$ s.t. $\forall x, y \in S$*

$$f(y) \geq f(x) + \frac{\partial f}{\partial x}(x)(y - x) + \frac{m}{2}\|y - x\|^2.$$

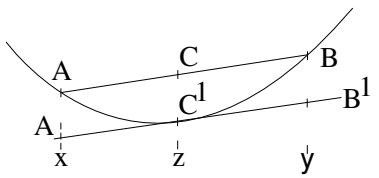
Proposition B.2 *If $f : V \to \mathbf{R}$ is strongly convex, it is strictly convex and, for any $x_0 \in V$, the sub-level set*

$$\{x : f(x) \leq f(x_0)\}$$

is bounded.

Proof. The first claim follows from Definition B.12 and the fact preceding it. Now, let h be an arbitrary *unit* vector in $\mathbf{R}^n$. Then

$$\begin{aligned} f(x_0 + h) \geq f(x_0) + \frac{\partial f}{\partial x}(x_0)h + \frac{1}{2}m\|h\|^2 & \\ &\geq f(x_0) - \left\|\frac{\partial f}{\partial x}(x_0)\right\| \|h\| + \frac{1}{2}m\|h\|^2 \\ &= f(x_0) + \left(\frac{m}{2}\|h\| - \left\|\frac{\partial f}{\partial x}(x_0)\right\|\right)\|h\| \\ &> f(x_0) \text{ whenever } \|h\| > (2/m)\left\|\frac{\partial f}{\partial x}(x_0)\right\|. \end{aligned}$$



A above A1
B above B1
⇒ C above C1

Figure B.3:

Hence $\{x : f(x) \leq f(x_0)\} \subseteq B\left(x_0, (2/m)\left\|\frac{\partial f}{\partial x}(x_0)\right\|\right)$, which is a bounded set.

■

Interpretation. The function f lies above the function

$$\tilde{f}(x) = f(x_0) + \frac{\partial f}{\partial x}(x_0)(x - x_0) + \frac{1}{2}m\|x - x_0\|^2$$

which grows without bound uniformly in all directions.

Proposition B.3 *Suppose that $f : V \to \mathbf{R}$ is twice continuously differentiable. Then f is convex if and only if $\frac{\partial^2 f}{\partial x^2} f(x)$ is positive semi-definite for all $x \in V$.*

Proof. We prove sufficiency. We can write, $\forall x, y \in V$

$$\begin{aligned} f(x+h) &= f(x) + \frac{\partial f}{\partial x}(x)h + \int_0^1 (1-t)\left(\frac{\partial^2 f}{\partial x^2}(x+th)h\right)h dt \\ &\Rightarrow \quad f(y) \geq f(x) + \frac{\partial f}{\partial x}(x)h \quad \forall x, y \in V \end{aligned}$$

and, from Proposition B.1, f is convex.

■

Exercise B.18 *Prove the necessity part of Proposition B.3.*

Exercise B.19 *Show that if f is twice continuously differentiable and $\nabla^2 f$ is positive definite on V, then f is strictly convex. Show by example that the converse does not hold in general.*

Exercise B.20 *Suppose $f : V \to \mathbf{R}$ is twice continuously differentiable. Then f is strongly convex if and only if there exists $m > 0$ such that for all $x, h \in V$,*

$$h^T \nabla^2 f(x) h \geq m||h||^2. \tag{B.27}$$

Exercise B.21 *Let $V = \mathbf{R}^n$. Show that (B.27) holds if, and only if, the eigenvalues of $\nabla^2 f(x)$ (they are real, why ?) are all positive and bounded away from zero, i.e., there exists $m > 0$ such that for all $x \in \mathbf{R}^n$ all eigenvalues of $\nabla^2 f(x)$ are larger than m.*

Exercise B.22 *Exhibit a function $f : \mathbf{R} \to \mathbf{R}$, twice continuously differentiable with Hessian everywhere positive definite, which is not strongly convex.*

Exercise B.23 *Exhibit a function $f : \mathbf{R}^2 \to \mathbf{R}$ such that for all $x, y \in \mathbf{R}$, $f(x, \cdot) : \mathbf{R} \to \mathbf{R}$ and $f(\cdot, y) : \mathbf{R} \to \mathbf{R}$ are strongly convex but f is not even convex. Hint: consider the Hessian matrix.*

Separation of Convex Sets [23, 25]

Here we consider subsets of a Hilbert space H.

Definition B.13 *Let $a \in H,\ a \neq 0$, and $\alpha \in \mathbf{R}$. The set*

$$P(a,\alpha) := \{x \in H \ : \ \langle a, x\rangle = \alpha\}$$

is called a hyperplane. Thus, hyperplanes are level sets of linear functionals.

Let us check, for $n = 2$, that this corresponds to the intuitive notion we have of a hyperplane, i.e., a straight line in this case. Such straight line can be expressed, for a given scalar ξ and vector a orthogonal to the straight line, as the set of all x of the form $x = \xi a + v$ for some $v \perp a$; see Figure B.4, where x should be v, which should be orthogonal to a, and H should be $P(a,\alpha)$. Hence

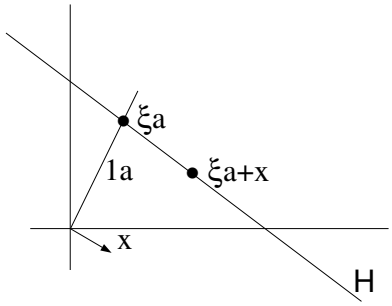


Figure B.4:

$$\langle a, x\rangle = \xi||a||^2 + \langle a, v\rangle = \xi||a||^2 \quad \forall x \in P(a,\alpha)$$

which corresponds to the above definition with $\alpha = \xi||a||^2$.

$\{x \ : \ \langle a, x\rangle \leq \alpha\}$ and $\{x \ : \ \langle a, x\rangle \geq \alpha\}$ are closed *half spaces*;
$\{x \ : \ \langle a, x\rangle < \alpha\}$ and $\{x \ : \ \langle a, x\rangle > \alpha\}$ are open *half spaces*.

Definition B.14 *Let $X, Y \subset H$. X and Y are (strictly) separated by $P(a,\alpha)$ if*

$$\langle a, x\rangle \ \geq (>)\alpha \quad \forall x \in X$$

$$\langle a, y\rangle \ \leq (<)\alpha \quad \forall y \in Y$$

This means that X is entirely in one of the half spaces and Y entirely in the other. Examples are given in Figure B.5.

Note. X and Y can be separated without being disjoint. Any hyperplane is even separated from itself (by itself!)

The idea of separability is strongly related to that of convexity. For instance, it can be shown that 2 disjoint convex sets can be separated. We will present here only one of the numerous separation theorems; this theorem will be used in connection with constrained optimization.

Theorem B.3 *Suppose $X \subset H$ is nonempty, closed and convex and suppose that $0 \notin X$. Then there exists $a \in H$, $a \neq 0$, and $\alpha > 0$ such that*

$$\langle a, x \rangle > \alpha \quad \forall x \in X \tag{B.28}$$

(i.e., X and $\{0\}$ are strictly separated by $P(a, \alpha)$).

Proof (see Figure B.6). We first prove the result for the case $H = \mathbf{R}^n$. Let $\|\cdot\|$ denote the norm derived from the underlying inner product. We first show that there exists $\hat{x} \in X$ such that

$$\|\hat{x}\| \leq \|x\| \quad \forall x \in X. \tag{B.29}$$

Choose $\rho > 0$ such that $\bar{B}(0, \rho) \cap X \neq \emptyset$. $\bar{B}(0, \rho) \cap X$ is bounded and closed (intersection of closed sets), hence compact. Since $||\cdot||$ is continuous, there exists $\hat{x} \in X$ such that

$$||\hat{x}|| \leq ||x|| \quad \forall x \in \bar{B}(0, \rho) \cap X \tag{B.30}$$

hence (B.29) holds since $||\hat{x}|| \leq \rho$ and $||x|| > \rho$ for all $x \notin \bar{B}(0, \rho)$. We now show that $\hat{x}$ is normal to a hyperplane that strictly separates X from the origin. We first show that $P(\hat{x}, \|\hat{x}\|^2)$, which containts $\hat{x}$, (non-strictly) separates X from θ. Clearly, $\langle \theta, \hat{x} \rangle < \|\hat{x}\|^2$. We show by contradiction that

$$\langle x, \hat{x} \rangle \geq \|\hat{x}\|^2 \quad \forall x \in X, \tag{B.31}$$

proving the separation. Thus suppose there exists $x \in X$ such that

$$\langle x, \hat{x} \rangle \;=\; ||\hat{x}||^2 - \epsilon, \tag{B.32}$$

where $\epsilon > 0$. Since X is convex,

$$x_\lambda := \lambda x + (1 - \lambda)\hat{x} = \hat{x} + \lambda(x - \hat{x}) \in X \quad \forall \lambda \in [0, 1].$$

Further

$$||x_\lambda||^2 = ||\hat{x}||^2 + \lambda^2 ||x - \hat{x}||^2 + 2\lambda(\langle \hat{x}, x \rangle - ||\hat{x}||^2) \quad \forall \lambda \in [0, 1]$$

i.e. using (B.32),

$$||x_\lambda||^2 = ||\hat{x}||^2 + \lambda^2 ||x - \hat{x}||^2 - 2\lambda\, \epsilon,$$

so that

$$||x_\lambda||^2 < ||\hat{x}||^2 \tag{B.33}$$

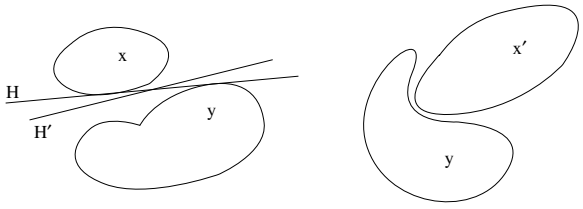


Figure B.5: X and Y are separated by H and strictly separated by H′. X' and Y' are not separated by any hyperplane (i.e., cannot be separated).

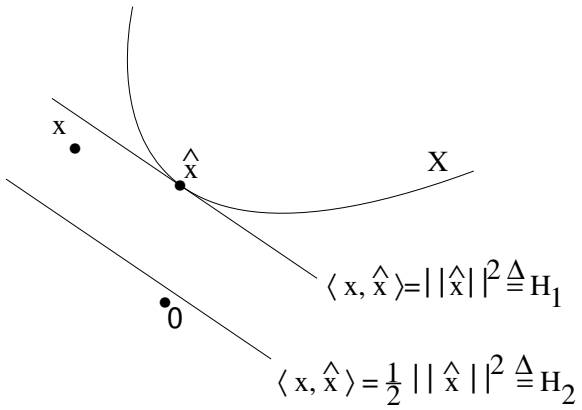


Figure B.6:

holds for $\lambda > 0$ small enough (since $\epsilon > 0$), which contradicts (B.29). Hence, (B.31) must hold. Since $\hat{x} \neq \theta$, it follows that, with $\alpha := \|\hat{x}\|^2/2$,

$$\langle \theta, \hat{x} \rangle = 0 < \alpha < \langle x, \hat{x} \rangle \quad \forall x \in X,$$

concluding the proof for the case $H = \mathbf{R}^n$.
Note that this proof in fact applies to the case of a general Hilbert space, except for the use of a compactness argument to prove (B.29): in general, $\bar{B}(0, \rho)$ may not be compact. We borrow the general proof from [23, Section 3.12]. Thus let $\delta := \inf_{x \in X} \|x\|$, so that

$$\|x\| \geq \delta \quad \forall x \in X,$$

and let $\{x_i\} \subset X$ be such that

$$\|x_i\| \to \delta \quad \text{as } i \to \infty. \tag{B.34}$$

Since X is convex, $(x_i + x_j)/2$ also belongs to X, so that $\|(x_i + x_j)/2\| \geq \delta$, implying

$$\|x_i + x_j\|^2 \geq 4\delta. \tag{B.35}$$

Now the parallelogram law gives

$$\|x_i - x_j\|^2 + \|x_i + x_j\|^2 = 2(\|x_i\|^2 + \|x_j\|^2),$$

i.e, in view of (B.34) and (B.35),

$$\|x_i - x_j\|^2 \leq 2(\|x_i\|^2 + \|x_j\|^2) - 4\delta \to 0 \text{ as } i, j \to \infty.$$

The sequence $\{x_i\}$ is Cauchy, hence (since H is complete) convergent, to some $\hat{x} \in X$ since X is closed. From continuity of the norm, we conclude that $\|\hat{x}\| = \delta$, concluding the proof. ■

Corollary B.3 *If X is closed and convex and $b \notin X$, then b and X are strictly separated.*

Remark B.3 Theorem B.3 also holds more generally on Banach spaces V; see, e.g., [23, Section 5.12]. In this context (there is no inner product), hyperplanes are more generally defined by

$$P(\ell, \alpha) := \{x \in V \ : \ \ell x = \alpha\}$$

where $\ell : V \to V$ is a continuous linear map. The proof is based on the celebrated Hahn-Banach theorem.

Fact. If X and Y are nonempty disjoint and convex, with X compact and Y closed, then X and Y are strictly separated.

Exercise B.24 *Prove the Fact. Hint: first show that $Y - X$, defined as $\{z|z = y - x, y \in Y, x \in X\}$, is closed, convex and does not contain 0.*

Exercise B.25 *Show by an example that if in the above theorem, X is merely closed, X and Y may not be strictly separated. (In particular, the difference of 2 closed sets, defined as above, may not be closed.)*

Exercise B.26 *Prove that, if C is a closed convex cone and $x \notin C$, then there exists h such that $h^T x > 0$ and $h^T v \leq 0$ for all $v \in C$.*

Acknowledgment

The author wishes to thank the numerous students who have contributed constructive comments towards improving these notes over the years. In addition, special thanks are addressed to Ji-Woong Lee, who used these notes when he taught an optimal control course at Penn State University in the Spring 2008 semester and, after the semester was over, provided many helpful comments towards improving the notes.

Bibliography

[1] B.D.O. Anderson and J.B. Moore. Optimal Control: Linear Quadratic Methods. Prentice Hall, 1990.

[2] A. Avez. Differential Calculus. J. Wiley and Sons, 1986.

[3] D.P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Academic Press, New York, 1982.

[4] D.P. Bertsekas. Dynamic Programming and Optimal Control, Vol. 1. Athena Scientific, Belmont, Massachusetts, 2017.

[5] S.P. Bertsekas. Nonlinear Programming. Athena Scientific, Belmont, Massachusetts, 1996.

[6] S.P. Bertsekas, A. Nedic, and A.E. Ozdaglar. Convex Analysis and Optimization. Athena Scientific, Belmont, Massachusetts, 2003.

[7] R.W. Brockett. Finite Dimensional Linear Systems. J. Wiley and Sons, 1970.

[8] F.M. Callier and C.A. Desoer. Linear System Theory. Springer-Verlag, New York, 1991.

[9] M.D. Canon, JR. C.D. Cullum, and E. Polak. Theory of Optimal Control and Mathematical Programming. McGraw-Hill Book Company, New York, 1970.

[10] M.D. Canon, C.D. Cullum, and E.Polak. Theory of Optimal Control and Mathematical Programming. McGraw-Hill, 1970.

[11] F.H. Clarke. Optimization and Nonsmooth Analysis. Wiley Interscience, 1983.

[12] V.F. Dem'yanov and L.V. Vasil'ev. Nondifferentiable Optimization. Translations Series in Mathematics and Engineering. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1985.

[13] P.M. Fitzpatrick. Advanced Calculus. Thomson, 2006.

[14] W.H. Fleming and R.W. Rishel. Deterministic and Stochastic Optimal Control. Springer-Verlag, New York, 1975.

[15] F.J. Gould and J.W. Tolle. A necessary and sufficient qualification for constrained optimization. SIAM J. on Applied Mathematics, 20, 1971.

[16] J.K. Hale. Ordinary Differential Equations. Wiley-Interscience, 1969.

[17] H.K. Khalil. Nonlinear Systems. Prentice Hall, 2002. Third Edition.

[18] Peter D. Lax. Functional Analysis. J. Wiley & Sons Inc., 2002.

[19] E.B. Lee and L. Markus. Foundations of Optimal Control Theory. Wiley, New York, 1967.

[20] G. Leitman. The Calculus of Variations and OptimalControl. Plenum Press, 1981.

[21] D. Liberzon. Calculus of Variations and Optimal Control Theory: A Concise Introduction. Princeton University Press, 2011.

[22] D. G. Luenberger. Introduction to Linear and Nonlinear Programming. Addison-Wesley, Reading, Mass., 1973.

[23] D.G. Luenberger. Optimization by Vector Space Methods. J. Wiley and Sons, 1969.

[24] J. Nocedal and S.J. Wright. Numerical Optimization. Second edition. Springer-Verlag, 2006.

[25] J. Ortega and W. Rheinboldt. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York, 1970.

[26] E. Polak. Computational Methods in Optimization. Academic Press, New York, N.Y., 1971.

[27] L.S. Pontryagin, V.G. Boltyansky, R.V. Gamkrelidze, and E.F. Mi shchenko. The Mathematical Theory of Optimal Processes. Interscience, 1962.

[28] W. Rudin. Real and Complex Analysis. McGraw-Hill, New York, N.Y., 1974. second edition.

[29] W.J. Rugh. Linear System Theory. Prentice Hall, 1993.

[30] E.D. Sontag. Mathematical Control Theory. Deterministic Finite Dimensional Systems. Springer-Verlag, 1990.

[31] H.J. Sussmann and J.C. Willems. 300 years of optimal control: From the brachystrochrone to the maximum principle. IEEE CSS Magazine, 17, 1997.

[32] P.P. Varaiya. Notes in Optimization. Van Nostrand Reinhold, 1972.

[33] R. Vinter. Optimal Control. Springer Verlag, 2010.

[34] K. Zhou, J.C. Doyle, and K. Glover. Robust and Optimal Control. Prentice Hall, New Jersey, 1996.

