A
Mathematical Background

Version: date: April 5, 2019

A.1 Introduction

In this appendix we give a brief review of some concepts that we need. It is assumed that the reader has had at least a first course on linear systems and has some familiarity with linear algebra and analysis. The appendices of Polak (1997); Nocedal and Wright (2006); Boyd and Vandenberghe (2004) provide useful summaries of the results we require. The material presented in Sections A.2–A.14 follows closely Polak (1997) and earlier lecture notes of Professor Polak.

A.2 Vector Spaces

The Euclidean space $\mathbb{R}^n$ is an example of a vector space that satisfies a set of axioms the most significant being: if x and z are two elements of a vector space $\mathcal{V}$, then $\alpha x + \beta z$ is also an element of $\mathcal{V}$ for all α, $\beta \in \mathbb{R}$. This definition presumes addition of two elements of $\mathcal{V}$ and multiplication of any element of $\mathcal{V}$ by a scalar are defined. Similarly, $S \subset \mathcal{V}$ is a linear subspace[1] of $\mathcal{V}$ if any two elements of x and z of S satisfy $\alpha x + \beta z \in S$ for all $\alpha, \beta \in \mathbb{R}$. Thus, in $\mathbb{R}^3$, the origin, a line or a plane passing through the origin, the whole set $\mathbb{R}^3$, and even the empty set are all subspaces.

A.3 Range and Nullspace of Matrices

Suppose $A \in \mathbb{R}^{m \times n}$. Then $\mathcal{R}(A)$, the *range* of A, is the set $\{Ax \mid x \in \mathbb{R}^n\}$; $\mathcal{R}(A)$ is a subspace of $\mathbb{R}^m$ and its dimension, i.e., the number of linearly independent vectors that span $\mathcal{R}(A)$, is the rank of A. For

[1] All of the subspaces used in this text are linear subspaces, so we often omit the adjective linear.

example, if A is the column vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$, then $\mathcal{R}(A)$ is the subspace spanned by the vector $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and the rank of A is 1. The *nullspace* $\mathcal{N}(A)$ is the set of vectors in $\mathbb{R}^n$ that are mapped to zero by A so that $\mathcal{N}(A) = \{x \mid Ax = 0\}$. The nullspace $\mathcal{N}(A)$ is a subspace of $\mathbb{R}^n$. For the example above, $\mathcal{N}(A)$ is the subspace spanned by the vector $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$. It is an important fact that $\mathcal{R}(A') \oplus \mathcal{N}(A) = \mathbb{R}^n$ or, equivalently, that $\mathcal{N}(A) = (\mathcal{R}(A'))^\perp$ where $A' \in \mathbb{R}^{n \times m}$ is the transpose of A and $S^\perp$ denotes the orthogonal complement of any subspace S; a consequence is that the sum of the dimensions $\mathcal{R}(A)$ and $\mathcal{N}(A)$ is n. If A is square and invertible, then $n = m$ and the dimension of $\mathcal{R}(A)$ is n so that the dimension of $\mathcal{N}(A)$ is 0, i.e., the nullspace contains only the zero vector, $\mathcal{N}(A) = \{0\}$.

A.4 Linear Equations — Existence and Uniqueness

Let $A \in \mathbb{R}^{m \times n}$ be a real-valued matrix with m rows and n columns. We are often interested in solving linear equations of the type

$$Ax = b$$

in which $b \in \mathbb{R}^m$ is given, and $x \in \mathbb{R}^n$ is the unknown. The fundamental theorem of linear algebra gives a complete characterization of the existence and uniqueness of solutions to $Ax = b$ (Strang, 1980, pp.87–88). Every matrix A decomposes the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ into the four fundamental subspaces depicted in Figure A.1. A solution to $Ax = b$ *exists* for every b if and only if the *rows* of A are linearly independent. A solution to $Ax = b$ is *unique* if and only if the *columns* of A are linearly independent.

A.5 Pseudo-Inverse

The solution of $Ax = y$ when A is invertible is $x = A^{-1}y$ where A^{-1} is the inverse of A. Often an approximate inverse of $y = Ax$ is required when A is *not* invertible. This is yielded by the pseudo-inverse $A^\dagger$ of A; if $A \in \mathbb{R}^{m \times n}$, then $A^\dagger \in \mathbb{R}^{n \times m}$. The properties of the pseudo-inverse are illustrated in Figure A.2 for the case when $A \in \mathbb{R}^{2 \times 2}$ where both $\mathcal{R}(A)$ and $\mathcal{N}(A)$ have dimension 1. Suppose we require a solution to the equation $Ax = y$. Since every $x \in \mathbb{R}^2$ is mapped into $\mathcal{R}(A)$, we see that a solution may only be obtained if $y \in \mathcal{R}(A)$. Suppose this is not the case, as in Figure A.2. Then the closest point, in the Euclidean sense, to y in $\mathcal{R}(A)$ is the point y^* which is the orthogonal projection

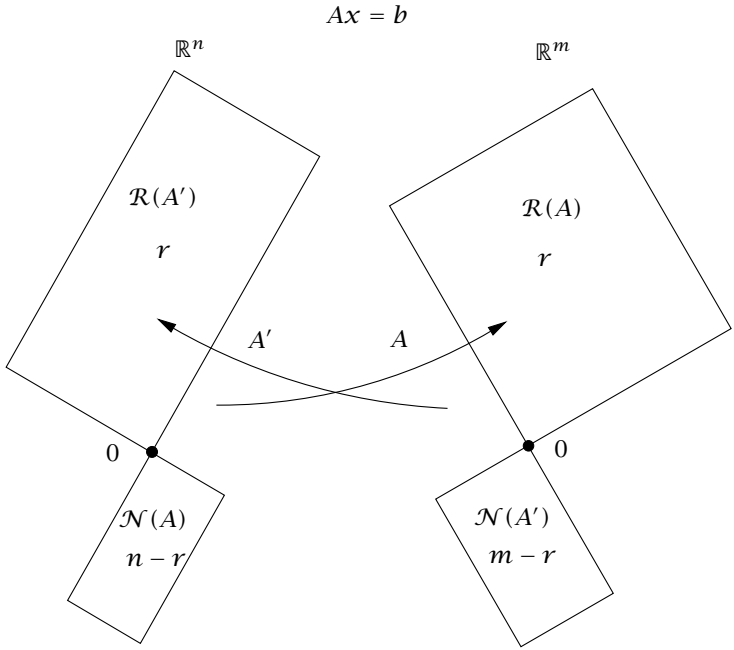


Figure A.1: The four fundamental subspaces of matrix A (after (Strang, 1980, p.88)). The dimension of the range of A and A' is r, the rank of matrix A. The nullspace of A and range of A' are orthogonal as are the nullspace of A' and range of A. Solutions to $Ax = b$ exist for all b if and only if $m = r$ (rows independent). A solution to $Ax = b$ is unique if and only if $n = r$ (columns independent).

of y onto $\mathcal{R}(A)$, i.e., $y - y^*$ is orthogonal to $\mathcal{R}(A)$. Since $y^* \in \mathcal{R}(A)$, there exists a point in $\mathbb{R}^2$ that A maps into y^*. Now A maps any point of the form $x + h$ where $h \in \mathcal{N}(A)$ into $A(x + h) = Ax + Ah = Ax$ so that there must exist a point $x^* \in (\mathcal{N}(A))^\perp = \mathcal{R}(A')$ such that $Ax^* = y^*$, as shown in Figure A.2. All points of the form $x = x^* + h$ where $h \in \mathcal{N}(A)$ are also mapped into y^*; x^* is the point of least norm that satisfies $Ax^* = y^*$ where y^* is that point in $\mathcal{R}(A)$ closest, in the Euclidean sense, to y.

The pseudo-inverse $A^\dagger$ of a matrix $A \in \mathbb{R}^{m \times n}$ is a matrix in $\mathbb{R}^{n \times m}$ that maps every $y \in \mathbb{R}^m$ to that point $x \in \mathcal{R}(A')$ of least Euclidean norm that minimizes $|y - Ax|_2$. The operation of $A^\dagger$ is illustrated in

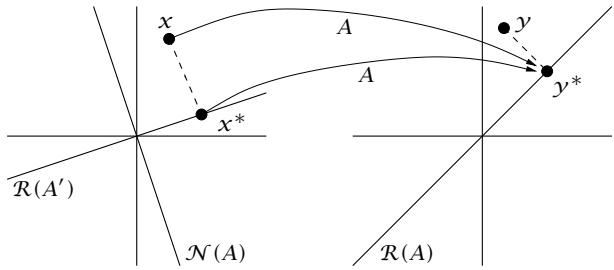


Figure A.2: Matrix A maps into $\mathcal{R}(A)$.

Figure A.3. Hence $AA^\dagger$ projects any point $y \in \mathbb{R}^m$ orthogonally onto $\mathcal{R}(A)$, i.e., $AA^\dagger y = y^*$, and $A^\dagger A$ projects any $x \in \mathbb{R}^n$ orthogonally onto $\mathcal{R}(A')$, i.e., $A^\dagger A x = x^*$.

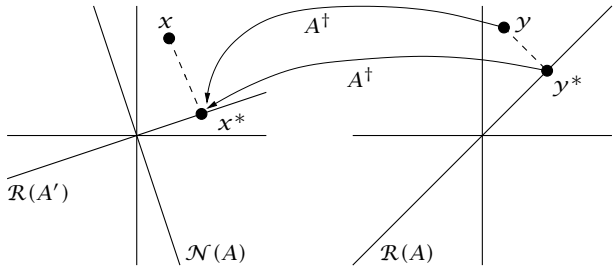


Figure A.3: Pseudo-inverse of A maps into $\mathcal{R}(A')$.

If $A \in \mathbb{R}^{m\times n}$ where $m < n$ has maximal rank m, then $AA' \in \mathbb{R}^{m\times m}$ is invertible and $A^\dagger = A'(AA')^{-1}$; in this case, $\mathcal{R}(A) = \mathbb{R}^m$ and every $y \in \mathbb{R}^m$ lies in $\mathcal{R}(A)$. Similarly, if $n < m$ and A has maximal rank n, then $A'A \in \mathbb{R}^{n\times n}$ is invertible and $A^\dagger = (A'A)^{-1}A'$; in this case, $\mathcal{R}(A') = \mathbb{R}^n$ and every $x \in \mathbb{R}^n$ lies in $\mathcal{R}(A')$. More generally, if $A \in \mathbb{R}^{m\times n}$ has rank r, then A has the *singular-value decomposition* $A = U\Sigma V'$ where $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{r\times n}$ are orthogonal matrices, i.e., $U'U = I_r$ and $V'V = I_r$, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$ where $\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$. The pseudo-inverse of A is then

$$A^\dagger = V\Sigma^{-1}U'$$

A.6 Partitioned Matrix Inversion Theorem

Let matrix Z be partitioned into

$$Z = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

and assume Z^{-1}, B^{-1} and E^{-1} exist. Performing row elimination gives

$$Z^{-1} = \begin{bmatrix} B^{-1} + B^{-1}C(E - DB^{-1}C)^{-1}DB^{-1} & -B^{-1}C(E - DB^{-1}C)^{-1} \\ -(E - DB^{-1}C)^{-1}DB^{-1} & (E - DB^{-1}C)^{-1} \end{bmatrix}$$

Note that this result is still valid if E is singular. Performing column elimination gives

$$Z^{-1} = \begin{bmatrix} (B - CE^{-1}D)^{-1} & -(B - CE^{-1}D)^{-1}CE^{-1} \\ -E^{-1}D(B - CE^{-1}D)^{-1} & E^{-1} + E^{-1}D(B - CE^{-1}D)^{-1}CE^{-1} \end{bmatrix}$$

Note that this result is still valid if B is singular. A host of other useful control-related inversion formulas follow from these results. Equating the (1,1) or (2,2) entries of Z^{-1} gives the identity

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1}$$

A useful special case of this result is

$$(I + X^{-1})^{-1} = I - (I + X)^{-1}$$

Equating the (1,2) or (2,1) entries of Z^{-1} gives the identity

$$(A + BCD)^{-1}BC = A^{-1}B(DA^{-1}B + C^{-1})^{-1}$$

Determinants. We require some results on determinants of partitioned matrices when using normal distributions in the discussion of probability. If E is nonsingular

$$\det(A) = \det(E)\det(B - CE^{-1}D)$$

If B is nonsingular

$$\det(A) = \det(B)\det(E - DB^{-1}C)$$

A.7 Quadratic Forms

Positive definite and positive semidefinite matrices show up often in LQ problems. Here are some basic facts about them. In the following Q is real and symmetric and R is real.

The matrix Q is positive definite ($Q > 0$), if

$$x'Qx > 0, \quad \forall \text{ nonzero } x \in \mathbb{R}^n$$

The matrix Q is positive semidefinite ($Q \geq 0$), if

$$x'Qx \geq 0, \quad \forall x \in \mathbb{R}^n$$

You should be able to prove the following facts.

1. $Q > 0$ if and only if $\lambda(Q) > 0, \quad \lambda \in \text{eig}(Q)$.
2. $Q \geq 0$ if and only if $\lambda(Q) \geq 0, \quad \lambda \in \text{eig}(Q)$.
3. $Q \geq 0 \Rightarrow R'QR \geq 0 \quad \forall R$.
4. $Q > 0$ and R nonsingular $\Rightarrow R'QR > 0$.
5. $Q > 0$ and R full column rank $\Rightarrow R'QR > 0$.
6. $Q_1 > 0, Q_2 \geq 0 \Rightarrow Q = Q_1 + Q_2 > 0$.
7. $Q > 0 \Rightarrow z^*Qz > 0 \quad \forall$ nonzero $z \in \mathbb{C}^n$.
8. Given $Q \geq 0$, $x'Qx = 0$ if and only if $Qx = 0$.

You may want to use the Schur decomposition (Schur, 1909) of a matrix in establishing some of these eigenvalue results. Golub and Van Loan (1996, p.313) provide the following theorem

Theorem A.1 (Schur decomposition)**.** *If $A \in \mathbb{C}^{n\times n}$ then there exists a unitary $Q \in \mathbb{C}^{n\times n}$ such that*

$$Q^*AQ = T$$

in which T is upper triangular.

Note that because T is upper triangular, its diagonal elements are the eigenvalues of A. Even if A is a real matrix, T can be complex because the eigenvalues of a real matrix may come in complex conjugate pairs. Recall a matrix Q is unitary if $Q^*Q = I$. You should also be able to prove the following facts (Horn and Johnson, 1985).

1. If $A \in \mathbb{C}^{n\times n}$ and $BA = I$ for some $B \in \mathbb{C}^{n\times n}$, then
   (a) A is nonsingular
   (b) B is unique
   (c) $AB = I$
2. The matrix Q is unitary if and only if
   (a) Q is nonsingular and $Q^* = Q^{-1}$
   (b) $QQ^* = I$
   (c) Q^* is unitary
   (d) The rows of Q form an orthonormal set
   (e) The columns of Q form an orthonormal set
3. If A is real and symmetric, then T is real and diagonal and Q can be chosen real and orthogonal. It does not matter if the eigenvalues of A are repeated.

For real, but not necessarily symmetric, A you can restrict yourself to real matrices, by using the real Schur decomposition (Golub and Van Loan, 1996, p.341), but the price you pay is that you can achieve only block upper triangular T, rather than strictly upper triangular T.

Theorem A.2 (Real Schur decomposition)**.** *If $A \in \mathbb{R}^{n\times n}$ then there exists an orthogonal $Q \in \mathbb{R}^{n\times n}$ such that*

$$Q'AQ = \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{mm} \end{bmatrix}$$

in which each R_{ii} is either a real scalar or a 2×2 real matrix having complex conjugate eigenvalues; the eigenvalues of R_{ii} are the eigenvalues of A.

If the eigenvalues of R_{ii} are disjoint (i.e., the eigenvalues are not repeated), then R can be taken block diagonal instead of block triangular (Golub and Van Loan, 1996, p.366).

A.8 Norms in $\mathbb{R}^n$

A norm in $\mathbb{R}^n$ is a function $|\cdot| : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ such that

(a) $|x| = 0$ if and only if $x = 0$;

(b) $|\lambda x| = |\lambda| \ |x|$, for all $\lambda \in \mathbb{R}, x \in \mathbb{R}^n$;

(c) $|x + y| \leq |x| + |y|$, for all $x, y \in \mathbb{R}^n$.

Let $\mathcal{B} := \{x \mid |x| \leq 1\}$ denote the *closed* ball of radius 1 centered at the origin. For any $x \in \mathbb{R}^n$ and $\rho > 0$, we denote by $x \oplus \rho\mathcal{B}$ or $B(x, \rho)$ the *closed* ball $\{z \mid |z - x| \leq \rho\}$ of radius ρ centered at x. Similarly $\{x \mid |x| < 1\}$ denotes the *open* ball of radius 1 centered at the origin and $\{z \mid |z - x| < \rho\}$ the *open* ball of radius ρ centered at x; closed and open sets are defined below.

A.9 Sets in $\mathbb{R}^n$

The complement of $S \subset \mathbb{R}^n$ in $\mathbb{R}^n$, is the set $S^c := \{x \in \mathbb{R}^n \mid x \notin S\}$. A set $X \subset \mathbb{R}^n$ is said to be *open*, if for every $x \in X$, there exists a $\rho > 0$ such that $B(x, \rho) \subseteq X$. A set $X \subset \mathbb{R}^n$ is said to be *closed* if X^c, its complement in $\mathbb{R}^n$, is open.

A set $X \subset \mathbb{R}^n$ is said to be *bounded* if there exists an $M < \infty$ such that $|x| \leq M$ for all $x \in X$. A set $X \subset \mathbb{R}^n$ is said to be *compact* if X is closed and bounded. An element $x \in S \subseteq \mathbb{R}^n$ is an *interior* point of the set S if there exists a $\rho > 0$ such that $z \in S$, for all $|z - x| < \rho$. The interior of a set $S \subset \mathbb{R}^n$, $\text{int}(S)$, is the set of all interior points of S; $\text{int}(S)$ is an open set, the *largest*[2] open subset of S. For example, if $S = [a, b] \subset \mathbb{R}$, then $\text{int}(S) = (a, b)$; as another example, $\text{int}(B(x, \rho)) = \{z \mid |z - x| < \rho\}$. The closure of a set $S \subset \mathbb{R}^n$, denoted $\bar{S}$, is the *smallest*[3] closed set containing S. For example, if $S = (a, b] \subset \mathbb{R}$, then $\bar{S} = [a, b]$. The boundary of $S \subset \mathbb{R}^n$, is the set $\delta S := \bar{S} \setminus \text{int}(S) = \{s \in \bar{S} \mid s \notin \text{int}(S)\}$. For example, if $S = (a, b] \subset \mathbb{R}$, then $\text{int}(S) = (a, b), \bar{S} = [a, b], \partial S = \{a, b\}$.

An *affine* set $S \subset \mathbb{R}^n$ is a set that can be expressed in the form $S = \{x\} \oplus \mathcal{V} := \{x + v \mid v \in \mathcal{V}\}$ for some $x \in \mathbb{R}^n$ and some subspace $\mathcal{V}$ of $\mathbb{R}^n$. An example is a line in $\mathbb{R}^n$ not passing through the origin. The *affine hull* of a set $S \subset \mathbb{R}^n$, denoted $\text{aff}(S)$, is the smallest[4] affine

[2] Largest in the sense that every open subset of S is a subset of $\text{int}(S)$.

[3] Smallest in the sense that $\bar{S}$ is a subset of any closed set containing S.

[4] In the sense that $\text{aff}(S)$ is a subset of any other affine set containing S.

set that contains S. That is equivalent to the intersection of all affine sets containing S.

Some sets S, such as a line in $\mathbb{R}^n, n \geq 2$, do not have an interior, but do have an interior *relative* to the smallest affine set in which S lies, which is $\text{aff}(S)$ defined above. The *relative interior* of S is the set $\{x \in S \mid \exists \rho > 0 \text{ such that } \text{int}(B(x,\rho)) \cap \text{aff}(S) \subset S\}$. Thus the line segment, $S := \{x \subset \mathbb{R}^2 \mid x = \lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (1-\lambda) \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda \in [0,1]\}$ does not have an interior, but does have an interior relative to the line containing it, $\text{aff}(S)$. The relative interior of S is the open line segment $\{x \in \mathbb{R}^2 \mid x = \lambda \begin{bmatrix} 1 \\ 0 \end{bmatrix} + (1-\lambda) \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \lambda \in (0,1)\}$.

A.10 Sequences

Let the set of nonnegative integers be denoted by $\mathbb{I}_{\geq 0}$. A *sequence* is a function from $\mathbb{I}_{\geq 0}$ into $\mathbb{R}^n$. We denote a sequence by its values, $(x_i)_{i \in \mathbb{I}_{\geq 0}}$. A *subsequence* of $(x_i)_{i \in \mathbb{I}_{\geq 0}}$ is a sequence of the form $(x_i)_{i \in K}$, where K is an infinite subset of $\mathbb{I}_{\geq 0}$.

A sequence $(x_i)_{i \in \mathbb{I}_{\geq 0}}$ in $\mathbb{R}^n$ is said to *converge* to a point $\hat{x}$ if $\lim_{i \to \infty} |x_i - \hat{x}| = 0$, i.e., if, for all $\delta > 0$, there exists an integer k such that $|x_i - \hat{x}| \leq \delta$ for all $i \geq k$; we write $x_i \to \hat{x}$ as $i \to \infty$ to denote the fact that the sequence (x_i) converges to $\hat{x}$. The point $\hat{x}$ is called a *limit* of the sequence (x_i). A point x^* is said to be an *accumulation* point of a sequence $(x_i)_{i \in \mathbb{I}_{\geq 0}}$ in $\mathbb{R}^n$, if there exists an infinite subset $K \subset \mathbb{I}_{\geq 0}$ such that $x_i \to x^*$ as $i \to \infty, i \in K$ in which case we say $x_i \overset{K}{\to} x^*$.[5]

Let (x_i) be a bounded infinite sequence in $\mathbb{R}$ and let the S be the set of all accumulation points of (x_i). Then S is compact and $\limsup x_i$ is the largest and $\liminf x_i$ the smallest accumulation point of (x_i):

$$\limsup_{i \to \infty} x_i := \max\{x \mid x \in S\}, \text{ and}$$
$$\liminf_{i \to \infty} x_i := \min\{x \mid x \in S\}$$

Theorem A.3 (Bolzano-Weierstrass)**.** *Suppose $X \subset \mathbb{R}^n$ is compact and $(x_i)_{i \in \mathbb{I}_{\geq 0}}$ takes its values in X. Then $(x_i)_{i \in \mathbb{I}_{\geq 0}}$ must have at least one accumulation point.*

From Exercise A.7, it follows that the accumulation point postulated by Theorem A.3 lies in X. In proving asymptotic stability we need the following property of monotone sequences.

[5] Be aware of inconsistent usage of the term *limit point.* Some authors use limit point as synonymous with limit. Others use limit point as synonymous with accumulation point. For this reason we avoid the term limit point.

Proposition A.4 (Convergence of monotone sequences)**.** *Suppose that* $(x_i)_{i\in\mathbb{I}_{\geq 0}}$ *is a sequence in* $\mathbb{R}$ *such that* $x_0 \geq x_1 \geq x_2 \geq \ldots$*, i.e., suppose the sequence is monotone nonincreasing. If* (x_i) *has an accumulation point* x^**, then* $x_i \to x^*$ *as* $i \to \infty$*, i.e.,* x^* *is a limit.*

Proof. For the sake of contradiction, suppose that $(x_i)_{i\in\mathbb{I}_{\geq 0}}$ does not converge to x^*. Then, for some $\rho > 0$, there exists a subsequence $(x_i)_{i\in K}$ such that $x_i \notin B(x^*, \rho)$ for all $i \in K$, i.e., $|x_i - x^*| > \rho$ for all $i \in K$. Since x^* is an accumulation point, there exists a subsequence $(x_i)_{i\in K^*}$ such that $x_i \overset{K^*}{\to} x^*$. Hence there is an $i_1 \in K^*$ such that $|x_i - x^*| \leq \rho/2$, for all $i \geq i_1, i \in K^*$. Let $i_2 \in K$ be such that $i_2 > i_1$. Then we must have that $x_{i_2} \leq x_{i_1}$ and $|x_{i_2} - x^*| > \rho$, which leads to the conclusion that $x_{i_2} < x^* - \rho$. Now let $i_3 \in K^*$ be such that $i_3 > i_2$. Then we must have that $x_{i_3} \leq x_{i_2}$ and hence that $x_{i_3} < x^* - \rho$ which implies that $|x_{i_3} - x^*| > \rho$. But this contradicts the fact that $|x_{i_3} - x^*| \leq \rho/2$, and hence we conclude that $x_i \to x^*$ as $i \to \infty$. ■

It follows from Proposition A.4 that if $(x_i)_{i\in\mathbb{I}_{\geq 0}}$ is a monotone decreasing sequence in $\mathbb{R}$ bounded below by b, then the sequence $(x_i)_{i\in\mathbb{I}_{\geq 0}}$ converges to some $x^* \in \mathbb{R}$ where $x^* \geq b$.

A.11 Continuity

We now summarize some essential properties of continuous functions.

1. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuous at a point* $x \in \mathbb{R}^n$, if for every $\delta > 0$ there exists a $\rho > 0$ such that

$$|f(x') - f(x)| < \delta \ \forall x' \in \text{int}(B(x, \rho))$$

   A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuous* if it is continuous at all $x \in \mathbb{R}^n$.

2. Let X be a closed subset of $\mathbb{R}^n$. A function $f : X \to \mathbb{R}^m$ is said to be *continuous at a point* x in X if for every $\delta > 0$ there exists a $\rho > 0$ such that

$$|f(x') - f(x)| < \delta \ \forall x' \in \text{int}(B(x, \rho)) \cap X$$

   A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuous* on X if it is continuous at all x in X.

3. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *upper semicontinuous at a point* $x \in \mathbb{R}^n$, if for every $\delta > 0$ there exists a $\rho > 0$ such that

$$f(x') - f(x) < \delta \ \forall x' \in \text{int}(B(x, \rho))$$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *upper semicontinuous* if it is upper semicontinuous at all $x \in \mathbb{R}^n$.

4. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *lower semicontinuous at a point* $x \in \mathbb{R}^n$, if for every $\delta > 0$ there exists a $\rho > 0$ such that

$$f(x') - f(x) > -\delta \ \forall x' \in \text{int}(B(x, \rho))$$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *lower semicontinuous* if it is lower semicontinuous at all $x \in \mathbb{R}^n$.

5. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *uniformly continuous* on a subset $X \subset \mathbb{R}^n$ if for any $\delta > 0$ there exists a $\rho > 0$ such that for any $x', x'' \in X$ satisfying $|x' - x''| < \rho$,

$$\left| f(x') - f(x'') \right| < \delta$$

Proposition A.5 (Uniform continuity)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous and that $X \subset \mathbb{R}^n$ is compact. Then f is uniformly continuous on X.*

Proof. For the sake of contradiction, suppose that f is *not* uniformly continuous on X. Then, for some $\delta > 0$, there exist sequences $\left(x_i'\right)$, $\left(x_i''\right)$ in X such that

$$\left| x_i' - x_i'' \right| < (1/i), \text{ for all } i \in \mathbb{I}_{\geq 0}$$

but

$$\left| f(x_i') - f(x_i'') \right| > \delta, \text{ for all } i \in \mathbb{I}_{\geq 0} \tag{A.1}$$

Since X is compact, there must exist a subsequence $\left(x_i'\right)_{i \in K}$ such that $x_i' \overset{K}{\to} x^* \in X$ as $i \to \infty$. Furthermore, because of (A.1), $x_i'' \overset{K}{\to} x^*$ also holds. Hence, since $f(\cdot)$ is continuous, we must have $f(x_i') \overset{K}{\to} f(x^*)$ and $f(x_i'') \overset{K}{\to} f(x^*)$. Therefore, there exists a $i_0 \in K$ such that for all $i \in K, i \geq i_0$

$$\left| f(x_i') - f(x_i'') \right| \leq \left| f(x_i') - f(x^*) \right| + \left| f(x^*) - f(x_i'') \right| < \delta/2$$

contradicting (A.1). This completes our proof. ∎

Proposition A.6 (Compactness of continuous functions of compact sets)**.** *Suppose that $X \subset \mathbb{R}^n$ is compact and that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous. Then the set*

$$f(X) \coloneqq \{f(x) \mid x \in X\}$$

is compact.

Proof.

(a) First we show that $f(X)$ is closed. Thus, let $(f(x_i) \mid i \in \mathbb{I}_{\geq 0})$, with $x_i \in X$, be any sequence in $f(X)$ such that $f(x_i) \to y$ as $i \to \infty$. Since (x_i) is in a compact set X, there exists a subsequence $(x_i)_{i \in K}$ such that $x_i \overset{K}{\to} x^* \in X$ as $i \to \infty$. Since $f(\cdot)$ is continuous, $f(x_i) \overset{K}{\to} f(x^*)$ as $i \to \infty$. But y is the limit of $(f(x_i))_{i \in \mathbb{I}_{\geq 0}}$ and hence it is the limit of any subsequence of $(f(x_i))$. We conclude that $y = f(x^*)$ and hence that $y \in f(X)$, i.e., $f(X)$ is closed.

(b) Next, we prove that $f(X)$ is bounded. Suppose $f(X)$ is not bounded. Then there exists a sequence (x_i) such that $|f(x_i)| \geq i$ for all $i \in \mathbb{I}_{\geq 0}$. Now, since (x_i) is in a compact set, there exists a subsequence $(x_i)_{i \in K}$ such that $x_i \overset{K}{\to} x^*$ with $x^* \in X$, and $f(x_i) \overset{K}{\to} f(x^*)$ by continuity of $f(\cdot)$. Hence there exists an i_0 such that for any $j > i > i_0$, $j, i \in K$

$$\left|f(x_j) - f(x_i)\right| \leq \left|f(x_j) - f(x^*)\right| + \left|f(x_i) - f(x^*)\right| < 1/2 \quad \text{(A.2)}$$

Let $i \geq i_0$ be given. By hypothesis there exists a $j \in K$, $j \geq i$ such that $\left|f(x_j)\right| \geq j \geq |f(x_i)| + 1$. Hence

$$\left|f(x_j) - f(x_i)\right| \geq \left|\, \left|f(x_j)\right| - |f(x_i)| \,\right| \geq 1$$

which contradicts (A.2). Thus $f(X)$ must be bounded, which completes the proof. ∎

Let $Y \subset \mathbb{R}$. Then $\inf(Y)$, the *infimum* of Y, is defined to be the greatest lower bound[6] of Y. If $\inf(Y) \in Y$, then $\min(Y) \coloneqq \min\{y \mid y \in Y\}$, the minimum of the set Y, exists and is equal to $\inf(Y)$. The infimum of a set Y always exists if Y is not empty and is bounded from below, in which case there always exist sequences $(y_i) \in Y$ such that $y_i \searrow \beta \coloneqq \inf(Y)$ as $i \to \infty$. Note that $\beta \coloneqq \inf(Y)$ does not necessarily lie in the set Y.

[6]The value $\alpha \in \mathbb{R}$ is the greatest lower bound of Y if $y \geq \alpha$ for all $y \in Y$, and $\beta > \alpha$ implies that β is *not* a lower bound for Y.

Proposition A.7 (Weierstrass)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and that $X \subset \mathbb{R}^n$ is compact. Then there exists an $\hat{x} \in X$ such that*

$$f(\hat{x}) = \inf_{x \in X} f(x)$$

i.e., $\min_{x \in X} f(x)$ *is well defined.*

Proof. Since X is compact, $f(X)$ is bounded. Hence $\inf_{x \in X} f(x) = \alpha$ is finite. Let (x_i) be an infinite sequence in X such that $f(x_i) \searrow \alpha$ as $i \to \infty$. Since X is compact, there exists a converging subsequence $(x_i)_{i \in K}$ such that $x_i \overset{K}{\to} \hat{x} \in X$. By continuity, $f(x_i) \overset{K}{\to} f(\hat{x})$ as $i \to \infty$. Because $(f(x_i))$ is a monotone nonincreasing sequence that has an accumulation point $f(\hat{x})$, it follows from Proposition A.4 that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. Since the limit of the sequence $(f(x_i))$ is unique, we conclude that $f(\hat{x}) = \alpha$. ∎

A.12 Derivatives

We first define some notation. If $f : \mathbb{R}^n \to \mathbb{R}$, then $(\partial/\partial x)f(x)$ is a *row* vector defined by

$$(\partial/\partial x)f(x) := [(\partial/\partial x_1)f(x), \ldots, (\partial/\partial x_n)f(x)]$$

provided the partial derivatives $(\partial/\partial x_i)f(x)$, $i = 1, 2, \ldots, n$ exist. Similarly, if $f : \mathbb{R}^n \to \mathbb{R}^m$, $(\partial/\partial x)f(x)$ is defined to be the matrix

$$(\partial/\partial x)f(x) := \begin{bmatrix} (\partial/\partial x_1)f_1(x) & (\partial/\partial x_2)f_1(x) & \ldots & (\partial/\partial x_n)f_1(x) \\ (\partial/\partial x_1)f_2(x) & (\partial/\partial x_2)f_2(x) & \ldots & (\partial/\partial x_n)f_2(x) \\ \vdots & \vdots & \vdots & \vdots \\ (\partial/\partial x_1)f_m(x) & (\partial/\partial x_2)f_m(x) & \ldots & (\partial/\partial x_n)f_m(x) \end{bmatrix}$$

where x_i and f_i denote, respectively, the ith component of the vectors x and f. We sometimes use $f_x(x)$ in place of $(\partial/\partial x)f(x)$. If $f : \mathbb{R}^n \to \mathbb{R}$, then its *gradient* $\nabla f(x)$ is a *column* vector defined by

$$\nabla f(x) := \begin{bmatrix} (\partial/\partial x_1)f(x) \\ (\partial/\partial x_2)f(x) \\ \vdots \\ (\partial/\partial x_n)f(x) \end{bmatrix}$$

and its *Hessian* is $\nabla^2 f(x) = (\partial^2/\partial x^2) f(x) = f_{xx}(x)$ defined by

$$\nabla^2 f(x) \coloneqq \begin{bmatrix} (\partial^2/\partial x_1^2) f(x) & (\partial^2/\partial x_1 \partial x_2) f(x) & \dots & (\partial^2/\partial x_1 \partial x_n) f(x) \\ (\partial^2/\partial x_2 \partial x_1) f(x) & (\partial x_2^2) f(x) & \dots & (\partial^2/\partial x_2 \partial x_n) f(x) \\ \vdots & \vdots & \ddots & \vdots \\ (\partial^2/\partial x_n \partial x_1) f(x) & (\partial^2/\partial x_n \partial x_2) f(x) & \dots & (\partial^2/\partial x_n^2) f(x) \end{bmatrix}$$

We note that $\nabla f(x) = [(\partial/\partial x) f(x)]' = f_x'(x)$.

We now define what we mean by the derivative of $f(\cdot)$. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a continuous function with domain $\mathbb{R}^n$. We say that $f(\cdot)$ is differentiable at $\hat{x}$ if there exists a matrix $Df(\hat{x}) \in \mathbb{R}^{m \times n}$ (the Jacobian) such that

$$\lim_{h \to 0} \frac{|f(\hat{x} + h) - f(\hat{x}) - Df(\hat{x})h|}{|h|} = 0$$

in which case $Df(\cdot)$ is called the derivative of $f(\cdot)$ at $\hat{x}$. When $f(\cdot)$ is differentiable at all $x \in \mathbb{R}^n$, we say that f is *differentiable*.

We note that the affine function $h \mapsto f(\hat{x}) + Df(\hat{x})h$ is a first order approximation of $f(\hat{x} + h)$. The Jacobian can be expressed in terms of the partial derivatives of $f(\cdot)$.

Proposition A.8 (Derivative and partial derivative)**.** *Suppose that the function $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\hat{x}$. Then its derivative $Df(\hat{x})$ satisfies*

$$Df(\hat{x}) = f_x(\hat{x}) \coloneqq \partial f(\hat{x})/\partial x$$

Proof. From the definition of $Df(\hat{x})$ we deduce that for each $i \in \{1, 2, \dots, m\}$

$$\lim_{h \to 0} \frac{|f_i(\hat{x} + h) - f_i(\hat{x}) - Df_i(\hat{x})h|}{|h|} = 0$$

where f_i is the ith element of f and $(Df)_i$ the ith row of Df. Set $h = te_j$, where e_j is the j-th unit vector in $\mathbb{R}^n$ so that $|h| = t$. Then $(Df)_i(\hat{x})h = t(Df)_i(\hat{x})e_j = (Df)_{ij}(\hat{x})$, the ijth element of the matrix $Df(\hat{x})$. It then follows that

$$\lim_{t \searrow 0} \frac{\left| f^i(\hat{x} + te_j) - f(\hat{x}) - t(Df)_{ij}(\hat{x}) \right|}{t} = 0$$

which shows that $(Df)_{ij}(\hat{x}) = \partial f_i(\hat{x})/\partial x_j$. ∎

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *locally Lipschitz continuous* at $\hat{x}$ if there exist $L \in [0, \infty)$, $\hat{\rho} > 0$ such that

$$|f(x) - f(x')| \leq L\,|x - x'|\,, \text{ for all } x, x' \in B(\hat{x}, \hat{\rho})$$

The function f is globally Lipschitz continuous if the inequality holds for all $x, x' \in \mathbb{R}^n$. The constant L is called the *Lipschitz constant* of f. It should be noted that the existence of partial derivatives of $f(\cdot)$ does not ensure the existence of the derivative $Df(\cdot)$ of $f(\cdot)$; see e.g. Apostol (1974, p.103). Thus consider the function

$$f(x, y) = x + y \text{ if } x = 0 \text{ or } y = 0$$

$$f(x, y) = 1 \text{ otherwise}$$

In this case

$$\frac{\partial f(0,0)}{\partial x} = \lim_{t \to 0} \frac{f(t,0) - f(0,0)}{t} = 1$$
$$\frac{\partial f(0,0)}{\partial y} = \lim_{t \to 0} \frac{f(0,t) - f(0,0)}{t} = 1$$

but the function is not even continuous at $(0,0)$. In view of this, the following result is relevant.

Proposition A.9 (Continuous partial derivatives)**.** *Consider a function $f : \mathbb{R}^n \to \mathbb{R}^m$ such that the partial derivatives $\partial f^i(x)/dx^j$ exist in a neighborhood of $\hat{x}$, for $i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$. If these partial derivatives are continuous at $\hat{x}$, then the derivative $Df(\hat{x})$ exists and is equal to $f_x(\hat{x})$.*

The following *chain rule* holds.

Proposition A.10 (Chain rule)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ is defined by $f(x) = h(g(x))$ with both $h : \mathbb{R}^l \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^l$ differentiable. Then*

$$\frac{\partial f(\hat{x})}{\partial x} = \frac{\partial h(g(\hat{x}))}{\partial y} \frac{\partial g(\hat{x})}{\partial x}$$

The following result Dieudonne (1960), replaces, *inter alia*, the mean value theorem for functions $f : \mathbb{R}^n \to \mathbb{R}^m$ when $m > 1$.

Proposition A.11 (Mean value theorem for vector functions)**.**

(a) Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ has continuous partial derivatives at each point x of $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$,

$$f(y) = f(x) + \int_0^1 f_x(x + s(y - x))(y - x) ds$$

(b) Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ has continuous partial derivatives of order two at each point x of $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$,

$$f(y) = f(x)+f_x(x)(y-x)+\int_0^1 (1-s)(y-x)' f_{xx}(x+s(y-x))(y-x)ds$$

Proof.

(a) Consider the function $g(s) = f(x + s(y - x))$ where $f : \mathbb{R}^n \to \mathbb{R}^m$. Then $g(1) = f(y), g(0) = f(x)$ and

$$\begin{aligned} g(1) - g(0) &= \int_0^1 g'(s)ds \\ &= \int_0^1 Df(x + s(y - x))(y - x)ds \end{aligned}$$

which completes the proof for $p = 1$.

(b) Consider the function $g(s) = f(x + s(y - x))$ where $f : \mathbb{R}^n \to \mathbb{R}$. Then

$$\frac{d}{ds}[g'(s)(1 - s) + g(s)] = g''(s)(1 - s)$$

Integrating from 0 to 1 yields

$$g(1) - g(0) - g'(0) = \int_0^1 (1 - s)g''(s)ds$$

But $g''(s) = (y-x)' f_{xx}(x+s(y-x))(y-x)$ so that the last equation yields

$$f(y)-f(x) = f_x(x)(y-x)+\int_0^1 (1-s)(y-x)' f_{xx}(x+s(y-x))(y-x)ds$$

when $g(s)$ is replaced by $f(x + s(y - x))$.

■

Finally, we define directional derivatives which may exist even when a function fails to have a derivative. Let $f : \mathbb{R}^n \to \mathbb{R}^m$. We define the *directional derivative* of f at a point $\hat{x} \in \mathbb{R}^n$ in the direction $h \in \mathbb{R}^n (h \neq 0)$ by

$$df(\hat{x}; h) := \lim_{t \searrow 0} \frac{f(\hat{x} + th) - f(\hat{x})}{t}$$

if this limit exists (note that $t > 0$ is required). The directional derivative is positively homogeneous, i.e., $df(x; \lambda h) = \lambda df(x; h)$ for all $\lambda > 0$.

Not all the functions we discuss are differentiable everywhere. Examples include the max function $\psi(\cdot)$ defined by $\psi(x) := \max_i\{f^i(x) \mid i \in I\}$ where each function $f^i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable everywhere. The function $\psi(\cdot)$ is not differentiable at those x for which the active set $I^0(x) := \{i \in I \mid f^i(x) = \psi(x)\}$ has more than one element. The directional derivative $d(x;h)$ exists for all x, h in $\mathbb{R}^n$, however, and is given by

$$d\psi(x;h) = \max_i\{df_i(x;h) \mid i \in I^0(x)\} = \max_i\{\langle \nabla f_i(x), h\rangle \mid i \in I^0(x)\}$$

When, as in this example, the directional derivative exists for all x, h in $\mathbb{R}^n$ we can define a generalization, called the *subgradient*, of the conventional gradient. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has a directional derivative for all x, h in $\mathbb{R}^n$. The $f(\cdot)$ has a subgradient $\partial f(\cdot)$ defined by

$$\partial\psi(x) := \{g \in \mathbb{R}^n \mid df(x;h) \geq \langle g, h\rangle \ \forall h \in \mathbb{R}^n\}$$

The subgradient at a point x is, unlike the ordinary gradient, a set. For our max example ($f(x) = \psi(x) = \max_i\{f_i(x) \mid i \in I\}$) we have $d\psi(x;h) = \max_i\{\langle \nabla f^i(x), h\rangle \mid i \in I^0(x)\}$. In this case, it can be shown that

$$\partial\psi(x) = \text{co}\{\nabla f^i(x) \mid i \in I^0(x)\}$$

If the directional derivative $h \mapsto df(x;h)$ is convex, then the subgradient $\partial f(x)$ is nonempty and the directional derivative $df(x;h)$ may be expressed as

$$df(x;h) = \max_g\{\langle g, h\rangle \mid g \in \partial f(x)\}$$

Figure A.4 illustrates this for the case when $\psi(x) := \max\{f_1(x), f_2(x)\}$ and $I^0(x) = \{1, 2\}$.

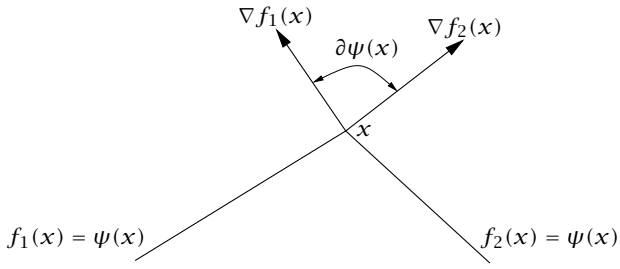


Figure A.4: Subgradient.

A.13 Convex Sets and Functions

Convexity is an enormous subject. We collect here only a few essential results that we will need in our study of optimization; for further details see Rockafellar (1970). We begin with convex sets.

A.13.1 Convex Sets

Definition A.12 (Convex set)**.** A set $S \in \mathbb{R}^n$ is said to be *convex* if, for any $x', x'' \in S$ and $\lambda \in [0,1]$, $(\lambda x' + (1-\lambda)x'') \in S$.

Let S be a subset of $\mathbb{R}^n$. We say that $\text{co}(S)$ is the *convex hull* of S if it is the smallest[7] convex set containing S.

Theorem A.13 (Caratheodory)**.** *Let S be a subset of $\mathbb{R}^n$. If $\bar{x} \in co(S)$, then it may be expressed as a convex combination of no more than $n+1$ points in S, i.e., there exist $m \leq n+1$ distinct points, $\{x_i\}_{i=1}^m$, in S such that $\bar{x} = \sum_{i=1}^m \mu^i x_i, \mu^i > 0, \sum_{i=1}^m \mu^i = 1$.*

Proof. Consider the set

$$C_s := \{x \mid x = \sum_{i=1}^{k_x} \mu^i x_i, x_i \in S, \mu^i \geq 0, \sum_{i=1}^{k_x} \mu^i = 1, k_x \in \mathbb{I}_{\geq 0}\}$$

First, it is clear that $S \subset C_s$. Next, since for any $x', x'' \in C_s, \lambda x' + (1-\lambda x'') \in C_s$, for $\lambda \in [0,1]$, it follows that C_s is convex. Hence we must have that $\text{co}(S) \subset C_s$. Because C_s consists of all the convex combinations of points in S, however, we must also have that $C_s \subset \text{co}(S)$. Hence $C_s = \text{co}(S)$. Now suppose that

$$\bar{x} = \sum_{i=1}^{\bar{k}} \bar{\mu}^i x_i$$

with $\bar{\mu}^i \geq 0, i = 1, 2, \ldots, \bar{k}, \sum_{i=1}^{\bar{k}} \bar{\mu}^i = 1$. Then the following system of equations is satisfied

$$\sum_{i=1}^{\bar{k}} \bar{\mu}^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \tag{A.3}$$

with $\bar{\mu}^i \geq 0$. Suppose that $\bar{k} > n+1$. Then there exist coefficients α^j, $j = 1, 2, \ldots, \bar{k}$, not all zero, such that

$$\sum_{i=1}^{\bar{k}} \alpha^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = 0 \tag{A.4}$$

[7] Smallest in the sense that any other convex set containing S also contains $\text{co}(S)$.

Adding (A.4) multiplied by θ to (A.3) we get

$$\sum_{i=1}^{\bar{k}} (\bar{\mu}^i + \theta\alpha^i) \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix}$$

Suppose, without loss of generality, that at least one $\alpha^i < 0$. Then there exists a $\bar{\theta} > 0$ such that $\bar{\mu}^j + \bar{\theta}\alpha^j = 0$ for some j while $\bar{\mu}^i + \bar{\theta}\alpha^i \geq 0$ for all other i. Thus we have succeeded in expressing $\bar{x}$ as a convex combination of $\bar{k} - 1$ vectors in S. Clearly, these reductions can go on as long as $\bar{x}$ is expressed in terms of more than $(n + 1)$ vectors in S. This completes the proof. ■

Let S_1, S_2 be any two sets in $\mathbb{R}^n$. We say that the hyperplane

$$H = \{x \in \mathbb{R}^n \mid \langle x, v \rangle = \alpha\}$$

separates S_1 and S_2 if

$$\langle x, v \rangle \geq \alpha \text{ for all } x \in S_1$$
$$\langle y, v \rangle \leq \alpha \text{ for all } y \in S_2$$

The separation is said to be *strong* if there exists an $\varepsilon > 0$ such that

$$\langle x, v \rangle \geq \alpha + \varepsilon \text{ for all } x \in S_1$$
$$\langle y, v \rangle \leq \alpha - \varepsilon \text{ for all } y \in S_2$$

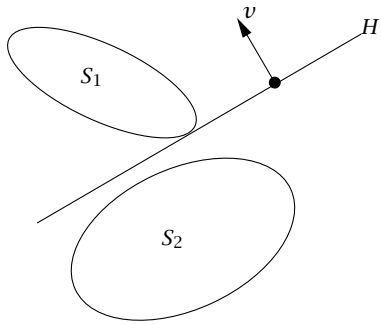


Figure A.5: Separating hyperplane.

Theorem A.14 (Separation of convex sets)**.** *Let S_1, S_2 be two convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$. Then there exists a hyperplane which separates S_1 and S_2. Furthermore, if S_1 and S_2 are closed and either S_1 or S_2 is compact, then the separation can be made strict.*

Theorem A.15 (Separation of convex set from zero)**.** *Suppose that* $S \subset \mathbb{R}^n$ *is closed and convex and* $0 \notin S$*. Let*

$$\hat{x} = \arg\min\{|x|^2 \mid x \in S\}$$

Then

$$H = \{x \mid \langle \hat{x}, x \rangle = |\hat{x}|^2\}$$

separates S *from 0, i.e.,* $\langle \hat{x}, x \rangle \geq |\hat{x}|^2$ *for all* $x \in S$*.*

Proof. Let $x \in S$ be arbitrary. Then, since S is convex, $[\hat{x} + \lambda(x - \hat{x})] \in S$ for all $\lambda \in [0, 1]$. By definition of $\hat{x}$, we must have

$$\begin{aligned} 0 < |\hat{x}|^2 &\leq |\hat{x} + \lambda(x - \hat{x})|^2 \\ &= |\hat{x}|^2 + 2\lambda\langle \hat{x}, x - \hat{x} \rangle + \lambda^2 |x - \hat{x}|^2 \end{aligned}$$

Hence, for all $\lambda \in (0, 1]$,

$$0 \leq 2\langle \hat{x}, x - \hat{x} \rangle + \lambda |x - \hat{x}|^2$$

Letting $\lambda \to 0$ we get the desired result. ∎

Theorem A.15 can be used to prove the following special case of Theorem A.14:

Corollary A.16 (Existence of separating hyperplane)**.** *Let* S_1, S_2 *be two compact convex sets in* $\mathbb{R}^n$ *such that* $S_1 \cap S_2 = \emptyset$*. Then there exists a hyperplane which separates* S_1 *and* S_2*.*

Proof. Let $C = S_1 - S_2 := \{x_1 - x_2 \mid x_1 \in S_1, x_2 \in S_2\}$. Then C is convex and compact and $0 \notin C$. Let $\hat{x} = (\hat{x}_1 - \hat{x}_2) = \arg\min\{|x|^2 \mid x \in C\}$, where $\hat{x}_1 \in S_1$ and $\hat{x}_2 \in S_2$. Then, by Theorem A.15

$$\langle x - \hat{x}, \hat{x} \rangle \geq 0, \text{ for all } x \in C \tag{A.5}$$

Let $x = x_1 - \hat{x}_2$, with $x_1 \in S_1$. Then (A.5) leads to

$$\langle x_1 - \hat{x}_2, \hat{x} \rangle \geq |\hat{x}|^2 \tag{A.6}$$

for all $x_1 \in S_1$. Similarly, letting $x = \hat{x}_1 - x_2$, in (A.5) yields

$$\langle \hat{x}_1 - x_2, \hat{x} \rangle \geq |\hat{x}|^2 \tag{A.7}$$

for all $x_2 \in S_2$. The inequality in (A.7) implies that

$$\langle \hat{x}_1 - \hat{x}_2 + \hat{x}_2 - x_2, \hat{x} \rangle \geq |\hat{x}|^2$$

Since $\hat{x}_1 - \hat{x}_2 = \hat{x}$, we obtain

$$\langle x_2 - \hat{x}_2, \hat{x} \rangle \leq 0 \tag{A.8}$$

for all $x_2 \in S_2$. The desired result follows from (A.6) and (A.8), the separating hyperplane H being $\{x \in \mathbb{R}^n \mid \langle \hat{x}, x - \hat{x}_2 \rangle = 0\}$. ■

Definition A.17 (Support hyperplane)**.** Suppose $S \subset \mathbb{R}^n$ is convex. We say that $H = \{x \mid \langle x - \bar{x}, v \rangle = 0\}$ is a *support hyperplane* to S through $\bar{x}$ with *inward (outward) normal* v if $\bar{x} \in \overline{S}$ and

$$\langle x - \bar{x}, v \rangle \geq 0 \ (\leq 0) \text{ for all } x \in S$$

Theorem A.18 (Convex set and halfspaces)**.** *A closed convex set is equal to the intersection of the halfspaces which contain it.*

Proof. Let C be a closed convex set and A the intersection of halfspaces containing C. Then clearly $C \subset A$. Now suppose $\bar{x} \notin C$. Then there exists a support hyperplane H which separates strictly $\bar{x}$ and C so that $\bar{x}$ does not belong to one halfspace containing C. It follows that $\bar{x} \notin A$. Hence $C^c \subset A^c$ which leads to the conclusion that $A \subset C$. ■

An important example of a convex set is a convex *cone*.

Definition A.19 (Convex cone)**.** A subset C of $\mathbb{R}^n$, $C \neq \emptyset$, is called a *cone* if $x \in C$ implies $\lambda x \in C$ for all $\lambda \geq 0$. A cone C is *pointed* if $C \cap -C = \{0\}$. A *convex* cone is a cone that is convex.

An example of a cone is a halfspaces with a boundary that is a hyperplane passing through the origin; an example of a pointed cone is the positive orthant. A polyhedron C defined by $C := \{x \mid \langle a_i, x \rangle \leq 0, i \in \mathcal{I}\}$ is a convex cone that is pointed

Definition A.20 (Polar cone)**.** Given a cone $C \subset \mathbb{R}^n$, the cone C^* defined by

$$C^* := \{h \mid \langle h, x \rangle \leq 0 \ \forall x \in C\}$$

is called the *polar cone* of C.

An illustration of this definition when C is a polyhedron containing the origin is given in Figure A.6. In this figure, H is the hyperplane with normal h passing through the origin.

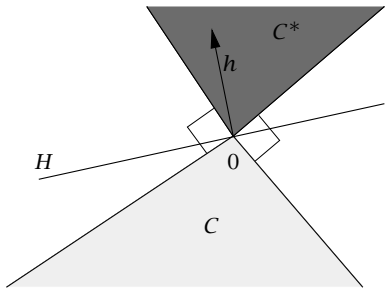


Figure A.6: Polar cone.

Definition A.21 (Cone generator)**.** A cone K is said to be *generated* by a set $\{a_i \mid i \in \mathcal{I}\}$ where $\mathcal{I}$ is an index set if

$$K = \left\{ \sum_{i \in \mathcal{I}} \mu_i a_i \mid \mu_i \geq 0, i \in \mathcal{I} \right\}$$

in which case we write $K = \text{cone}\{a_i \mid i \in \mathcal{I}\}$.

We make use of the following result:

Proposition A.22 (Cone and polar cone generator)**.**

(a) Suppose C is a convex cone containing the origin and defined by

$$C := \{x \in \mathbb{R}^n \mid \langle a_i, x \rangle \leq 0,\ i \in \mathcal{I}\}$$

Then

$$C^* = \text{cone}\{a_i \mid i \in \mathcal{I}\}$$

(b) If C is a closed convex cone, then $(C^*)^* = C$.

(c) If $C_1 \subset C_2$, then $C_2^* \subset C_1^*$.

Proof.

(a) Let the convex set K be defined by

$$K := \text{cone}\{a_i \mid i \in \mathcal{I}\}$$

We wish to prove $C^* = K$. To prove $K \subset C^*$, suppose h is an arbitrary point in $K := \text{cone}\{a_i \mid i \in \mathcal{I}\}$. Then $h = \sum_{i \in \mathcal{I}} \mu_i a_i$ where $\mu_i \geq 0$ for all $i \in \mathcal{I}$. Let x be an arbitrary point in C so that $\langle a_i, x \rangle \leq 0$ for all $i \in \mathcal{I}$. Hence

$$\langle h, x \rangle = \langle \sum_{i \in \mathcal{I}} \mu_i a_i, x \rangle = \sum_{i \in \mathcal{I}} \mu_i \langle a_i, x \rangle \leq 0$$

so that $h \in C^*$. This proves that $K \subset C^*$. To prove that $C^* \subset K$, assume that $h \in C^*$ but that, contrary to what we wish to prove, $h \notin K$. Hence $h = \sum_{i \in \mathcal{I}} \mu_i a_i + \widetilde{h}$ where either $\mu_j > 0$ for at least one $j \in \mathcal{I}$, or $\widetilde{h}$, which is orthogonal to $a_i, i \in \mathcal{I}$, is not zero, or both. If $\mu_j < 0$, let $x \in C$ be such that $\langle a_i, x \rangle = 0$ for all $i \in \mathcal{I}, i \neq j$ and $\langle a_j, x \rangle < 0$; if $\widetilde{h} \neq 0$, let $x \in C$ be such that $\langle \widetilde{h}, x \rangle > 0$ (both conditions can be satisfied). Then

$$\langle h, x \rangle = \langle \mu_j a_j, x \rangle + \langle \widetilde{h}, x \rangle = \mu_j \langle a_j, x \rangle + \langle \widetilde{h}, x \rangle > 0$$

since either both μ_j and $\langle a_j, x \rangle$ are strictly negative or $\widetilde{h} \neq 0$ or both. This contradicts the fact that $x \in C$ and $h \in C^*$ (so that $\langle h, x \rangle \leq 0$). Hence $h \in K$ so that $C^* \subset K$. It follows that $C^* = \text{cone}\{a_i \mid i \in \mathcal{I}\}$.

(b) That $(C^*)^* = C$ when C is a closed convex cone is given in Rockafellar and Wets (1998), Corollary 6.21.

(c) This result follows directly from the definition of a polar cone.

■

A.13.2 Convex Functions

Next we turn to convex functions. For an example see Figure A.7.

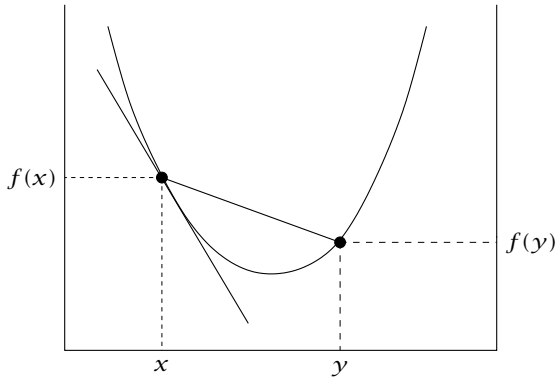


Figure A.7: A convex function.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if for any $x', x'' \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,

$$f(\lambda x' + (1 - \lambda)x'') \leq \lambda f(x') + (1 - \lambda) f(x'')$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *concave* if $-f$ is convex.

The *epigraph* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\text{epi}(f) := \{(x, y) \in \mathbb{R}^n \times \mathbb{R} \mid y \geq f(x)\}$$

Theorem A.23 (Convexity implies continuity)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. Then f is continuous in the interior of it domain.*

The following property is illustrated in Figure A.7.

Theorem A.24 (Differentiability and convexity)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then f is convex if and only if*

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle \text{ for all } x, y \in \mathbb{R}^n \tag{A.9}$$

Proof. $\Rightarrow$ Suppose f is convex. Then for any $x, y \in \mathbb{R}^n$, and $\lambda \in [0, 1]$

$$f(x + \lambda(y - x)) \leq (1 - \lambda) f(x) + \lambda f(y) \tag{A.10}$$

Rearranging (A.10) we get

$$\frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \leq f(y) - f(x) \text{ for all } \lambda \in [0, 1]$$

Taking the limit as $\lambda \to 0$ we get (A.9).

$\Leftarrow$ Suppose (A.9) holds. Let x and y be arbitrary points in $\mathbb{R}^n$ and let λ be an arbitrary point in $[0, 1]$. Let $z = \lambda x + (1 - \lambda) y$. Then

$$\begin{aligned} f(x) &\geq f(z) + f'(z)(x - z), \text{ and} \\ f(y) &\geq f(z) + f'(z)(y - z) \end{aligned}$$

Multiplying the first equation by λ and the second by $(1 - \lambda)$, adding the resultant equations, and using the fact that $z = \lambda x + (1 - \lambda) y$ yields

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(z) = f(\lambda x + (1 - \lambda) y)$$

Since x and y in $\mathbb{R}^n$ and λ in $[0, 1]$ are all arbitrary, the convexity of $f(\cdot)$ is established. ∎

Theorem A.25 (Second derivative and convexity)**.** *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then f is convex if and only if the Hessian (second derivative) matrix $\partial^2 f(x)/\partial x^2$ is positive semidefinite for all $x \in \mathbb{R}^n$, i.e., $\langle y, \partial^2 f(x)/\partial x^2 y \rangle \geq 0$ for all $x, y \in \mathbb{R}^n$.*

Proof. ⇒ Suppose f is convex. Then for any $x, y \in \mathbb{R}^n$, because of Theorem A.24 and Proposition A.11

$$\begin{aligned} 0 &\le f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= \int_0^1 (1-s) \left\langle y - x, \frac{\partial^2 f(x + s(y-x))}{\partial x^2}(y-x) \right\rangle ds \end{aligned} \tag{A.11}$$

Hence, dividing by $|y-x|^2$ and letting $y \to x$, we obtain that $\partial^2 f(x)/\partial x^2$ is positive semidefinite.

⇐ Suppose that $\partial^2 f(x)/\partial x^2$ is positive semidefinite for all $x \in \mathbb{R}$. Then it follows directly from the equality in (A.11) and Theorem A.24 that f is convex. ■

Definition A.26 (Level set)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$. A *level set* of f is a set of the form $\{x \mid f(x) = \alpha\}, \alpha \in \mathbb{R}$.

Definition A.27 (Sublevel set)**.** Suppose $f : \mathbb{R}^n \to \mathbb{R}$. A *sublevel set* $\mathbb{X}$ of f is a set of the form $\mathbb{X} = \{x \mid f(x) \le \alpha\}, \alpha \in \mathbb{R}$. We also write the sublevel set as $\mathbb{X} = \text{lev}_\alpha f$.

Definition A.28 (Support function)**.** Suppose $Q \subset \mathbb{R}^n$. The support function $\sigma_Q : \mathbb{R}^n \to \mathbb{R}_e = \mathbb{R} \cup \{+\infty\}$ is defined by:

$$\sigma_Q(p) = \sup_x \{\langle p, x \rangle \mid x \in Q\}$$

$\sigma_Q(p)$ measures how far Q extends in direction p.

Proposition A.29 (Set membership and support function)**.** *Suppose $Q \subset \mathbb{R}^n$ is a closed and convex set. Then $x \in Q$ if and only if $\sigma_Q(p) \ge \langle p, x \rangle$ for all $p \in \mathbb{R}^n$*

Proposition A.30 (Lipschitz continuity of support function)**.** *Suppose $Q \subset \mathbb{R}^n$ is bounded. Then σ_Q is bounded and Lipschitz continuous $|\sigma_Q(p) - \sigma_Q(q)| \le K |p - q|$ for all $p, q \in \mathbb{R}^n$, where $K := \sup\{|x| \mid x \in Q\} < \infty$.*

A.14 Differential Equations

Although difference equation models are employed extensively in this book, the systems being controlled are most often described by differential equations. Thus, if the system being controlled is described by

the differential equation $\dot{x} = f_c(x,u)$, as is often the case, and if it is decided to control the system using piecewise constant control with period Δ, then, at sampling instants $k\Delta$ where $k \in \mathbb{I}$, the system is described by the difference equation

$$x^+ = f(x,u)$$

then $f(\cdot)$ may be derived from $f_c(\cdot)$ as follows

$$f(x,u) = x + \int_0^\Delta f_c(\phi_c(s;x,u),u)ds$$

where $\phi_c(s;x,u)$ is the solution of $\dot{x} = f_c(x,u)$ at time s if its initial state at time 0 is x and the control has a constant value u in the interval $[0,\Delta]$. Thus x in the difference equation is the state at time k, say, u is the control in the interval $[0,\Delta]$, and x^+ is the state at time $k+1$.

Because the discrete time system is most often obtained by a continuous time system, we must be concerned with conditions which guarantee the existence and uniqueness of solutions of the differential equation describing the continuous time system. For excellent expositions of the theory of ordinary differential equations see the books by Hale (1980), McShane (1944), Hartman (1964), and Coddington and Levinson (1955).

Consider, first, the unforced system described by

$$(d/dt)x(t) = f(x(t),t) \text{ or } \dot{x} = f(x,t) \tag{A.12}$$

with initial condition

$$x(t_0) = x_0 \tag{A.13}$$

Suppose $f : D \to \mathbb{R}^n$, where D is an open set in $\mathbb{R}^n \times \mathbb{R}$, is continuous. A function $x : T \to \mathbb{R}^n$, where T is an interval in $\mathbb{R}$, is said to be a (conventional) solution of (A.12) with initial condition (A.13) (or passing through (x_0,t_0)) if:

(a) x is continuously differentiable and x satisfies (A.12) on T,

(b) $x(t_0) = x_0$,

and $(x(t),t) \in D$ for all t in T. It is easily shown, when f is continuous, that x satisfies (A.12) and (A.13) if and only if:

$$x(t) = x_0 + \int_{t_0}^{t} f(x(s),s)ds \tag{A.14}$$

Peano's existence theorem states that if f is continuous on D, then, for all $(x_0, t_0) \in D$ there exists at least one solution of (A.12)) passing through (x_0, t_0). The solution is not necessarily unique - a counter example being $\dot{x} = \sqrt{x}$ for $x \geq 0$. To proceed we need to be able to deal with systems for which $f(\cdot)$ is not necessarily continuous for the following reason. If the system is described by $\dot{x} = f(x, u, t)$ where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous, and the control $u : \mathbb{R} \to \mathbb{R}^m$ is continuous, then, for given $u(\cdot)$, the function $f^u : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ defined by:

$$f^u(x, t) := f(x, u(t), t)$$

is continuous in t. We often encounter controls that are not continuous, however, in which case $f^u(\cdot)$ is also not continuous. We need a richer class of controls. A suitable class is the class of *measurable* functions which, for the purpose of this book, we may take to be a class rich enough to include all controls, such as those that are merely piecewise continuous, that we may encounter. If the control $u(\cdot)$ is measurable and $f(\cdot)$ is continuous, then $f^u(\cdot)$, defined above, is continuous in x but measurable in t, so we are forced to study such functions. Suppose, as above, D is an open set in $\mathbb{R}^n \times \mathbb{R}$. The function $f : D \to \mathbb{R}^n$ is said to satisfy the *Caratheodory* conditions in D if:

(a) f is measurable in t for each fixed x,

(b) f is continuous in x for each fixed t,

(c) for each compact set F in D there exists a measurable function $t \mapsto m_F(t)$ such that

$$|f(x, t)| \leq m_F(t)$$

for all $(x, t) \in F$. We now make use of the fact that if $t \mapsto h(t)$ is measurable, its integral $t \mapsto H(t) \stackrel{\Delta}{=} \int_{t_0}^{t} h(s)ds$ is absolutely continuous and, therefore, has a derivative almost everywhere. Where $H(\cdot)$ is differentiable, its derivative is equal to $h(\cdot)$. Consequently, if $f(\cdot)$ satisfies the Caratheodory conditions, then the solution of (A.14), i.e., a function $\phi(\cdot)$ satisfying (A.14) everywhere does not satisfy (A.12) everywhere but only almost everywhere, at the points where $\phi(\cdot)$ is differentiable. In view of this, we may speak *either* of a solution of (A.14) *or* of a solution of (A.12) provided we interpret the latter as an absolutely continuous function which satisfies (A.12)) almost everywhere. The appropriate generalization of Peano's existence theorem is the following result due to Caratheodory:

Theorem A.31 (Existence of solution to differential equations)**.** *If D is an open set in $\mathbb{R}^n \times \mathbb{R}$ and $f(\cdot)$ satisfies the Caratheodory conditions on D, then, for any (x_0, t_0) in D, there exists a solution of* (A.14) *or* (A.12) *passing through (x_0, t_0).*

Two other classical theorems on ordinary differential equations that are relevant are:

Theorem A.32 (Maximal interval of existence)**.** *If D is an open set in $\mathbb{R}^n \times \mathbb{R}$, $f(\cdot)$ satisfies the Caratheodory conditions on D, and $\phi(\cdot)$ is a solution of* (A.10) *on some interval, then there is a continuation $\phi'(\cdot)$ of $\phi(\cdot)$ to a maximal interval (t_a, t_b) of existence. The solution $\phi'(\cdot)$, the continuation of $\phi(\cdot)$, tends to the boundary of D as $t \searrow t_a$ and $t \nearrow t_b$.*

Theorem A.33 (Continuity of solution to differential equation)**.** *Suppose D is an open set in $\mathbb{R}^n \times \mathbb{R}$, f satisfies the Caratheodory condition and, for each compact set U in D, there exists an integrable function $t \mapsto k_u(t)$ such that*

$$|f(x,t) - f(y,t)| \leq k_u(t)\,|x - y|$$

for all $(x,t), (y,t)$ in U. Then, for any (x_0, t_0) in U there exists a unique solution $\phi(\cdot; x_0, t_0)$ passing through (x_0, t_0). The function $(t, x_0, t_0) \mapsto \phi(t; x_0, t_0) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is continuous in its domain E which is open.

Note that D is often $\mathbb{R}^n \times \mathbb{R}$, in which case Theorem A.32 states that a solution $x(\cdot)$ of (A.14) escapes, i.e., $|x(t)| \to \infty$ as $t \searrow t_a$ or $t \nearrow t_b$ if t_a and t_b are finite; t_a and t_b are the escape times. An example of a differential equation with finite escape time is $\dot{x} = x^2$ which has, if $x_0 > 0, t_0 = 0$, a solution $x(t) = x_0[1 - (t - t_0)x_0]^{-1}$ and the maximal interval of existence is $(t_a, t_b) = (-\infty, t_0 + 1/x_0)$.

These results, apart from absence of a control u which is trivially corrected, do not go far enough. We require solutions on an interval $[t_0, t_f]$ given a priori. Further assumptions are needed for this. A useful tool in developing the required results is the Bellman-Gronwall lemma:

Theorem A.34 (Bellman-Gronwall)**.** *Suppose that $c \in (0, \infty)$ and that $\alpha : [0,1] \to \mathbb{R}_+$ is a bounded, integrable function, and that the integrable function $y : [0,1] \to \mathbb{R}$ satisfies the inequality*

$$y(t) \leq c + \int_0^t \alpha(s) y(s) ds \tag{A.15}$$

for all $t \in [0,1]$. *Then*

$$y(t) \leq ce^{\int_0^t \alpha(s)ds} \tag{A.16}$$

for all $t \in [0,1]$.

Note that, if the inequality in (A.15) were replaced by an equality, (A.15) could be integrated to yield (A.16).

Proof. Let the function $Y : [0,1] \to \mathbb{R}$ be defined by

$$Y(t) = \int_0^t \alpha(s)y(s)ds \tag{A.17}$$

so that $\dot{Y}(t) = \alpha(t)y(t)$ almost everywhere on $[0,1]$. It follows from (A.15) and (A.17) that:

$$y(t) \leq c + Y(t) \ \forall t \in [0,1]$$

Hence

$$\begin{aligned}(d/dt)[e^{-\int_0^t \alpha(s)ds}Y(t)] &= e^{-\int_0^t \alpha(s)ds}(\dot{Y}(t) - \alpha(t)Y(t)) \\ &= (e^{-\int_0^t \alpha(s)ds})\alpha(t)(y(t) - Y(t)) \\ &\leq c(e^{-\int_0^t \alpha(s)ds})\alpha(t) \end{aligned} \tag{A.18}$$

almost everywhere on $[0,1]$. Integrating both sides of (A.18) from 0 to t yields

$$e^{-\int_0^t \alpha(s)ds}Y(t) \leq c[1 - e^{-\int_0^t \alpha(s)ds}]$$

for all $t \in [0,1]$. Hence

$$Y(t) \leq c[e^{\int_0^t \alpha(s)ds} - 1]$$

and

$$y(t) \leq ce^{\int_0^t \alpha(s)ds}$$

for all $t \in [0,1]$. ∎

The interval $[0,1]$ may, of course, be replaced by $[t_0, t_f]$ for arbitrary $t_0, t_f \in (-\infty, \infty)$. Consider now the forced system described by

$$\dot{x}(t) = f(x(t), u(t), t) \text{ a.e} \tag{A.19}$$

with initial condition

$$x(0) = 0$$

The period of interest is now $T := [0,1]$ and "a.e." denotes "almost everywhere on T." Admissible controls $u(\cdot)$ are measurable and satisfy the control constraint

$$u(t) \in \Omega \text{ a.e.}$$

where $\Omega \subset \mathbb{R}^m$ is compact. For convenience, we denote the set of admissible controls by

$$\mathcal{U} := \{u : T \to \mathbb{R}^m \mid u(\cdot) \text{ is measurable}, u(t) \in \Omega \text{ a.e.}\}$$

Clearly $\mathcal{U}$ is a subset of L_∞. For simplicity we assume, in the sequel, that f is continuous; this is not restrictive. For each u in $\mathcal{U}$, x in $\mathbb{R}^n$, the function $t \mapsto f^u(x,t) := f(x,u(t),t)$ is measurable so that f^u satisfies the Caratheodory conditions and our previous results, Theorems A.31–A.33, apply. Our concern now is to show that, with additional assumptions, for each u in $\mathcal{U}$, a solution to (A.12) or (A.13) exists on T, rather than on some maximal interval that may be a subset of T, and that this solution is unique and bounded.

Theorem A.35 (Existence of solutions to forced systems)**.** *Suppose:*

(a) f is continuous and

(b) there exists a positive constant c such that

$$|f(x',u,t) - f(x,u,t)| \le c\,|x' - x|$$

for all $(x,u,t) \in \mathbb{R}^n \times \Omega \times T$. Then, for each u in $\mathcal{U}$, there exists a unique, absolutely continuous solution $x^u : T \to \mathbb{R}^n$ of (A.19) *on the interval T passing through $(x_0,0)$. Moreover, there exists a constant K such that*

$$|x^u(t)| \le K$$

for all $t \in T$, all $u \in \mathcal{U}$.

Proof. A direct consequence of (b) is the existence of a constant which, without loss of generality, we take to be c, satisfying

(c) $|f(x,u,t)| \le c(1 + |x|)$ for all $(x,u,t) \in \mathbb{R}^n \times \Omega \times T$.

Assumptions (a) and (b) and their corollary (c), a growth condition on $f(\cdot)$, ensure that $f^u(\cdot)$ satisfies the Caratheodory conditions stated earlier. Hence, our previous results apply, and there exists an interval $[0,t_b]$ on which a unique solution $x^u(\cdot)$ exists; moreover $|x^u(t)| \to \infty$ as $t \nearrow t_b$. Since $x^u(\cdot)$ satisfies

$$x^u(t) = x_0 + \int_0^t f(x^u(s),u(s),s)\,ds$$

it follows from the growth condition that

$$\begin{aligned} |x^u(t)| &\le |x_0| + \int_0^t |f(x^u(s), u(s), s)|\, ds \\ &\le |x_0| + c\int_0^t (1 + |x^u(s)|) ds \\ &\le (|x_0| + c) + c\int_0^t |x^u(s)|\, ds \end{aligned}$$

Applying the Bellman-Gronwall lemma yields

$$|x^u(t)| \le (c + |x_0|)e^{ct}$$

for all $t \in [0, t_b), u \in \mathcal{U}$. If follows that the escape time t_b cannot be finite, so that, for all u in $\mathcal{U}$, there exists a unique absolutely continuous solution $x^u(\cdot)$ on T passing through $(x_0, (0))$. Moreover, for all u in $\mathcal{U}$, all $t \in T$

$$|x^u(t)| \le K$$

where $K := (c + |x_0|)e^c$. ■

A.15 Random Variables and the Probability Density

Let ξ be a random variable taking values in the field of real numbers and the function $F_\xi(x)$ denote the **probability distribution function** of the random variable so that

$$F_\xi(x) = \Pr(\xi \le x)$$

i.e., $F_\xi(x)$ is the probability that the random variable ξ takes on a value less than or equal to x. F_ξ is obviously a nonnegative, nondecreasing function and has the following properties due to the axioms of probability

$$F_\xi(x_1) \le F_\xi(x_2) \quad \text{if } x_1 < x_2$$

$$\lim_{x\to-\infty} F_\xi(x) = 0$$
$$\lim_{x\to\infty} F_\xi(x) = 1$$

We next define the **probability density function**, denoted $p_\xi(x)$, such that

$$F_\xi(x) = \int_{-\infty}^{x} p_\xi(s) ds, \qquad -\infty < x < \infty \tag{A.20}$$

We can allow discontinuous F_ξ if we are willing to accept generalized functions (delta functions and the like) for p_ξ. Also, we can define the density function for discrete as well as continuous random variables if we allow delta functions. Alternatively, we can replace the integral in (A.20) with a sum over a discrete density function. The random variable may be a coin toss or a dice game, which takes on values from a discrete set contrasted to a temperature or concentration measurement, which takes on a values from a continuous set. The density function has the following properties

$$p_\xi(x) \geq 0$$

$$\int_{-\infty}^{\infty} p_\xi(x)dx = 1$$

and the interpretation in terms of probability

$$\Pr(x_1 \leq \xi \leq x_2) = \int_{x_1}^{x_2} p_\xi(x)dx$$

The **mean** or **expectation** of a random variable ξ is defined as

$$\mathcal{E}(\xi) = \int_{-\infty}^{\infty} x p_\xi(x)dx$$

The moments of a random variable are defined by

$$\mathcal{E}(\xi^n) = \int_{-\infty}^{\infty} x^n p_\xi(x)dx$$

and it is clear that the mean is the zeroth moment. Moments of ξ about the mean are defined by

$$\mathcal{E}((\xi - \mathcal{E}(\xi))^n) = \int_{-\infty}^{\infty} (x - \mathcal{E}(\xi))^n p_\xi(x)dx$$

and the variance is defined as the second moment about the mean

$$\text{var}(\xi) = \mathcal{E}((\xi - \mathcal{E}(\xi))^2) = \mathcal{E}(\xi^2) - \mathcal{E}^2(\xi)$$

The standard deviation is the square root of the variance

$$\sigma(\xi) = (\text{var}(\xi))^{1/2}$$

Normal distribution. The normal or Gaussian distribution is ubiquitous in applications. It is characterized by its mean, m and variance, σ^2, and is given by

$$p_\xi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right) \tag{A.21}$$

We proceed to check that the mean of this distribution is indeed m and the variance is σ^2 as claimed and that the density is normalized so that its integral is one. We require the definite integral formulas

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$
$$\int_0^{\infty} x^{1/2} e^{-x} dx = \Gamma(3/2) = \frac{\sqrt{\pi}}{2}$$

The first formula may also be familiar from the error function in transport phenomena

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$$
$$\mathrm{erf}(\infty) = 1$$

We calculate the integral of the normal density as follows

$$\int_{-\infty}^{\infty} p_\xi(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right) dx$$

Define the change of variable

$$u = \frac{1}{\sqrt{2}}\left(\frac{x-m}{\sigma}\right)$$

which gives

$$\int_{-\infty}^{\infty} p_\xi(x) dx = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \exp\left(-u^2\right) du = 1$$

and (A.21) does have unit area. Computing the mean gives

$$\mathcal{E}(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right) dx$$

using the same change of variables as before yields

$$\mathcal{E}(\xi) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} (\sqrt{2}u\sigma + m) e^{-u^2} du$$

The first term in the integral is zero because u is an odd function, and the second term produces

$$\mathcal{E}(\xi) = m$$

as claimed. Finally the definition of the variance of ξ gives

$$\text{var}(\xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (x-m)^2 \exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right) dx$$

Changing the variable of integration as before gives

$$\text{var}(\xi) = \frac{2}{\sqrt{\pi}}\sigma^2 \int_{-\infty}^{\infty} u^2 e^{-u^2} du$$

and because the integrand is an even function,

$$\text{var}(\xi) = \frac{4}{\sqrt{\pi}}\sigma^2 \int_{0}^{\infty} u^2 e^{-u^2} du$$

Now changing the variable of integration again using $s = u^2$ gives

$$\text{var}(\xi) = \frac{2}{\sqrt{\pi}\sigma^2} \int_{0}^{\infty} s^{1/2} e^{-s} ds$$

The second integral formula then gives

$$\text{var}(\xi) = \sigma^2$$

Shorthand notation for the random variable ξ having a normal distribution with mean m and variance σ^2 is

$$\xi \sim N(m, \sigma^2)$$

Figure A.8 shows the normal distribution with a mean of one and variances of 1/2, 1 and 2. Notice that a large variance implies that the random variable is likely to take on large values. As the variance shrinks to zero, the probability density becomes a delta function and the random variable approaches a deterministic value.

Central limit theorem.

> The central limit theorem states that if a set of n random variables $x_i, i = 1, 2, \ldots, n$ are independent, then under general conditions the density p_y of their sum
>
> $$y = x_1 + x_2 + \cdots + x_n$$
>
> properly normalized, tends to a normal density as $n \to \infty$. (Papoulis, 1984, p. 194).

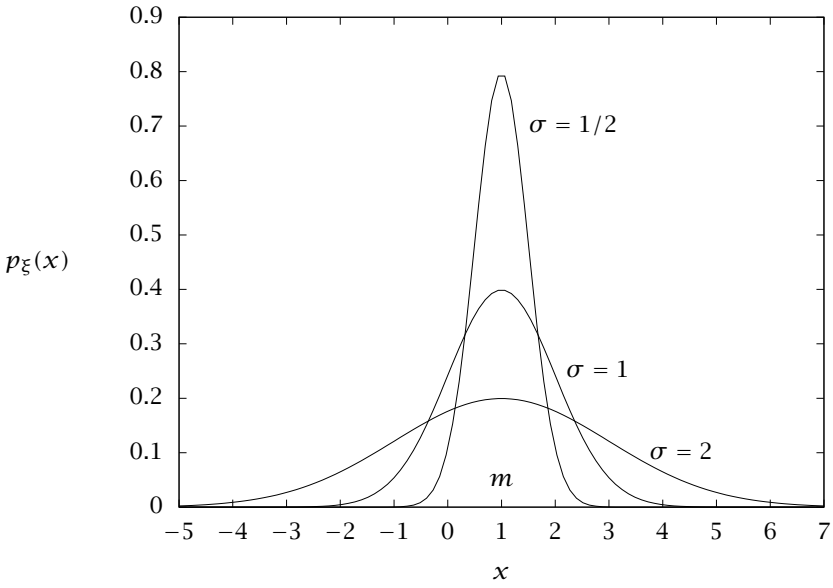


Figure A.8: Normal distribution, $p_\xi(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{1}{2}\dfrac{(x-m)^2}{\sigma^2}\right)$. Mean is one and standard deviations are 1/2, 1 and 2.

Notice that we require only mild restrictions on how the x_i themselves are distributed for the sum y to tend to a normal. See Papoulis (1984, p. 198) for one set of sufficient conditions and a proof of this theorem.

Fourier transform of the density function. It is often convenient to handle the algebra of density functions, particularly normal densities, by using the Fourier transform of the density function rather than the density itself. The transform, which we denote as $\varphi_\xi(u)$, is often called the characteristic function or generating function in the statistics literature. From the definition of the Fourier transform

$$\varphi_\xi(u) = \int_{-\infty}^{\infty} e^{iux} p_\xi(x) dx$$

The transform has a one-to-one correspondence with the density function, which can be seen from the inverse transform formula

$$p_\xi(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iux} \varphi_\xi(u) du$$

Example A.36: Fourier transform of the normal density.

Show the Fourier transform of the normal density is $\varphi_\xi(u) = \exp\left(ium - \frac{1}{2}u^2\sigma^2\right)$. □

A.16 Multivariate Density Functions

In applications we normally do not have a single random variable but a collection of random variables. We group these variables together in a vector and let random variable ξ now take on values in $\mathbb{R}^n$. The probability density function is still a nonnegative scalar function

$$p_\xi(x) : \mathbb{R}^n \to \mathbb{R}^+$$

which is sometimes called the **joint density function**. As in the scalar case, the probability that the n-dimensional random variable ξ takes on values between a and b is given by

$$\Pr(a \le \xi \le b) = \int_{a_n}^{b_n} \cdots \int_{a_1}^{b_1} p_\xi(x) dx_1 \cdots dx_n$$

Marginal density functions. We are often interested in only some subset of the random variables in a problem. Consider two vectors of random variables, $\xi \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$. We can consider the joint distribution of both of these random variables $p_{\xi,\eta}(x, y)$ or we may only be interested in the ξ variables, in which case we can integrate out the m η variables to obtain the marginal density of ξ

$$p_\xi(x) = \int_{-\infty}^{\infty} \cdots \int p_{\xi,\eta}(x, y) dy_1 \cdots dy_m$$

Analogously to produce the marginal density of η we use

$$p_\eta(y) = \int_{-\infty}^{\infty} \cdots \int p_{\xi,\eta}(x, y) dx_1 \cdots dx_n$$

Multivariate normal density. We define the multivariate normal density of the random variable $\xi \in \mathbb{R}^n$ as

$$p_\xi(x) = \frac{1}{(2\pi)^{n/2}(\det P)^{1/2}} \exp\left[-\frac{1}{2}(x - m)' P^{-1}(x - m)\right] \quad \text{(A.22)}$$

in which $m \in \mathbb{R}^n$ is the mean and $P \in \mathbb{R}^{n \times n}$ is the covariance matrix. The notation $\det P$ denotes determinant of P. As noted before, P is a

$$p(x) = \exp\left(-1/2\left(3.5x_1^2 + 2(2.5)x_1x_2 + 4.0x_2^2\right)\right)$$

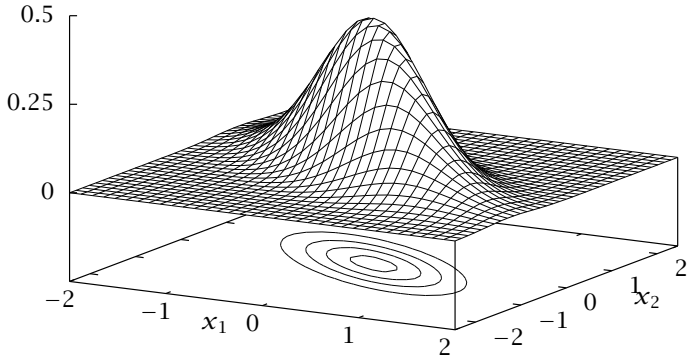


Figure A.9: Multivariate normal in two dimensions.

real, symmetric matrix. The multivariate normal density is well-defined only for $P > 0$. The singular, or degenerate, case $P \geq 0$ is discussed subsequently. Shorthand notation for the random variable ξ having a normal distribution with mean m and covariance P is

$$\xi \sim N(m, P)$$

The matrix P is a real, symmetric matrix. Figure A.9 displays a multivariate normal for

$$P^{-1} = \begin{bmatrix} 3.5 & 2.5 \\ 2.5 & 4.0 \end{bmatrix} \qquad m = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

As displayed in Figure A.9, lines of constant probability in the multivariate normal are lines of constant

$$(x - m)' P^{-1} (x - m)$$

To understand the geometry of lines of constant probability (ellipses in two dimensions, ellipsoids or hyperellipsoids in three or more dimensions) we examine the eigenvalues and eigenvectors of the P^{-1} matrix.

$$x'Ax = b$$
$$Av_i = \lambda_i v_i$$

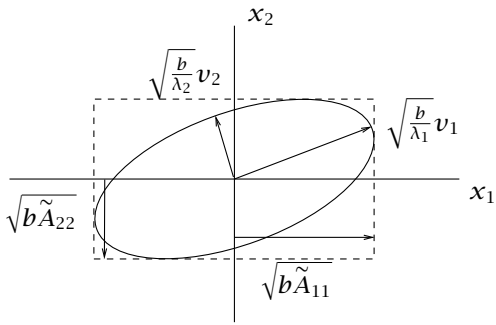


Figure A.10: The geometry of quadratic form $x'Ax = b$.

Consider the quadratic function $x'Ax$ depicted in Figure A.10. Each eigenvector of A points along one of the axes of the ellipse $x'Ax = b$. The eigenvalues show us how stretched the ellipse is in each eigenvector direction. If we want to put simple bounds on the ellipse, then we draw a box around it as shown in Figure A.10. Notice the box contains much more area than the corresponding ellipse and we have lost the correlation between the elements of x. This loss of information means we can put different tangent ellipses of quite different shapes inside the same box. The size of the bounding box is given by

$$\text{length of } i\text{th side} = \sqrt{b\widetilde{A}_{ii}}$$

in which

$$\widetilde{A}_{ii} = (i,i) \text{ element of } A^{-1}$$

See Exercise A.45 for a derivation of the size of the bounding box. Figure A.10 displays these results: the eigenvectors are aligned with the ellipse axes and the eigenvalues scale the lengths. The lengths of the sides of the box that are tangent to the ellipse are proportional to the square root of the diagonal elements of A^{-1}.

Singular or degenerate normal distributions. It is often convenient to extend the definition of the normal distribution to admit positive *semidefinite* covariance matrices. The distribution with a semidefinite covariance is known as a singular or degnerate normal distribution (An-

derson, 2003, p. 30). Figure A.11 shows a nearly singular normal distribution.

To see how the singular normal arises, let the scalar random variable ξ be distributed normally with zero mean and positive definite covariance, $\xi \sim N(0, P_x)$, and consider the simple linear transformation

$$\eta = A\xi \qquad A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

in which we have created two identical copies of ξ for the two components η_1 and η_2 of η. Now consider the density of η. If we try to use the standard formulas for transformation of a normal, we would have

$$\eta \sim N(0, P_y) \qquad P_y = AP_xA' = \begin{bmatrix} P_x & P_x \\ P_x & P_x \end{bmatrix}$$

and P_y is singular since its rows are linearly dependent. Therefore one of the eigenvalues of P_y is zero and P_y is positive semidefinite and not positive definite. Obviously we cannot use (A.22) for the density in this case because the inverse of P_y does not exist. To handle these cases, we first provide an interpretation that remains valid when the covariance matrix is singular and semidefinite.

Definition A.37 (Density of a singular normal)**.** A singular joint normal density of random variables (ξ_1, ξ_2), $\xi_1 \in \mathbb{R}^{n_1}$, $\xi_2 \in \mathbb{R}^{n_2}$, is denoted

$$\begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \sim N\left[\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \right]$$

with $\Lambda_1 > 0$. The density is defined by

$$p_\xi(x_1, x_2) = \frac{1}{(2\pi)^{\frac{n_1}{2}} (\det \Lambda_1)^{\frac{1}{2}}} \exp\left[-\frac{1}{2} |x_1 - m_1)|^2_{\Lambda_1^{-1}} \right] \delta(x_2 - m_2) \tag{A.23}$$

In this limit, the "random" variable ξ_2 becomes deterministic and equal to its mean m_2. For the case $n_1 = 0$, we have the completely degenerate case in which $p_{\xi_2}(x_2) = \delta(x_2 - m_2)$, which describes the completely deterministic case $\xi_2 = m_2$ and there is no random component ξ_1. This expanded definition enables us to generalize the important result that the linear transformation of a normal is normal, so that it holds for *any* linear transformation, including rank deficient transformations such as the A matrix given above in which the rows

are not independent (see Exercise 1.40). Starting with the definition of a singular normal, we can obtain the density for $\xi \sim N(m_x, P_x)$ for any positive semidefinite $P_x \geq 0$. The result is

$$p_\xi(x) = \frac{1}{(2\pi)^{\frac{r}{2}}(\det \Lambda_1)^{\frac{1}{2}}} \exp\left[-\frac{1}{2} |(x - m_x)|^2_{Q_1} \right] \delta(Q_2'(x - m_x)) \tag{A.24}$$

in which matrices $\Lambda \in \mathbb{R}^{r \times r}$ and orthonormal $Q \in \mathbb{R}^{n \times n}$ are obtained from the eigenvalue decomposition of P_x

$$P_x = Q\Lambda Q' = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix}$$

and $\Lambda_1 > 0 \in \mathbb{R}^{r \times r}$, $Q_1 \in \mathbb{R}^{n \times r}$, $Q_2 \in \mathbb{R}^{n \times (n-r)}$. This density is nonzero for x satisfying $Q_2'(x - m_x) = 0$. If we let $N(Q_2')$ denote the r dimensional nullspace of Q_2', we have that the density is nonzero for $x \in N(Q_2') \oplus \{m_x\}$ in which $\oplus$ denotes set addition.

Example A.38: Marginal normal density

Given that ξ and η are jointly, normally distributed with mean

$$m = \begin{bmatrix} m_x \\ m_y \end{bmatrix}$$

and covariance matrix

$$P = \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}$$

show that the marginal density of ξ is normal with the following parameters

$$\xi \sim N(m_x, P_x) \tag{A.25}$$

Solution

As a first approach to establish (A.25), we directly integrate the y variables. Let $\bar{x} = x - m_x$ and $\bar{y} = y - m_y$, and n_x and n_y be the dimension of the ξ and η variables, respectively, and $n = n_x + n_y$. Then the definition of the marginal density gives

$$p_\xi(x) = \frac{1}{(2\pi)^{n/2}(\det P)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2} \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix}' \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}^{-1} \begin{bmatrix} \bar{x} \\ \bar{y} \end{bmatrix} \right] d\bar{y}$$

Let the inverse of P be denoted as $\widetilde{P}$ and partition $\widetilde{P}$ as follows

$$\begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}^{-1} = \begin{bmatrix} \widetilde{P}_x & \widetilde{P}_{xy} \\ \widetilde{P}_{yx} & \widetilde{P}_y \end{bmatrix} \tag{A.26}$$

Substituting (A.26) into the definition of the marginal density and expanding the quadratic form in the exponential yields

$$(2\pi)^{n/2}(\det P)^{1/2} p_\xi(x) = \exp\left(-(1/2)\bar{x}'\widetilde{P}_x\bar{x}\right)\int_{-\infty}^{\infty}\exp\left(-(1/2)(2\bar{y}'\widetilde{P}_{yx}\bar{x} + \bar{y}'\widetilde{P}_y\bar{y})\right)d\bar{y}$$

We complete the square on the term in the integral by noting that

$$(\bar{y}+\widetilde{P}_y^{-1}\widetilde{P}_{yx}\bar{x})'\widetilde{P}_y(\bar{y}+\widetilde{P}_y^{-1}\widetilde{P}_{yx}\bar{x}) = \bar{y}'\widetilde{P}_y\bar{y}+2\bar{y}'\widetilde{P}_{yx}\bar{x}+\bar{x}'\widetilde{P}_{yx}'\widetilde{P}_y^{-1}\widetilde{P}_{yx}\bar{x}$$

Substituting this relation into the previous equation gives

$$(2\pi)^{n/2}(\det P)^{1/2} p_\xi(x) = \exp\left(-(1/2)\bar{x}'(\widetilde{P}_x - \widetilde{P}_{yx}'\widetilde{P}_y^{-1}\widetilde{P}_{yx})\bar{x}\right)\int_{-\infty}^{\infty}\exp\left(-(1/2)(\bar{y}+a)'\widetilde{P}_y(\bar{y}+a)\right)d\bar{y}$$

in which $a = \widetilde{P}_y^{-1}\widetilde{P}_{yx}\bar{x}$. Using (A.22) to evaluate the integral gives

$$p_\xi(x) = \frac{1}{(2\pi)^{n_x/2}\left(\det(P)\det(\widetilde{P}_y)\right)^{1/2}}\exp\left(-(1/2)\bar{x}'(\widetilde{P}_x - \widetilde{P}_{yx}'\widetilde{P}_y^{-1}\widetilde{P}_{yx})\bar{x}\right)$$

From the matrix inversion formula we conclude

$$\widetilde{P}_x - \widetilde{P}_{xy}'\widetilde{P}_y^{-1}\widetilde{P}_{yx} = P_x^{-1}$$

and

$$\det(P) = \det(P_x)\det(P_y - P_{yx}P_x^{-1}P_{xy}) = \det P_x \det\widetilde{P}_y^{-1} = \frac{\det P_x}{\det\widetilde{P}_y}$$

Substituting these results into the previous equation gives

$$p_\xi(x) = \frac{1}{(2\pi)^{n_x/2}(\det P_x)^{1/2}}\exp\left(-(1/2)\bar{x}'P_x^{-1}\bar{x}\right)$$

Therefore

$$\xi \sim N(m_x, P_x)$$

□

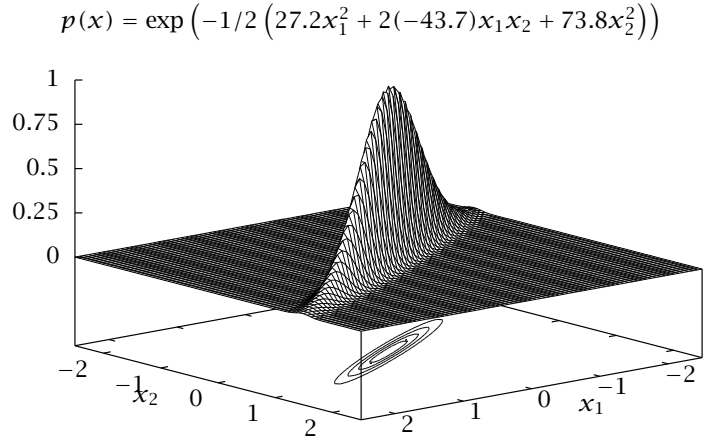


Figure A.11: A nearly singular normal density in two dimensions.

Functions of random variables. In stochastic dynamical systems we need to know how the density of a random variable is related to the density of a function of that random variable. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a mapping of the random variable ξ into the random variable η and assume that the inverse mapping also exits

$$\eta = f(\xi), \qquad \xi = f^{-1}(\eta)$$

Given the density of $\xi, p_\xi(x)$, we wish to compute the density of η, $p_\eta(y)$, induced by the function f. Let S denote an arbitrary region of the field of the random variable ξ and define the set S' as the transform of this set under the function f

$$S' = \{y \mid y = f(x), x \in S\}$$

Then we seek a function $p_\eta(y)$ such that

$$\int_S p_\xi(x)dx = \int_{S'} p_\eta(y)dy \tag{A.27}$$

for every admissible set S. Using the rules of calculus for transforming a variable of integration we can write

$$\int_S p_\xi(x)dx = \int_{S'} p_\xi(f^{-1}(y)) \left| \det\left(\frac{\partial f^{-1}(y)}{\partial y}\right) \right| dy \tag{A.28}$$

in which $\left|\det(\partial f^{-1}(y)/\partial y)\right|$ is the absolute value of the determinant of the Jacobian matrix of the transformation from η to ξ. Subtracting (A.28) from (A.27) gives

$$\int_{S'} \left(p_\eta(y) - p_\xi(f^{-1}(y)) \left|\det(\partial f^{-1}(y)/\partial y)\right| \right) dy = 0 \tag{A.29}$$

Because (A.29) must be true for any set S', we conclude (a proof by contradiction is immediate)[8]

$$\boxed{p_\eta(y) = p_\xi(f^{-1}(y)) \left|\det(\partial f^{-1}(y)/\partial y)\right|} \tag{A.30}$$

Example A.39: Nonlinear transformation

Show that

$$p_\eta(y) = \frac{1}{3\sqrt{2\pi}\sigma y^{2/3}} \exp\left[-\frac{1}{2}\left(\frac{y^{1/3} - m}{\sigma} \right)^2 \right]$$

is the density function of the random variable η under the transformation

$$\eta = \xi^3$$

for $\xi \sim N(m, \sigma^2)$. Notice that the density p_η is singular at $y = 0$. □

Noninvertible transformations. Given n random variables $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ with joint density p_ξ and k random variables $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ defined by the transformation $\eta = f(\xi)$

$$\eta_1 = f_1(\xi) \quad \eta_2 = f_2(\xi) \quad \cdots \quad \eta_k = f_k(\xi)$$

We wish to find p_η in terms of p_ξ. Consider the region generated in $\mathbb{R}^n$ by the vector inequality

$$f(x) \leq c$$

[8] Some care should be exercised if one has generalized functions in mind for the conditional density.

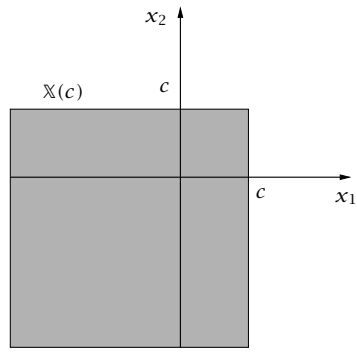


Figure A.12: The region $\mathbb{X}(c)$ for $y = \max(x_1, x_2) \leq c$.

Call this region $\mathbb{X}(c)$, which is by definition

$$\mathbb{X}(c) = \{x \mid f(x) \leq c\}$$

Note $\mathbb{X}$ is not necessarily simply connected. The probability distribution (not density) for η then satisfies

$$P_\eta(y) = \int_{\mathbb{X}(y)} p_\xi(x)dx \tag{A.31}$$

If the density p_η is of interest, it can be obtained by differentiating P_η.

Example A.40: Maximum of two random variables

Given two independent random variables, ξ_1, ξ_2 and the new random variable defined by the noninvertible, nonlinear transformation

$$\eta = \max(\xi_1, \xi_2)$$

Show that η's density is given by

$$p_\eta(y) = p_{\xi_1}(y) \int_{-\infty}^{y} p_{\xi_2}(x)dx + p_{\xi_2}(y) \int_{-\infty}^{y} p_{\xi_1}(x)dx$$

Solution

The region $\mathbb{X}(c)$ generated by the inequality $y = \max(x_1, x_2) \leq c$ is sketched in Figure A.12. Applying (A.31) then gives

$$\begin{aligned} P_\eta(y) &= \int_{-\infty}^{y} \int_{-\infty}^{y} p_\xi(x_1, x_2)dx_1dx_2 \\ &= P_\xi(y, y) \\ &= P_{\xi_1}(y)P_{\xi_2}(y) \end{aligned}$$

which has a clear physical interpretation. It says the probability that the *maximum* of two independent random variables is less than some value is equal to the probability that *both* random variables are less than that value. To obtain the density, we differentiate

$$\begin{aligned} p_\eta(y) &= p_{\xi_1}(y)P_{\xi_2}(y) + P_{\xi_1}(y)p_{\xi_2}(y) \\ &= p_{\xi_1}(y)\int_{-\infty}^{y} p_{\xi_2}(x)dx + p_{\xi_2}(y)\int_{-\infty}^{y} p_{\xi_1}(x)dx \end{aligned}$$

□

A.16.1 Statistical Independence and Correlation

We say two random variables ξ, η are **statistically independent** or simply independent if

$$p_{\xi,\eta}(x,y) = p_\xi(x)p_\eta(y), \qquad \text{all } x, y$$

The covariance of two random variables ξ, η is defined as

$$\text{cov}(\xi,\eta) = \mathcal{E}\left((\xi - \mathcal{E}(\xi))\,(\eta - \mathcal{E}(\eta))\right)$$

The covariance of the vector-valued random variable ξ with components $\xi_i, i = 1,\ldots n$ can be written as

$$P_{ij} = \text{cov}(\xi_i,\xi_j)$$

$$P = \begin{bmatrix} \text{var}(\xi_1) & \text{cov}(\xi_1,\xi_2) & \cdots & \text{cov}(\xi_1,\xi_n) \\ \text{cov}(\xi_2,\xi_1) & \text{var}(\xi_2) & \cdots & \text{cov}(\xi_2,\xi_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\xi_n,\xi_1) & \text{cov}(\xi_n,\xi_2) & \cdots & \text{var}(\xi_n) \end{bmatrix}$$

We say two random variables, ξ and η, are **uncorrelated** if

$$\text{cov}(\xi,\eta) = 0$$

Example A.41: Independent implies uncorrelated

Prove that if ξ and η are statistically independent, then they are uncorrelated.

Solution

The definition of covariance gives

$$\begin{aligned}\operatorname{cov}(\xi,\eta) &= \mathcal{E}((\xi-\mathcal{E}(\xi))(\eta-\mathcal{E}(\eta)))\\ &= \mathcal{E}(\xi\eta-\xi\mathcal{E}(\eta)-\eta\mathcal{E}(\xi)+\mathcal{E}(\xi)\mathcal{E}(\eta))\\ &= \mathcal{E}(\xi\eta)-\mathcal{E}(\xi)\mathcal{E}(\eta)\end{aligned}$$

Taking the expectation of the product $\xi\eta$ and using the fact that ξ and η are independent gives

$$\begin{aligned}\mathcal{E}(\xi\eta) &= \int_{-\infty}^{\infty}\!\!\int xyp_{\xi,\eta}(x,y)dxdy\\ &= \int_{-\infty}^{\infty}\!\!\int xyp_{\xi}(x)p_{\eta}(y)dxdy\\ &= \int_{-\infty}^{\infty} xp_{\xi}(x)dx\int_{-\infty}^{\infty} yp_{\eta}(y)dy\\ &= \mathcal{E}(\xi)\mathcal{E}(\eta)\end{aligned}$$

Substituting this fact into the covariance equation gives

$$\operatorname{cov}(\xi,\eta)=0$$

□

Example A.42: Does uncorrelated imply independent?

Let ξ and η be jointly distributed random variables with probability density function

$$p_{\xi,\eta}(x,y)=\begin{cases}\frac{1}{4}[1+xy(x^2-y^2)], & |x|<1, \quad |y|<1\\ 0, & \text{otherwise}\end{cases}$$

(a) Compute the marginals $p_{\xi}(x)$ and $p_{\eta}(y)$. Are ξ and η independent?

(b) Compute $\operatorname{cov}(\xi,\eta)$. Are ξ and η uncorrelated?

(c) What is the relationship between independent and uncorrelated? Are your results on this example consistent with this relationship? Why or why not?

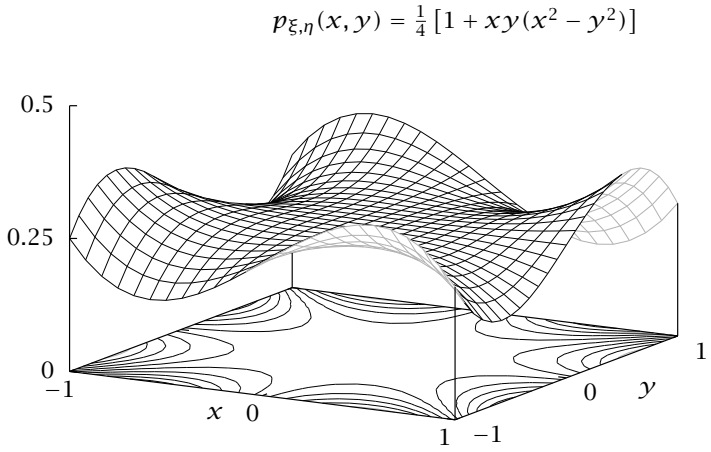


Figure A.13: A joint density function for the two uncorrelated random variables in Example A.42.

Solution

The joint density is shown in Figure A.13.

(a) Direct integration of the joint density produces

$$p_\xi(x) = (1/2), \quad |x| < 1 \qquad \mathcal{E}(\xi) = 0$$
$$p_\eta(y) = (1/2), \quad |y| < 1 \qquad \mathcal{E}(\eta) = 0$$

and we see that both marginals are zero mean, uniform densities. Obviously ξ and η are not independent because the joint density is not the product of the marginals.

(b) Performing the double integral for the expectation of the product term gives

$$\mathcal{E}(\xi\eta) = \int_{-1}^{1}\int xy + (xy)^2(x^2 - y^2)dxdy$$
$$= 0$$

and the covariance of ξ and η is therefore

$$\begin{aligned}\operatorname{cov}(\xi,\eta) &= \mathcal{E}(\xi\eta) - \mathcal{E}(\xi)\mathcal{E}(\eta)\\ &= 0\end{aligned}$$

and ξ and η are uncorrelated.

(c) We know independent implies uncorrelated. This example does not contradict that relationship. This example shows uncorrelated does not imply independent, in general, but see the next example for normals.

□

Example A.43: Independent and uncorrelated are equivalent for normals

If two random variables are jointly normally distributed,

$$\begin{bmatrix}\xi\\ \eta\end{bmatrix} \sim N\left(\begin{bmatrix}m_x\\ m_y\end{bmatrix}, \begin{bmatrix}P_x & P_{xy}\\ P_{yx} & P_y\end{bmatrix}\right)$$

Prove ξ and η are statistically independent if and only if ξ and η are uncorrelated, or, equivalently, P is block diagonal.

Solution

We have already shown that independent implies uncorrelated for any density, so we now show that, *for normals*, uncorrelated implies independent. Given $\operatorname{cov}(\xi,\eta) = 0$, we have

$$P_{xy} = P'_{yx} = 0 \qquad \det P = \det P_x \det P_y$$

so the density can be written

$$p_{\xi,\eta}(x,y) = \frac{\exp\left(-\frac{1}{2}\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}'\begin{bmatrix}P_x & 0\\ 0 & P_y\end{bmatrix}^{-1}\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}\right)}{(2\pi)^{(n_x+n_y)/2}\left(\det P_x \det P_y\right)^{1/2}} \tag{A.32}$$

For any joint normal, we know the marginals are simply

$$\xi \sim N(m_x, P_x) \qquad \eta \sim N(m_y, P_y)$$

so we have

$$p_\xi(x) = \frac{1}{(2\pi)^{n_x/2}(\det P_x)^{1/2}} \exp\left(-(1/2)\bar{x}' P_x^{-1}\bar{x}\right)$$
$$p_\eta(y) = \frac{1}{(2\pi)^{n_y/2}(\det P_y)^{1/2}} \exp\left(-(1/2)\bar{y}' P_y^{-1}\bar{y}\right)$$

Forming the product and combining terms gives

$$p_\xi(x)p_\eta(y) = \frac{\exp\left(-\frac{1}{2}\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}'\begin{bmatrix}P_x^{-1} & 0\\ 0 & P_y^{-1}\end{bmatrix}\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}\right)}{(2\pi)^{(n_x+n_y)/2}\left(\det P_x \det P_y\right)^{1/2}}$$

Comparing this equation to (A.32), and using the inverse of a block-diagonal matrix, we have shown that ξ and η are statistically independent. □

A.17 Conditional Probability and Bayes's Theorem

Let ξ and η be jointly distributed random variables with density $p_{\xi,\eta}(x, y)$. We seek the density function of ξ given a specific realization y of η has been observed. We define the conditional density function as

$$p_{\xi|\eta}(x|y) = \frac{p_{\xi,\eta}(x, y)}{p_\eta(y)}$$

Consider a roll of a single die in which η takes on values E or O to denote whether the outcome is even or odd and ξ is the integer value of the die. The twelve values of the joint density function are simply computed

$$\begin{array}{rclrcl}
p_{\xi,\eta}(1,\mathrm{E}) &=& 0 & p_{\xi,\eta}(1,\mathrm{O}) &=& 1/6\\
p_{\xi,\eta}(2,\mathrm{E}) &=& 1/6 & p_{\xi,\eta}(2,\mathrm{O}) &=& 0\\
p_{\xi,\eta}(3,\mathrm{E}) &=& 0 & p_{\xi,\eta}(3,\mathrm{O}) &=& 1/6\\
p_{\xi,\eta}(4,\mathrm{E}) &=& 1/6 & p_{\xi,\eta}(4,\mathrm{O}) &=& 0\\
p_{\xi,\eta}(5,\mathrm{E}) &=& 0 & p_{\xi,\eta}(5,\mathrm{O}) &=& 1/6\\
p_{\xi,\eta}(6,\mathrm{E}) &=& 1/6 & p_{\xi,\eta}(6,\mathrm{O}) &=& 0
\end{array} \tag{A.33}$$

The marginal densities are then easily computed; we have for ξ

$$p_\xi(x) = \sum_{y=\mathrm{O}}^{\mathrm{E}} p_{\xi,\eta}(x, y)$$

which gives by summing across rows of (A.33)

$$p_\xi(x) = 1/6, \qquad x = 1, 2, \ldots 6$$

Similarly, we have for η

$$p_\eta(y) = \sum_{x=1}^{6} p_{\xi,\eta}(x, y)$$

which gives by summing down the columns of (A.33)

$$p_\eta(y) = 1/2, \qquad y = \mathrm{E}, \mathrm{O}$$

These are both in accordance of our intuition on the rolling of the die: uniform probability for each value 1 to 6 and equal probability for an even or an odd outcome. Now the conditional density is a different concept. The conditional density $p_{\xi|\eta}(x, y)$ tells us the density of x given that $\eta = y$ has been observed. So consider the value of this function

$$p_{\xi|\eta}(1|\mathrm{O})$$

which tells us the probability that the die has a 1 given that we know that it is odd. We expect that the additional information on the die being odd causes us to revise our probability that it is 1 from 1/6 to 1/3. Applying the defining formula for conditional density indeed gives

$$p_{\xi|\eta}(1|\mathrm{O}) = p_{\xi,\eta}(1, \mathrm{O})/p_\eta(\mathrm{O}) = \frac{1/6}{1/2} = 1/3$$

Consider the reverse question, the probability that we have an odd given that we observe a 1. The definition of conditional density gives

$$p_{\eta,\xi}(\mathrm{O}|1) = p_{\eta,\xi}(\mathrm{O}, 1)/p_\xi(1) = \frac{1/6}{1/6} = 1$$

i.e., we are sure the die is odd if it is 1. Notice that the arguments to the conditional density do not commute as they do in the joint density.

This fact leads to a famous result. Consider the definition of conditional density, which can be expressed as

$$p_{\xi,\eta}(x, y) = p_{\xi|\eta}(x|y)p_\eta(y)$$

or

$$p_{\eta,\xi}(y, x) = p_{\eta|\xi}(y|x)p_\xi(x)$$

Because $p_{\xi,\eta}(x,y) = p_{\eta,\xi}(y,x)$, we can equate the right-hand sides and deduce

$$p_{\xi|\eta}(x|y) = \frac{p_{\eta|\xi}(y|x)p_{\xi}(x)}{p_{\eta}(y)}$$

which is known as Bayes's theorem (Bayes, 1763). Notice that this result comes in handy whenever we wish to switch the variable that is known in the conditional density, which we will see is a key step in state estimation problems.

Example A.44: Conditional normal density

Show that if ξ and η are jointly normally distributed as

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} \sim N\left(\begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}\right)$$

then the conditional density of ξ given η is also normal

$$(\xi|\eta) \sim N(m,P)$$

in which the mean is

$$m = m_x + P_{xy}P_y^{-1}(y - m_y) \tag{A.34}$$

and the covariance is

$$P = P_x - P_{xy}P_y^{-1}P_{yx} \tag{A.35}$$

Solution

The definition of conditional density gives

$$p_{\xi|\eta}(x|y) = \frac{p_{\xi,\eta}(x,y)}{p_{\eta}(y)}$$

Because (ξ,η) is jointly normal, we know from Example A.38

$$p_{\eta}(y) = \frac{1}{(2\pi)^{n_\eta/2}(\det P_y)^{1/2}} \exp\left(-(1/2)(y - m_y)'P_y^{-1}(y - m_y)\right)$$

and therefore

$$p_{\xi|\eta}(x|y) = \frac{(\det P_y)^{1/2}}{(2\pi)^{n_\xi/2}\left(\det\begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}\right)^{1/2}} \exp(-1/2a) \tag{A.36}$$

in which the argument of the exponent is

$$a = \begin{bmatrix} x - m_x \\ y - m_y \end{bmatrix}' \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}^{-1} \begin{bmatrix} x - m_x \\ y - m_y \end{bmatrix} - (y - m_y)' P_y^{-1} (y - m_y)$$

If we use $P = P_x - P_{xy} P_y^{-1} P_{yx}$ as defined in (A.35) then we can use the partitioned matrix inversion formula to express the matrix inverse in the previous equation as

$$\begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix}^{-1} = \begin{bmatrix} P^{-1} & -P^{-1} P_{xy} P_y^{-1} \\ -P_y^{-1} P_{yx} P^{-1} & P_y^{-1} + P_y^{-1} P_{yx} P^{-1} P_{xy} P_y^{-1} \end{bmatrix}$$

Substituting this expression and multiplying out terms yields

$$\begin{aligned} a = {} & (x - m_x)' P^{-1} (x - m_x) - 2(y - m_y)' (P_y^{-1} P_{yx} P^{-1}) (x - m_x) \\ & + (y - m_y)' (P_y^{-1} P_{yx} P^{-1} P_{xy} P_y^{-1}) (y - m_y) \end{aligned}$$

which is the expansion of the following quadratic term

$$a = \left[(x - m_x) - P_{xy} P_y^{-1} (y - m_y) \right]' P^{-1} \left[(x - m_x) - P_{xy} P_y^{-1} (y - m_y) \right]$$

in which we use the fact that $P_{xy} = P_{yx}'$. Substituting (A.34) into this expression yields

$$a = (x - m)' P^{-1} (x - m) \tag{A.37}$$

Finally noting that for the partitioned matrix

$$\det \begin{bmatrix} P_x & P_{xy} \\ P_{yx} & P_y \end{bmatrix} = \det P_y \det P \tag{A.38}$$

and substitution of equations (A.38) and (A.37) into (A.36) yields

$$p_{\xi|\eta}(x|y) = \frac{1}{(2\pi)^{n_\xi/2} (\det P)^{1/2}} \exp\left(-\frac{1}{2} (x - m)' P^{-1} (x - m) \right)$$

which is the desired result. □

Example A.45: More normal conditional densities

Let the joint conditional of random variables a and b given c be a normal distribution with

$$p(a, b|c) \sim N\left(\begin{bmatrix} m_a \\ m_b \end{bmatrix}, \begin{bmatrix} P_a & P_{ab} \\ P_{ba} & P_b \end{bmatrix} \right) \tag{A.39}$$

Then the conditional density of a given b and c is also normal

$$p(a|b,c) \sim N(m,P)$$

in which the mean is

$$m = m_a + P_{ab}P_b^{-1}(b - m_b)$$

and the covariance is

$$P = P_a - P_{ab}P_b^{-1}P_{ba}$$

Solution

From the definition of joint density we have

$$p(a|b,c) = \frac{p(a,b,c)}{p(b,c)}$$

Multiplying the top and bottom of the fraction by $p(c)$ yields

$$p(a|b,c) = \frac{p(a,b,c)}{p(c)}\frac{p(c)}{p(b,c)}$$

or

$$p(a|b,c) = \frac{p(a,b|c)}{p(b|c)}$$

Substituting the distribution given in (A.39) and using the result in Example A.38 to evaluate $p(b|c)$ yields

$$p(a|b,c) = \frac{N\left(\begin{bmatrix} m_a \\ m_b \end{bmatrix}, \begin{bmatrix} P_a & P_{ab} \\ P_{ba} & P_b \end{bmatrix}\right)}{N(m_b, P_b)}$$

And now applying the methods of Example A.44 this ratio of normal distributions reduces to the desired expression. □

Adjoint operator. Given a linear operator $\mathcal{G} : \mathbb{U} \to \mathbb{V}$ and inner products for the spaces $\mathbb{U}$ and $\mathbb{V}$, the adjoint of $\mathcal{G}$, denoted by $\mathcal{G}^*$ is the linear operator $\mathcal{G}^* : \mathbb{V} \to \mathbb{U}$ such that

$$\langle u, \mathcal{G}^* v\rangle = \langle \mathcal{G}u, v\rangle, \qquad \forall u \in \mathbb{U}, v \in \mathbb{V} \tag{A.40}$$

Dual dynamic system (Callier and Desoer, 1991). The dynamic system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k), \qquad k = 0, \ldots, N-1 \\ y(k) &= Cx(k) + Du(k) \end{aligned}$$

maps an initial condition and input sequence $(x(0), u(0), \ldots, u(N-1))$ into a final condition and an output sequence $(x(N), y(0), \ldots, y(N-1))$. Call this linear operator $\mathcal{G}$

$$\begin{bmatrix} x(N) \\ y(0) \\ \vdots \\ y(N-1) \end{bmatrix} = \mathcal{G} \begin{bmatrix} x(0) \\ u(0) \\ \vdots \\ u(N-1) \end{bmatrix}$$

The dual dynamic system represents the adjoint operator $\mathcal{G}^*$

$$\begin{bmatrix} \overline{x}(0) \\ \overline{y}(1) \\ \vdots \\ \overline{y}(N) \end{bmatrix} = \mathcal{G}^* \begin{bmatrix} \overline{x}(N) \\ \overline{u}(1) \\ \vdots \\ \overline{u}(N) \end{bmatrix}$$

We define the usual inner product, $\langle a, b \rangle = a'b$, and substitute into (A.40) to obtain

$$\underbrace{x(0)'\overline{x}(0) + u(0)'\overline{y}(1) + \cdots + u(N-1)'\overline{y}(N)}_{\langle u, \mathcal{G}^* v \rangle} - \underbrace{x(N)'\overline{x}(N) + y(0)'\overline{u}(1) + \cdots + y(N-1)'\overline{u}(N)}_{\langle \mathcal{G}u, v \rangle} = 0$$

If we express the $y(k)$ in terms of $x(0)$ and $u(k)$ and collect terms we obtain

$$\begin{aligned} 0 = {} & x(0)' \left[\overline{x}(0) - C'\overline{u}(1) - A'C'\overline{u}(2) - \cdots - A'^N \overline{x}(N) \right] \\ & + u(0)' \left[\overline{y}(1) - D'\overline{u}(1) - B'C'\overline{u}(2) - \cdots - B'A'^{(N-2)}C'\overline{u}(N) - B'A'^{(N-1)}\overline{x}(N) \right] \\ & + \cdots \\ & + u(N-2)' \left[\overline{y}(N-1) - D'\overline{u}(N-1) - B'C'\overline{u}(N) - B'A'\overline{x}(N) \right] \\ & + u(N-1)' \left[\overline{y}(N) - D'\overline{u}(N) - B'\overline{x}(N) \right] \end{aligned}$$

Since this equation must hold for all $(x(0), u(0), \ldots, u(N-1))$, each term in brackets must vanish. From the $u(N-1)$ term we conclude

$$\overline{y}(N) = B'\overline{x}(N) + D'\overline{u}(N)$$

Using this result, the $u(N-2)$ term gives

$$B'\left(\overline{x}(N-1) - \left(A'\overline{x}(N) + C'\overline{u}(N)\right)\right) = 0$$

From which we find the state recursion for the dual system

$$\overline{x}(N-1) = A'\overline{x}(N) + C'\overline{u}(N)$$

Passing through each term then yields the dual state space description of the adjoint operator $\mathcal{G}^*$

$$\begin{aligned}\overline{x}(k-1) &= A'\overline{x}(k) + C'\overline{u}(k), \qquad k = N,\ldots,1\\ \overline{y}(k) &= B'\overline{x}(k) + D'\overline{u}(k)\end{aligned}$$

So the primal and dual dynamic systems change matrices in the following way

$$(A,B,C,D) \longrightarrow (A',C',B',D')$$

Notice this result produces the duality variables listed in Table A.1 if we first note that we have also renamed the regulator's input matrix B to G in the estimation problem. We also note that time runs in the opposite directions in the dynamic system and the dual dynamic system, which corresponds to the fact that the Riccati equation iterations run in opposite directions in the regulation and estimation problems.

A.18 Exercises

Exercise A.1: Norms in $\mathbb{R}^n$

Show that the following three functions are all norms in $\mathbb{R}^n$

$$\begin{aligned}|x|_2 &:= \left(\sum_{i=1}^{n}(x^i)^2\right)^{1/2}\\ |x|_\infty &:= \max\{\left|x^1\right|, \left|x^2\right|, \ldots, |x^n|\}\\ |x|_1 &:= \sum_{i=1}^{n}\left|x^j\right|\end{aligned}$$

where x^j denotes the jth component of the vector x.

Exercise A.2: Equivalent norms

Show that there are finite constants $K_{ij}, i,j = 1,2,\infty$ such that

$$|x|_i \le K_{ij}\,|x|_j\,, \text{ for all } i,j \in \{1,2,\infty\}.$$

This result shows that the norms are *equivalent* and may be used interchangeably for establishing that sequences are convergent, sets are open or closed, etc.

Regulator	Estimator
A	A'
B	C'
C	G'
k	$l = N - k$
$\Pi(k)$	$P^-(l)$
$\Pi(k-1)$	$P^-(l+1)$
Π	P^-
Q	Q
R	R
$Q(N)$	$Q(0)$
K	$-\widetilde{L}'$
$A + BK$	$(A - \widetilde{L}C)'$
x	ε

Regulator	Estimator
$R > 0, \quad Q > 0$	$R > 0, \quad Q > 0$
(A, B) stabilizable	(A, C) detectable
(A, C) detectable	(A, G) stabilizable

Table A.1: Duality variables and stability conditions for linear quadratic regulation and linear estimation.

Exercise A.3: Open and closed balls

Let $x \in \mathbb{R}^n$ and $\rho > 0$ be given. Show that $\{z \mid |z - x| < \rho\}$ is open and that $B(x, \rho)$ is closed.

Exercise A.4: Condition for closed set

Show that $X \subset \mathbb{R}^n$ is closed if and only if $\text{int}(B(x, \rho)) \cap X \neq \emptyset$ for all $\rho > 0$ implies $x \in X$.

Exercise A.5: Convergence

Suppose that $x_i \to \hat{x}$ as $i \to \infty$; show that for every $\rho > 0$ there exists an $i_p \in \mathbb{I}_{\geq 0}$ such that $x_i \in B(\hat{x}, \rho)$ for all $i \geq i_p$.

Exercise A.6: Limit is unique

Suppose that $\hat{x}, \hat{x}'$ are limits of a sequence $(x_i)_{i \in \mathbb{I}_{\geq 0}}$. Show that $\hat{x} = \hat{x}'$.

Exercise A.7: Open and closed sets

(a) Show that a set $X \subset \mathbb{R}^n$ is open if and only if, for any $\hat{x} \in X$ and any sequence $(x_i) \subset \mathbb{R}^n$ such that $x_i \to \hat{x}$ as $i \to \infty$, there exists a $q \in \mathbb{I}_{\geq 0}$ such that $x_i \in X$ for all $i \geq q$.

(b) Show that a set $X \subset \mathbb{R}^n$ is closed if and only if for all $(x_i) \subset X$, if $x_i \to \hat{x}$ as $i \to \infty$, then $\hat{x} \in X$, i.e., a set X is closed if and only if it contains the limit of every convergent sequences lying in X.

Exercise A.8: Decreasing and bounded below

Prove the observation at the end of Section A.10 that a monotone decreasing sequence that is bounded below converges.

Exercise A.9: Continuous function

Show that $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\hat{x}$ implies $f(x_i) \to f(\hat{x})$ for any sequence (x_i) satisfying $x_i \to \hat{x}$ as $i \to \infty$.

Exercise A.10: Alternative proof of existence of minimum of continuous function on compact set

Prove Proposition A.7 by making use of the fact that $f(X)$ is compact.

Exercise A.11: Differentiable implies Lipschitz

Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ has a continuous derivative $f_x(\cdot)$ in a neighborhood of $\hat{x}$. Show that f is locally Lipschitz continuous at $\hat{x}$.

Exercise A.12: Continuous, Lipschitz continuous, and differentiable

Provide examples of functions meeting the following conditions.

1. Continuous but not Lipschitz continuous.
2. Lipschitz continuous but not differentiable.

Exercise A.13: Differentiating quadratic functions and time-varying matrix inverses

(a) Show that $\nabla f(x) = Qx$ if $f(x) = (1/2)x'Qx$ and Q is symmetric.

(b) Show that $(d/dt)A^{-1}(t) = -A^{-1}(t)\dot{A}(t)A^{-1}(t)$ if $A : \mathbb{R} \to \mathbb{R}^{n\times n}$, $A(t)$ is invertible for all $t \in \mathbb{R}$, and $\dot{A}(t) := (d/dt)A(t)$.

Exercise A.14: Directional derivative

Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ has a derivative $f_x(\hat{x})$ at $\hat{x}$. Show that for any h, the directional derivative $df(\hat{x}; h)$ exists and is given by

$$df(\hat{x}; h) = f_x(\hat{x})h = (\partial f(x)/\partial x)h.$$

Exercise A.15: Convex combination

Suppose $S \subset \mathbb{R}^n$ is convex. Let $\{x_i\}_{i=1}^k$ be points in S and let $\{\mu^i\}_{i=1}^k$ be scalars such that $\mu^i \geq 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^k \mu^i = 1$. Show that

$$\left(\sum_{i=1}^{k} \mu^i x_i\right) \in S.$$

Exercise A.16: Convex epigraph

Show that $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if its epigraph is convex.

Exercise A.17: Bounded second derivative and minimum

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable and that for some $\infty > M \geq m > 0$, $M \left|y\right|^2 \geq \langle y, \partial^2 f / \partial x^2(x) y \rangle \geq m \left|y\right|^2$ for all $x, y \in \mathbb{R}^n$. Show that the sublevel sets of f are convex and compact and that $f(\cdot)$ attains its infimum.

Exercise A.18: Sum and max of convex functions are convex

Suppose that $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, 2, \ldots, m$ are convex. Show that

$$\psi^1(x) \coloneqq \max_i \{f_i(x) \mid i \in \{1, 2, \ldots, m\}\},$$

$$\psi^2(x) \coloneqq \sum_{i=1}^{m} f_i(x)$$

are both convex.

Exercise A.19: Einige kleine Mathprobleme

(a) Prove that if λ is an eigenvalue and v is an eigenvector of A ($Av = \lambda v$), then λ is also an eigenvalue of T in which T is upper triangular and given by the Schur decomposition of A

$$Q^* A Q = T$$

What is the corresponding eigenvector?

(b) Prove statement 1 on positive definite matrices (from Section A.7). Where is this fact needed?

(c) Prove statement 6 on positive definite matrices. Where is this fact needed?

(d) Prove statement 5 on positive definite matrices.

(e) Prove statement 8 on positive semidefinite matrices.

(f) Derive the two expressions for the partitioned A^{-1}.

Exercise A.20: Positive definite but not symmetric matrices

Consider redefining the notation $A > 0$ for $A \in \mathbb{R}^{n \times n}$ to mean $x'Ax > 0$ for all $x \in \mathbb{R}^n \neq 0$. In other words, the restriction that A is symmetric in the usual definition of positive definiteness is removed. Consider also $B \coloneqq (A + A')/2$. Show the following hold for all A. (a) $A > 0$ if and only if B is positive definite. (b) $\mathrm{tr}(A) = \mathrm{tr}(B)$. (Johnson, 1970; Johnson and Hillar, 2002)

Exercise A.21: Trace of a matrix function

Derive the following formula for differentiating the trace of a function of a square matrix

$$\frac{d \,\mathrm{tr}(f(A))}{dA} = g(A') \qquad g(x) = \frac{df(x)}{dx}$$

in which g is the usual scalar derivative of the scalar function f. This result proves useful in evaluating the change in the expectation of the stage cost in stochastic control problems.

Exercise A.22: Some matrix differentiation

Derive the following formulas (Bard, 1974). $A, B \in \mathbb{R}^{n\times n}, a, x \in \mathbb{R}^n$.

(a)
$$\frac{\partial x' Ax}{\partial x} = Ax + A'x$$

(b)
$$\frac{\partial Axa'Bx}{\partial x'} = (a'Bx)A + Axa'B$$

(c)
$$\frac{\partial a'Ab}{\partial A} = ab'$$

Exercise A.23: Partitioned matrix inversion formula

In deriving the partitioned matrix inversion formula we assumed A is partitioned into

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

and that A^{-1}, B^{-1} and E^{-1} exist. In the final formula, the term

$$(E - DB^{-1}C)^{-1}$$

appears, but we did not assume this matrix is invertible. Did we leave out an assumption or can the existence of this matrix inverse be proven given the other assumptions? If we left out an assumption, provide an example in which this matrix is not invertible. If it follows from the other assumptions, prove this inverse exists.

Exercise A.24: Partitioned positive definite matrices

Consider the partitioned positive definite, symmetric matrix

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}$$

Prove that the following matrices are also positive definite

1. H_{11}
2. H_{22}
3. $\overline{H}$ in which
$$\overline{H} = \begin{bmatrix} H_{11} & -H_{12} \\ -H_{21} & H_{22} \end{bmatrix}$$
4. $H_{11} - H_{12}H_{22}^{-1}H_{21}$ and $H_{22} - H_{21}H_{11}^{-1}H_{12}$

Exercise A.25: Properties of the matrix exponential

Prove that the following properties of the matrix exponential, which are useful for dealing with continuous time linear systems. The matrix A is a real-valued $n \times n$ matrix, and t is real.

(a)
$$\text{rank}\left(e^{At}\right) = n \quad \forall t$$

(b)

$$\text{rank}\left(\int_0^t e^{A\tau}d\tau\right) = n \quad \forall t > 0$$

Exercise A.26: Controllability in continuous time

A linear, time-invariant, continuous time system

$$\begin{aligned}\frac{dx}{dt} &= Ax + Bu\\ x(0) &= x_0\end{aligned} \tag{A.41}$$

is **controllable** if there exists an input $u(t), 0 \le t \le t_1, t_1 > 0$ that takes the system from any x_0 at time zero to any x_1 at some finite time t_1.

(a) Prove that the system in (A.41) is controllable if and only if

$$\text{rank}\,(C) = n$$

in which C is, remarkably, the same controllability matrix that was defined for discrete time systems 1.16

$$C = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

(b) Describe a calculational procedure for finding this required input.

Exercise A.27: Reachability Gramian in continuous time

Consider the symmetric, $n \times n$ matrix W defined by

$$W(t) = \int_0^t e^{(t-\tau)A} BB' e^{(t-\tau)A'} d\tau$$

The matrix W is known as the reachability Gramian of the linear, time-invariant system. The reachability Gramian proves useful in analyzing controllability and reachability. Prove the following important properties of the reachability Gramian.

(a) The reachability Gramian satisfies the following matrix differential equation

$$\begin{aligned}\frac{dW}{dt} &= BB' + AW + WA'\\ W(0) &= 0\end{aligned}$$

which provides one useful way to calculate its values.

(b) The reachability Gramian $W(t)$ is full rank for all $t > 0$ if and only if the system is controllable.

Exercise A.28: Differences in continuous time and discrete time systems

Consider the definition that a system is controllable if there exists an input that takes the system from any x_0 at time zero to any x_1 at some finite time t_1.

(a) Show that x_1 can be taken as zero without changing the meaning of controllability for a linear continuous time system.

(b) In linear discrete time systems, x_1 cannot be taken as zero without changing the meaning of controllability. Why not? Which A require a distinction in discrete time. What are the eigenvalues of the corresponding A in continuous time?

Exercise A.29: Observability in continuous time

Consider the linear time-invariant continuous time system

$$\begin{aligned} \frac{dx}{dt} &= Ax \\ x(0) &= x_0 \\ y &= Cx \end{aligned} \tag{A.42}$$

and let $y(t;x_0)$ represent the solution to (A.42) as a function of time t given starting state value x_0 at time zero. Consider the output from two different initial conditions $y(t;w)$, $y(t;z)$ on the time interval $0 \le t \le t_1$ with $t_1 > 0$.

The system in (A.42) is **observable** if

$$y(t;w) = y(t;z), \quad 0 \le t \le t_1 \Longrightarrow w = z$$

In other words, if two output measurement trajectories agree, the initial conditions that generated the output trajectories must agree, and hence, the initial condition is unique. This uniqueness of the initial condition allows us to consider building a state estimator to reconstruct $x(0)$ from $y(t;x_0)$. After we have found the unique $x(0)$, solving the model provides the rest of the state trajectory $x(t)$. We will see later that this procedure is not the preferred way to build a state estimator; it simply shows that if the system is observable, the goal of state estimation is reasonable.

Show that the system in (A.42) is observable if and only if

$$\text{rank}\,(\mathcal{O}) = n$$

in which $\mathcal{O}$ is, again, the same observability matrix that was defined for discrete time systems 1.36

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

Hint: what happens if you differentiate $y(t;w) - y(t;z)$ with respect to time? How many times is this function differentiable?

Exercise A.30: Observability Gramian in continuous time

Consider the symmetric, $n \times n$ matrix W_o defined by

$$W_o(t) = \int_0^t e^{A'\tau} C' C e^{A\tau} d\tau$$

The matrix W_o is known as the observability Gramian of the linear, time-invariant system. Prove the following important properties of the observability Gramian.

(a) The observability Gramian $W_o(t)$ is full rank for all $t > 0$ if and only if the system is observable.

(b) Consider an observable linear time invariant system with $u(t) = 0$ so that $y(t) = Ce^{At}x_0$. Use the observability Gramian to solve this equation for x_0 as a function of $y(t), 0 \le t \le t_1$.

(c) Extend your result from the previous part to find x_0 for an arbitrary $u(t)$.

Exercise A.31: Detectability of (A, C) and output penalty

Given a system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) \end{aligned}$$

Suppose (A, C) is detectable and an input sequence has been found such that

$$u(k) \to 0 \qquad y(k) \to 0$$

Show that $x(k) \to 0$.

Exercise A.32: Prove your favorite Hautus lemma

Prove the Hautus lemma for controllability, Lemma 1.2, or observability, Lemma 1.4.

Exercise A.33: Positive semidefinite Q penalty and its square root

Consider the linear quadratic problem with system

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Q^{1/2}x(k) \end{aligned}$$

and infinite horizon cost function

$$\begin{aligned} \Phi &= \sum_{k=0}^{\infty} x(k)'Qx(k) + u(k)'Ru(k) \\ &= \sum_{k=0}^{\infty} y(k)'y(k) + u(k)'Ru(k) \end{aligned}$$

with $Q \geq 0$, $R > 0$, and (A, B) stabilizable. In Exercise A.31 we showed that if $(A, Q^{1/2})$ is detectable and an input sequence has been found such that

$$u(k) \to 0 \qquad y(k) \to 0$$

then $x(k) \to 0$.

(a) Show that if $Q \geq 0$, then $Q^{1/2}$ is a well defined, real, symmetric matrix and $Q^{1/2} \geq 0$.

Hint: apply Theorem A.1 to Q, using the subsequent fact 3.

(b) Show that $(A, Q^{1/2})$ is detectable (observable) if and only if (A, Q) is detectable (observable). So we can express one of the LQ existence, uniqueness, and stability conditions using detectability of (A, Q) instead of $(A, Q^{1/2})$.

Exercise A.34: Probability density of the inverse function

Consider a scalar random variable $\xi \in \mathbb{R}$ and let the random variable η be defined by the inverse function

$$\eta = \xi^{-1}$$

(a) If ξ is distributed uniformly on $[a, 1]$ with $0 < a < 1$, what is the density of η?

(b) Is η's density well defined if we allow $a = 0$? Explain your answer.

Exercise A.35: Expectation as a linear operator

(a) Consider the random variable x to be defined as a linear combination of the random variables a and b

$$x = a + b$$

Show that

$$\mathcal{E}(x) = \mathcal{E}(a) + \mathcal{E}(b)$$

Do a and b need to be statistically independent for this statement to be true?

(b) Next consider the random variable x to be defined as a scalar multiple of the random variable a

$$x = \alpha a$$

Show that

$$\mathcal{E}(x) = \alpha \mathcal{E}(a)$$

(c) What can you conclude about $\mathcal{E}(x)$ if x is given by the linear combination

$$x = \sum_i \alpha_i v_i$$

in which v_i are random variables and α_i are scalars.

Exercise A.36: Minimum of two random variables

Given two independent random variables, ξ_1, ξ_2 and the random variable defined by the minimum operator

$$\eta = \min(\xi_1, \xi_2)$$

(a) Sketch the region $\mathbb{X}(c)$ for the inequality $\min(x_1, x_2) \leq c$.

(b) Find η's probability density in terms of the probability densities of ξ_1, ξ_2.

Exercise A.37: Maximum of n normally distributed random variables

Given n independent, identically distributed normal random variables, $\xi_1, \xi_2, \ldots, \xi_n$ and the random variable defined by the maximum operator

$$\eta = \max(\xi_1, \xi_2, \ldots \xi_n)$$

(a) Derive a formula for η's density.

(b) Plot p_η for $\xi_i \sim N(0,1)$ and $n = 1, 2, \ldots 5$. Describe the trend in p_η as n increases.

Exercise A.38: Another picture of mean

Consider a scalar random variable ξ with probability distribution P_ξ shown in Figure A.14. Consider the inverse probability distribution, P_ξ^{-1}, also shown in Figure A.14.

(a) Show that the expectation of ξ is equal to the following integral of the probability distribution (David, 1981, p. 38)

$$\mathcal{E}(\xi) = -\int_{-\infty}^{0} P_\xi(x)dx + \int_0^{\infty} (1 - P_\xi(x))dx \tag{A.43}$$

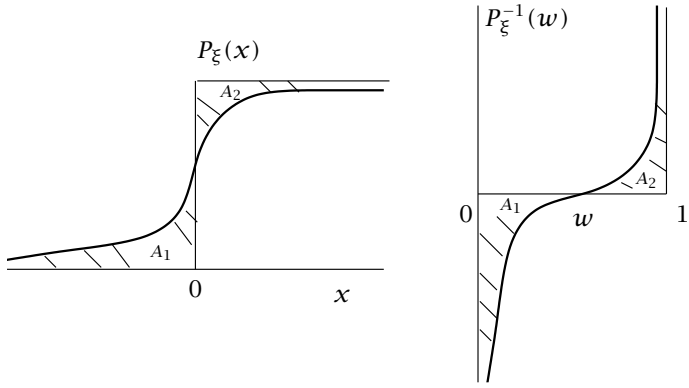


Figure A.14: The probability distribution and inverse distribution for random variable ξ. The mean of ξ is given by the difference in the hatched areas, $\mathcal{E}(\xi) = A_2 - A_1$.

(b) Show that the expectation of ξ is equal to the following integral of the inverse probability distribution

$$\mathcal{E}(\xi) = \int_0^1 P_\xi^{-1}(w)dw \tag{A.44}$$

These interpretations of mean are shown as the hatched areas in Figure A.14, $\mathcal{E}(\xi) = A_2 - A_1$.

Exercise A.39: Ordering random variables

We can order two random variables A and B if they obey an inequality such as $A \geq B$. The frequency interpretation of the probability distribution, $P_A(c) = \Pr(A \leq c)$, then implies that $P_A(c) \leq P_B(c)$ for all c.

If $A \geq B$, show that

$$\mathcal{E}(A) \geq \mathcal{E}(B)$$

Exercise A.40: Max of the mean and mean of the max

Given two random variables A and B, establish the following inequality

$$\max(\mathcal{E}(A), \mathcal{E}(B)) \leq \mathcal{E}(\max(A, B))$$

In other words, the max of the mean is an underbound for the mean of the max.

Exercise A.41: Observability

Consider the linear system with zero input

$$x(k+1) = Ax(k)$$
$$y(k) = Cx(k)$$

with

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

(a) What is the observability matrix for this system? What is its rank?

(b) Consider a string of data measurements

$$y(0) = y(1) = \cdots = y(n-1) = 0$$

Now $x(0) = 0$ is clearly consistent with these data. Is this $x(0)$ unique? If yes, prove it. If no, characterize the set of all $x(0)$ that are consistent with these data.

Exercise A.42: Nothing is revealed

An agitated graduate student shows up at your office. He begins, "I am afraid I have discovered a deep contradiction in the foundations of systems theory." You ask him to calm down and tell you about it. He continues, "Well, we have the pole placement theorem that says if (A, C) is observable, then there exists a matrix L such that the eigenvalues of an observer

$$A - ALC$$

can be assigned arbitrarily."

You reply, "Well, they do have to be conjugate pairs because the matrices A, L, C are real-valued, but yeah, sure, so what?"

He continues, "Well we also have the Hautus lemma that says (A, C) is observable if and only if

$$\operatorname{rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n \qquad \forall \lambda \in \mathbb{C}$$

"You know, the Hautus lemma has always been one of my favorite lemmas; I don't see a problem," you reply.

"Well," he continues, "isn't the innovations form of the system, $(A - ALC, C)$, observable if and only if the original system, (A, C), is observable?"

"Yeah ...I seem to recall something like that," you reply, starting to feel a little uncomfortable.

"OK, how about if I decide to put all the observer poles at zero?" he asks, innocently.

You object, "Wait a minute, I guess you can do that, but that's not going to be a very good observer, so I don't think it matters if ...."

"Well," he interrupts, "how about we put all the eigenvalues of $A - ALC$ at zero, like I said, and then we check the Hautus condition at $\lambda = 0$? I get

$$\operatorname{rank} \begin{bmatrix} \lambda I - (A - ALC) \\ C \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 0 \\ C \end{bmatrix} \qquad \lambda = 0$$

"So tell me, how is that matrix on the right ever going to have rank n with that big, fat zero sitting there?" At this point, you start feeling a little dizzy.

What's causing the contradiction here: the pole placement theorem, the Hautus lemma, the statement about equivalence of observability in innovations form, something else? How do you respond to this student?

Exercise A.43: The sum of throwing two dice

Using (A.30), what is the probability density for the sum of throwing two dice? On what number do you want to place your bet? How often do you expect to win if you bet on this outcome?

Make the standard assumptions: the probability density for each die is uniform over the integer values from one to six, and the outcome of each die is independent of the other die.

Exercise A.44: The product of throwing two dice

Using (A.30), what is the probability density for the product of throwing two dice? On what number do you want to place your bet? How often do you expect to win if you bet on this outcome?

Make the standard assumptions: the probability density for each die is uniform over the integer values from one to six, and the outcome of each die is independent of the other die.

Exercise A.45: The size of an ellipse's bounding box

Here we derive the size of the bounding box depicted in Figure A.10. Consider a real, positive definite, symmetric matrix $A \in \mathbb{R}^{n \times n}$ and a real vector $x \in \mathbb{R}^n$. The set of x for which the scalar $x'Ax$ is constant are n-dimensional ellipsoids. Find the length of the sides of the smallest box that contains the ellipsoid defined by

$$x'Ax = b$$

Hint: Consider the equivalent optimization problem to minimize the value of $x'Ax$ such that the ith component of x is given by $x_i = c$. This problem defines the ellipsoid that is tangent to the plane $x_i = c$, and can be used to answer the original question.

Exercise A.46: The tangent points of an ellipse's bounding box

Find the tangent points of an ellipsoid defined by $x'Ax = b$, and its bounding box as depicted in Figure A.10 for $n = 2$. For $n = 2$, draw the ellipse, bounding box and compute the tangent points for the following parameters taken from Figure A.10

$$A = \begin{bmatrix} 3.5 & 2.5 \\ 2.5 & 4.0 \end{bmatrix} \qquad b = 1$$

Exercise A.47: Let's make a deal!

Consider the following contest of the American television game show of the 1960s, Let's Make a Deal. In the show's grand finale, a contestant is presented with three doors. Behind one of the doors is a valuable prize such as an all-expenses-paid vacation to Hawaii or a new car. Behind the other two doors are goats and donkeys. The contestant selects a door, say door number one. The game show host, Monty Hall, then says,

"Before I show you what is behind your door, let's reveal what is behind door number three!" Monty always chooses a door that has one of the booby prizes behind it. As the goat or donkey is revealed, the audience howls with laughter. Then Monty asks innocently,

"Before I show you what is behind your door, I will allow you one chance to change your mind. Do you want to change doors?" While the contestant considers this option, the audience starts screaming out things like,

"Stay with your door! No, switch, switch!" Finally the contestant chooses again, and then Monty shows them what is behind their chosen door.

Let's analyze this contest to see how to *maximize* the chance of winning. Define

$$p(i, j, y), \qquad i, j, y = 1, 2, 3$$

to be the probability that you chose door i, the prize is behind door j and Monty showed you door y (named after the data!) after your initial guess. Then you would want to

$$\max_j p(j|i, y)$$

for your optimal choice after Monty shows you a door.

(a) Calculate this conditional density and give the probability that the prize is behind door i, your original choice, and door $j \neq i$.

(b) You will need to specify a model of Monty's behavior. Please state the one that is appropriate to Let's Make a Deal.

(c) For what other model of Monty's behavior is the answer that it doesn't matter if you switch doors. Why is this a poor model for the game show?

Exercise A.48: Norm of an extended state

Consider $x \in \mathbb{R}^n$ with a norm denoted $|\cdot|_\alpha$, and $u \in \mathbb{R}^m$ with a norm denoted $|\cdot|_\beta$. Now consider a proposed norm for the extended state (x, u)

$$|(x, u)|_\gamma := |x|_\alpha + |u|_\beta$$

Show that this proposal satisfies the definition of a norm given in Section A.8.

If the α and β norms are chosen to be p-norms, is the γ norm also a p-norm? Show why or why not.

Exercise A.49: Distance of an extended state to an extended set

Let $x \in \mathbb{R}^n$ and $\mathbb{X}$ a set of elements in $\mathbb{R}^n$, and $u \in \mathbb{R}^m$ and $\mathbb{U}$ a set of elements in $\mathbb{R}^m$. Denote distances from elements to their respective sets as

$$|x|_\mathbb{X} := \inf_{y \in \mathbb{X}} |x - y|_\alpha \qquad |u|_\mathbb{U} := \inf_{v \in \mathbb{U}} |u - v|_\beta$$

$$|(x, u)|_{\mathbb{X}\times\mathbb{U}} := \inf_{(y,v) \in \mathbb{X}\times\mathbb{U}} |(x, u) - (y, v)|_\gamma$$

Use the norm of the extended state defined in Exercise A.48 to show that

$$|(x, u)|_{\mathbb{X}\times\mathbb{U}} = |x|_\mathbb{X} + |u|_\mathbb{U}$$

Bibliography

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New York, third edition, 2003.

T. M. Apostol. *Mathematical analysis*. Addison-Wesley, 1974.

Y. Bard. *Nonlinear Parameter Estimation*. Academic Press, New York, 1974.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 53:370–418, 1763. Reprinted in *Biometrika*, 35:293–315, 1958.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

F. M. Callier and C. A. Desoer. *Linear System Theory*. Springer-Verlag, New York, 1991.

E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. McGraw Hill, 1955.

H. A. David. *Order Statistics*. John Wiley & Sons, Inc., New York, second edition, 1981.

J. Dieudonne. *Foundations of modern analysis*. Academic Press, 1960.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, Maryland, third edition, 1996.

J. Hale. *Ordinary Differential Equations*. Robert E. Krieger Publishing Company, second edition, 1980.

P. Hartman. *Ordinary Differential Equations*. John Wiley and Sons, 1964.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

C. R. Johnson. Positive definite matrices. *Amer. Math. Monthly*, 77(3):259–264, March 1970.

C. R. Johnson and C. J. Hillar. Eigenvalues of words in two positive definite letters. *SIAM J. Matrix Anal. and Appl.*, 23(4):916–928, 2002.

E. J. McShane. *Integration*. Princeton University Press, 1944.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, second edition, 2006.

A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, Inc., second edition, 1984.

E. Polak. *Optimization: Algorithms and Consistent Approximations*. Springer Verlag, New York, 1997. ISBN 0-387-94971-2.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, N.J., 1970.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer-Verlag, 1998.

I. Schur. On the characteristic roots of a linear substitution with an application to the theory of integral equations (German). *Math Ann.*, 66:488–510, 1909.

G. Strang. *Linear Algebra and its Applications*. Academic Press, New York, second edition, 1980.

B
Stability Theory

Version: date: April 5, 2019

B.1 Introduction

In this appendix we consider stability properties of discrete time systems. A good general reference for stability theory of continuous time systems is Khalil (2002). There are not many texts for stability theory of discrete time systems; a useful reference is LaSalle (1986). Recently stability theory for discrete time systems has received more attention in the literature. In the notes below we draw on Jiang and Wang (2001, 2002); Kellett and Teel (2004a,b).

We consider systems of the form

$$x^+ = f(x,u)$$

where the state x lies in $\mathbb{R}^n$ and the control (input) u lies in $\mathbb{R}^m$; in this formulation x and u denote, respectively, the current state and control, and x^+ the successor state. We assume in the sequel that the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous. Let $\phi(k;x,\mathbf{u})$ denote the solution of $x^+ = f(x,u)$ at time k if the initial state is $x(0) = x$ and the control sequence is $\mathbf{u} = (u(0),u(1),u(2),\ldots)$; the solution exists and is unique. If a state-feedback control law $u = \kappa(x)$ has been chosen, the closed-loop system is described by $x^+ = f(x,\kappa(x))$, which has the same form $x^+ = f_c(x)$ where $f_c(\cdot)$ is defined by $f_c(x) := f(x, \kappa(x))$. Let $\phi(k;x,\kappa(\cdot))$ denote the solution of this difference equation at time k if the initial state at time 0 is $x(0) = x$; the solution exists and is unique (even if $\kappa(\cdot)$ is discontinuous). If $\kappa(\cdot)$ is not continuous, as may be the case when $\kappa(\cdot)$ is an implicit model predictive control (MPC) law, then $f_c(\cdot)$ may not be continuous. In this case we assume that $f_c(\cdot)$ is *locally bounded*.[1]

[1] A function $f : X \to X$ is locally bounded if, for any $x \in X$, there exists a neighborhood $\mathcal{N}$ of x such that $f(\mathcal{N})$ is a bounded set, i.e., if there exists a $M > 0$ such that $|f(x)| \leq M$ for all $x \in \mathcal{N}$.

We would like to be sure that the controlled system is "stable", i.e., that small perturbations of the initial state do not cause large variations in the subsequent behavior of the system, and that the state converges to a desired state or, if this is impossible due to disturbances, to a desired set of states. These objectives are made precise in Lyapunov stability theory; in this theory, the system $x^+ = f(x)$ is assumed given and conditions ensuring the stability, or asymptotic stability of a specified state or set are sought; the terms *stability* and *asymptotic stability* are defined below. If convergence to a specified state, x^* say, is sought, it is desirable for this state to be an *equilibrium* point:

Definition B.1 (Equilibrium point)**.** A point x^* is an equilibrium point of $x^+ = f(x)$ if $x(0) = x^*$ implies $x(k) = \phi(k; x^*) = x^*$ for all $k \geq 0$. Hence x^* is an equilibrium point if it satisfies

$$x^* = f(x^*)$$

An equilibrium point x^* is isolated if there are no other equilibrium points in a sufficiently small neighborhood of x^*. A linear system $x^+ = Ax + b$ has a single equilibrium point $x^* = (I - A)^{-1}b$ if $I - A$ is invertible; if not, the linear system has a continuum $\{x \mid (I - A)x = b\}$ of equilibrium points. A nonlinear system, unlike a linear system, may have several isolated equilibrium points.

In other situations, for example when studying the stability properties of an oscillator, convergence to a specified closed set $\mathcal{A} \subset \mathbb{R}^n$ is sought. In the case of a linear oscillator with state dimension 2, this set is an ellipse. If convergence to a set $\mathcal{A}$ is sought, it is desirable for the set $\mathcal{A}$ to be *positive invariant*:

Definition B.2 (Positive invariant set)**.** A closed set $\mathcal{A}$ is positive invariant for the system $x^+ = f(x)$ if $x \in \mathcal{A}$ implies $f(x) \in \mathcal{A}$.

Clearly, any solution of $x^+ = f(x)$ with initial state in $\mathcal{A}$, remains in $\mathcal{A}$. The closed set $\mathcal{A} = \{x^*\}$ consisting of a (single) equilibrium point is a special case; $x \in \mathcal{A}$ $(x = x^*)$ implies $f(x) \in \mathcal{A}$ $(f(x) = x^*)$. Define $|x|_{\mathcal{A}} := \inf_{z \in \mathcal{A}} |x - z|$ to be the distance of a point x from the set $\mathcal{A}$; if $\mathcal{A} = \{x^*\}$, then $|x|_{\mathcal{A}} = |x - x^*|$ which reduces to $|x|$ when $x^* = 0$.

Before introducing the concepts of stability and asymptotic stability and their characterization by Lyapunov functions, it is convenient to make a few definitions.

Definition B.3 ($\mathcal{K}$, $\mathcal{K}_\infty$, $\mathcal{KL}$, and $\mathcal{PD}$ functions)**.** A function $\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ belongs to class $\mathcal{K}$ if it is continuous, zero at zero, and strictly increasing; $\sigma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ belongs to class $\mathcal{K}_\infty$ if it is a class $\mathcal{K}$ and unbounded ($\sigma(s) \to \infty$ as $s \to \infty$). A function $\beta : \mathbb{R}_{\geq 0} \times \mathbb{I}_{\geq 0} \to \mathbb{R}_{\geq 0}$ belongs to class $\mathcal{KL}$ if it is continuous and if, for each $t \geq 0$, $\beta(\cdot, t)$ is a class $\mathcal{K}$ function and for each $s \geq 0$, $\beta(s, \cdot)$ is nonincreasing and satisfies $\lim_{t \to \infty} \beta(s, t) = 0$. A function $\gamma : \mathbb{R} \to \mathbb{R}_{\geq 0}$ belongs to class $\mathcal{PD}$ (is positive definite) if it is zero at zero and positive everywhere else.[2]

The following useful properties of these functions are established in Khalil (2002, Lemma 4.2): if $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are $\mathcal{K}$ functions ($\mathcal{K}_\infty$ functions), then $\alpha_1^{-1}(\cdot)$ and $(\alpha_1 \circ \alpha_2)(\cdot)$[3] are $\mathcal{K}$ functions[4] ($\mathcal{K}_\infty$ functions). Moreover, if $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are $\mathcal{K}$ functions and $\beta(\cdot)$ is a $\mathcal{KL}$ function, then $\sigma(r, s) = \alpha_1(\beta(\alpha_2(r), s))$ is a $\mathcal{KL}$ function.

The following properties prove useful when analyzing the robustness of perturbed systems.

1. For $\gamma(\cdot) \in \mathcal{K}$, the following holds for all $a_i \in \mathbb{R}_{\geq 0}$, $i \in \mathbb{I}_{1:n}$

$$\frac{1}{n}(\gamma(a_1) + \cdots + \gamma(a_n)) \leq \gamma(a_1 + \cdots + a_n) \leq \gamma(na_1) + \cdots + \gamma(na_n) \quad \text{(B.1)}$$

2. Similarly, for $\beta(\cdot) \in \mathcal{KL}$ the following holds for all $a_i \in \mathbb{R}_{\geq 0}$, $i \in \mathbb{I}_{1:n}$, and all $t \in \mathbb{R}_{\geq 0}$

$$\frac{1}{n}(\beta(a_1, t) + \cdots + \beta(a_n, t)) \leq \beta((a_1 + \cdots + a_n), t) \leq \beta(na_1, t) + \beta(na_2, t) + \cdots + \beta(na_n, t) \quad \text{(B.2)}$$

3. If $\alpha_i(\cdot) \in \mathcal{K}(\mathcal{K}_\infty)$ for $i \in \mathbb{I}_{1:n}$, then

$$\min_i \{\alpha_i(\cdot)\} := \underline{\alpha}(\cdot) \in \mathcal{K}(\mathcal{K}_\infty) \quad \text{(B.3)}$$

$$\max_i \{\alpha_i(\cdot)\} := \overline{\alpha}(\cdot) \in \mathcal{K}(\mathcal{K}_\infty) \quad \text{(B.4)}$$

[2] Be aware that the existing stability literature sometimes includes continuity in the definition of a positive definite function. We used such a definition in the first edition of this text, for example. But in the second edition, we remove continuity and retain only the requirement of positivity in the definition of positive definite function.

[3] $(\alpha_1 \circ \alpha_2)(\cdot)$ is the composition of the two functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ and is defined by $(\alpha_1 \circ \alpha_2)(s) := \alpha_1(\alpha_2(s))$.

[4] Note, however, that the domain of $\alpha^{-1}(\cdot)$ may be restricted from $\mathbb{R}_{\geq 0}$ to $[0, a)$ for some $a > 0$.

4. Let $v_i \in \mathbb{R}^{n_i}$ for $i \in \mathbb{I}_{1:n}$, and $v := (v_1, \ldots, v_n) \in \mathbb{R}^{\sum n_i}$. If $\alpha_i(\cdot) \in \mathcal{K}(\mathcal{K}_\infty)$ for $i \in \mathbb{I}_{1:n}$, then there exist $\underline{\alpha}(\cdot), \overline{\alpha}(\cdot) \in \mathcal{K}(\mathcal{K}_\infty)$ such that
$$\underline{\alpha}(|v|) \leq \alpha_1(|v_1|) + \cdots + \alpha_n(|v_n|) \leq \overline{\alpha}(|v|) \tag{B.5}$$

See (Rawlings and Ji, 2012) for short proofs of (B.1) and (B.2), and (Allan, Bates, Risbeck, and Rawlings, 2017, Proposition 23) for a short proof of (B.3). The result (B.4) follows similarly to (B.3). Result (B.5) follows from (B.1) and (B.3)-(B.4). See also Exercise B.9.

B.2 Stability and Asymptotic Stability

In this section we consider the stability properties of the autonomous system $x^+ = f(x)$; we assume that $f(\cdot)$ is locally bounded, and that the set $\mathcal{A}$ is closed and positive invariant for $x^+ = f(x)$ unless otherwise stated.

Definition B.4 (Local stability)**.** The (closed, positive invariant) set $\mathcal{A}$ is *locally stable* for $x^+ = f(x)$ if, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that $|x|_{\mathcal{A}} < \delta$ implies $|\phi(i;x)|_{\mathcal{A}} < \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

See Figure B.1 for an illustration of this definition when $\mathcal{A} = \{0\}$; in this case we speak of stability of the origin.

Remark. Stability of the origin, as defined above, is equivalent to continuity of the map $x \mapsto \mathbf{x} := (x, \phi(1;x), \phi(2;x), \ldots)$, $\mathbb{R} \to \ell_\infty$ at the origin so that $\|\mathbf{x}\| \to 0$ as $x \to 0$ (a small perturbation in the initial state causes a small perturbation in the subsequent motion).

Definition B.5 (Global attraction)**.** The (closed, positive invariant) set $\mathcal{A}$ is *globally attractive* for the system $x^+ = f(x)$ if $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$ for all $x \in \mathbb{R}^n$.

Definition B.6 (Global asymptotic stability)**.** The (closed, positive invariant) set $\mathcal{A}$ is *globally asymptotically stable* (GAS) for $x^+ = f(x)$ if it is locally stable and globally attractive.

It is possible for the origin to be globally attractive but *not* locally stable. Consider a second order system

$$x^+ = Ax + \phi(x)$$

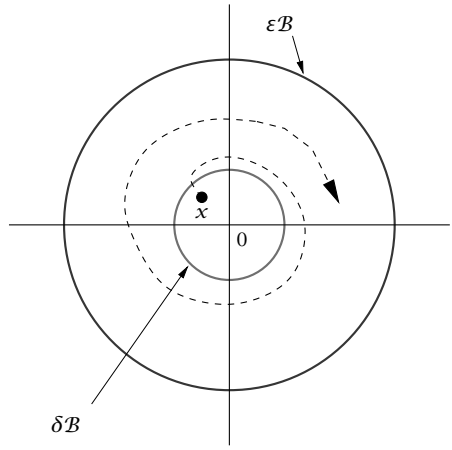


Figure B.1: Stability of the origin. $\mathcal{B}$ denotes the unit ball.

where A has eigenvalues $\lambda_1 = 0.5$ and $\lambda_2 = 2$ with associated eigenvectors w_1 and w_2, shown in Figure B.2; w_1 is the "stable" and w_2 the "unstable" eigenvector; the smooth function $\phi(\cdot)$ satisfies $\phi(0) = 0$ and $(\partial/\partial x)\phi(0) = 0$ so that $x^+ = Ax + \phi(x)$ behaves like $x^+ = Ax$ near the origin. If $\phi(x) \equiv 0$, the motion corresponding to an initial state αw_1, $\alpha \neq 0$, converges to the origin, whereas the motion corresponding to an initial state αw_2 diverges. If $\phi(\cdot)$ is such that it steers nonzero states toward the horizontal axis, we get trajectories of the form shown in Figure B.2. All trajectories converge to the origin but the motion corresponding to an initial state αw_2, *no matter how small*, is similar to that shown in Figure B.2 and cannot satisfy the ε, δ definition of local stability. The origin is globally attractive but not stable. A trajectory that joins an equilibrium point to itself, as in Figure B.2, is called a homoclinic orbit.

We collect below a set of useful definitions:

Definition B.7 (Various forms of stability)**.** The (closed, positive invariant) set $\mathcal{A}$ is

(a) locally stable if, for each $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $|x|_{\mathcal{A}} < \delta$ implies $|\phi(i;x)|_{\mathcal{A}} < \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

(b) unstable, if it is not locally stable.

(c) locally attractive if there exists $\eta > 0$ such that $|x|_{\mathcal{A}} < \eta$ implies $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$.

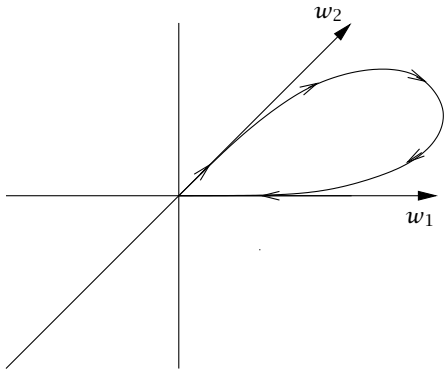


Figure B.2: An attractive but unstable origin.

(d) globally attractive if $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$ for all $x \in \mathbb{R}^n$.

(e) locally asymptotically stable if it is locally stable and locally attractive.

(f) globally asymptotically stable if it is locally stable and globally attractive.

(g) locally exponentially stable if there exist $\eta > 0$, $c > 0$, and $\gamma \in (0,1)$ such that $|x|_{\mathcal{A}} < \eta$ implies $|\phi(i;x)|_{\mathcal{A}} \leq c\,|x|_{\mathcal{A}}\,\gamma^i$ for all $i \in \mathbb{I}_{\geq 0}$.

(h) globally exponentially stable if there exists a $c > 0$ and a $\gamma \in (0,1)$ such that $|\phi(i;x)|_{\mathcal{A}} \leq c\,|x|_{\mathcal{A}}\,\gamma^i$ for all $x \in \mathbb{R}^n$, all $i \in \mathbb{I}_{\geq 0}$.

The following stronger definition of GAS has recently started to become popular.

Definition B.8 (Global asymptotic stability (KL version))**.** The (closed, positive invariant) set $\mathcal{A}$ is *globally asymptotically stable* (GAS) for $x^+ = f(x)$ if there exists a $\mathcal{KL}$ function $\beta(\cdot)$ such that, for each $x \in \mathbb{R}^n$

$$|\phi(i;x)|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i) \qquad \forall\, i \in \mathbb{I}_{\geq 0} \tag{B.6}$$

Proposition B.9 (Connection of classical and KL global asymptotic stability)**.** *Suppose $\mathcal{A}$ is compact (and positive invariant) and that $f(\cdot)$ is continuous. Then the classical and KL definitions of global asymptotic stability of $\mathcal{A}$ for $x^+ = f(x)$ are equivalent.*

The KL version of global asymptotic stability implies the classical version from (B.6) and the definition of a $\mathcal{KL}$ function. The converse

is harder to prove but is established in Jiang and Wang (2002) where Proposition 2.2 establishes the equivalence of the existence of a $\mathcal{KL}$ function satisfying (2) with UGAS (uniform global asymptotic stability), and Corollary 3.3 which establishes the equivalence, *when* $\mathcal{A}$ is compact, of uniform global asymptotic stability and global asymptotic stability. Note that $f(\cdot)$ must be continuous for the two definitions to be equivalent. See Exercise B.8 for an example with discontinuous $f(\cdot)$ where the system is GAS in the classical sense but does not satisfy (B.6), i.e., is not GAS in the KL sense.

For a KL version of exponential stability, one simply restricts the form of the KL function $\beta(\cdot)$ of asymptotic stability to $\beta(|x|_{\mathcal{A}}, i) = c\,|x|_{\mathcal{A}}\,\lambda^i$ with $c > 0$ and $\lambda \in (0,1)$, but, as we see, that is exactly the classical definition so there is no distinction between the two forms for exponential stability.

In practice, global asymptotic stability of $\mathcal{A}$ often cannot be achieved because of state constraints. Hence we have to extend slightly the definitions given above. In the following, let $\mathcal{B}$ denote a unit ball in $\mathbb{R}^n$ with center at the origin.

Definition B.10 (Various forms of stability (constrained))**.** Suppose $X \subset \mathbb{R}^n$ is positive invariant for $x^+ = f(x)$, that $\mathcal{A} \subseteq X$ is closed and positive invariant for $x^+ = f(x)$. Then $\mathcal{A}$ is

(a) locally stable in X if, for each $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $x \in X \cap (\mathcal{A} \oplus \delta\mathcal{B})$, implies $|\phi(i;x)|_{\mathcal{A}} < \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

(b) locally attractive in X if there exists a $\eta > 0$ such that $x \in X \cap (\mathcal{A} \oplus \eta\mathcal{B})$ implies $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$.

(c) attractive in X if $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$ for all $x \in X$.

(d) locally asymptotically stable in X if it is locally stable in X and locally attractive in X.

(e) asymptotically stable in X if it is locally stable in X and attractive in X.

(f) locally exponentially stable in X if there exist $\eta > 0$, $c > 0$, and $\gamma \in (0,1)$ such that $x \in X \cap (\mathcal{A} \oplus \eta\mathcal{B})$ implies $|\phi(i;x)|_{\mathcal{A}} \leq c\,|x|_{\mathcal{A}}\,\gamma^i$ for all $i \in \mathbb{I}_{\geq 0}$.

(g) exponentially stable in X if there exists a $c > 0$ and a $\gamma \in (0,1)$ such that $|\phi(i;x)|_{\mathcal{A}} \leq c\,|x|_{\mathcal{A}}\,\gamma^i$ for all $x \in X$, all $i \in \mathbb{I}_{\geq 0}$.

The assumption that X is positive invariant for $x^+ = f(x)$ ensures

that $\phi(i;x) \in X$ for all $x \in X$, all $i \in \mathbb{I}_{\geq 0}$. The KL version of asymptotic stability in X is the following.

Definition B.11 (Asymptotic stability (constrained, KL version))**.** Suppose that X is positive invariant and the set $\mathcal{A} \subseteq X$ is closed and positive invariant for $x^+ = f(x)$. The set $\mathcal{A}$ is *asymptotically stable in* X for $x^+ = f(x)$ if there exists a $\mathcal{KL}$ function $\beta(\cdot)$ such that, for each $x \in X$

$$|\phi(i;x)|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i) \qquad \forall i \in \mathbb{I}_{\geq 0} \tag{B.7}$$

Finally, we define the *domain of attraction* of an asymptotically stable set $\mathcal{A}$ for the system $x^+ = f(x)$ to be the set of all initial states x such that $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$. We use the term *region of attraction* to denote any set of initial states x such that $|\phi(i;x)|_{\mathcal{A}} \to 0$ as $i \to \infty$. From these definitions, if $\mathcal{A}$ is attractive *in* X, then X is a region of attraction of set $\mathcal{A}$ for the system $x^+ = f(x)$.

B.3 Lyapunov Stability Theory

Energy in a passive electrical or mechanical system provides a useful analogy to Lyapunov stability theory. In a lumped mechanical system, the total mechanical energy is the sum of the potential and kinetic energies. As time proceeds, this energy is dissipated by friction into heat and the total mechanical energy decays to zero at which point the system is in equilibrium. To establish stability or asymptotic stability, Lyapunov theory follows a similar path. If a real-valued function can be found that is positive and decreasing if the state does not lie in the set $\mathcal{A}$, then the state converges to this set as time tends to infinity. We now make this intuitive idea more precise.

B.3.1 Time-Invariant Systems

First we consider the time-invariant (autonomous) model $x^+ = f(x)$.

Definition B.12 (Lyapunov function (unconstrained and constrained))**.** Suppose that X is positive invariant and the set $\mathcal{A} \subseteq X$ is closed and positive invariant for $x^+ = f(x)$, and $f(\cdot)$ is locally bounded. A function $V : X \to \mathbb{R}_{\geq 0}$ is said to be a Lyapunov function in X for the system $x^+ = f(x)$ and set $\mathcal{A}$ if there exist functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, and *contin-*

uous function $\alpha_3 \in \mathcal{PD}$ such that for any $x \in X$

$$V(x) \geq \alpha_1(|x|_{\mathcal{A}}) \tag{B.8}$$

$$V(x) \leq \alpha_2(|x|_{\mathcal{A}}) \tag{B.9}$$

$$V(f(x)) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}}) \tag{B.10}$$

If $X = \mathbb{R}^n$, then we drop the restrictive phrase "in X."

Remark (Discontinuous f and V)**.** In MPC, the value function for the optimal control problem solved online is often employed as a Lyapunov function. The reader should be aware that many similar but different definitions of Lyapunov functions are in use in many different branches of the science and engineering literature. To be of the most use in MPC analysis, we do not assume here that $f(\cdot)$ or $V(\cdot)$ is continuous. We assume only that $f(\cdot)$ is locally bounded; $V(\cdot)$ is also locally bounded due to (B.9), and continuous on the set $\mathcal{A}$ (but not necessarily on a neighborhood of $\mathcal{A}$) due to (B.8)–(B.9).

Remark (Continuous (and positive definite) α_3)**.** One may wonder why $\alpha_3(\cdot)$ is assumed continuous in addition to positive definite in the definition of the Lyapunov function, when much of the classical literature leaves out continuity; see for example the autonomous case given in Kalman and Bertram (1960). Again, most of this classical literature assumes instead that $f(\cdot)$ is continuous, which we do not assume here. See Exercise B.7 for an example from Lazar, Heemels, and Teel (2009) with discontinuous $f(\cdot)$ for which removing continuity of $\alpha_3(\cdot)$ in Definition B.12 would give a Lyapunov function that fails to imply asymptotic stability.

For making connections to the wide body of existing stability literature, which mainly uses the classical definition of asymptotic stability, and because the proof is instructive, we first state and prove the classical version of the Lyapunov stability theorem.

Theorem B.13 (Lyapunov function and GAS (classical definition))**.** *Suppose that X is positive invariant and the set $\mathcal{A} \subseteq X$ is closed and positive invariant for $x^+ = f(x)$, and $f(\cdot)$ is locally bounded. Suppose $V(\cdot)$ is a Lyapunov function for $x^+ = f(x)$ and set $\mathcal{A}$. Then $\mathcal{A}$ is globally asymptotically stable (classical definition).*

Proof.

(a) Stability. Let $\varepsilon > 0$ be arbitrary and let $\delta := \alpha_2^{-1}(\alpha_1(\varepsilon))$. Suppose $|x|_{\mathcal{A}} < \delta$ so that, by (B.9), $V(x) \leq \alpha_2(\delta) = \alpha_1(\varepsilon)$. From (B.10),

$(V(x(i)))_{i\in\mathbb{I}_{\geq 0}}$, $x(i) := \phi(i;x)$, is a nonincreasing sequence so that, for all $i \in \mathbb{I}_{\geq 0}$, $V(x(i)) \leq V(x)$. From (B.8), $|x(i)|_{\mathcal{A}} \leq \alpha_1^{-1}(V(x)) \leq \alpha_1^{-1}(\alpha_1(\varepsilon)) = \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

(b) Attractivity. Let $x \in \mathbb{R}^n$ be arbitrary. From (B.9) $V(x)$ is finite, and from (B.8) and (B.10), the sequence $(V(x(i)))_{i\in\mathbb{I}_{\geq 0}}$ is nonincreasing and bounded below by zero and therefore converges to $\bar{V} \geq 0$ as $i \to \infty$. We next show that $\bar{V} = 0$. From (B.8) and (B.9) and the properties of K_∞ functions, we have that for all $i \geq 0$,

$$\alpha_2^{-1}(V(x(i))) \leq |x(i)|_{\mathcal{A}} \leq \alpha_1^{-1}(V(x(i))) \tag{B.11}$$

Assume for contradiction that $\bar{V} > 0$. Since $\alpha_3(\cdot)$ is continuous and positive definite and interval $\mathcal{I} := [\alpha_2^{-1}(\bar{V}), \alpha_1^{-1}(\bar{V})]$ is compact, the following optimization has a positive solution

$$\rho := \min_{|x|_{\mathcal{A}}\in\mathcal{I}} \alpha_3(|x|_{\mathcal{A}}) > 0$$

From repeated use of (B.10), we have that for all $i \geq 0$

$$V(x(i)) \leq V(x) - \sum_{j=0}^{i-1} \alpha_3(\,|x(j)|_{\mathcal{A}})$$

Since $|x(i)|_{\mathcal{A}}$ converges to interval $\mathcal{I}$ where $\alpha_3(|x(i)|_{\mathcal{A}})$ is underbounded by $\rho > 0$, $\alpha_3(\cdot)$ is continuous, and $V(x)$ is finite, the inequality above implies that $V(x(i)) \to -\infty$ as $i \to \infty$, which is a contradiction. Therefore $V(x(i))$ converges to $\bar{V} = 0$ and (B.11) implies $x(i)$ converges to $\mathcal{A}$ as $i \to \infty$. ■

Next we establish the analogous Lyapunov stability theorem using the stronger KL definition of GAS, Definition B.8. Before establishing the Lyapunov stability theorem, it is helpful to present the following lemma established by Jiang and Wang (2002, Lemma 2.8) that enables us to assume when convenient that $\alpha_3(\cdot)$ in (B.10) is a $\mathcal{K}_\infty$ function rather than just a continuous $\mathcal{PD}$ function.

Lemma B.14 (From $\mathcal{PD}$ to $\mathcal{K}_\infty$ function (Jiang and Wang (2002))). *Assume $V(\cdot)$ is a Lyapunov function for system $x^+ = f(x)$ and set $\mathcal{A}$, and $f(\cdot)$ is locally bounded. Then there exists a smooth function*[5] *$\rho(\cdot) \in \mathcal{K}_\infty$ such that $W(\cdot) := \rho \circ V(\cdot)$ is also a Lyapunov function for system $x^+ = f(x)$ and set $\mathcal{A}$ that satisfies for all $x \in \mathbb{R}^n$*

$$W(f(x)) - W(x) \leq -\alpha(|x|_{\mathcal{A}})$$

[5] A smooth function has derivatives of all orders.

with $\alpha(\cdot) \in \mathcal{K}_\infty$.

Note that Jiang and Wang (2002) prove this lemma under the assumption that both $f(\cdot)$ and $V(\cdot)$ are continuous, but their proof remains valid if both $f(\cdot)$ and $V(\cdot)$ are only locally bounded.

We next establish the Lyapunov stability theorem in which we add the parenthetical (KL definition) purely for emphasis and to distinguish this result from the previous classical result, but we discontinue this emphasis after this theorem, and use exclusively the KL definition.

Theorem B.15 (Lyapunov function and global asymptotic stability (KL definition))**.** *Suppose that X is positive invariant and the set $\mathcal{A} \subseteq X$ is closed and positive invariant for $x^+ = f(x)$, and $f(\cdot)$ is locally bounded. Suppose $V(\cdot)$ is a Lyapunov function for $x^+ = f(x)$ and set $\mathcal{A}$. Then $\mathcal{A}$ is globally asymptotically stable (KL definition).*

Proof. Due to Lemma B.14 we assume without loss of generality that $\alpha_3 \in \mathcal{K}_\infty$. From (B.10) we have that

$$V(\phi(i+1;x)) \leq V(\phi(i;x)) - \alpha_3(\left|\phi(i;x)\right|_{\mathcal{A}}) \qquad \forall x \in \mathbb{R}^n \quad i \in \mathbb{I}_{\geq 0}$$

Using (B.9) we have that

$$\alpha_3(|x|_{\mathcal{A}}) \geq \alpha_3 \circ \alpha_2^{-1}(V(x)) \qquad \forall x \in \mathbb{R}^n$$

Combining these we have that

$$V(\phi(i+1;x)) \leq \sigma_1(V(\phi(i;x))) \qquad \forall x \in \mathbb{R}^n \quad i \in \mathbb{I}_{\geq 0}$$

in which

$$\sigma_1(s) := s - \alpha_3 \circ \alpha_2^{-1}(s)$$

We have that $\sigma_1(\cdot)$ is continuous on $\mathbb{R}_{\geq 0}$, $\sigma_1(0) = 0$, and $\sigma_1(s) < s$ for $s > 0$. But $\sigma_1(\cdot)$ may not be increasing. We modify σ_1 to achieve this property in two steps. First define

$$\sigma_2(s) := \max_{s' \in [0,s]} \sigma_1(s') \qquad s \in \mathbb{R}_{\geq 0}$$

in which the maximum exists for each $s \in \mathbb{R}_{\geq 0}$ because $\sigma_1(\cdot)$ is continuous. By its definition, $\sigma_2(\cdot)$ is nondecreasing, $\sigma_2(0) = 0$, and $0 \leq \sigma_2(s) < s$ for $s > 0$, and we next show that $\sigma_2(\cdot)$ is continuous on $\mathbb{R}_{\geq 0}$. Assume that $\sigma_2(\cdot)$ is discontinuous at a point $c \in \mathbb{R}_{\geq 0}$. Because it is a nondecreasing function, there is a positive jump in the

function $\sigma_2(\cdot)$ at c (Bartle and Sherbert, 2000, p. 150). Define [6]

$$a_1 \coloneqq \lim_{s \nearrow c} \sigma_2(s) \qquad a_2 \coloneqq \lim_{s \searrow c} \sigma_2(s)$$

We have that $\sigma_1(c) \leq a_1 < a_2$ or we violate the limit of σ_2 from below. Since $\sigma_1(c) < a_2$, $\sigma_1(s)$ must achieve value a_2 for some $s < c$ or we violate the limit from above. But $\sigma_1(s) = a_2$ for $s < c$ also violates the limit from below, and we have a contradiction and $\sigma_2(\cdot)$ is continuous. Finally, define

$$\sigma(s) \coloneqq (1/2)(s + \sigma_2(s)) \qquad s \in \mathbb{R}_{\geq 0}$$

and we have that $\sigma(\cdot)$ is a continuous, strictly increasing, and unbounded function satisfying $\sigma(0) = 0$. Therefore, $\sigma(\cdot) \in \mathcal{K}_\infty$, $\sigma_1(s) < \sigma(s) < s$ for $s > 0$ and therefore

$$V(\phi(i+1;x)) \leq \sigma(V(\phi(i;x))) \qquad \forall x \in \mathbb{R}^n \quad i \in \mathbb{I}_{\geq 0} \tag{B.12}$$

Repeated use of (B.12) and then (B.9) gives

$$V(\phi(i;x)) \leq \sigma^i \circ \alpha_2(|x|_{\mathcal{A}}) \qquad \forall x \in \mathbb{R}^n \quad i \in \mathbb{I}_{\geq 0}$$

in which σ^i represents the composition of σ with itself i times. Using (B.8) we have that

$$\left|\phi(i;x)\right|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i) \qquad \forall x \in \mathbb{R}^n \quad i \in \mathbb{I}_{\geq 0}$$

in which

$$\beta(s,i) \coloneqq \alpha_1^{-1} \circ \sigma^i \circ \alpha_2(s) \qquad \forall s \in \mathbb{R}_{\geq 0} \quad i \in \mathbb{I}_{\geq 0}$$

For all $s \geq 0$, the sequence $w_i \coloneqq \sigma^i(\alpha_2(s))$ is nonincreasing with i, bounded below (by zero), and therefore converges to a, say, as $i \to \infty$. Therefore, both $w_i \to a$ and $\sigma(w_i) \to a$ as $i \to \infty$. Since $\sigma(\cdot)$ is continuous we also have that $\sigma(w_i) \to \sigma(a)$ so $\sigma(a) = a$, which implies that $a = 0$, and we have shown that for all $s \geq 0$, $\alpha_1^{-1} \circ \sigma^i \circ \alpha_2(s) \to 0$ as $i \to \infty$. Since $\alpha_1^{-1}(\cdot)$ also is a $\mathcal{K}$ function, we also have that for all $s \geq 0$, $\alpha_1^{-1} \circ \sigma^i \circ \alpha_2(s)$ is nonincreasing with i. We have from the properties of $\mathcal{K}$ functions that for all $i \geq 0$, $\alpha_1^{-1} \circ \sigma^i \circ \alpha_2(s)$ is a $\mathcal{K}$ function, and can therefore conclude that $\beta(\cdot)$ is a $\mathcal{KL}$ function and the proof is complete. ■

[6] The limits from above and below exist because $\sigma_2(\cdot)$ is nondecreasing (Bartle and Sherbert, 2000, p. 149). If the point $c = 0$, replace the limit from below by $\sigma_2(0)$.

Theorem B.15 provides merely a sufficient condition for global asymptotic stability that might be thought to be conservative. Next we establish a *converse* stability theorem that demonstrates necessity. In this endeavor we require a useful preliminary result on $\mathcal{KL}$ functions (Sontag, 1998b, Proposition 7)

Proposition B.16 (Improving convergence (Sontag (1998b))). *Assume that $\beta(\cdot) \in \mathcal{KL}$. Then there exists $\theta_1(\cdot), \theta_2(\cdot) \in \mathcal{K}_\infty$ so that*

$$\beta(s,t) \leq \theta_1(\theta_2(s)e^{-t}) \qquad \forall s \geq 0, \quad \forall t \geq 0 \tag{B.13}$$

Theorem B.17 (Converse theorem for global asymptotic stability). *Suppose that the (closed, positive invariant) set $\mathcal{A}$ is globally asymptotically stable for the system $x^+ = f(x)$. Then there exists a Lyapunov function for the system $x^+ = f(x)$ and set $\mathcal{A}$.*

Proof. Since the set $\mathcal{A}$ is GAS we have that for each $x \in \mathbb{R}^n$ and $i \in \mathbb{I}_{\geq 0}$

$$\left| \phi(i;x)_{\mathcal{A}} \right| \leq \beta(|x|_{\mathcal{A}}, i)$$

in which $\beta(\cdot) \in \mathcal{KL}$. Using (B.13) then gives for each $x \in \mathbb{R}^n$ and $i \in \mathbb{I}_{\geq 0}$

$$\theta_1^{-1}\left(\left| \phi(i;x) \right|_{\mathcal{A}} \right) \leq \theta_2(|x|_{\mathcal{A}})e^{-i}$$

in which $\theta_1^{-1}(\cdot) \in \mathcal{K}_\infty$. Propose as Lyapunov function

$$V(x) = \sum_{i=0}^{\infty} \theta_1^{-1}\left(\left| \phi(i;x) \right|_{\mathcal{A}} \right)$$

Since $\phi(0;x) = x$, we have that $V(x) \geq \theta_1^{-1}(|x|_{\mathcal{A}})$ and we choose $\alpha_1(\cdot) = \theta_1^{-1}(\cdot) \in \mathcal{K}_\infty$. Performing the sum gives

$$V(x) = \sum_{i=0}^{\infty} \theta_1^{-1}\left(\left| \phi(i;x)_{\mathcal{A}} \right| \right) \leq \theta_2(|x|_{\mathcal{A}}) \sum_{i=0}^{\infty} e^{-i} = \theta_2(|x|_{\mathcal{A}}) \frac{e}{e-1}$$

and we choose $\alpha_2(\cdot) = (e/(e-1))\theta_2(\cdot) \in \mathcal{K}_\infty$. Finally, noting that $f(\phi(i;x)) = \phi(i+1;x)$ for each $x \in \mathbb{R}^n$, $i \in \mathbb{I}_{\geq 0}$, we have that

$$\begin{aligned} V(f(x)) - V(x) &= \sum_{i=0}^{\infty} \theta_1^{-1}\left(\left| f(\phi(i;x)) \right|_{\mathcal{A}} \right) - \theta_1^{-1}\left(\left| \phi(i;x) \right|_{\mathcal{A}} \right) \\ &= -\theta_1^{-1}\left(\left| \phi(0;x) \right|_{\mathcal{A}} \right) \\ &= -\theta_1^{-1}(|x|_{\mathcal{A}}) \end{aligned}$$

and we choose $\alpha_3(\cdot) = \theta_1^{-1}(\cdot) \in \mathcal{K}_\infty$, and the result is established. ■

The appropriate generalization of Theorem B.15 for the constrained case is:

Theorem B.18 (Lyapunov function for asymptotic stability (constrained))**.** *If there exists a Lyapunov function in X for the system $x^+ = f(x)$ and set $\mathcal{A}$, then $\mathcal{A}$ is asymptotically stable in X for $x^+ = f(x)$.*

The proof of this result is similar to that of Theorem B.15 and is left as an exercise.

Theorem B.19 (Lyapunov function for exponential stability)**.** *If there exists $V : X \to \mathbb{R}_{\geq 0}$ satisfying the following properties for all $x \in X$*

$$a_1 |x|_{\mathcal{A}}^{\sigma} \leq V(x) \leq a_2 |x|_{\mathcal{A}}^{\sigma}$$
$$V(f(x)) - V(x) \leq -a_3 |x|_{\mathcal{A}}^{\sigma}$$

in which $a_1, a_2, a_3, \sigma > 0$, then $\mathcal{A}$ is exponentially stable in X for $x^+ = f(x)$.

Linear time-invariant systems. We review some facts involving the discrete matrix Lyapunov equation and stability of the linear system

$$x^+ = Ax$$

in which $x \in \mathbb{R}^n$. The discrete time system is asymptotically stable if and only if the magnitudes of the eigenvalues of A are strictly less than unity. Such an A matrix is called stable, convergent, or discrete time Hurwitz.

In the following, $A, S, Q \in \mathbb{R}^{n \times n}$. The following matrix equation is known as a discrete matrix Lyapunov equation,

$$A'SA - S = -Q$$

The properties of solutions to this equation allow one to draw conclusions about the stability of A without computing its eigenvalues. Sontag (1998a, p. 231) provides the following lemma

Lemma B.20 (Lyapunov function for linear systems)**.** *The following statements are equivalent (Sontag, 1998a).*

(a) A is stable.

(b) For each $Q \in \mathbb{R}^{n \times n}$, there is a unique solution S of the discrete matrix Lyapunov equation

$$A'SA - S = -Q$$

and if $Q > 0$ then $S > 0$.

(c) There is some $S > 0$ such that $A'SA - S < 0$.

(d) There is some $S > 0$ such that $V(x) = x'Sx$ is a Lyapunov function for the system $x^+ = Ax$.

Exercise B.1 asks you to establish the equivalence of (a) and (b).

B.3.2 Time-Varying, Constrained Systems

Following the discussion in Rawlings and Risbeck (2017), we consider the nonempty sets $X(i) \subseteq \mathbb{R}^n$ indexed by $i \in \mathbb{I}_{\geq 0}$. We define the time-varying system

$$x^+ = f(x, i)$$

with $f(\cdot, i) : X(i) \to X(i+1)$. We assume that $f(\cdot, i)$ is locally bounded for all $i \in \mathbb{I}_{\geq 0}$. Note from the definition of f that the sets $X(i)$ satisfy positive invariance in the following sense: $x \in X(i)$ for any $i \geq 0$ implies $x(i+1) := f(x, i) \in X(i+1)$. We say that the set sequence $(X(i))_{i \geq 0}$ is *sequentially* positive invariant to denote this form of invariance.

Definition B.21 (Sequential positive invariance)**.** A sequence of sets $(X(i))_{i \geq 0}$ is sequentially positive invariant for the system $x^+ = f(x, i)$ if for any $i \geq 0$, $x \in X(i)$ implies $f(x, i) \in X(i+1)$.

We again assume that $\mathcal{A}$ is closed and positive invariant for the time-varying system, i.e, $x \in \mathcal{A}$ at any time $i \geq 0$ implies $f(x, i) \in \mathcal{A}$. We also assume that $\mathcal{A} \subseteq X(i)$ for all $i \geq 0$. We next define asymptotic stability of $\mathcal{A}$.

Definition B.22 (Asymptotic stability (time-varying, constrained))**.** Suppose that the sequence $(X(i))_{i \geq 0}$ is sequentially positive invariant and the set $\mathcal{A} \subseteq X(i)$ for all $i \geq 0$ is closed and positive invariant for $x^+ = f(x, i)$. The set $\mathcal{A}$ is *asymptotically stable* in $X(i)$ at each time $i \geq 0$ for $x^+ = f(x, i)$ if the following holds for all $i \geq i_0 \geq 0$, and $x \in X(i_0)$

$$\left| \phi(i; x, i_0) \right|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i - i_0) \tag{B.14}$$

in which $\beta \in \mathcal{KL}$ and $\phi(i; x, i_0)$ is the solution to $x^+ = f(x, i)$ at time $i \geq i_0$ with initial condition x at time $i_0 \geq 0$.

This stability definition is somewhat restrictive because $\phi(i; x, i_0)$ is bounded by a function depending on $i - i_0$ rather than on i. For example, to be more general we could define a time-dependent set of

$\mathcal{KL}$ functions, $\beta_j(\cdot)$, $j \geq 0$, and replace (B.14) with $|\phi(i;x,i_0)|_{\mathcal{A}} \leq \beta_{i_0}(|x|_{\mathcal{A}}, i)$ for all $i \geq i_0 \geq 0$.

We define a time-varying Lyapunov function for this system as follows.

Definition B.23 (Lyapunov function: time-varying, constrained case)**.** Let the sequence $(X(i))_{i\geq 0}$ be sequentially positive invariant, and the set $\mathcal{A} \subseteq X(i)$ for all $i \geq 0$ be closed and positive invariant. Let $V(\cdot, i) : X(i) \to \mathbb{R}_{\geq 0}$ satisfy for all $x \in X(i), i \in \mathbb{I}_{\geq 0}$

$$\begin{aligned} \alpha_1(|x|_{\mathcal{A}}) \leq V(x,i) \leq \quad \alpha_2(|x|_{\mathcal{A}}) \\ V(f(x,i),i+1) - V(x,i) \leq -\alpha_3(|x|_{\mathcal{A}}) \end{aligned}$$

with $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{K}_\infty$. Then $V(\cdot,\cdot)$ is a time-varying Lyapunov function in the sequence $(X(i))_{i\geq 0}$ for $x^+ = f(x,i)$ and set $\mathcal{A}$.

Note that $f(x,i) \in X(i+1)$ since $x \in X(i)$ which verifies that $V(f(x,i),i+1)$ is well defined for all $x \in X(i), i \geq 0$. We then have the following asymptotic stability result for the time-varying, constrained case.

Theorem B.24 (Lyapunov theorem for asymptotic stability (time-varying, constrained))**.** *Let the sequence $(X(i))_{i\geq 0}$ be sequentially positive invariant, and the set $\mathcal{A} \subseteq X(i)$ for all $i \geq 0$ be closed and positive invariant, and $V(\cdot,\cdot)$ be a time-varying Lyapunov function in the sequence $(X(i))_{i\geq 0}$ for $x^+ = f(x,i)$ and set $\mathcal{A}$. Then $\mathcal{A}$ is asymptotically stable in $X(i)$ at each time $i \geq 0$ for $x^+ = f(x,i)$.*

Proof. For $x \in X(i_0)$, we have that $(\phi(i;x,i_0),i) \in X(i)$ for all $i \geq i_0$. From the first and second inequalities we have that for all $i \geq i_0$ and $x \in X(i_0)$

$$\begin{aligned} V(\phi(i+1;x,i_0),i+1) &\leq V(\phi(i;x,i_0),i) - \alpha_3(|\phi(i;x,i_0)|_{\mathcal{A}}) \\ &\leq \sigma_1(V(\phi(i;x,i_0),i)) \end{aligned}$$

with $\sigma_1(s) \coloneqq s - \alpha_3 \circ \alpha_2^{-1}(s)$. Note that $\sigma_1(\cdot)$ may not be $\mathcal{K}_\infty$ because it may not be increasing. But given this result we can find, as in the proof of Theorem B.15, $\sigma(\cdot) \in \mathcal{K}_\infty$ satisfying $\sigma_1(s) < \sigma(s) < s$ for all $s \in \mathbb{R}_{>0}$ such that $V(\phi(i+1;x,i_0),i+1) \leq \sigma(V(\phi(i;x,i_0),i))$. We then have that

$$|\phi(i;x,i_0)|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i-i_0) \qquad \forall x \in X(i_0), \quad i \geq i_0$$

in which $\beta(s,i) \coloneqq \alpha_1^{-1} \circ \sigma^i \circ \alpha_2(s)$ for $s \in \mathbb{R}_{\geq 0}, i \geq 0$ is a $\mathcal{KL}$ function, and the result is established. ∎

B.3.3 Upper bounding $\mathcal{K}$ functions

In using Lyapunov functions for stability analysis, we often have to establish that the upper bound inequality holds on some closed set. The following result proves useful in such situations.

Proposition B.25 (Global K function overbound)**.** *Let $X \subseteq \mathbb{R}^n$ be closed and suppose that a function $V : X \to \mathbb{R}_{\geq 0}$ is continuous at $x_0 \in X$ and locally bounded on X, i.e., bounded on every compact subset of X. Then, there exists a K function α such that*

$$|V(x) - V(x_0)| \leq \alpha(|x - x_0|) \quad \textit{for all } x \in X$$

A proof is given in Rawlings and Risbeck (2015).

B.4 Robust Stability

We now turn to the task of obtaining stability conditions for discrete time systems subject to disturbances. There are two separate questions that should be addressed. The first is *nominal* robustness; is asymptotic stability of a set $\mathcal{A}$ for a (nominal) system $x^+ = f(x)$ maintained in the presence of arbitrarily small disturbances? The second question is the determination of conditions for asymptotic stability of a set $\mathcal{A}$ for a system perturbed by disturbances lying in a given compact set.

B.4.1 Nominal Robustness

Here we follow Teel (2004). The nominal system is $x^+ = f(x)$. Consider the perturbed system

$$x^+ = f(x + e) + w \tag{B.15}$$

where e is the state error and w the additive disturbance. Let $\mathbf{e} := (e(0), e(1), \ldots)$ and $\mathbf{w} := (w(0), w(1), \ldots)$ denote the disturbance sequences with norms $\|\mathbf{e}\| := \sup_{i \geq 0} |e(i)|$ and $\|\mathbf{w}\| := \sup_{i \geq 0} |w(i)|$. Let $M_\delta := \{(\mathbf{e}, \mathbf{w}) \mid \|\mathbf{e}\| \leq \delta, \|\mathbf{w}\| \leq \delta\}$ and, for each $x \in \mathbb{R}^n$, let S_δ denote the set of solutions $\phi(\cdot; x, \mathbf{e}, \mathbf{w})$ of (B.15) with initial state x (at time 0) and perturbation sequences $(\mathbf{e}, \mathbf{w}) \in M_\delta$. A closed, compact set $\mathcal{A}$ is *nominally* robustly asymptotically stable for the (nominal) system $x^+ = f(x)$ if a small neighborhood of $\mathcal{A}$ is locally stable and attractive for all sufficiently small perturbation sequences. We use the adjective *nominal* to indicate that we are examining how a system $x^+ = f(x)$ for which $\mathcal{A}$ is known to be asymptotically stable behaves when subjected to small disturbances. More precisely Teel (2004):

Definition B.26 (Nominal robust global asymptotic stability)**.** The closed, compact set $\mathcal{A}$ is said to be nominally robustly globally asymptotically stable (nominally RGAS) for the system $x^+ = f(x)$ if there exists a $\mathcal{KL}$ function $\beta(\cdot)$ and, for each $\varepsilon > 0$ and each compact set X, there exists a $\delta > 0$ such that, for each $x \in X$ and each solution $\phi(\cdot)$ of the perturbed system lying in S_δ, $|\phi(i)|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i) + \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

Thus, for each $\varepsilon > 0$, there exists a $\delta > 0$ such that each solution $\phi(\cdot)$ of $x^+ = f(x+e)+w$ starting in a δ neighborhood of $\mathcal{A}$ remains in a $\beta(\delta, 0) + \varepsilon$ neighborhood of $\mathcal{A}$, and each solution starting anywhere in $\mathbb{R}^n$ converges to a ε neighborhood of $\mathcal{A}$. These properties are a necessary relaxation (because of the perturbations) of local stability and global attractivity.

Remark. What we call "nominally robustly globally asymptotically stable" in the above definition is called "robustly globally asymptotically stable" in Teel (2004); we use the term "nominal" to indicate that we are concerned with the effect of perturbations e and w on the stability properties of a "nominal" system $x^+ = f(x)$ for which asymptotic stability of a set $\mathcal{A}$ has been established (in the absence of perturbations). We use the expression "$\mathcal{A}$ is globally asymptotically stable for $x^+ = f(x + e) + w$" to refer to the case when asymptotic stability of a set $\mathcal{A}$ has been established for the perturbed system $x^+ = f(x+e)+w$.

The following result, where we add the adjective "nominal", is established in (Teel, 2004, Theorem 2):

Theorem B.27 (Nominal robust global asymptotic stability and Lyapunov function)**.** *Suppose set $\mathcal{A}$ is closed and compact and $f(\cdot)$ is locally bounded. Then the set $\mathcal{A}$ is nominally robustly globally asymptotically stable for the system $x^+ = f(x)$ if and only if there exists a continuous (in fact, smooth) Lyapunov function for $x^+ = f(x)$ and set $\mathcal{A}$.*

The significance of this result is that while a nonrobust system, for which $\mathcal{A}$ is globally asymptotically stable, has a Lyapunov function, that function is *not* continuous. For the globally asymptotically stable example $x^+ = f(x)$ discussed in Section 3.2 of Chapter 3, where $f(x) = (0, |x|)$ when $x_1 \neq 0$ and $f(x) = (0, 0)$ otherwise, one Lyapunov function $V(\cdot)$ is $V(x) = 2|x|$ if $x_1 \neq 0$ and $V(x) = |x|$ if $x_1 = 0$. That $V(\cdot)$ is a Lyapunov function follows from the fact that it satisfies $V(x) \geq |x|$, $V(x) \leq 2|x|$ and $V(f(x)) - V(x) = -|x|$ for all $x \in \mathbb{R}^2$.

It follows immediately from its definition that $V(\cdot)$ is not continuous; but we can also deduce from Theorem B.27 that every Lyapunov function for this system is not continuous since, as shown in Section 3.2 of Chapter 3, global asymptotic stability for this system is not robust. Theorem B.27 shows that existence of a continuous Lyapunov function guarantees nominal robustness. Also, it follows from Theorem B.17 that there exists a smooth Lyapunov function for $x^+ = f(x)$ if $f(\cdot)$ is continuous and $\mathcal{A}$ is GAS for $x^+ = f(x)$. Since $f(\cdot)$ is locally bounded if it is continuous, it then follows from Theorem B.27 that $\mathcal{A}$ is nominally robust GAS for $x^+ = f(x)$ if it is GAS and $f(\cdot)$ is continuous.

B.4.2 Robustness

We turn now to stability conditions for systems subject to bounded disturbances (not vanishingly small) and described by

$$x^+ = f(x, w) \tag{B.16}$$

where the disturbance w lies in the compact set $\mathbb{W}$. This system may equivalently be described by the difference inclusion

$$x^+ \in F(x) \tag{B.17}$$

where the set $F(x) := \{f(x, w) \mid w \in \mathbb{W}\}$. Let $S(x)$ denote the set of all solutions of (B.16) or (B.17) with initial state x. We require, in the sequel, that the closed set $\mathcal{A}$ is positive invariant for (B.16) (or for $x^+ \in F(x)$):

Definition B.28 (Positive invariance with disturbances)**.** The closed set $\mathcal{A}$ is positive invariant for $x^+ = f(x, w)$, $w \in \mathbb{W}$ if $x \in \mathcal{A}$ implies $f(x, w) \in \mathcal{A}$ for all $w \in \mathbb{W}$; it is positive invariant for $x^+ \in F(x)$ if $x \in \mathcal{A}$ implies $F(x) \subseteq \mathcal{A}$.

Clearly the two definitions are equivalent; $\mathcal{A}$ is positive invariant for $x^+ = f(x, w)$, $w \in \mathbb{W}$, if and only if it is positive invariant for $x^+ \in F(x)$.

Remark. In the MPC literature, but not necessarily elsewhere, the term robust positive invariant is often used in place of positive invariant to emphasize that positive invariance is maintained despite the presence of the disturbance w. However, since the uncertain system $x^+ = f(x, w)$, $w \in \mathbb{W}$ is specified ($x^+ = f(x, w)$, $w \in \mathbb{W}$ or $x^+ \in F(x)$) in the assertion that a closed set $\mathcal{A}$ is positive invariant, the word "robust"

appears to be unnecessary. In addition, in the systems literature, the closed set $\mathcal{A}$ is said to be robust positive invariant for $x^+ \in F(x)$ if it satisfies conditions similar to those of Definition B.26 with $x^+ \in F(x)$ replacing $x^+ = f(x)$; see Teel (2004), Definition 3.

In Definitions B.29–B.31, we use "positive invariant" to denote "positive invariant for $x^+ = f(x, w)$, $w \in \mathbb{W}$" or for $x^+ \in F(x)$.

Definition B.29 (Local stability (disturbances))**.** The closed, positive invariant set $\mathcal{A}$ is *locally stable* for $x^+ = f(x, w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$) if, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that, for each x satisfying $|x|_{\mathcal{A}} < \delta$, each solution $\phi \in S(x)$ satisfies $|\phi(i)|_{\mathcal{A}} < \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$.

Definition B.30 (Global attraction (disturbances))**.** The closed, positive invariant set $\mathcal{A}$ is *globally attractive* for the system $x^+ = f(x, w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$) if, for each $x \in \mathbb{R}^n$, each solution $\phi(\cdot) \in S(x)$ satisfies $|\phi(i)|_{\mathcal{A}} \to 0$ as $i \to \infty$.

Definition B.31 (GAS (disturbances))**.** The closed, positive invariant set $\mathcal{A}$ is *globally asymptotically stable* for $x^+ = f(x, w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$) if it is locally stable and globally attractive.

An alternative definition of global asymptotic stability of closed set $\mathcal{A}$ for $x^+ = f(x, w)$, $w \in \mathbb{W}$, if $\mathcal{A}$ is compact, is the existence of a $\mathcal{KL}$ function $\beta(\cdot)$ such that for each $x \in \mathbb{R}^n$, each $\phi \in S(x)$ satisfies $|\phi(i)|_{\mathcal{A}} \leq \beta(|x|_{\mathcal{A}}, i)$ for all $i \in \mathbb{I}_{\geq 0}$. To cope with disturbances we require a modified definition of a Lyapunov function.

Definition B.32 (Lyapunov function (disturbances))**.** A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is said to be a Lyapunov function for the system $x^+ = f(x, w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$) and closed set $\mathcal{A}$ if there exist functions $\alpha_i \in \mathcal{K}_\infty$, $i = 1, 2, 3$ such that for any $x \in \mathbb{R}^n$,

$$V(x) \geq \alpha_1(|x|_{\mathcal{A}}) \tag{B.18}$$

$$V(x) \leq \alpha_2(|x|_{\mathcal{A}}) \tag{B.19}$$

$$\sup_{z \in F(x)} V(z) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}}) \tag{B.20}$$

Remark. Without loss of generality, we can choose the function $\alpha_3(\cdot)$ in (B.20) to be a class $\mathcal{K}_\infty$ function if $f(\cdot)$ is continuous (see Jiang and Wang (2002), Lemma 2.8).

Inequality B.20 ensures $V(f(x,w)) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})$ for all $w \in \mathbb{W}$. The existence of a Lyapunov function for the system $x^+ \in F(x)$ and closed set $\mathcal{A}$ is a sufficient condition for $\mathcal{A}$ to be globally asymptotically stable for $x^+ \in F(x)$ as shown in the next result.

Theorem B.33 (Lyapunov function for global asymptotic stability (disturbances))**.** *Suppose $V(\cdot)$ is a Lyapunov function for $x^+ = f(x,w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$) and closed set $\mathcal{A}$ with $\alpha_3(\cdot)$ a $\mathcal{K}_\infty$ function. Then $\mathcal{A}$ is globally asymptotically stable for $x^+ = f(x,w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x)$).*

Proof. (i) Local stability: Let $\varepsilon > 0$ be arbitrary and let $\delta := \alpha_2^{-1}(\alpha_1(\varepsilon))$. Suppose $|x|_{\mathcal{A}} < \delta$ so that, by (B.19), $V(x) \leq \alpha_2(\delta) = \alpha_1(\varepsilon)$. Let $\phi(\cdot)$ be any solution in $S(x)$ so that $\phi(0) = x$. From (B.20), $(V(\phi(i)))_{i \in \mathbb{I}_{\geq 0}}$ is a nonincreasing sequence so that, for all $i \in \mathbb{I}_{\geq 0}$, $V(\phi(i)) \leq V(x)$. From (B.18), $|\phi(i)|_{\mathcal{A}} \leq \alpha_1^{-1}(V(x)) \leq \alpha_1^{-1}(\alpha_1(\varepsilon)) = \varepsilon$ for all $i \in \mathbb{I}_{\geq 0}$. (ii) Global attractivity: Let $x \in \mathbb{R}^n$ be arbitrary. Let $\phi(\cdot)$ be any solution in $S(x)$ so that $\phi(0) = x$. From Equations B.18 and B.20, since $\phi(i+1) \in F(\phi(i))$, the sequence $(V(\phi(i)))_{i \in \mathbb{I}_{\geq 0}}$ is nonincreasing and bounded from below by zero. Hence both $V(\phi(i))$ and $V(\phi(i+1))$ converge to $\bar{V} \geq 0$ as $i \to \infty$. But $\phi(i+1) \in F(\phi(i))$ so that, from (B.20), $\alpha_3(|\phi(i)|_{\mathcal{A}}) \to 0$ as $i \to \infty$. Since $|\phi(i)|_{\mathcal{A}} = \alpha_3^{-1}(\alpha_3(|\phi(i)|_{\mathcal{A}}))$ where $\alpha_3^{-1}(\cdot)$ is a $\mathcal{K}_\infty$ function, $|\phi(i)|_{\mathcal{A}} \to 0$ as $i \to \infty$. ■

B.5 Control Lyapunov Functions

A control Lyapunov function is a useful generalization, due to Sontag (1998a, pp.218–233), of a Lyapunov function; while a Lyapunov function is relevant for a system $x^+ = f(x)$ and provides conditions for the (asymptotic) stability of a set for this system, a control Lyapunov function is relevant for a control system $x^+ = f(x,u)$ and provides conditions for the existence of a controller $u = \kappa(x)$ that ensures (asymptotic) stability of a set for the controlled system $x^+ = f(x, \kappa(x))$. Consider the control system

$$x^+ = f(x,u)$$

where the control u is subject to the constraint

$$u \in \mathbb{U}$$

Our standing assumptions in this section are that $f(\cdot)$ is continuous and $\mathbb{U}$ is compact.

Definition B.34 (Global control Lyapunov function (CLF))**.** A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is a global control Lyapunov function for the system $x^+ = f(x,u)$ and closed set $\mathcal{A}$ if there exist $\mathcal{K}_\infty$ functions $\alpha_1(\cdot), \alpha_2(\cdot)$, $\alpha_3(\cdot)$ satisfying for all $x \in \mathbb{R}^n$:

$$\alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}})$$
$$\inf_{u \in \mathbb{U}} V(f(x,u)) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})$$

Definition B.35 (Global stabilizability)**.** Let set $\mathcal{A}$ be compact. The set $\mathcal{A}$ is globally stabilizable for the system $x^+ = f(x,u)$ if there exists a state-feedback function $\kappa : \mathbb{R}^n \to \mathbb{U}$ such that $\mathcal{A}$ is globally asymptotically stable for $x^+ = f(x, \kappa(x))$.

Remark. Given a global control Lyapunov function $V(\cdot)$, one can choose a control law $\kappa : \mathbb{R}^n \to \mathbb{U}$ satisfying

$$V(f(x, \kappa(x))) \leq V(x) - \alpha_3(|x|_{\mathcal{A}})/2$$

for all $x \in \mathbb{R}^n$ (seeTeel (2004)). Since $\mathbb{U}$ is compact, $\kappa(\cdot)$ is locally bounded and, hence, so is $x \mapsto f(x, \kappa(x))$. Thus we may use Theorem B.13 to deduce that $\mathcal{A}$ is globally asymptotically stable for $x^+ = f(x, \kappa(x))$. If $V(\cdot)$ is continuous, one can also establish nominal robustness properties.

In a similar fashion one can extend the concept of control Lyapunov functions to the case when the system is subject to disturbances. Consider the system

$$x^+ = f(x,u,w)$$

where the control u is constrained to lie in $\mathbb{U}$ and the disturbance takes values in the set $\mathbb{W}$. We assume that $f(\cdot)$ is continuous and that $\mathbb{U}$ and $\mathbb{W}$ are compact. The system may be equivalently defined by

$$x^+ \in F(x,u)$$

where the set-valued function $F(\cdot)$ is defined by

$$F(x,u) := \{f(x,u,w) \mid w \in \mathbb{W}\}$$

We can now make the obvious generalizations of the definitions in Section B.4.2.

Definition B.36 (Positive invariance (disturbance and control))**.** The closed set $\mathcal{A}$ is positive invariant for $x^+ = f(x,u,w)$, $w \in \mathbb{W}$ (or for $x^+ \in F(x,u)$) if for all $x \in \mathcal{A}$ there exists a $u \in \mathbb{U}$ such that $f(x,u,w) \in \mathcal{A}$ for all $w \in \mathbb{W}$ (or $F(x,u) \subseteq \mathcal{A}$).

Definition B.37 (CLF (disturbance and control))**.** A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is said to be a control Lyapunov function for the system $x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$ (or $x^+ \in F(x,u)$, $u \in \mathbb{U}$) and set $\mathcal{A}$ if there exist functions $\alpha_i \in \mathcal{K}_\infty$, $i = 1,2,3$ such that for any $x \in \mathbb{R}^n$,

$$\begin{aligned} \alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}}) \\ \inf_{u \in \mathbb{U}} \sup_{z \in F(x,u)} V(z) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}}) \end{aligned} \tag{B.21}$$

Remark (CLF implies control law)**.** Given a global control Lyapunov function $V(\cdot)$, one can choose a control law $\kappa : \mathbb{R}^n \to \mathbb{U}$ satisfying

$$\sup_{z \in F(x,\kappa(x))} V(z) \leq V(x) - \alpha_3(|x|_{\mathcal{A}})/2$$

for all $x \in \mathbb{R}^n$. Since $\mathbb{U}$ is compact, $\kappa(\cdot)$ is locally bounded and, hence, so is $x \mapsto f(x,\kappa(x))$. Thus we may use Theorem B.33 to deduce that $\mathcal{A}$ is globally asymptotically stable for $x^+ = f(x,\kappa(x),w)$, $w \in \mathbb{W}$ (for $x^+ \in F(x,\kappa(x))$).

These results can be further extended to deal with the constrained case. First, we generalize the definitions of positive invariance of a set.

Definition B.38 (Positive invariance (constrained))**.** The closed set $\mathcal{A}$ is control invariant for $x^+ = f(x,u)$, $u \in \mathbb{U}$ if, for all $x \in \mathcal{A}$, there exists a $u \in \mathbb{U}$ such that $f(x,u) \in \mathcal{A}$.

Suppose that the state x is required to lie in the closed set $\mathbb{X} \subset \mathbb{R}^n$. In order to show that it is possible to ensure a decrease of a Lyapunov function, as in (B.21), in the presence of the state constraint $x \in \mathbb{X}$, we assume that there exists a control invariant set $X \subseteq \mathbb{X}$ for $x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$. This enables us to obtain a control law that keeps the state in X and, hence, in $\mathbb{X}$, and, under suitable conditions, to satisfy a variant of (B.21).

Definition B.39 (CLF (constrained))**.** Suppose the set X and closed set $\mathcal{A}$, $\mathcal{A} \subset X$, are control invariant for $x^+ = f(x,u)$, $u \in \mathbb{U}$. A function $V : X \to \mathbb{R}_{\geq 0}$ is said to be a control Lyapunov function in X for the

system $x^+ = f(x,u)$, $u \in \mathbb{U}$, and closed set $\mathcal{A}$ in X if there exist functions $\alpha_i \in \mathcal{K}_\infty$, $i = 1,2,3$, defined on X, such that for any $x \in X$,

$$\alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}})$$
$$\inf_{u \in \mathbb{U}} \{V(f(x,u)) \mid f(x,u) \in X\} - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})$$

Remark. Again, if $V(\cdot)$ is a control Lyapunov function in X for $x^+ = f(x,u)$, $u \in \mathbb{U}$ and closed set $\mathcal{A}$ in X, one can choose a control law $\kappa : \mathbb{R}^n \to \mathbb{U}$ satisfying

$$V(f(x,\kappa(x))) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})/2$$

for all $x \in X$. Since $\mathbb{U}$ is compact, $\kappa(\cdot)$ is locally bounded and, hence, so is $x \mapsto f(x,\kappa(x))$. Thus, when $\alpha_3(\cdot)$ is a $\mathcal{K}_\infty$ function, we may use Theorem B.18 to deduce that $\mathcal{A}$ is asymptotically stable for $x^+ = f(x, \kappa(x))$, $u \in \mathbb{U}$ in X; also $\phi(i;x) \in X \subset \mathbb{X}$ for all $x \in X$, all $i \in \mathbb{I}_{\geq 0}$.

Finally we consider the constrained case in the presence of disturbances. First we define control invariance in the presence of disturbances.

Definition B.40 (Control invariance (disturbances, constrained))**.** The closed set $\mathcal{A}$ is control invariant for $x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$ if, for all $x \in \mathcal{A}$, there exists a $u \in \mathbb{U}$ such that $f(x,u,w) \in \mathcal{A}$ for all $w \in \mathbb{W}$ (or $F(x,u) \subseteq \mathcal{A}$ where $F(x,u) := \{f(x,u,w) \mid w \in \mathbb{W}\}$).

Next, we define what we mean by a control Lyapunov function in this context.

Definition B.41 (CLF (disturbances, constrained))**.** Suppose the set X and closed set $\mathcal{A}$, $\mathcal{A} \subset X$, are control invariant for $x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$. A function $V : X \to \mathbb{R}_{\geq 0}$ is said to be a control Lyapunov function in X for the system $x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$ and set $\mathcal{A}$ if there exist functions $\alpha_i \in \mathcal{K}_\infty$, $i = 1,2,3$, defined on X, such that for any $x \in X$,

$$\alpha_1(|x|_{\mathcal{A}}) \leq V(x) \leq \alpha_2(|x|_{\mathcal{A}})$$
$$\inf_{u \in \mathbb{U}} \sup_{z \in F(x,u) \cap X} V(z) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})$$

Suppose now that the state x is required to lie in the closed set $\mathbb{X} \subset \mathbb{R}^n$. Again, in order to show that there exists a condition similar to (B.21), we assume that there exists a control invariant set $X \subseteq \mathbb{X}$ for

$x^+ = f(x,u,w)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$. This enables us to obtain a control law that keeps the state in X and, hence, in $\mathbb{X}$, and, under suitable conditions, to satisfy a variant of (B.21).

Remark. If $V(\cdot)$ is a control Lyapunov function in X for $x^+ = f(x,u)$, $u \in \mathbb{U}$, $w \in \mathbb{W}$ and set $\mathcal{A}$ in X, one can choose a control law $\kappa : X \to \mathbb{U}$ satisfying

$$\sup_{z \in F(x,\kappa(x))} V(z) - V(x) \leq -\alpha_3(|x|_{\mathcal{A}})/2$$

for all $x \in X$. Since $\mathbb{U}$ is compact, $\kappa(\cdot)$ is locally bounded and, hence, so is $x \mapsto f(x,\kappa(x))$. Thus, when $\alpha_3(\cdot)$ is a $\mathcal{K}_\infty$ function, we may use Theorem B.18 to deduce that $\mathcal{A}$ is asymptotically stable in X for $x^+ = f(x,\kappa(x),w)$, $w \in \mathbb{W}$ (or, equivalently, for $x^+ \in F(x,\kappa(x))$); also $\phi(i) \in X \subset \mathbb{X}$ for all $x \in X$, all $i \in \mathbb{I}_{\geq 0}$, all $\phi \in S(x)$.

B.6 Input-to-State Stability

We consider, as in the previous section, the system

$$x^+ = f(x,w)$$

where the disturbance w takes values in $\mathbb{R}^p$. In input-to-state stability (Sontag and Wang, 1995; Jiang and Wang, 2001) we seek a bound on the state in terms of a uniform bound on the disturbance sequence $\mathbf{w} := (w(0), w(1), \ldots)$. Let $\|\cdot\|$ denote the usual ℓ_∞ norm for sequences, i.e., $\|\mathbf{w}\| := \sup_{k \geq 0} |w(k)|$.

Definition B.42 (Input-to-state stable (ISS))**.** The system $x^+ = f(x,w)$ is (globally) input-to-state stable (ISS) if there exists a $\mathcal{KL}$ function $\beta(\cdot)$ and a $\mathcal{K}$ function $\sigma(\cdot)$ such that, for each $x \in \mathbb{R}^n$, and each disturbance sequence $\mathbf{w} = (w(0), w(1), \ldots)$ in ℓ_∞

$$\left| \phi(i; x, \mathbf{w}_i) \right| \leq \beta(|x|, i) + \sigma(\|\mathbf{w}_i\|)$$

for all $i \in \mathbb{I}_{\geq 0}$, where $\phi(i; x, \mathbf{w}_i)$ is the solution, at time i, if the initial state is x at time 0 and the input sequence is $\mathbf{w}_i := (w(0), w(1), \ldots, w(i-1))$.

We note that this definition implies the origin is globally asymptotically stable if the input sequence is identically zero. Also, the norm of the state is asymptotically bounded by $\sigma(\|\mathbf{w}\|)$ where $\mathbf{w} := (w(0), w(1), \ldots)$. As before, we seek a Lyapunov function that ensures input-to-state stability.

Definition B.43 (ISS-Lyapunov function)**.** A function $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is an ISS-Lyapunov function for system $x^+ = f(x, w)$ if there exist $\mathcal{K}_\infty$ functions $\alpha_1(\cdot), \alpha_2(\cdot), \alpha_3(\cdot)$ and a $\mathcal{K}$ function $\sigma(\cdot)$ such that for all $x \in \mathbb{R}^n$, $w \in \mathbb{R}^p$

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|)$$
$$V(f(x, w)) - V(x) \leq -\alpha_3(|x|) + \sigma(|w|)$$

The following result appears in Jiang and Wang (2001, Lemma 3.5)

Lemma B.44 (ISS-Lyapunov function implies ISS)**.** *Suppose $f(\cdot)$ is continuous and that there exists a continuous ISS-Lyapunov function for $x^+ = f(x, w)$. Then the system $x^+ = f(x, w)$ is ISS.*

The converse, i.e., input-to-state stability implies the existence of a smooth ISS-Lyapunov function for $x^+ = f(x, w)$ is also proved in Jiang and Wang (2002, Theorem 1). We now consider the case when the state satisfies the constraint $x \in \mathbb{X}$ where $\mathbb{X}$ is a closed subset of $\mathbb{R}^n$. Accordingly, we assume that the disturbance w satisfies $w \in \mathbb{W}$ where $\mathbb{W}$ is a compact set containing the origin and that $X \subset \mathbb{X}$ is a closed robust positive invariant set for $x^+ = f(x, w)$, $w \in \mathbb{W}$ or, equivalently, for $x^+ \in F(x, u)$.

Definition B.45 (ISS (constrained))**.** Suppose that $\mathbb{W}$ is a compact set containing the origin and that $X \subset \mathbb{X}$ is a closed robust positive invariant set for $x^+ = f(x, w)$, $w \in \mathbb{W}$. The system $x^+ = f(x, w)$, $w \in \mathbb{W}$ is ISS in X if there exists a class $\mathcal{KL}$ function $\beta(\cdot)$ and a class $\mathcal{K}$ function $\sigma(\cdot)$ such that, for all $x \in X$, all $\mathbf{w} \in \mathcal{W}$ where $\mathcal{W}$ is the set of infinite sequences $\mathbf{w}$ satisfying $w(i) \in \mathbb{W}$ for all $i \in \mathbb{I}_{\geq 0}$

$$\left| \phi(i; x, \mathbf{w}_i) \right| \leq \beta(|x|, i) + \sigma(\|\mathbf{w}_i\|)$$

Definition B.46 (ISS-Lyapunov function (constrained))**.** A function $V : X \to \mathbb{R}_{\geq 0}$ is an ISS-Lyapunov function in X for system $x^+ = f(x, w)$ if there exist $\mathcal{K}_\infty$ functions $\alpha_1(\cdot), \alpha_2(\cdot), \alpha_3(\cdot)$ and a $\mathcal{K}$ function $\sigma(\cdot)$ such that for all $x \in X$, all $w \in \mathbb{W}$

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|)$$
$$V(f(x, w)) - V(x) \leq -\alpha_3(|x|) + \sigma(|w|)$$

The following result is a minor generalization of Lemma 3.5 in Jiang and Wang (2001).

Lemma B.47 (ISS-Lyapunov function implies ISS (constrained))**.** *Suppose that $\mathbb{W}$ is a compact set containing the origin and that $X \subset \mathbb{X}$ is a closed robust positive invariant set for $x^+ = f(x,w)$, $w \in \mathbb{W}$. If $f(\cdot)$ is continuous and there exists a continuous ISS-Lyapunov function in X for the system $x^+ = f(x,w)$, $w \in \mathbb{W}$, then the system $x^+ = f(x,w)$, $w \in \mathbb{W}$ is ISS in X.*

B.7 Output-to-State Stability and Detectability

We present some definitions and results that are discrete time versions of results due to Sontag and Wang (1997) and Krichman, Sontag, and Wang (2001). The output-to-state (OSS) property corresponds, informally, to the statement that "no matter what the initial state is, if the observed outputs are small, then the state must eventually be small". It is therefore a natural candidate for the concept of nonlinear (zero-state) detectability. We consider first the autonomous system

$$x^+ = f(x) \qquad y = h(x) \tag{B.22}$$

where $f(\cdot) : \mathbb{X} \to \mathbb{X}$ is locally Lipschitz continuous and $h(\cdot)$ is continuously differentiable where $\mathbb{X} = \mathbb{R}^n$ for some n. We assume $x = 0$ is an equilibrium state, i.e., $f(0) = 0$. We also assume $h(0) = 0$. We use $\phi(k; x_0)$ to denote the solution of (B.22) with initial state x_0, and $y(k; x_0)$ to denote $h(\phi(k; x_0))$. The function $y_{x_0}(\cdot)$ is defined by

$$y_{x_0}(k) := y(k; x_0)$$

We use $|\cdot|$ and $\|\cdot\|$ to denote, respectively the Euclidean norm of a vector and the sup norm of a sequence; $\|\cdot\|_{0:k}$ denotes the max norm of a sequence restricted to the interval $[0,k]$. For conciseness, $\mathbf{u}$, $\mathbf{y}$ denote, respectively, the sequences $(u(j))$, $(y(j))$.

Definition B.48 (Output-to-state stable (OSS))**.** The system (B.22) is output-to-state stable (OSS) if there exist functions $\beta(\cdot) \in \mathcal{KL}$ and $\gamma(\cdot) \in \mathcal{K}$ such that for all $x_0 \in \mathbb{R}^n$ and all $k \geq 0$

$$|x(k; x_0)| \leq \max\left\{\beta(|x_0|, k),\ \gamma(\|\mathbf{y}\|_{0:k})\right\}$$

Definition B.49 (OSS-Lyapunov function)**.** An OSS-Lyapunov function for system (B.22) is any function $V(\cdot)$ with the following properties

(a) There exist $\mathcal{K}_\infty$ functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ such that

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|)$$

for all x in $\mathbb{R}^n$.

(b) There exist $\mathcal{K}_\infty$ functions $\alpha(\cdot)$ and $\sigma(\cdot)$ such that for all $x \in \mathbb{R}^n$ either

$$V(x^+) \leq V(x) - \alpha(|x|) + \sigma(|y|)$$

or

$$V(x^+) \leq \rho V(x) + \sigma(|y|) \tag{B.23}$$

with $x^+ = f(x)$, $y = h(x)$, and $\rho \in (0,1)$.

Inequality (B.23) corresponds to an exponential-decay OSS-Lyapunov function.

Theorem B.50 (OSS and OSS-Lyapunov function)**.** *The following properties are equivalent for system* (B.22)*:*

(a) The system is OSS.

(b) The system admits an OSS-Lyapunov function.

(c) The system admits an exponential-decay OSS-Lyapunov function.

B.8 Input/Output-to-State Stability

Consider now a system with both inputs and outputs

$$x^+ = f(x,u) \qquad y = h(x) \tag{B.24}$$

Input/output-to-state stability corresponds roughly to the statement that, no matter what the initial state is, if the input and the output converge to zero, so does the state. We assume $f(\cdot)$ and $h(\cdot)$ are continuous. We also assume $f(0,0) = 0$ and $h(0) = 0$. Let $x(\cdot, x_0, \mathbf{u})$ denote the solution of (B.24) which results from initial state x_0 and control $\mathbf{u} = (u(j))_{j\geq 0}$ and let $y_{x_0,\mathbf{u}}(k) := y(k; x_0, \mathbf{u})$ denote $h(x(k; x_0, \mathbf{u}))$.

Definition B.51 (Input/output-to-state stable (IOSS))**.** The system (B.24) is input/output-to-state stable (IOSS) if there exist functions $\beta(\cdot) \in \mathcal{KL}$ and $\gamma_1(\cdot), \gamma_2(\cdot) \in \mathcal{K}$ such that

$$|x(k; x_0)| \leq \max\left\{\beta(|x_0|, k), \gamma_1\left(\|\mathbf{u}\|_{0:k-1}\right), \gamma_2\left(\|\mathbf{y}\|_{0:k}\right)\right\}$$

for every initial state $x_0 \in \mathbb{R}^n$, every control sequence $\mathbf{u} = (u(j))$, and all $k \geq 0$.

Definition B.52 (IOSS-Lyapunov function)**.** An IOSS-Lyapunov function for system (B.24) is any function $V(\cdot)$ with the following properties:

(a) There exist $\mathcal{K}_\infty$ functions $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ such that

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|)$$

for all $x \in \mathbb{R}^n$.

(b) There exist $\mathcal{K}_\infty$ functions $\alpha(\cdot)$, $\sigma_1(\cdot)$, and $\sigma_2(\cdot)$ such that for every x and u either

$$V(x^+) \leq V(x) - \alpha(|x|) + \sigma_1(|u|) + \sigma_2(|y|)$$

or

$$V(x^+) \leq \rho V(x) + \sigma_1(|u|) + \sigma_2(|y|)$$

with $x^+ = f(x,u)$, $y = h(x)$, and $\rho \in (0,1)$.

The following result proves useful when establishing that MPC employing cost functions based on the inputs and outputs rather than inputs and states is stabilizing for IOSS systems. Consider the system $x^+ = f(x,u), y = h(x)$ with stage cost $\ell(y,u)$ and constraints $(x, u) \in \mathbb{Z}$. The stage cost satisfies $\ell(0,0) = 0$ and $\ell(y,u) \geq \alpha(|(y,u)|)$ for all $(y,u) \in \mathbb{R}^p \times \mathbb{R}^m$ with α_1 a $\mathcal{K}_\infty$ function. Let $\mathbb{X} := \{x \mid \exists u \text{ with } (x,u) \in \mathbb{Z}\}$.

Theorem B.53 (Modified IOSS-Lyapunov function)**.** *Assume that there exists an IOSS-Lyapunov function $V : \mathbb{X} \to \mathbb{R}_{\geq 0}$ for the constrained system $x^+ = f(x,u)$ such that the following holds for all $(x,u) \in \mathbb{Z}$ for which $f(x,u) \in \mathbb{X}$*

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|)$$
$$V(f(x,u)) - V(x) \leq -\alpha_3(|x|) + \sigma(\ell(y,u))$$

with $\alpha_1, \alpha_2, \alpha_3 \in \mathcal{K}_\infty$ and $\sigma \in \mathcal{K}$.

For any $\overline{\alpha}_4 \in \mathcal{K}_\infty$, there exists another IOSS-Lyapunov function $\Lambda : \mathbb{X} \to \mathbb{R}_{\geq 0}$ for the constrained system $x^+ = f(x,u)$ such that the following holds for all $(x,u) \in \mathbb{Z}$ for which $f(x,u) \in \mathbb{X}$

$$\overline{\alpha}_1(|x|) \leq \Lambda(x) \leq \overline{\alpha}_2(|x|)$$
$$\Lambda(f(x,u)) - \Lambda(x) \leq -\rho(|x|) + \overline{\alpha}_4(\ell(y,u))$$

with $\overline{\alpha}_1, \overline{\alpha}_2 \in \mathcal{K}_\infty$ and continuous function $\rho \in \mathcal{PD}$. Note that $\Lambda = \gamma \circ V$ for some $\gamma \in \mathcal{K}$.

Conjecture B.54 (IOSS and IOSS-Lyapunov function)**.** *The following properties are equivalent for system* (B.24)*:*

(a) The system is IOSS.

(b) The system admits a smooth IOSS-Lyapunov function.

(c) The system admits an exponential-decay IOSS-Lyapunov function.

As discussed in the Notes section of Chapter 2, Grimm, Messina, Tuna, and Teel (2005) use a storage function like $\Lambda(\cdot)$ in Theorem B.53 to treat a semidefinite stage cost. Cai and Teel (2008) provide a discrete time converse theorem for IOSS that holds for all $\mathbb{R}^n$. Allan and Rawlings (2018) provide the converse theorem on closed positive invariant sets (Theorem 36), and also provide a lemma for changing the supply rate function (Theorem 38).

B.9 Incremental-Input/Output-to-State Stability

Definition B.55 (Incremental input/output-to-state stable)**.** The system (B.24) is incrementally input/output-to-state stable (i-IOSS) if there exists some $\beta(\cdot) \in \mathcal{KL}$ and $\gamma_1(\cdot)$, $\gamma_2(\cdot) \in \mathcal{K}$ such that, for every two initial states z_1 and z_2 and any two control sequences $\mathbf{u}_1 = (u_1(j))$ and $\mathbf{u}_2 = (u_2(j))$

$$
\begin{aligned}
&|x(k; z_1, \mathbf{u}_1) - x(k; z_2, \mathbf{u}_2)| \leq \\
&\max\left\{\beta(|z_1 - z_2|, k),\ \gamma_1(\|\mathbf{u}_1 - \mathbf{u}_2\|_{0:k-1}),\ \gamma_2(\left\|\mathbf{y}_{z_1,\mathbf{u}_1} - \mathbf{y}_{z_2,\mathbf{u}_2}\right\|_{0:k})\right\}
\end{aligned}
$$

B.10 Observability

Definition B.56 (Observability)**.** The system (B.24) is (uniformly) observable if there exists a positive integer N and an $\alpha(\cdot) \in \mathcal{K}$ such that

$$
\sum_{j=0}^{k-1} \left| h(x(j; x, u)) - h(x(j; z, u)) \right| \geq \alpha(|x - z|) \tag{B.25}
$$

for all x, z, all $k \geq N$ and all control sequences u; here $x(j; z, u) = \phi(j; z, u)$, the solution of (B.24) when the initial state is z at time 0 and the control sequence is u.

When the system is linear, i.e., $f(x, u) = Ax + Bu$ and $h(x) = Cx$, this assumption is equivalent to assuming the observability Gramian $\sum_{j=0}^{n-1} CA^j (A^j)' C'$ is positive definite. Consider the system described by

$$
z^+ = f(z, u) + w \qquad y + v = h(z) \tag{B.26}
$$

with output $y_w = y + v$. Let $z(k; z, u, w)$ denote the solution, at time k of (B.26) if the state at time 0 is z, the control sequence is u and the disturbance sequence is w. We assume, in the sequel, that

Assumption B.57 (Lipschitz continuity of model)**.**

(a) The function $f(\cdot)$ is globally Lipschitz continuous in $\mathbb{R}^n \times \mathbf{U}$ with Lipschitz constant c.

(b) The function $h(\cdot)$ is globally Lipschitz continuous in $\mathbb{R}^n$ with Lipschitz constant c.

Lemma B.58 (Lipschitz continuity and state difference bound)**.** *Suppose Assumption B.57 is satisfied (with Lipschitz constant c). Then,*

$$|x(k; x, u) - z(k; z, u, w)| \leq c^k |x - z| + \sum_{i=0}^{k-1} c^{k-i-1} |w(i)|$$

Proof. Let $\delta(k) := |x(k; x, u) - z(k; z, u, w)|$. Then

$$\begin{aligned} \delta(k+1) &= \left| f(x(k; x, u), u(k)) - f(z(k; z, u, w), u(k)) - w(k) \right| \\ &\leq c\, |\delta(k)| + |w(k)| \end{aligned}$$

Iterating this equation yields the desired result. ∎

Theorem B.59 (Observability and convergence of state)**.** *Suppose* (B.24) *is (uniformly) observable and that Assumption B.57 is satisfied. Then, $w(k) \to 0$ and $v(k) \to 0$ as $k \to \infty$ imply $|x(k; x, u) - z(k; z, u, w)| \to 0$ as $k \to \infty$.*

Proof. Let $x(k)$ and $z(k)$ denote $x(k; x, u)$ and $z(k; z, u, w)$, respectively, in the sequel. Since (B.24) is observable, there exists an integer N satisfying (B.25). Consider the sum

$$\begin{aligned} S(k) = \sum_{j=k}^{k+N} v(k) &= \sum_{j=k}^{k+N} \left| h(x(j; x, u)) - h(z(j; z, u, w)) \right| \\ &\geq \sum_{j=k}^{k+N} \left| h(x(j; x(k), u)) - h(x(j; z(k), u)) \right| \\ &\quad - \sum_{j=k}^{k+N} \left| h(x(j; z(k), u)) - h(z(j; z(k), u, w)) \right| \end{aligned} \tag{B.27}$$

where we have used the fact that $|a+b| \geq |a| - |b|$. By the assumption of observability

$$\sum_{j=k}^{k+N} \left| h(x(j;x(k),u)) - h(x(j;z(k),u)) \right| \geq \alpha(|x(k) - z(k)|)$$

for all k. From Lemma B.58 and the Lipschitz assumption on $h(\cdot)$

$$\left| h(x(j;z(k),u)) - h(z(j;z(k),u,w)) \right| \leq$$
$$c \left| x(j;z(k),u) - z(j;z(k),u,w) \right| \leq c \sum_{i=k}^{j-1} c^{j-1-i} |w(i)|$$

for all j in $\{k+1, k+2, \ldots k+N\}$. Hence there exists a $d \in (0,\infty)$ such that the last term in (B.27) satisfies

$$\sum_{j=k}^{k+N} \left| h(x(j;x(k),u)) - h(x(j;z(k),u)) \right| \leq d \|w\|_{k-N:k}$$

Hence, (B.27) becomes

$$\alpha(|x(k) - z(k)|) \leq N \|v\|_{k-N:k} + d \|w\|_{k-N:k}$$

Since, by assumption, $w(k) \to 0$ and $v(k) \to 0$ as $k \to \infty$, and $\alpha(\cdot) \in \mathcal{K}$, it follows that $|x(k) - z(k)| \to 0$ as $k \to \infty$. ■

B.11 Exercises

Exercise B.1: Lyapunov equation and linear systems

Establish the equivalence of (a) and (b) in Lemma B.20.

Exercise B.2: Lyapunov function for exponential stability

Let $V : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be a Lyapunov function for the system $x^+ = f(x)$ with the following properties. For all $x \in \mathbb{R}^n$

$$a_1 |x|^\sigma \leq V(x) \leq a_2 |x|^\sigma$$
$$V(f(x)) - V(x) \leq -a_3 |x|^\sigma$$

in which $a_1, a_2, a_3, \sigma > 0$. Show that the origin of the system $x^+ = f(x)$ is globally exponentially stable.

Exercise B.3: A converse theorem for exponential stability

(a) Assume that the origin is globally exponentially stable (GES) for the system

$$x^+ = f(x)$$

in which $f(\cdot)$ is continuous. Show that there exists a continuous Lyapunov function $V(\cdot)$ for the system satisfying for all $x \in \mathbb{R}^n$

$$a_1 |x|^\sigma \le V(x) \le a_2 |x|^\sigma$$
$$V(f(x)) - V(x) \le -a_3 |x|^\sigma$$

in which $a_1, a_2, a_3, \sigma > 0$.

Hint: Consider summing the solution $|\phi(i;x)|^\sigma$ on i as a candidate Lyapunov function $V(x)$.

(b) Establish that in the Lyapunov function defined above, any $\sigma > 0$ is valid, and also that the constant a_3 can be chosen as large as one wishes.

Exercise B.4: Revisit Lemma 1.3 in Chapter 1

Establish Lemma 1.3 in Chapter 1 using the Lyapunov function tools established in this appendix. Strengthen the conclusion and establish that the closed-loop system is globally exponentially stable.

Exercise B.5: Continuity of Lyapunov function for asymptotic stability

Let X be a compact subset of $\mathbb{R}^n$ containing the origin in its interior that is positive invariant for the system $x^+ = f(x)$. If $f(\cdot)$ is continuous on X and the origin is asymptotically stable with a region of attraction X, show that the Lyapunov function suggested in Theorem B.17 is continuous on X.

Exercise B.6: A Lipschitz continuous converse theorem for exponential stability

Consider the system $x^+ = f(x)$, $f(0) = 0$, with function $f : D \to \mathbb{R}^n$ Lipschitz continuous on compact set $D \subset \mathbb{R}^n$ containing the origin in its interior. Choose $R > 0$ such that $B_R \subseteq D$. Assume that there exist scalars $c > 0$ and $\lambda \in (0,1)$ such that

$$|\phi(k;x)| \le c\,|x|\,\lambda^k \qquad \text{for all } |x| \le r, \quad k \ge 0$$

with $r := R/c$.

Show that there exists a *Lipschitz continuous* Lyapunov function $V(\cdot)$ satisfying for all $x \in B_r$

$$a_1 |x|^2 \le V(x) \le a_2 |x|^2$$
$$V(f(x)) - V(x) \le -a_3 |x|^2$$

with $a_1, a_2, a_3 > 0$.

Hint: Use the proposed Lyapunov function of Exercise B.3 with $\sigma = 2$. See also (Khalil, 2002, Exercise 4.68).

Exercise B.7: Lyapunov function requirements: continuity of α_3

Consider the following scalar system $x^+ = f(x)$ with piecewise affine and discontinuous $f(\cdot)$ (Lazar et al., 2009)

$$f(x) = \begin{cases} 0, & x \in (-\infty, 1] \\ (1/2)(x+1), & x \in (1, \infty) \end{cases}$$

Note that the origin is a steady state

(a) Consider $V(x) = |x|$ as a candidate Lyapunov function. Show that this V satisfies (B.8)–(B.10) of Definition B.12, in which $\alpha_3(x)$ is positive definite but *not* continuous.

(b) Show by direction calculation that the origin is not globally asymptotically stable. Show that for initial conditions $x_0 \in (1, \infty)$, $x(k; x_0) \to 1$ as $k \to \infty$.

The conclusion here is that one cannot leave out continuity of α_3 in the definition of a Lyapunov function when allowing discontinuous system dynamics.

Exercise B.8: Difference between classical and KL stability definitions (Teel)

Consider the *discontinuous* nonlinear scalar example $x^+ = f(x)$ with

$$f(x) = \begin{cases} \dfrac{1}{2}x & |x| \in [0, 1] \\ \dfrac{2x}{2 - |x|} & |x| \in (1, 2) \\ 0 & |x| \in [2, \infty) \end{cases}$$

Is this system GAS under the classical definition? Is this system GAS under the KL definition? Discuss why or why not.

Exercise B.9: Combining $\mathcal{K}$ functions

Establish (B.5) starting from (B.3) and (B.4) and then using (B.1).

Bibliography

D. A. Allan and J. B. Rawlings. An input/output-to-state stability converse theorem for closed positive invariant sets. Technical Report 2018–01, TWCCC Technical Report, December 2018.

D. A. Allan, C. N. Bates, M. J. Risbeck, and J. B. Rawlings. On the inherent robustness of optimal and suboptimal nonlinear MPC. *Sys. Cont. Let.*, 106: 68–78, August 2017.

R. G. Bartle and D. R. Sherbert. *Introduction to Real Analysis*. John Wiley & Sons, Inc., New York, third edition, 2000.

C. Cai and A. R. Teel. Input–output-to-state stability for discrete-time systems. *Automatica*, 44(2):326 – 336, 2008.

G. Grimm, M. J. Messina, S. E. Tuna, and A. R. Teel. Model predictive control: For want of a local control Lyapunov function, all is not lost. *IEEE Trans. Auto. Cont.*, 50(5):546–558, 2005.

Z.-P. Jiang and Y. Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37:857–869, 2001.

Z.-P. Jiang and Y. Wang. A converse Lyapunov theorem for discrete-time systems with disturbances. *Sys. Cont. Let.*, 45:49–58, 2002.

R. E. Kalman and J. E. Bertram. Control system analysis and design via the "Second method" of Lyapunov, Part II: Discrete–time systems. *ASME J. Basic Engr.*, pages 394–400, June 1960.

C. M. Kellett and A. R. Teel. Discrete-time asymptotic controllability implies smooth control-Lyapunov function. *Sys. Cont. Let.*, 52:349–359, 2004a.

C. M. Kellett and A. R. Teel. Smooth Lyapunov functions and robustness of stability for difference inclusions. *Sys. Cont. Let.*, 52:395–405, 2004b.

H. K. Khalil. *Nonlinear Systems*. Prentice-Hall, Upper Saddle River, NJ, third edition, 2002.

M. Krichman, E. D. Sontag, and Y. Wang. Input-output-to-state stability. *SIAM J. Cont. Opt.*, 39(6):1874–1928, 2001.

J. P. LaSalle. *The stability and control of discrete processes*, volume 62 of *Applied Mathematical Sciences*. Springer-Verlag, 1986.

M. Lazar, W. P. M. H. Heemels, and A. R. Teel. Lyapunov functions, stability and input-to-state stability subtleties for discrete-time discontinuous systems. *IEEE Trans. Auto. Cont.*, 54(10):2421–2425, 2009.

J. B. Rawlings and L. Ji. Optimization-based state estimation: Current status and some new results. *J. Proc. Cont.*, 22:1439–1444, 2012.

J. B. Rawlings and M. J. Risbeck. On the equivalence between statements with epsilon-delta and K-functions. Technical Report 2015–01, TWCCC Technical Report, December 2015.

J. B. Rawlings and M. J. Risbeck. Model predictive control with discrete actuators: Theory and application. *Automatica*, 78:258–265, 2017.

E. D. Sontag. *Mathematical Control Theory*. Springer-Verlag, New York, second edition, 1998a.

E. D. Sontag. Comments on integral variants of ISS. *Sys. Cont. Let.*, 34:93–100, 1998b.

E. D. Sontag and Y. Wang. On the characterization of the input to state stability property. *Sys. Cont. Let.*, 24:351–359, 1995.

E. D. Sontag and Y. Wang. Output-to-state stability and detectability of nonlinear systems. *Sys. Cont. Let.*, 29:279–290, 1997.

A. R. Teel. Discrete time receding horizon control: is the stability robust. In Marcia S. de Queiroz, Michael Malisoff, and Peter Wolenski, editors, *Optimal control, stabilization and nonsmooth analysis*, volume 301 of *Lecture notes in control and information sciences*, pages 3–28. Springer, 2004.

C
Optimization

Version: date: April 5, 2019

C.1 Dynamic Programming

The name *dynamic programming* dates from the 1950s when it was coined by Richard Bellman for a technique for solving dynamic optimization problems, i.e., optimization problems associated with deterministic or stochastic systems whose behavior is governed by differential or difference equations. Here we review some of the basic ideas behind dynamic programming (DP) Bellman (1957); Bertsekas, Nedic, and Ozdaglar (2001).

To introduce the topic in its simplest form, consider the simple routing problem illustrated in Figure C.1. To maintain connection with optimal control, each node in the graph can be regarded as a point (x, t) in a subset S of $X \times T$ where both the state space $X = \{a, b, c, \ldots, g\}$ and the set of times $T = \{0, 1, 2, 3\}$ are discrete. The set of permissible control actions is $\mathbb{U} = \{U, D\}$, i.e., to go "up" or "down." The control problem is to choose the lowest cost path from event $(d, 0)$ (state d at $t = 0$) to any of the states at $t = 3$; the cost of going from one event to the next is indicated on the graph. This problem is equivalent to choosing an open-loop control, i.e., a sequence $(u(0), u(1), u(2))$ of admissible control actions. There are 2^N controls where N is the number of stages, 3 in this example. The cost of each control can, in this simple example, be evaluated and is given in Table C.1.

There are two different *open-loop* optimal controls, namely (U, D, U) and (D, D, D), each incurring a cost of 16. The corresponding

control	UUU	UUD	UDU	UDD	DUU	DUD	DDU	DDD
cost	20	24	16	24	24	32	20	16

Table C.1: Control Cost.

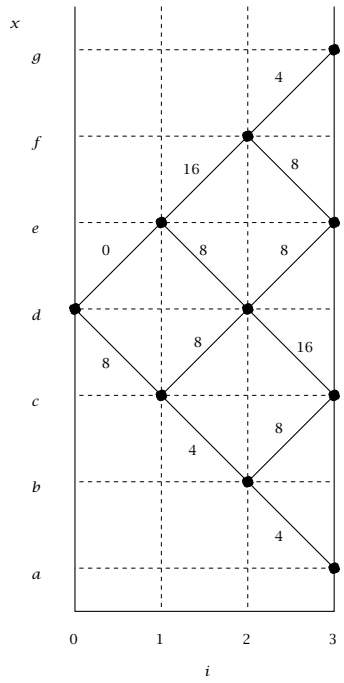


Figure C.1: Routing problem.

state trajectories are (d, e, d, e) and (d, c, b, a).

In discrete problems of this kind, DP replaces the N-stage problem by M single stage problems, where M is the total number of nodes, i.e., the number of elements in $S \subset X \times T$. The first set of optimization problems deals with the states b, d, f at time $N - 1 = 2$. The optimal decision at event $(f, 2)$, i.e., state f at time 2, is the control U and gives rise to a cost of 4. The optimal cost and control for node $(f, 2)$ are recorded; see Table C.2. The procedure is then repeated for states d and b at time $t = 2$ (nodes $(d, 2)$ and $(b, 2)$) and recorded as shown in Table C.2. Attention is next focused on the states e and c at $t = 1$ (nodes $(e, 1)$ and $(c, 1)$). The lowest cost that can be achieved at node $(e, 1)$ *if* control U is chosen, is $16 + 4$, the sum of the path cost 16 associated with the control U, and the *optimal* cost 4 associated with the node $(f, 2)$ that results from using control U at node $(e, 1)$. Similarly the lowest possible cost, if control D is chosen, is $8 + 8$. Hence

t	0	1		2		
state	d	e	c	f	d	b
control	U or D	D	D	U	U	D
optimal cost	16	16	8	4	8	4

Table C.2: Optimal Cost and Control

the optimal control and cost for node $(e, 1)$ are, respectively, D and 16. The procedure is repeated for the remaining state d at $t = 1$ (node $(d, 1)$). A similar calculation for the state d at $t = 0$ (node $(d, 0)$), where the optimal control is U or D, completes this backward recursion; this backward recursion provides the optimal cost and control for each (x, t), as recorded in Table C.2. The procedure therefore yields an optimal *feedback* control that is a function of $(x, t) \in S$. To obtain the optimal open-loop control for the initial node $(d, 0)$, the feedback law is obeyed, leading to control U or D at $t = 0$; if U is chosen, the resultant state at $t = 1$ is e. From Table C.2, the optimal control at $(e, 1)$ is D, so that the successor node is $(d, 2)$. The optimal control at node $(d, 2)$ is U. Thus the optimal open-loop control sequence (U, D, U) is re-obtained. On the other hand, if the decision at $(d, 0)$ is chosen to be D, the optimal sequence (D, D, D) is obtained. This simple example illustrates the main features of DP that we will now examine in the context of discrete time optimal control.

C.1.1 Optimal Control Problem

The discrete time system we consider is described by

$$x^+ = f(x, u) \tag{C.1}$$

where $f(\cdot)$ is continuous. The system is subject to the mixed state-control constraint

$$(x, u) \in \mathbb{Z}$$

where $\mathbb{Z}$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^m$ and $\mathcal{P}_u(\mathbb{Z})$ is compact where $\mathcal{P}_u$ is the projection operator $(x, u) \mapsto u$. Often $\mathbb{Z} = \mathbb{X} \times \mathbb{U}$ in which case the constraint $(x, u) \in \mathbb{Z}$ becomes $x \in \mathbb{X}$ and $u \in \mathbb{U}$ and $\mathcal{P}_u(\mathbb{Z}) = \mathbb{U}$ so that $\mathbb{U}$ is compact. In addition there is a constraint on the terminal state $x(N)$:

$$x(N) \in \mathbb{X}_f$$

where $\mathbb{X}_f$ is closed. In this section we find it easier to express the value function and the optimal control in terms of the current state and current time i rather than using time-to-go k. Hence we replace time-to-go k by time i where $k = N - i$, replace $V_k^0(x)$ (the optimal cost at state x when the time-to-go is k) by $V^0(x, i)$ (the optimal cost at state x, time i) and replace X_k by $X(i)$ where $X(i)$ is the domain of $V^0(\cdot, i)$).

The cost associated with an initial state x at time 0 and a control sequence $\mathbf{u} := (u(0), u(1), \ldots, u(N-1))$ is

$$V(x, 0, \mathbf{u}) = V_f(x(N)) + \sum_{i=1}^{N-1} \ell(x(i), u(i)) \tag{C.2}$$

where $\ell(\cdot)$ and $V_f(\cdot)$ are continuous and, for each i, $x(i) = \phi(i; (x, 0), \mathbf{u})$ is the solution at time i of (C.1) if the initial state is x at time 0 and the control sequence is $\mathbf{u}$. The optimal control problem $\mathbb{P}(x, 0)$ is defined by

$$V^0(x, 0) = \min_{\mathbf{u}} V(x, 0, \mathbf{u}) \tag{C.3}$$

subject to the constraints $(x(i), u(i)) \in \mathbb{Z}$, $i = 0, 1, \ldots, N-1$ and $x(N) \in \mathbb{X}_f$. Equation (C.3) may be rewritten in the form

$$V^0(x, 0) = \min_{\mathbf{u}} \{V(x, 0, \mathbf{u}) \mid \mathbf{u} \in \mathcal{U}(x, 0)\} \tag{C.4}$$

where $\mathbf{u} := (u(0), u(1), \ldots, u(N-1))$,

$$\mathcal{U}(x, 0) := \{\mathbf{u} \in \mathbb{R}^{Nm} \mid (x(i), u(i)) \in \mathbb{Z}, i = 0, 1, \ldots, N-1; x(N) \in \mathbb{X}_f\}$$

and $x(i) := \phi(i; (x, 0), \mathbf{u})$. Thus $\mathcal{U}(x, 0)$ is the set of admissible control sequences[1] if the initial state is x at time 0. It follows from the continuity of $f(\cdot)$ that for all $i \in \{0, 1, \ldots, N-1\}$ and all $x \in \mathbb{R}^n$, $\mathbf{u} \mapsto \phi(i; (x, 0), \mathbf{u})$ is continuous, $\mathbf{u} \mapsto V(x, 0, \mathbf{u})$ is continuous and $\mathcal{U}(x, 0)$ is compact. Hence the minimum in (C.4) exists at all $x \in \{x \in \mathbb{R}^n \mid \mathcal{U}(x, 0) \neq \emptyset\}$.

DP embeds problem $\mathbb{P}(x, 0)$ for a given state x in a whole family of problems $P(x, i)$ where, for each (x, i), problem $\mathbb{P}(x, i)$ is defined by

$$V^0(x, i) = \min_{\mathbf{u}^i} \{V(x, i, \mathbf{u}^i) \mid \mathbf{u}^i \in \mathcal{U}(x, i)\}$$

where

$$\mathbf{u}^i := (u(i), u(i+1), \ldots, u(N-1))$$

[1] An admissible control sequence satisfies all constraints.

$$V(x,i,\mathbf{u}^i) := V_f(x(N)) + \sum_{j=i}^{N-1} \ell(x(j),u(j)) \tag{C.5}$$

and

$$\mathcal{U}(x,i) := \{\mathbf{u}^i \in \mathbb{R}^{(N-i)m} \mid (x(j),u(j)) \in \mathbb{Z}, j = i, i+1, \ldots, N-1 \\ x(N) \in \mathbb{X}_f\} \tag{C.6}$$

In (C.5) and (C.6), $x(j) = \phi(j;(x,i),\mathbf{u}^i)$, the solution at time j of (C.1) if the initial state is x at time i and the control sequence is $\mathbf{u}^i$. For each i, $X(i)$ denotes the domain of $V^0(\cdot,i)$ and $\mathcal{U}(\cdot,i)$ so that

$$X(i) = \{x \in \mathbb{R}^n \mid \mathcal{U}(x,i) \neq \emptyset\}. \tag{C.7}$$

C.1.2 Dynamic Programming

One way to approach DP for discrete time control problems is the simple observation that for all (x,i)

$$\begin{aligned} V^0(x,i) &= \min_{\mathbf{u}^i}\{V(x,i,\mathbf{u}^i) \mid \mathbf{u}^i \in \mathcal{U}(x,i)\} \\ &= \min_{u}\{\ell(x,u) + \min_{\mathbf{u}^{i+1}} V(f(x,u),i+1,\mathbf{u}^{i+1}) \mid \\ &\qquad \{u,\mathbf{u}^{i+1}\} \in \mathcal{U}(x,i)\} \end{aligned} \tag{C.8}$$

where $\mathbf{u}^i = (u,u(i+1),\ldots,u(N-1)) = \left(u,\mathbf{u}^{i+1}\right)$. We now make use of the fact that $\{u,\mathbf{u}^{i+1}\} \in \mathcal{U}(x,i)$ if and only if $(x,u) \in \mathbb{Z}$, $f(x,u) \in X(i+1)$, and $\mathbf{u}^{i+1} \in \mathcal{U}(f(x,u),i+1)$ since $f(x,u) = x(i+1)$. Hence we may rewrite (C.8) as

$$\begin{aligned} V^0(x,i) = \min_{u}\{&\ell(x,u) + V^0(f(x,u),i+1) \mid \\ &(x,u) \in \mathbb{Z}, f(x,u) \in X(i+1)\} \end{aligned} \tag{C.9}$$

for all $x \in X(i)$ where

$$X(i) = \{x \in \mathbb{R}^n \mid \exists u \text{ such that } (x,u) \in \mathbb{Z} \text{ and } f(x,u) \in X(i+1)\} \tag{C.10}$$

Equations (C.9) and (C.10), together with the boundary condition

$$V^0(x,N) = V_f(x) \ \forall x \in X(N), \quad X(N) = \mathbb{X}_f$$

constitute the DP recursion for constrained discrete time optimal control problems. If there are no state constraints, i.e., if $\mathbb{Z} = \mathbb{R}^n \times \mathbb{U}$ where

$\mathbb{U} \subset \mathbb{R}^m$ is compact, then $X(i) = \mathbb{R}^n$ for all $i \in \{0, 1, \ldots, N\}$ and the DP equations revert to the familiar DP recursion:

$$V^0(x, i) = \min_u \{\ell(x, u) + V^0(f(x, u), i + 1)\} \ \forall x \in \mathbb{R}^n$$

with boundary condition

$$V^0(x, N) = V_f \ \forall x \in \mathbb{R}^n$$

We now prove some basic facts; the first is the well known *principle of optimality*.

Lemma C.1 (Principle of optimality)**.** *Let $x \in X_N$ be arbitrary, let $\mathbf{u} := (u(0), u(1), \ldots, u(N-1)) \in \mathcal{U}(x, 0)$ denote the solution of $\mathbb{P}(x, 0)$ and let $(x, x(1), x(2), \ldots, x(N))$ denote the corresponding optimal state trajectory so that for each i, $x(i) = \phi(i; (x, 0), \mathbf{u})$. Then, for any $i \in \{0, 1, \ldots, N-1\}$, the control sequence $\mathbf{u}^i := (u(i), u(i+1), \ldots, u(N-1))$ is optimal for $\mathbb{P}(x(i), i)$ (any portion of an optimal trajectory is optimal).*

Proof. Since $\mathbf{u} \in \mathcal{U}(x, 0)$, the control sequence $\mathbf{u}^i \in \mathcal{U}(x(i), i)$. If $\mathbf{u}^i = (u(i), u(i+1), \ldots, u(N-1))$ is not optimal for $\mathbb{P}(x(i), i)$, there exists a control sequence $\mathbf{u}' = (u'(i), u'(i+1), \ldots, u(N-1)') \in \mathcal{U}(x(i), i)$ such that $V(x(i), i, \mathbf{u}') < V(x(i), \mathbf{u})$. Consider now the control sequence $\tilde{\mathbf{u}} := (u(0), u(1), \ldots, u(i-1), u'(i), u'(i+1), \ldots, u(N-1)')$. It follows that $\tilde{\mathbf{u}} \in \mathcal{U}(x, 0)$ and $V(x, 0, \tilde{\mathbf{u}}) < V(x, 0, \mathbf{u}) = V^0(x, 0)$, a contradiction. Hence $\mathbf{u}(x(i), i)$ is optimal for $\mathbb{P}(x(i), i)$. ∎

The most important feature of DP is the fact that the DP recursion yields the optimal value $V^0(x, i)$ and the optimal control $\kappa(x, i) = \arg\min_u \{\ell(x, u) + V^0(f(x, u), i + 1) \mid (x, u) \in \mathbb{Z}, f(x, u) \in X(i + 1)\}$ for each $(x, i) \in X(i) \times \{0, 1, \ldots, N-1\}$.

Theorem C.2 (Optimal value function and control law from DP)**.** *Suppose that the function $\Psi : \mathbb{R}^n \times \{0, 1, \ldots, N\} \to \mathbb{R}$, satisfies, for all $i \in \{1, 2, \ldots, N-1\}$, all $x \in X(i)$, the DP recursion*

$$\Psi(x, i) = \min\{\ell(x, u) + \Psi(f(x, u), i + 1) \mid (x, u) \in \mathbb{Z}, f(x, u) \in X(i + 1)\}$$
$$X(i) = \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^m \text{ such that } (x, u) \in \mathbb{Z}, \ f(x, u) \in X(i + 1)\}$$

with boundary conditions

$$\Psi(x, N) = V_f(x) \ \forall x \in \mathbb{X}_f, \quad X(N) = \mathbb{X}_f$$

Then $\Psi(x, i) = V^0(x, i)$ for all $(x, i) \in X(i) \times \{0, 1, 2, \ldots, N\}$; the DP recursion yields the optimal value function and the optimal control law.

Proof. Let $(x,i) \in X(i) \times \{0,1,\ldots,N\}$ be arbitrary. Let $\mathbf{u} = (u(i), u(i+1), \ldots, u(N-1))$ be an arbitrary control sequence in $\mathcal{U}(x, i)$ and let $\mathbf{x} = (x, x(i+1), \ldots, x(N))$ denote the corresponding trajectory starting at (x,i) so that for each $j \in \{i, i+1, \ldots, N\}$, $x(j) = \phi(j; x, i, \mathbf{u})$. For each $j \in \{i, i+1, \ldots, N-1\}$, let $\mathbf{u}^j := (u(j), u(j+1), \ldots, u(N-1))$; clearly $\mathbf{u}^j \in \mathcal{U}(x(j), j)$. The cost due to initial event $(x(j), j)$ and control sequence $\mathbf{u}^j$ is $\Phi(x(j), j)$ defined by

$$\Phi(x(j), j) := V(x(j), j, \mathbf{u}^j)$$

Showing that $\Psi(x,i) \le \Phi(x,i)$ proves that $\Psi(x,i) = V^0(x,i)$ since $\mathbf{u}$ is an arbitrary sequence in $\mathcal{U}(x,i)$; because $(x,i) \in X(i) \times \{0,1,\ldots,N\}$ is arbitrary, that fact that $\Psi(x,i) = V^0(x,i)$ proves that DP yields the optimal value function.

To prove that $\Psi(x,i) \le \Phi(x,i)$, we compare $\Psi(x(j), j)$ and $\Phi(x(j), j)$ for each $j \in \{i, i+1, \ldots, N\}$, i.e., we compare the costs yielded by the DP recursion and by the arbitrary control $\mathbf{u}$ along the corresponding trajectory $\mathbf{x}$. By definition, $\Psi(x(j), j)$ satisfies for each j

$$\Psi(x(j), j) = \min_u \{\ell(x(j), u) + \Psi(f(x(j), u), j+1) \mid (x(j), u) \in \mathbb{Z}, f(x(j), u) \in X(j+1)\} \quad \text{(C.11)}$$

To obtain $\Phi(x(j), j)$ for each j we solve the following recursive equation

$$\Phi(x(j), j) = \ell(x(j), u(j)) + \Phi(f(x(j), u(j)), j+1) \quad \text{(C.12)}$$

The boundary conditions are

$$\Psi(x(N), N) = \Phi(x(N), N) = V_f(x(N)) \quad \text{(C.13)}$$

Since $u(j)$ satisfies $(x(j), u(j)) \in \mathbb{Z}$ and $f(x(j), u(j)) \in X(j+1)$ but is not necessarily a minimizer in (C.11), we deduce that

$$\Psi(x(j), j) \le \ell(x(j), u(j)) + \Psi(f(x(j), u(j)), j+1) \quad \text{(C.14)}$$

For each j, let $E(j)$ be defined by

$$E(j) := \Psi(x(j), j) - \Phi(x(j), j)$$

Subtracting (C.12) from (C.14) and replacing $f(x(j), u(j))$ by $x(j+1)$ yields

$$E(j) \le E(j+1) \ \forall j \in \{i, i+1, \ldots N\}$$

Since $E(N) = 0$ by virtue of (C.13), we deduce that $E(j) \leq 0$ for all $j \in \{i, i+1, \ldots, N\}$; in particular, $E(i) \leq 0$ so that

$$\Psi(x, i) \leq \Phi(x, i) = V(x, i, \mathbf{u})$$

for all $\mathbf{u} \in \mathcal{U}(x, i)$. Hence $\Psi(x, i) = V^0(x, i)$ for all $(x, i) \in X(i) \times \{0, 1, \ldots, N\}$. ■

Example C.3: DP applied to linear quadratic regulator

A much used example is the familiar linear quadratic regulator problem. The system is defined by

$$x^+ = Ax + Bu$$

There are no constraints. The cost function is defined by (C.2) where

$$\ell(x, u) := (1/2)x'Qx + (1/2)u'Ru$$

and $V_f(x) = 0$ for all x; the horizon length is N. We assume that Q is symmetric and positive semidefinite and that R is symmetric and positive definite. The DP recursion is

$$V^0(x, i) = \min_u \{\ell(x, u) + V^0(Ax + Bu, i + 1)\} \ \forall x \in \mathbb{R}^n$$

with terminal condition

$$V^0(x, N) = 0 \ \forall x \in \mathbb{R}^n$$

Assume that $V^0(\cdot, i + 1)$ is quadratic and positive semidefinite and, therefore, has the form

$$V^0(x, i + 1) = (1/2)x'P(i + 1)x$$

where $P(i + 1)$ is symmetric and positive semidefinite. Then

$$V^0(x, i) = (1/2) \min_u \{x'Qx + u'Ru + (Ax + Bu)'P(i + 1)(Ax + Bu)\}$$

The right-hand side of the last equation is a positive definite function of u for all x, so that it has a unique minimizer given by

$$\kappa(x, i) = K(i)x \qquad K(i) := -(B'P(i + 1)B + R)^{-1}B'P(i + 1)$$

Substituting $u = K(i)x$ in the expression for $V^0(x, i)$ yields

$$V^0(x, i) = (1/2)x'P(i)x$$

where $P(i)$ is given by:

$$P(i) = Q + K(i)'RK(i) - A'P(i+1)B(B'P(i+1)B + R)^{-1}B'P(i+1)A$$

Hence $V^0(\cdot, i)$ is quadratic and positive semidefinite if $V^0(\cdot, i+1)$ is. But $V^0(\cdot, N)$, defined by

$$V^0(x, N) := (1/2)x'P(N)x = 0 \qquad P(N) := 0$$

is symmetric and positive semidefinite. By induction $V^0(\cdot, i)$ is quadratic and positive semidefinite (and $P(i)$ is symmetric and positive semidefinite) for all $i \in \{0, 1, \ldots, N\}$. Substituting $K(i) = -(B'P(i+1)B + R)^{-1}B'P(i+1)A$ in the expression for $P(i)$ yields the more familiar matrix Riccati equation

$$P(i) = Q + A'P(i+1)A - A'P(i+1)B(B'P(i+1)B + R)^{-1}BP(i+1)A$$

□

C.2 Optimality Conditions

In this section we obtain optimality conditions for problems of the form

$$f^0 = \inf_u \{f(u) \mid u \in U\}$$

In these problems, $u \in \mathbb{R}^m$ is the *decision* variable, $f(u)$ the cost to be minimized by appropriate choice of u and $U \subset \mathbb{R}^m$ the constraint set. The value of the problem is f^0. Some readers may wish to read only Section C.2.2, which deals with convex optimization problems and Section C.2.3 which deals with convex optimization problems in which the constraint set U is polyhedral. These sections require some knowledge of tangent and normal cones discussed in Section C.2.1; Proposition C.7 in particular derives the normal cone for the case when U is convex.

C.2.1 Tangent and Normal Cones

In determining conditions of optimality, it is often convenient to employ approximations to the cost function $f(\cdot)$ and the constraint set U. Thus the cost function $f(\cdot)$ may be approximated, in the neighborhood of a point $\bar{u}$, by the first order expansion $f(\bar{u}) + \langle \nabla f(\bar{u}), (u - \bar{u}) \rangle$ or by the second order expansion $f(\bar{u}) + \langle \nabla f(\bar{u}), (u - \bar{u}) \rangle + (1/2)((u - \bar{u})'\nabla^2 f(\bar{x})(u - \bar{u}))$ if the necessary derivatives exist. Thus we see that

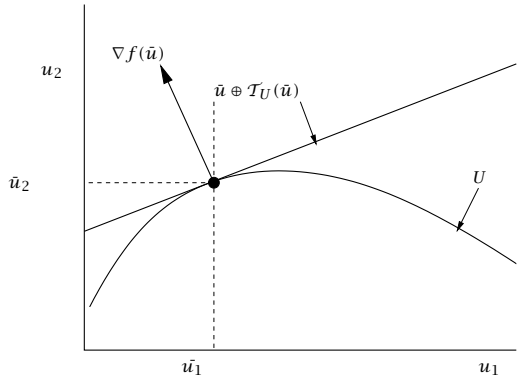


Figure C.2: Approximation of the set U.

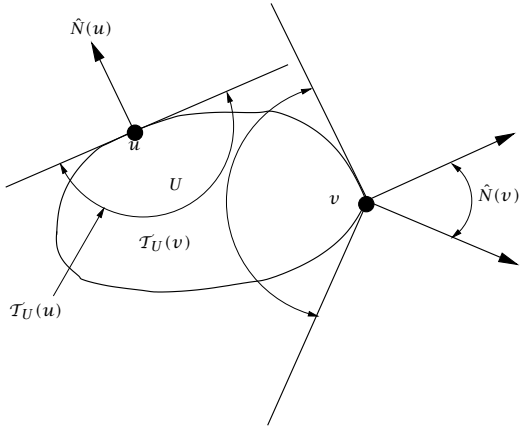


Figure C.3: Tangent cones.

in the unconstrained case, a necessary condition for the optimality of $\bar{u}$ is $\nabla f(\bar{u}) = 0$. To obtain necessary conditions of optimality for constrained optimization problems, we need to approximate the constraint set as well; this is more difficult. An example of U and its approximation is shown in Figure C.2; here the set $U = \{u \in \mathbb{R}^2 \mid g(u) = 0\}$ where $g : \mathbb{R} \to \mathbb{R}$ is approximated in the neighborhood of a point $\bar{u}$ satisfying $g(\bar{u}) = 0$ by the set $\bar{u} \oplus \mathcal{T}_U(\bar{u})$ where[2] the tangent cone $\mathcal{T}_U(\bar{u}) := \{h \in \mathbb{R}^2 \mid \nabla g(\bar{u}), u - \bar{u}\rangle = 0\}$. In general, a set U is approx-

[2]If A and B are two subsets of $\mathbb{R}^n$, say, then $A \oplus B := \{a + b \mid a \in A, b \in B\}$ and $a \oplus B := \{a + b \mid b \in B\}$.

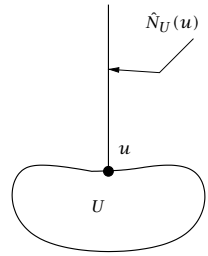


Figure C.4: Normal at u.

imated, near a point $\bar{u}$, by $\bar{u} \oplus \mathcal{T}_U(\bar{u})$ where its *tangent cone* $\mathcal{T}_U(\bar{u})$ is defined below. Following Rockafellar and Wets (1998), we use $u \xrightarrow[U]{\nu} v$ to denote that the sequence $\{u^\nu \mid \nu \in \mathbb{I}_{\geq 0}\}$ converges to v as $\nu \to \infty$ while satisfying $u^\nu \in U$ for all $\nu \in \mathbb{I}_{\geq 0}$.

Definition C.4 (Tangent vector)**.** A vector $h \in \mathbb{R}^m$ is tangent to the set U at $\bar{u}$ if there exist sequences $u \xrightarrow[U]{\nu} \bar{u}$ and $\lambda^\nu \searrow 0$ such that

$$[u^\nu - \bar{u}]/\lambda^\nu \to h$$

$\mathcal{T}_U(u)$ is the set of all tangent vectors.

Equivalently, a vector $h \in \mathbb{R}^m$ is tangent to the set U at $\bar{u}$ if there exist sequences $h^\nu \to h$ and $\lambda^\nu \searrow 0$ such that $\bar{u} + \lambda^\nu h^\nu \in U$ for all $\nu \in \mathbb{I}_{\geq 0}$. This equivalence can be seen by identifying u^ν with $\bar{u} + \lambda^\nu h^\nu$.

Proposition C.5 (Tangent vectors are closed cone)**.** *The set $\mathcal{T}_U(u)$ of all tangent vectors to U at any point $u \in U$ is a closed cone.*

See Rockafellar and Wets (1998), Proposition 6.2. That $\mathcal{T}_U(\bar{u})$ is a cone may be seen from its definition; if h is a tangent, so is αh for any $\alpha \geq 0$. Two examples of a tangent cone are illustrated in Figure C.3.

Associated with each tangent cone $\mathcal{T}_U(u)$ is a normal cone $\hat{N}(u)$ defined as follows Rockafellar and Wets (1998):

Definition C.6 (Regular normal)**.** A vector $g \in \mathbb{R}^m$ is a regular normal to a set $U \subset \mathbb{R}^m$ at $\bar{u} \in U$ if

$$\langle g, u - \bar{u} \rangle \leq o(|u - \bar{u}|) \ \forall u \in U \tag{C.15}$$

where $o(\cdot)$ has the property that $o(|u - \bar{u}|)/|u - \bar{u}| \to 0$ as $u \xrightarrow[U]{} \bar{u}$ with $u \neq \bar{u}$; $\hat{N}_U(u)$ is the set of all regular normal vectors.

Some examples of normal cones are illustrated in Figure C.3; here the set $\hat{N}_U(u) = \{\lambda g \mid \lambda \geq 0\}$ is a cone generated by a single vector g, say, while $\hat{N}_U(v) = \{\lambda_1 g_1 + \lambda_2 g_2 \mid \lambda_1 \geq 0, \lambda_2 \geq 0\}$ is a cone generated by two vectors g_1 and g_2, say. The term $o(|u - \bar{u}|)$ may be replaced by 0 if U is convex as shown in Proposition C.7(b) below but is needed in general since U may not be locally convex at $\bar{u}$ as illustrated in Figure C.4.

The tangent cone $\mathcal{T}_U(\bar{u})$ and the normal cone $\hat{N}_U(\bar{u})$ at a point $\bar{u} \in U$ are related as follows.

Proposition C.7 (Relation of normal and tangent cones)**.**

(a) At any point $\bar{u} \in U \subset \mathbb{R}^m$,

$$\hat{N}_U(\bar{u}) = \mathcal{T}_U(\bar{u})^* := \{g \mid \langle g, h\rangle \leq 0 \ \forall h \in \mathcal{T}_U(\bar{u})\}$$

where, for any cone V, $V^* := \{g \mid \langle g, h\rangle \leq 0 \ \forall h \in V\}$ denotes the polar cone of V.

(b) If U is convex, then, at any point $\bar{u} \in U$

$$\hat{N}_U(\bar{u}) = \{g \mid \langle g, u - \bar{u}\rangle \leq 0 \ \forall u \in U\} \tag{C.16}$$

Proof.

(a) To prove $\hat{N}_U(\bar{u}) \subset \mathcal{T}_U(\bar{u})^*$, we take an arbitrary point g in $\hat{N}_U(\bar{u})$ and show that $\langle g, h\rangle \leq 0$ for all $h \in \mathcal{T}(\bar{u})$ implying that $g \in \mathcal{T}_U^*(\bar{u})$. For, if h is tangent to U at $\bar{u}$, there exist, by definition, sequences $u^\nu \underset{U}{\longrightarrow} \bar{u}$ and $\lambda^\nu \searrow 0$ such that

$$h^\nu := (u^\nu - \bar{u})/\lambda^\nu \to h$$

Since $g \in \hat{N}_U(\bar{u})$, it follows from (C.15) that $\langle g, h^\nu\rangle \leq o(|(u^\nu - \bar{u})|) = o(\lambda^\nu |h^\nu|)$; the limit as $\nu \to \infty$ yields $\langle g, h\rangle \leq 0$, so that $g \in \mathcal{T}_U^*(\bar{u})$. Hence $\hat{N}_U(\bar{u}) \subset \mathcal{T}_U(\bar{u})^*$. The proof of this result, and the more subtle proof of the converse, that $\mathcal{T}_U(\bar{u})^* \subset \hat{N}_U(\bar{u})$, are given in Rockafellar and Wets (1998), Proposition 6.5.

(b) This part of the proposition is proved in (Rockafellar and Wets, 1998, Theorem 6.9). ■

Remark. A consequence of (C.16) is that for each $g \in \hat{N}_U(\bar{u})$, the half-space $H_g := \{u \mid \langle g, u - \bar{u}\rangle \leq 0\}$ supports the convex set U at $\bar{u}$, i.e., $U \subset H_g$ and $\bar{u}$ lies on the boundary of the half-space H_g.

We wish to derive optimality conditions for problems of the form $\mathbb{P} : \inf_u \{f(u) \mid u \in U\}$. The *value* of the problem is defined to be

$$f^0 := \inf_u \{f(u) \mid u \in U\}$$

There may not exist a $u \in U$ such that $f(u) = f^0$. If, however, $f(\cdot)$ is continuous and U is compact, there exists a minimizing u in U, i.e.,

$$f^0 = \inf_u \{f(u) \mid u \in U\} = \min_u \{f(u) \mid u \in U\}$$

The minimizing u, if it exists, may not be unique so

$$u^0 := \arg\min_u \{f(u) \mid u \in U\}$$

may be a set. We say u is feasible if $u \in U$. A point u is *globally optimal* for problem $\mathbb{P}$ if u is feasible and $f(v) \geq f(u)$ for all $v \in U$. A point u is *locally optimal* for problem $\mathbb{P}$ if u is feasible and there exists a $\varepsilon > 0$ such that $f(v) \geq f(u)$ for all v in $(u \oplus \varepsilon\mathcal{B}) \cap U$ where $\mathcal{B}$ is the closed unit ball $\{u \mid \min |u| \leq 1\}$.

C.2.2 Convex Optimization Problems

The optimization problem $\mathbb{P}$ is convex if the function $f : \mathbb{R}^m \to \mathbb{R}$ and the set $U \subset \mathbb{R}^m$ are convex. In convex optimization problems, U often takes the form $\{u \mid g_j(u) \leq 0, j \in \mathcal{J}\}$ where $\mathcal{J} := \{1, 2, \ldots, J\}$ and each function $g_j(\cdot)$ is convex. A useful feature of convex optimization problems is the following result:

Proposition C.8 (Global optimality for convex problems)**.** *Suppose the function $f(\cdot)$ is convex and differentiable and the set U is convex. Any locally optimal point of the convex optimization problem* $\inf_u \{f(u) \mid u \in U\}$ *is globally optimal.*

Proof. Suppose u is locally optimal so that there exists an $\varepsilon > 0$ such that $f(v) \geq f(u)$ for all $v \in (u \oplus \varepsilon\mathcal{B}) \cap U$. If, contrary to what we wish to prove, u is *not* globally optimal, there exists a $w \in U$ such that $f(w) < f(u)$. For any $\lambda \in [0, 1]$, the point $w_\lambda := \lambda w + (1 - \lambda)u$ lies in $[u, w]$ (the line joining u and w). Then $w_\lambda \in U$ (because U is convex) and $f(w_\lambda) \leq \lambda f(w) + (1 - \lambda) f(u) < f(u)$ for all $\lambda \in (0, 1]$ (because $f(\cdot)$ is convex and $f(w) < f(u)$). We can choose $\lambda > 0$ so that $w_\lambda \in (u \oplus \varepsilon\mathcal{B}) \cap U$ and $f(w_\lambda) < f(u)$. This contradicts the local optimality of u. Hence u is globally optimal. ∎

On the assumption that $f(\cdot)$ is differentiable, we can obtain a simple necessary and sufficient condition for the (global) optimality of a point u.

Proposition C.9 (Optimality conditions—normal cone)**.** *Suppose the function $f(\cdot)$ is convex and differentiable and the set U is convex. The point u is optimal for problem $\mathbb{P}$ if and only if $u \in U$ and*

$$df(u; v-u) = \langle \nabla f(u), v-u \rangle \geq 0 \ \forall v \in U \tag{C.17}$$

or, equivalently

$$-\nabla f(u) \in \hat{N}_U(u) \tag{C.18}$$

Proof. Because $f(\cdot)$ is convex, it follows from Theorem 7 in Appendix A1 that

$$f(v) \geq f(u) + \langle \nabla f(u), v-u \rangle \tag{C.19}$$

for all u, v in U. To prove sufficiency, suppose $u \in U$ and that the condition in (C.17) is satisfied. It then follows from (C.19) that $f(v) \geq f(u)$ for all $v \in U$ so that u is globally optimal. To prove necessity, suppose that u is globally optimal but that, contrary to what we wish to prove, the condition on the right-hand side of (C.17) is not satisfied so that there exists a $v \in U$ such that

$$df(u; h) = \langle \nabla f(u), v-u \rangle = -\delta < 0$$

where $h := v - u$. For all $\lambda \in [0,1]$, let $v_\lambda := \lambda v + (1-\lambda)u = u + \lambda h$; because U is convex, each v_λ lies in U. Since

$$df(u; h) = \lim_{\lambda \searrow 0} \frac{f(u+\lambda h) - f(u)}{\lambda} = \lim_{\lambda \searrow 0} \frac{f(v_\lambda) - f(u)}{\lambda} = -\delta$$

there exists a $\lambda \in (0,1]$ such that $f(v_\lambda) - f(u) \leq -\lambda\delta/2 < 0$ which contradicts the optimality of u. Hence the condition in (C.17) must be satisfied. That (C.17) is equivalent to (C.18) follows from Proposition C.7(b). ∎

Remark. The condition (C.17) implies that the linear approximation $\hat{f}(v) := f(u) + \langle \nabla f(u), v-u \rangle$ to $f(v)$ achieves its minimum over U at u.

It is an interesting fact that U in Proposition C.9 may be replaced by its approximation $u \oplus \mathcal{T}_U(u)$ at u yielding

Proposition C.10 (Optimality conditions—tangent cone)**.** *Suppose the function $f(\cdot)$ is convex and differentiable and the set U is convex. The point u is optimal for problem $\mathbb{P}$ if and only if $u \in U$ and*

$$df(u; v-u) = \langle \nabla f(u), h \rangle \geq 0 \ \forall h \in \mathcal{T}_U(u)$$

or, equivalently

$$-\nabla f(u) \in \hat{N}_U(u) = \mathcal{T}_U^*(u).$$

Proof. It follows from Proposition C.9 that u is optimal for problem $\mathbb{P}$ if and only if $u \in U$ and $-\nabla f(u) \in \hat{N}_U(u)$. But, by Proposition C.7, $\hat{N}_U(u) = \{g \mid \langle g, h \rangle \leq 0 \ \forall h \in \mathcal{T}_U(u)\}$ so that $-\nabla f(u) \in \hat{N}_U(u)$ is equivalent to $\langle \nabla f(u), h \rangle \geq 0$ for all $h \in \mathcal{T}_U(u)$. ∎

C.2.3 Convex Problems: Polyhedral Constraint Set

The definitions of tangent and normal cones given above may appear complex but this complexity is necessary for proper treatment of the general case when U is not necessarily convex. When U is polyhedral, i.e., when U is defined by a set of linear inequalities

$$U := \{u \in \mathbb{R}^m \mid Au \leq b\}$$

where $A \in \mathbb{R}^{p \times m}$ and $b \in \mathbb{R}^p$, $\mathcal{I} := \{1, 2, \ldots, p\}$, then the normal and tangent cones are relatively simple. We first note that U is equivalently defined by

$$U := \{u \in \mathbb{R}^m \mid \langle a_i, u \rangle \leq b_i, \ i \in \mathcal{I}\}$$

where a_i is the ith row of A and b_i is the ith element of b. For each $u \in U$, let

$$\mathcal{I}^0(u) := \{i \in \mathcal{I} \mid \langle a_i, u \rangle = b_i\}$$

denote the index set of constraints *active* at u. Clearly $\mathcal{I}^0(u) = \emptyset$ if u lies in the interior of U. An example of a polyhedral constraint set is shown in Figure C.5. The next result shows that in this case, the tangent cone is the set of h in $\mathbb{R}^m$ that satisfy $\langle a_i, h \rangle \leq 0$ for all i in $\mathcal{I}^0(u)$ and the normal cone is the cone generated by the vectors a_i, $i \in \mathcal{I}^0(u)$; each normal h in the normal cone may be expressed as $\sum_{i \in \mathcal{I}^0(u)} \mu_i a_i$ where each $\mu_i \geq 0$.

Proposition C.11 (Representation of tangent and normal cones)**.** *Let $U := \{u \in \mathbb{R}^m \mid \langle a_i, u \rangle \leq b_i, \ i \in \mathcal{I}\}$. Then, for any $u \in U$:*

$$\mathcal{T}_U(u) = \{h \mid \langle a_i, h \rangle \leq 0, \ i \in \mathcal{I}^0(u)\}$$
$$\hat{N}_U(u) = \mathcal{T}_U^*(u) = \text{cone}\{a_i \mid i \in \mathcal{I}^0(u)\}$$

Proof. (i) Suppose h is any vector in $\{h \mid \langle a_i, h\rangle \le 0,\ i \in \mathcal{I}^0(u)\}$. Let the sequences u^ν and λ^ν satisfy $u^\nu = u + \lambda^\nu h$ and $\lambda^\nu \searrow 0$ with λ^0, the first element in the sequence λ^ν, satisfying $u + \lambda^0 h \in U$. It follows that $[u^\nu - u]/\lambda^\nu \equiv h$ so that from Definition C.4, h is tangent to U at u. Hence $\{h \mid \langle a_i, h\rangle \le 0,\ i \in \mathcal{I}^0(u)\} \subset \mathcal{T}_U(u)$. (ii) Conversely, if $h \in \mathcal{T}_U(u)$, then there exist sequences $\lambda^\nu \searrow 0$ and $h^\nu \to h$ such that $\langle a_i, u + \lambda^\nu h^\nu\rangle \le b_i$ for all $i \in \mathcal{I}$, all $\nu \in \mathbb{I}_{\ge 0}$. Since $\langle a_i, u\rangle = b_i$ for all $i \in \mathcal{I}^0(u)$, it follows that $\langle a_i, h^\nu\rangle \le 0$ for all $i \in \mathcal{I}^0(u)$, all $\nu \in \mathbb{I}_{\ge 0}$; taking the limit yields $\langle a_i, h\rangle \le 0$ for all $i \in \mathcal{I}^0(u)$ so that $h \in \{h \mid \langle a_i, h\rangle \le 0,\ i \in \mathcal{I}^0(u)\}$ which proves $\mathcal{T}_U(u) \subset \{h \mid \langle a_i, h\rangle \le 0,\ i \in \mathcal{I}^0(u)\}$. We conclude from (i) and (ii) that $\mathcal{T}_U(u) = \{h \mid \langle a_i, h\rangle \le 0,\ i \in \mathcal{I}^0(u)\}$. That $\hat{N}_U(u) = \mathcal{T}_U^*(u) = \text{cone}\{a_i \mid i \in \mathcal{I}^0(u)\}$ then follows from Proposition C.7 above and Proposition 9 in Appendix A1. ∎

The next result follows from Proposition C.5 and Proposition C.7.

Proposition C.12 (Optimality conditions—linear inequalities)**.** *Suppose the function $f(\cdot)$ is convex and differentiable and U is the convex set $\{u \mid Au \le b\}$. Then u is optimal for $\mathbb{P} : \min_u\{f(u) \mid u \in U\}$ if and only if $u \in U$ and*

$$-\nabla f(u) \in \hat{N}_U(u) = \text{cone}\{a_i \mid i \in \mathcal{I}^0(u)\}$$

Corollary C.13 (Optimality conditions—linear inequalities)**.** *Suppose the function $f(\cdot)$ is convex and differentiable and $U = \{u \mid Au \le b\}$. Then u is optimal for $\mathbb{P} : \min_u\{f(u) \mid u \in U\}$ if and only if $Au \le b$ and there exist multipliers $\mu_i \ge 0, i \in \mathcal{I}^0(u)$ satisfying*

$$\nabla f(u) + \sum_{i \in \mathcal{I}^0(u)} \mu_i \nabla g_i(u) = 0 \tag{C.20}$$

where, for each i, $g_i(u) := \langle a_i, u\rangle - b_i$ so that $g_i(u) \le 0$ is the constraint $\langle a_i, u\rangle \le b_i$ and $\nabla g_i(u) = a_i$.

Proof. Since any point $g \in \text{cone}\{a_i \mid i \in \mathcal{I}^0(u)\}$ may be expressed as $g = \sum_{i \in \mathcal{I}^0(u)} \mu_i a_i$ where, for each i, $\mu_i \ge 0$, the condition $-\nabla f(u) \in \text{cone}\{a_i \mid i \in \mathcal{I}^0(u)\}$ is equivalent to the existence of multipliers $\mu_i \ge 0$, $i \in \mathcal{I}^0(u)$ satisfying (C.20). ∎

The above results are easily extended if U is defined by linear equality and inequality constraints, i.e., if

$$U := \{\langle a_i, u\rangle \le b_i,\ i \in \mathcal{I},\ \langle c_i, u\rangle = d_i,\ i \in \mathcal{E}\}$$

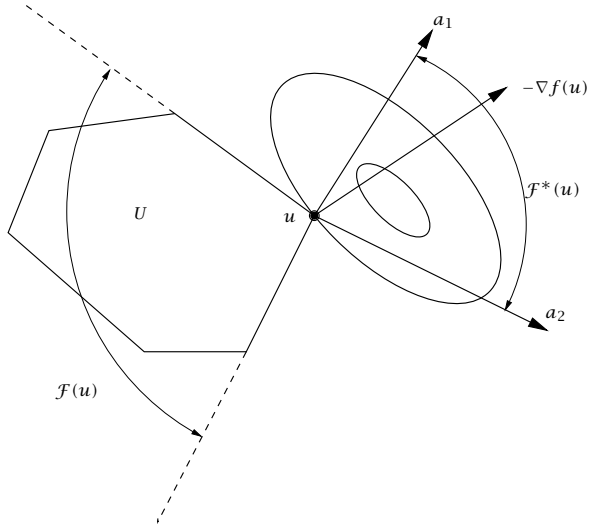


Figure C.5: Condition of optimality.

In this case, at any point $u \in U$, the tangent cone is

$$\mathcal{T}_U(u) = \{h \mid \langle a_i, h\rangle \leq 0,\ i \in \mathcal{I}^0(u),\ \langle c_i, h\rangle = 0,\ i \in \mathcal{E}\}$$

and the normal cone is

$$\hat{N}_U(u) = \{\sum_{i\in\mathcal{I}^0(u)} \lambda_i a_i + \sum_{i\in\mathcal{E}} \mu_i c_i \mid \lambda_i \geq 0\ \forall i \in \mathcal{I}^0(u), \mu_i \in \mathbb{R}\ \forall i \in \mathcal{E}\}$$

With U defined this way, u is optimal for $\min_u\{f(u) \mid u \in U\}$ where $f(\cdot)$ is convex and differentiable if and only if

$$-\nabla f(u) \in \hat{N}_U(u)$$

For each $i \in \mathcal{I}$ let $g_i(u) := \langle a_i, u\rangle - b_i$ and for each $i \in \mathcal{E}$, let $h_i(u) := \langle c_i, u\rangle - d_i$ so that $\nabla g(u_i) = a_i$ and $\nabla h_i = c_i$. It follows from the characterization of $\hat{N}_U(u)$ that u is optimal for $\min_u\{f(u) \mid u \in U\}$ if and only if there exist multipliers $\lambda_i \geq 0,\ i \in \mathcal{I}^0(u)$ and $\mu_i \in \mathbb{R},\ i \in \mathcal{E}$ such that

$$\nabla f(u) + \sum_{i\in\mathcal{I}^0(u)} \mu_i \nabla g_i(u) + \sum_{i\in\mathcal{E}} h_i(u) = 0 \tag{C.21}$$

C.2.4 Nonconvex Problems

We first obtain a necessary condition of optimality for the problem $\min\{f(u) \mid u \in U\}$ where $f(\cdot)$ is differentiable but not necessarily

convex and $U \subset \mathbb{R}^m$ is not necessarily convex; this result generalizes the necessary condition of optimality in Proposition C.9.

Proposition C.14 (Necessary condition for nonconvex problem)**.** *A necessary condition for u to be locally optimal for the problem of minimizing a differentiable function $f(\cdot)$ over the set U is*

$$df(u;h) = \langle \nabla f(u), h \rangle \geq 0, \ \forall h \in \mathcal{T}_U(u)$$

which is equivalent to the condition

$$-\nabla f(u) \in \hat{N}_U(u)$$

Proof. Suppose, contrary to what we wish to prove, that there exists a $h \in \mathcal{T}_U(u)$ and a $\delta > 0$ such that $\langle \nabla f(u), h \rangle = -\delta < 0$. Because $h \in \mathcal{T}_U(u)$, there exist sequences $h^\nu \underset{U}{\longrightarrow} h$ and $\lambda^\nu \searrow 0$ such that $u^\nu := u + \lambda^\nu h^\nu$ converges to u and satisfies $u^\nu \in U$ for all $\nu \in \mathbb{I}_{\geq 0}$. Then

$$f(u^\nu) - f(u) = \langle \nabla f(u), \lambda^\nu h^\nu \rangle + o(\lambda^\nu |h^\nu|)$$

Hence

$$[f(u^\nu) - f(u)]/\lambda^\nu = \langle \nabla f(u), h^\nu \rangle + o(\lambda^\nu)/\lambda^\nu$$

where we make use of the fact that $|h^\nu|$ is bounded for ν sufficiently large. It follows that

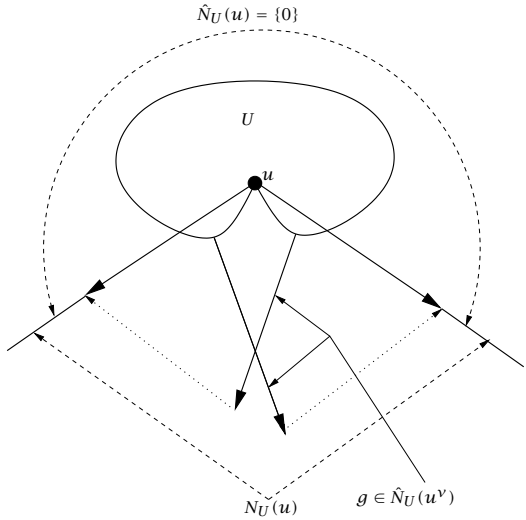
$$[f(u^\nu) - f(u)]/\lambda^\nu \to \langle \nabla f(u), h \rangle = -\delta$$

so that there exists a finite integer j such that $f(u^j) - f(u) \leq -\lambda^j \delta/2 < 0$ which contradicts the local optimality of u. Hence $\langle \nabla f(u), h \rangle \geq 0$ for all $h \in \mathcal{T}_U(u)$. That $-\nabla f(u) \in \hat{N}_U(u)$ follows from Proposition C.7. ∎

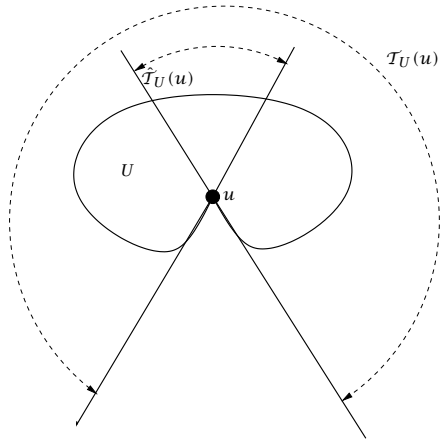
A more concise proof proceeds as follows Rockafellar and Wets (1998). Since $f(v) - f(u) = \langle \nabla f(u), v - u \rangle + o(|v - u|)$ it follows that $\langle -\nabla f(u), v - u \rangle = o(|v - u|) - (f(v) - f(u))$. Because u is locally optimal, $f(v) - f(u) \geq 0$ for all v in the neighborhood of u so that $\langle -\nabla f(u), v - u \rangle \leq o(|v - u|)$ which, by (C.15), is the definition of a normal vector. Hence $-\nabla f(u) \in \hat{N}_U(u)$.

C.2.5 Tangent and Normal Cones

The material in this section is *not* required for Chapters 1-7; it is presented merely to show that alternative definitions of tangent and normal cones are useful in more complex situations than those considered



(a) Normal cones.



(b) Tangent cones.

Figure C.6: Tangent and normal cones.

above. Thus, the normal and tangent cones defined in C.2.1 have some limitations when U is not convex or, at least, not similar to the constraint set illustrated in Figure C.4. Figure C.6 illustrates the type of difficulty that may occur. Here the tangent cone $\mathcal{T}_U(u)$ is not convex, as shown in Figure C.6(b), so that the associated normal cone

$\hat{N}_U(u) = \mathcal{T}_U(u)^* = \{0\}$. Hence the necessary condition of optimality of u for the problem of minimizing a differentiable function $f(\cdot)$ over U is $\nabla f(u) = 0$; the only way a *differentiable* function $f(\cdot)$ can achieve a minimum over U at u is for the condition $\nabla f(u) = 0$ to be satisfied. Alternative definitions of normality and tangency are sometimes necessary. In Rockafellar and Wets (1998), a vector $g \in \hat{N}_U(u)$ is normal in the *regular* sense; a normal in the *general* sense is then defined by:

Definition C.15 (General normal)**.** A vector g is normal to U at u in the general sense if there exist sequences $u^\nu \xrightarrow[U]{} u$ and $g^\nu \to g$ where $g^\nu \in \hat{N}_U(u^\nu)$ for all ν; $N_U(u)$ is the set of all general normal vectors.

The cone $N_U(u)$ of general normal vectors is illustrated in Figure C.6(a); here the cone $N_U(u)$ is the union of two distinct cones each having form $\{\alpha g \mid \alpha \geq 0\}$. Also shown in Figure C.6(a) are single elements of two sequences g^ν in $\hat{N}_U(u^\nu)$ converging to $N_U(u)$. Counter intuitively, the general normal vectors in this case point into the interior of U. Associated with $N_U(u)$ is the set $\hat{\mathcal{T}}_U(u)$ of regular tangents to U at u defined, when U is locally closed,[3] in (Rockafellar and Wets, 1998, Theorem 6.26) by:

Definition C.16 (General tangent)**.** Suppose U is locally closed at u. A vector h is tangent to U at u in the regular sense if, for all sequences $u^\nu \xrightarrow[U]{} u$, there exists a sequence $h^\nu \to h$ that satisfies $h^\nu \in \mathcal{T}_u(u^\nu)$ for all ν; $\hat{\mathcal{T}}_U(u)$ is the set of all regular tangent vectors to U at u.

Alternatively, a vector h is tangent to U at u in the regular sense if, for all sequences $u^\nu \xrightarrow[U]{} u$ and $\lambda^\nu \searrow 0$, there exists a sequence $h^\nu \to h$ satisfying $u^\nu + \lambda^\nu h^\nu \in U$ for all $\nu \in \mathbb{I}_{\geq 0}$. The cone of regular tangent vectors for the example immediately above is shown in Figure C.6(b). The following result is proved in Rockafellar and Wets (1998), Theorem 6.26:

Proposition C.17 (Set of regular tangents is closed convex cone)**.** *At any $u \in U$, the set $\hat{\mathcal{T}}_U(u)$ of regular tangents to U at u is a closed convex cone with $\hat{\mathcal{T}}_U(u) \subset \mathcal{T}_U(u)$. Moreover, if U is locally closed at u, then $\hat{\mathcal{T}}_U(u) = N_U(u)^*$.*

[3] A set U is locally closed at a point u if there exists a closed neighborhood $\mathcal{N}$ of u such that $U \cap \mathcal{N}$ is closed; U is locally closed if it is locally closed at all u.

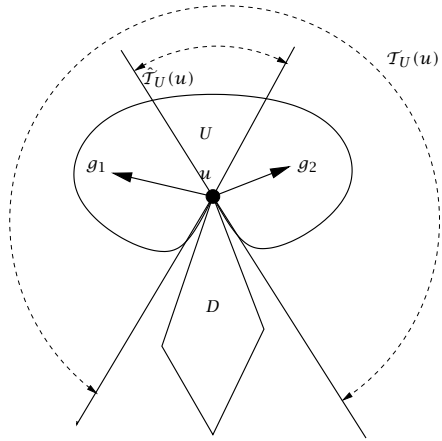


Figure C.7: Condition of optimality.

Figure C.7 illustrates some of these results. In Figure C.7, the constant cost contour $\{v \mid f(v) = f(u)\}$ of a *nondifferentiable* cost function $f(\cdot)$ is shown together with a sublevel set D passing through the point u: $f(v) \leq f(u)$ for all $v \in D$. For this example, $df(u;h) = \max\{\langle g_1, h\rangle, \langle g_2, h\rangle\}$ where g_1 and g_2 are normals to the level set of $f(\cdot)$ at u so that $df(u;h) \geq 0$ for all $h \in \hat{\mathcal{T}}_U(u)$, a necessary condition of optimality; on the other hand, there exist $h \in \mathcal{T}_U(u)$ such that $df(u;h) < 0$. The situation is simpler if the constraint set U is *regular* at u.

Definition C.18 (Regular set)**.** A set U is regular at a point $u \in U$ in the sense of Clarke if it is locally closed at u and if $N_U(u) = \hat{N}_U(u)$ (all normal vectors at u are regular).

The following consequences of Clarke regularity are established in Rockafellar and Wets (1998), Corollary 6.29:

Proposition C.19 (Conditions for regular set)**.** *Suppose U is locally closed at $u \in U$. Then U is regular at u is equivalent to each of the following.*

(a) $N_U(u) = \hat{N}_U(u)$ *(all normal vectors at u are regular).*

(b) $\mathcal{T}_U(u) = \hat{\mathcal{T}}_U(u)$ *(all tangent vectors at u are regular).*

(c) $N_U(u) = \mathcal{T}_U(u)^*$.

(d) $\mathcal{T}_U(u) = N_U(u)^*$.

(e) $\langle g, h\rangle \leq 0$ *for all* $h \in \mathcal{T}_U(u)$, *all* $g \in N_U(u)$.

It is shown in Rockafellar and Wets (1998) that if U is regular at u and a constraint qualification is satisfied, then a necessary condition of optimality, similar to (C.21), may be obtained. To obtain this result, we pursue a slightly different route in Sections C.2.6 and C.2.7.

C.2.6 Constraint Set Defined by Inequalities

We now consider the case when the set U is specified by a set of differentiable inequalities:

$$U := \{u \mid g_i(u) \leq 0 \ \forall i \in \mathcal{I}\} \tag{C.22}$$

where, for each $i \in \mathcal{I}$, the function $g_i : \mathbb{R}^m \to \mathbb{R}$ is differentiable. For each $u \in U$

$$\mathcal{I}^0(u) := \{i \in \mathcal{I} \mid g_i(u) = 0\}$$

is the index set of active constraints. For each $u \in U$, the set $\mathcal{F}_U(u)$ of feasible variations for the *linearized* set of inequalities; $\mathcal{F}_U(u)$ is defined by

$$\mathcal{F}_U(u) := \{h \mid \langle \nabla g_i(u), h\rangle \leq 0 \ \forall i \in \mathcal{I}^0(u)\} \tag{C.23}$$

The set $\mathcal{F}_U(u)$ is a closed, convex cone and is called a cone of first order feasible variations in Bertsekas (1999) because h is a descent direction for $g_i(u)$ for all $i \in \mathcal{I}^0(u)$, i.e., $g_i(u + \lambda h) \leq 0$ for all λ sufficiently small. When U is polyhedral, the case discussed in C.2.3, $g_i(u) = \langle a_i, u\rangle - b_i$ and $\nabla g_i(u) = a_i$ so that $\mathcal{F}_U(u) = \{h \mid \langle a_i, h\rangle \leq 0 \ \forall i \in \mathcal{I}^0(u)\}$ which was shown in Proposition C.11 to be the tangent cone $\mathcal{T}_U(u)$. An important question whether $\mathcal{F}_U(u)$ is the tangent cone $\mathcal{T}_U(u)$ for a wider class of problems because, if $\mathcal{F}_U(u) = \mathcal{T}_U(u)$, a condition of optimality of the form in (C.20) may be obtained. In the example in Figure C.8, $\mathcal{F}_U(u)$ is the horizontal axis $\{h \in \mathbb{R}^2 \mid h_2 = 0\}$ whereas $\mathcal{T}_U(u)$ is the half-line $\{h \in \mathbb{R}^2 \mid h_1 \geq 0,\ h_2 = 0\}$ so that in this case, $\mathcal{F}_U(u) \neq \mathcal{T}_U(u)$. While $\mathcal{F}_U(u)$ is always convex, being the intersection of a set of half-spaces, the tangent cone $\mathcal{T}_U(u)$ is not necessarily convex as Figure C.6b shows. The set U is said to be *quasiregular* at $u \in U$ if $\mathcal{F}_U(u) = \mathcal{T}_U(u)$ is which case u is said to be a quasiregular point Bertsekas (1999). The next result, due to Bertsekas (1999), shows that $\mathcal{F}_U(u) = \mathcal{T}_U(u)$, i.e., U is quasiregular at u, when a certain constraint qualification is satisfied.

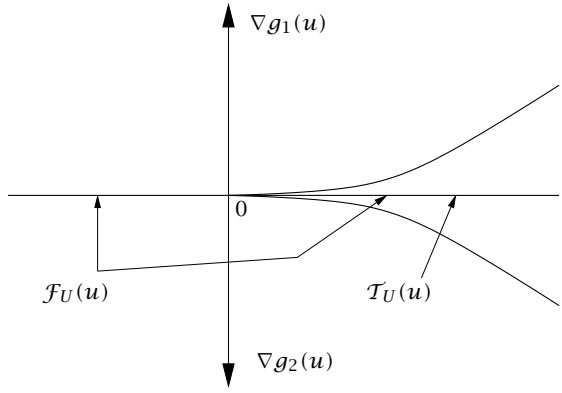


Figure C.8: $\mathcal{F}_U(u) \neq \mathcal{T}_U(u)$.

Proposition C.20 (Quasiregular set)**.** *Suppose* $U := \{u \mid g_i(u) \leq 0 \ \forall i \in \mathcal{I}\}$ *where, for each* $i \in \mathcal{I}$*, the function* $g_i : \mathbb{R}^m \to \mathbb{R}$ *is differentiable. Suppose also that* $u \in U$ *and that there exists a vector* $\bar{h} \in \mathcal{F}_U(u)$ *such that*

$$\langle \nabla g_i(u), \bar{h} \rangle < 0, \ \forall \ i \in \mathcal{I}^0(u) \tag{C.24}$$

Then

$$\mathcal{T}_U(u) = \mathcal{F}_U(u)$$

i.e., U *is quasiregular at* u*.*

Equation (C.24) is the constraint qualification; it can be seen that it precludes the situation shown in Figure C.8.

Proof. It follows from the definition (C.23) of $\mathcal{F}_U(u)$ and the constraint qualification (C.24) that:

$$\langle \nabla g_i(u), h + \alpha(\bar{h} - h) \rangle < 0, \ \forall h \in \mathcal{F}_U(u), \alpha \in (0,1], i \in \mathcal{I}^0(u)$$

Hence, for all $h \in \mathcal{F}_U(u)$, all $\alpha \in (0,1]$, there exists a vector $h_\alpha := h + \alpha(\bar{h} - h)$, in $\mathcal{F}_U(u)$ satisfying $\langle \nabla g_i(u), h_\alpha \rangle < 0$ for all $i \in \mathcal{I}^0(u)$. Assuming for the moment that $h_\alpha \in \mathcal{T}_U(u)$ for all $\alpha \in (0,1]$, it follows, since $h_\alpha \to h$ as $\alpha \to 0$ and $\mathcal{T}_U(u)$ is closed, that $h \in \mathcal{T}_U(u)$, thus proving $\mathcal{F}_U(u) \subset \mathcal{T}_U(u)$. It remains to show that h_α is tangent to U at u. Consider the sequences h^ν and $\lambda^\nu \searrow 0$ where $h^\nu := h_\alpha$ for all $\nu \in \mathbb{I}_{\geq 0}$. There exists a $\delta > 0$ such that $\langle \nabla g_i(u), h_\alpha \rangle \leq -\delta$ for all $i \in \mathcal{I}^0(u)$ and $g_i(u) \leq -\delta$ for all $i \in \mathcal{I} \setminus \mathcal{I}^0(u)$. Since

$$g_i(u + \lambda^\nu h^\nu) = g_i(u) + \lambda^\nu \langle \nabla g_i(u), h_\alpha \rangle + o(\lambda^\nu) \leq -\lambda^\nu \delta + o(\lambda^\nu)$$

for all $i \in \mathcal{I}^0(u)$, it follows that there exists a finite integer N such that $g_i(u + \lambda^\nu h^\nu) \leq 0$ for *all* $i \in \mathcal{I}$, all $\nu \geq N$. Since the sequences $\{h^\nu\}$ and $\{\lambda^\nu\}$ for all $\nu \geq N$ satisfy $h^\nu \to h_\alpha$, $\lambda^\nu \searrow 0$ and $u + \lambda^\nu h^\nu \in U$ for all $i \in \mathcal{I}$, it follows that $h_\alpha \in \mathcal{T}_U(u)$, thus completing the proof that $\mathcal{F}_U(u) \subset \mathcal{T}_U(u)$.

Suppose now that $h \in \mathcal{T}_U(u)$. There exist sequences $h^\nu \to h$ and $\lambda^\nu \to 0$ such that $u + \lambda^\nu h^\nu \in U$ so that $g(u + \lambda^\nu h^\nu) \leq 0$ for all $\nu \in \mathbb{I}_{\geq 0}$. Since $g(u + \lambda^\nu h^\nu) = g(u) + \langle \nabla g_j(u), \lambda^\nu h^\nu \rangle + o(\lambda^\nu |h^\nu|) \leq 0$, it follows that $\langle \nabla g_j(u), \lambda^\nu h^\nu \rangle + o(\lambda^\nu) \leq 0$ for all $j \in \mathcal{I}^0(u)$, all $\nu \in \mathbb{I}_{\geq 0}$. Hence $\langle \nabla g_j(u), h^\nu \rangle + o(\lambda^\nu)/\lambda^\nu \leq 0$ for all $j \in \mathcal{I}^0(u)$, all $\nu \in \mathbb{I}_{\geq 0}$. Taking the limit yields $\langle \nabla g_j(u), h^\nu \rangle \leq 0$ for all $j \in \mathcal{I}^0(u)$ so that $h \in \mathcal{F}_U(u)$ which proves $\mathcal{T}_U(u) \subset \mathcal{F}_U(u)$. Hence $\mathcal{T}_U(u) = \mathcal{F}_U(u)$. ∎

The existence of a $\bar{h}$ satisfying (C.24) is, as we have noted above, a constraint qualification. If u is locally optimal for the inequality constrained optimization problem of minimizing a differentiable function $f(\cdot)$ over the set U defined in (C.22) and, if (C.24) is satisfied thereby ensuring that $\mathcal{T}_U(u) = \mathcal{F}_U(u)$, then a condition of optimality of the form (C.20) may be easily obtained as shown in the next result.

Proposition C.21 (Optimality conditions nonconvex problem)**.** *Suppose u is locally optimal for the problem of minimizing a differentiable function $f(\cdot)$ over the set U defined in* (C.22) *and that $\mathcal{T}_U(u) = \mathcal{F}_U(u)$. Then*

$$-\nabla f(u) \in \text{cone}\{\nabla g_i(u) \mid i \in \mathcal{I}^0(u)\}$$

and there exist multipliers $\mu_i \geq 0$, $i \in \mathcal{I}^0(u)$ satisfying

$$\nabla f(u) + \sum_{i \in \mathcal{I}^0(u)} \mu_i \nabla g_i(u) = 0 \tag{C.25}$$

Proof. It follows from Proposition C.14 that $-\nabla f(u) \in \hat{\mathcal{N}}_U(u)$ and from Proposition C.7 that $\hat{\mathcal{N}}_U(u) = \mathcal{T}_U^*(u)$. But, by hypothesis, $\mathcal{T}_U(u) = \mathcal{F}_U(u)$ so that $\hat{\mathcal{N}}_U(u) = \mathcal{F}_U^*(u)$, the polar cone of $\mathcal{F}_U(u)$. It follows from (C.23) and the definition of a polar cone, given in Appendix A1, that

$$\mathcal{F}_U^*(u) = \text{cone}\{\nabla g_i(u) \mid i \in| I^0(u)\}$$

Hence

$$-\nabla f(u) \in \text{cone}\{\nabla g_i(u) \mid i \in \mathcal{I}^0(u)\}$$

The existence of multipliers μ_i satisfying (C.25) follows from the definition of a cone generated by $\{\nabla g_i(u) \mid i \in \mathcal{I}^0(u)\}$. ∎

C.2.7 Constraint Set Defined by Equalities and Inequalities

Finally, we consider the case when the set U is specified by a set of differentiable equalities *and* inequalities:

$$U := \{u \mid g_i(u) \leq 0 \ \forall i \in \mathcal{I},\ h_i(u) = 0 \ \forall i \in \mathcal{E}\}$$

where, for each $i \in \mathcal{I}$, the function $g_i : \mathbb{R}^m \to \mathbb{R}$ is differentiable and for each $i \in \mathcal{E}$, the function $h_i : \mathbb{R}^m \to \mathbb{R}$ is differentiable. For each $u \in U$

$$\mathcal{I}^0(u) := \{i \in \mathcal{I} \mid g_i(u) = 0\}$$

the index set of active inequality constraints is defined as before. We wish to obtain necessary conditions for the problem of minimizing a differentiable function $f(\cdot)$ over the set U. The presence of equality constraints makes this objective more difficult than for the case when U is defined merely by differentiable inequalities. The result we wish to prove is a natural extension of Proposition C.21 in which the equality constraints are included in the set of active constraints:

Proposition C.22 (Fritz-John necessary conditions)**.** *Suppose u is a local minimizer for the problem of minimizing $f(u)$ subject to the constraint $u \in U$ where U is defined in* (C.22)*. Then there exist multipliers μ_0, $\mu_i,\ i \in \mathcal{I}$ and $\lambda_i,\ i \in \mathcal{E}$, not all zero, such that*

$$\mu_0 \nabla f(u) + \sum_{i \in \mathcal{I}} \mu_i \nabla g_i(u) + \sum_{j \in \mathcal{E}} \lambda_j \nabla h_j(u) = 0 \tag{C.26}$$

and

$$\mu_i g_i(u) = 0 \ \forall i \in \mathcal{I}$$

where $\mu_0 \geq 0$ and $\mu_i \geq 0$ for all $i \in \mathcal{I}^0$.

The condition $\mu_i g_i(u) = 0$ for all $i \in \mathcal{I}$ is known as the *complementarity* conditions and implies $\mu_i = 0$ for all $i \in \mathcal{I}$ such that $g_i(u) < 0$. If $\mu_0 > 0$, then (C.26) may be normalized by dividing each term by μ_0 yielding the more familiar expression

$$\nabla f(u) + \sum_{i \in \mathcal{I}} \mu_i \nabla g_i(u) + \sum_{j \in \mathcal{E}} \nabla h_j(u) = 0$$

We return to this point later. Perhaps the simplest method for proving Proposition C.22 is the penalty approach adopted by Bertsekas (1999), Proposition 3.3.5. We merely give an outline of the proof. The constrained problem of minimizing $f(v)$ over U is approximated, for each

$k \in \mathbb{I}_{\geq 0}$ by a penalized problem defined below; as k increases the penalized problem becomes a closer approximation to the constrained problem. For each $i \in \mathcal{I}$, we define

$$g_i^+(v) := \max\{g_i(v), 0\}$$

For each k, the penalized problem $\mathbb{P}^k$ is then defined as the problem of minimizing $F^k(v)$ defined by

$$F^k(v) := f(v) + (k/2)\sum_{i\in\mathcal{I}}(g_i^+(v))^2 + (k/2)\sum_{j\in\mathcal{E}}(h_j(v))^2 + (1/2)|v-u|^2$$

subject to the constraint

$$S := \{v \mid |v-u| \leq \varepsilon\}$$

where $\epsilon > 0$ is such that $f(u) \leq f(v)$ for all v in $S \cap U$. Let v^k denote the solution of $\mathbb{P}^k$. Bertsekas shows that $v^k \to u$ as $k \to \infty$ so that for all k sufficiently large, v^k lies in the interior of S and is, therefore, the unconstrained minimizer of $F^k(v)$. Hence for each k sufficiently large, v^k satisfies $\nabla F^k(v^k) = 0$, or

$$\nabla f(v^k) + \sum_{i\in\mathcal{I}} \bar{\mu}_i^k \nabla g(v^k) + \sum_{i\in\mathcal{E}} \bar{\lambda}_i^k \nabla h(v^k) = 0 \qquad \text{(C.27)}$$

where

$$\bar{\mu}_i^k := k g_i^+(v^k), \quad \bar{\lambda}_i^k := k h_i(v^k)$$

Let μ^k denote the vector with elements μ_i^k, $i \in \mathcal{I}$ and λ^k the vector with elements λ_i^k, $k \in \mathcal{E}$. Dividing (C.27) by δ^k defined by

$$\delta^k := [1 + |\mu^k|^2 + |\lambda^k|^2]^1/2$$

yields

$$\mu_0^k \nabla f(v^k) + \sum_{i\in\mathcal{I}} \mu_i^k \nabla g(v^k) + \sum_{j\in\mathcal{E}} \lambda_j^k \nabla h(v^k) = 0$$

where

$$\mu_0^k := \bar{\mu}_i^k/\delta^k, \quad \mu_i^k := \bar{\mu}_i^k/\delta^k, \quad \lambda_j^k := \bar{\lambda}_i^k/\delta^k$$

and

$$(\mu_0^k)^2 + |\mu^k|^2 + |\lambda^k|^2 = 1$$

Because of the last equation, the sequence $\left(\mu_0^k, \mu^k, \lambda^k\right)$ lies in a compact set, and therefore has a subsequence, indexed by $K \subset \mathbb{I}_{\geq 0}$, converging to some limit (μ_0, μ, λ) where μ and λ are vectors whose elements are,

respectively, $\mu_i,\ i \in \mathcal{I}$ and $\lambda_j,\ j \in \mathcal{E}$. Because $v^k \to u$ as $k \in K$ tends to infinity, it follows from (C.27) that

$$\mu_0 \nabla f(u) + \sum_{i \in \mathcal{I}} \mu_i \nabla g_i(u) + \sum_{j \in \mathcal{E}} \lambda_j \nabla h_j(u) = 0$$

To prove the complementarity condition, suppose, contrary to what we wish to prove, that there exists a $i \in \mathcal{I}$ such that $g_i(u) < 0$ but $\mu_i > 0$. Since $\mu_i^k \to \mu_i > 0$ and $g_i(v^k) \to g_i(u)$ as $k \to \infty$, $k \in K$, it follows that $\mu_i \mu_i^k > 0$ for all $k \in K$ sufficiently large. But $\mu_i^k = \bar{\mu}_i^k / \delta^k = k g_i^+(v^k)/\delta^k$ so that $\mu_i \mu_i^k > 0$ implies $\mu_i g_i^+(v^k) > 0$ which in turn implies $g_i^+(v^k) = g_i(v^k) > 0$ for all $k \in K$ sufficiently large. This contradicts the fact that $g_i(v^k) \to g_i(u) < 0$ as $k \to \infty$, $k \in K$. Hence we must have $g_i(u) = 0$ for all $i \in \mathcal{I}$ such that $\mu_i > 0$.

The Fritz-John condition in Proposition C.22 is known as the Karush-Kuhn-Tucker (KKT) condition if $\mu_0 > 0$; if this is the case, μ_0 may be normalized to $\mu_0 = 1$. A constraint qualification is required for the Karush-Kuhn-Tucker condition to be a necessary condition of optimality for the optimization problem considered in this section. A simple constraint qualification is linear independence of $\{\nabla g_i(u),\ i \in \mathcal{I}^0(u),\ \nabla h_j(u),\ j \in \mathcal{E}\}$ at a local minimizer u. For, if u is a local minimizer and $\mu_0 = 0$, then the Fritz-John condition implies that $\sum_{i \in \mathcal{I}^0(u)} \mu_i \nabla g_i(u) + \sum_{j \in \mathcal{E}} \lambda_j \nabla h_j(u) = 0$ which contradicts the linear independence of $\{\nabla g_i(u),\ i \in \mathcal{I}^0(u),\ \nabla h_j(u),\ j \in \mathcal{E}\}$ since not all the multipliers are zero. Another constraint qualification, used in Propositions C.20 and C.21 for an optimization problem in which the constraint set is $U := \{u \mid g_i(u) \le 0,\ i \in \mathcal{I}\}$, is the existence of a vector $\bar{h}(u) \in \mathcal{F}_U(u)$ such that $\langle \nabla g_i(u), \bar{h} \rangle < 0$ for all $i \in \mathcal{I}^0(u)$; this condition ensures $\mu_0 = 1$ in C.25. Many other constraint qualifications exist; see, for example, Bertsekas (1999), Chapter 3.

C.3 Set-Valued Functions and Continuity of Value Function

A set-valued function $U(\cdot)$ is one for which, for each value of x, $U(x)$ is a set; these functions are encountered in parametric programming. For example, in the problem $\mathbb{P}(x) : \inf_u \{f(x,u) \mid u \in U(x)\}$ (which has the same form as an optimal control problem in which x is the state and u is a control sequence), the constraint set U is a set-valued function of the state. The solution to the problem $\mathbb{P}(x)$ (the value of u that achieves the minimum) can also be set-valued. It is important to

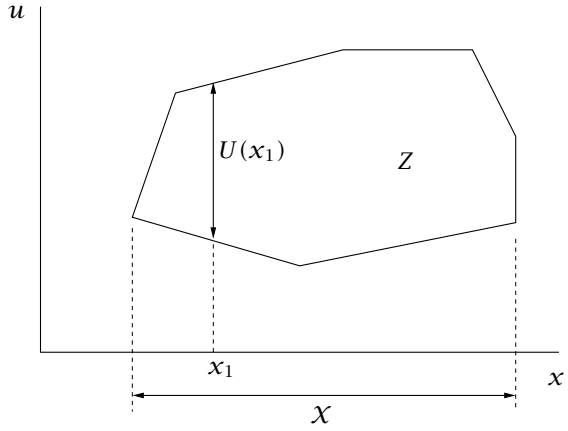
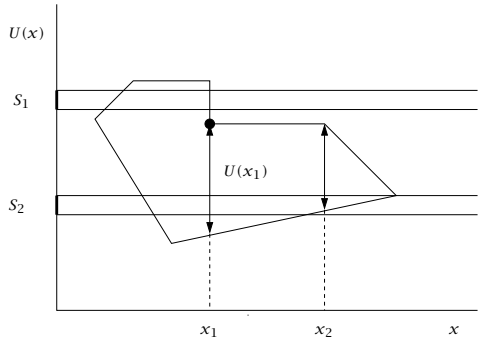


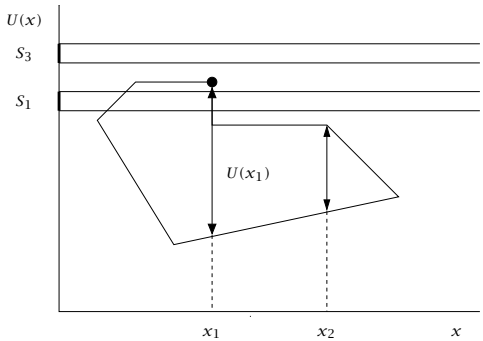
Figure C.9: Graph of set-valued function $U(\cdot)$.

know how smoothly these set-valued functions vary with the parameter x. In particular, we are interested in the continuity properties of the value function $x \mapsto f^0(x) = \inf_u\{f(x,u) \mid u \in U(x)\}$ since, in optimal control problems we employ the value function as a Lyapunov function and robustness depends, as we have discussed earlier, on the continuity of the Lyapunov function. Continuity of the value function depends, in turn, on continuity of the set-valued constraint set $U(\cdot)$. We use the notation $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ to denote the fact that $U(\cdot)$ maps points in $\mathbb{R}^n$ into subsets of $\mathbb{R}^m$.

The *graph* of a set-valued functions is often a useful tool. The graph of $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is defined to be the set $Z := \{(x,u) \in \mathbb{R}^n \times \mathbb{R}^m \mid u \in U(x)\}$; the *domain* of the set-valued function U is the set $\mathcal{X} := \{x \in \mathbb{R}^n \mid U(x) \neq \emptyset\} = \{x \in \mathbb{R}^n \mid \exists u \in \mathbb{R}^m \text{ such that } (x,u) \in Z\}$; clearly $\mathcal{X} \subset \mathbb{R}^n$. Also $\mathcal{X}$ is the *projection* of the set $Z \subset \mathbb{R}^n \times \mathbb{R}^m$ onto $\mathbb{R}^n$, i.e., $(x,u) \in Z$ implies $x \in \mathcal{X}$. An example is shown in Figure C.9. In this example, $U(x)$ varies continuously with x. Examples in which $U(\cdot)$ is discontinuous are shown in Figure C.10. In Figure C.10(a), the set $U(x)$ varies continuously if x increases from its initial value of x_1, but jumps to a much larger set if x decreases an infinitesimal amount (from its initial value of x_1); this is an example of a set-valued function that is inner semicontinuous at x_1. In Figure C.10(b), the set $U(x)$ varies continuously if x decreases from its initial value of x_1, but jumps to a much smaller set if x increases an infinitesimal amount (from its initial value of x_1); this is an example of a set-valued function that is



(a) Inner semicontinuous set-valued function.



(b) Outer semicontinuous set-valued function.

Figure C.10: Graphs of discontinuous set-valued functions.

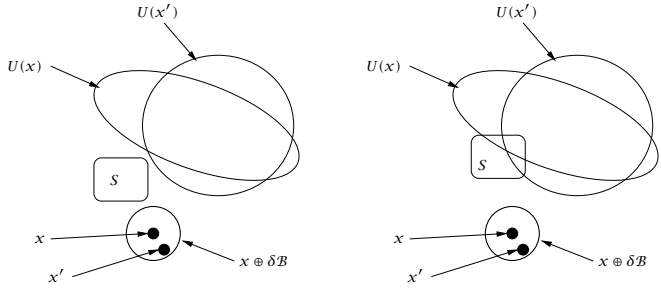
outer semicontinuous at x_1. The set-valued function is continuous at x_2 where it is both outer and inner semicontinuous.

We can now give precise definitions of inner and outer semicontinuity.

C.3.1 Outer and Inner Semicontinuity

The concepts of inner and outer semicontinuity were introduced by Rockafellar and Wets (1998, p. 144) to replace earlier definitions of lower and upper semicontinuity of set-valued functions. This section is based on the useful summary provided by Polak (1997, pp. 676-682).

Definition C.23 (Outer semicontinuous function)**.** A set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is said to be outer semicontinuous (osc) at x if $U(x)$



(a) Outer semicontinuity. (b) Inner semicontinuity.

Figure C.11: Outer and inner semicontinuity of $U(\cdot)$.

is closed and if, for every compact set S such that $U(x) \cap S = \emptyset$, there exists a $\delta > 0$ such that $U(x') \cap S = \emptyset$ for all $x' \in x \oplus \delta\mathcal{B}$.[4] The set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is outer semicontinuous if it is outer semicontinuous at each $x \in \mathbb{R}^n$.

Definition C.24 (Inner semicontinuous function)**.** A set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is said to be inner semicontinuous (isc) at x if, for every open set S such that $U(x) \cap S \neq \emptyset$, there exists a $\delta > 0$ such that $U(x') \cap S \neq \emptyset$ for all $x' \in x \oplus \delta\mathcal{B}$. The set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is inner semicontinuous if it is inner semicontinuous at each $x \in \mathbb{R}^n$.

These definitions are illustrated in Figure C.11. Roughly speaking, a set-valued function that is outer semicontinuous at x cannot explode as x changes to x' arbitrarily close to x; similarly, a set-valued function that is inner semicontinuous at x cannot collapse as x changes to x' arbitrarily close to x.

Definition C.25 (Continuous function)**.** A set-valued function is continuous (at x) if it is both outer and inner continuous (at x).

If we return to Figure C.10(a) we see that $S_1 \cap U(x_1) = \emptyset$ but $S_1 \cap U(x) \neq \emptyset$ for x infinitesimally less than x_1 so that $U(\cdot)$ is not outer semicontinuous at x_1. For all S_2 such that $S_2 \cap U(x_1) \neq \emptyset$, however, $S_2 \cap U(x) \neq \emptyset$ for all x in a sufficiently small neighborhood of x_1 so that $U(\cdot)$ is inner semicontinuous at x_1. If we turn to Figure C.10(b) we see that $S_1 \cap U(x_1) \neq \emptyset$ but $S_1 \cap U(x) = \emptyset$ for x infinitesimally greater than x_1 so that in this case $U(\cdot)$ is not inner semicontinuous at x_1. For all S_3 such that $S_3 \cap U(x_1) = \emptyset$, however, $S_3 \cap U(x) = \emptyset$ for

[4] Recall that $\mathcal{B} := \{x \mid |x| \leq 1\}$ is the closed unit ball in $\mathbb{R}^n$.

all x in a sufficiently small neighborhood of x_1 so that $U(\cdot)$ is outer semicontinuous at x_1.

The definitions of outer and inner semicontinuity may be interpreted in terms of infinite sequences (Rockafellar and Wets, 1998, p. 152), (Polak, 1997, pp. 677-678).

Theorem C.26 (Equivalent conditions for outer and inner semicontinuity)**.**

(a) A set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is outer semicontinuous at x if and only if for every infinite sequence (x_i) converging to x, any accumulation point[5] u of any sequence (u_i), satisfying $u_i \in U(x_i)$ for all i, lies in $U(x)$ ($u \in U(x)$).

(b) A set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is inner semicontinuous at x if and only if for every $u \in U(x)$ and for every infinite sequence (x_i) converging to x, there exists an infinite sequence (u_i), satisfying $u_i \in U(x_i)$ for all i, that converges to u.

Proofs of these results may be found in Rockafellar and Wets (1998); Polak (1997). Another result that we employ is:

Proposition C.27 (Outer semicontinuity and closed graph)**.** *A set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is outer semicontinuous in its domain if and only if its graph Z is closed in $\mathbb{R}^n \times \mathbb{R}^m$.*

Proof. Since $(x, u) \in Z$ is equivalent to $u \in U(x)$, this result is a direct consequence of the Theorem C.26. ∎

In the above discussion we have assumed, as in Polak (1997), that $U(x)$ is defined everywhere in $\mathbb{R}^n$; in constrained parametric optimization problems, however, $U(x)$ is defined on $\mathcal{X}$, a closed subset of $\mathbb{R}^n$; see Figure C.9. Only minor modifications of the above definitions are then required. In definitions C.23 and C.24 we replace the closed set $\delta\mathcal{B}$ by $\delta\mathcal{B} \cap \mathcal{X}$ and in Theorem C.26 we replace "every infinite sequence (in $\mathbb{R}^n$)" by "every infinite sequence in $\mathcal{X}$." In effect, we are replacing the topology of $\mathbb{R}^n$ by its topology relative to $\mathcal{X}$.

C.3.2 Continuity of the Value Function

Our main reason for introducing set-valued functions is to provide us with tools for analyzing the continuity properties of the value function and optimal control law in constrained optimal control problems.

[5] Recall, u is the limit of (u_i) if $u_i \to u$ as $i \to \infty$; u is an accumulation point of (u_i) if it is the limit of a subsequence of (u_i).

These problems have the form

$$V^0(x) = \min\{V(x,u) \mid u \in U(x)\} \tag{C.28}$$
$$u^0(x) = \arg\min\{V(x,u) \mid u \in U(x)\} \tag{C.29}$$

where $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is a set-valued function and $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous; in optimal control problems arising from MPC, u should be replaced by $\mathbf{u} = (u(0), u(1), \ldots, u(N-1))$ and m by Nm. We are interested in the continuity properties of the value function $V^0 : \mathbb{R}^n \to \mathbb{R}$ and the control law $u^0 : \mathbb{R}^n \to \mathbb{R}^m$; the latter may be set-valued (if the minimizer in (C.28) is not unique).

The following max problem has been extensively studied in the literature

$$\phi^0(x) = \max\{\phi(x,u) \mid u \in U(x)\}$$
$$\mu^0(x) = \arg\max\{\phi(x,u) \mid u \in U(x)\}$$

If we define $\phi(\cdot)$ by $\phi(x,u) := -V(x,u)$, we see that $\phi^0(x) = -V^0(x)$ and $\mu^0(x) = u^0(x)$ so that we can obtain the continuity properties of $V^0(\cdot)$ and $u^0(\cdot)$ from those of $\phi^0(\cdot)$ and $\mu^0(\cdot)$ respectively. Using this transcription and Corollary 5.4.2 and Theorem 5.4.3 in Polak (1997) we obtain the following result:

Theorem C.28 (Minimum theorem)**.** *Suppose that $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous, that $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is continuous, compact-valued and satisfies $U(x) \subset \mathbb{U}$ for all $x \in \mathcal{X}$ where $\mathbb{U}$ is compact. Then $V^0(\cdot)$ is continuous and $u^0(\cdot)$ is outer semicontinuous. If, in addition, $u^0(x) = \{\mu^0(x)\}$ (there is a unique minimizer $\mu^0(x)$), then $\mu^0(\cdot)$ is continuous.*

It is unfortunately the case, however, that due to state constraints, $U(\cdot)$ is often not continuous in constrained optimal control problems. If $U(\cdot)$ is constant, which is the case in optimal control problem if state or mixed control-state constraints are absent, then, from the above results, the value function $V^0(\cdot)$ is continuous. Indeed, under slightly stronger assumptions, the value function is Lipschitz continuous.

Lipschitz continuity of the value function. If we assume that $V(\cdot)$ is Lipschitz continuous and that $U(x) \equiv U$, we can establish Lipschitz continuity of $V^0(\cdot)$. Interestingly the result does not require, nor does it imply, Lipschitz continuity of the minimizer $u^0(\cdot)$.

Theorem C.29 (Lipschitz continuity of the value function, constant U)**.** *Suppose that $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous on bounded sets*[6] *and that $U(x) \equiv U$ where U is a compact subset of $\mathbb{R}^m$. Then $V^0(\cdot)$ is Lipschitz continuous on bounded sets.*

Proof. Let S be an arbitrary bounded set in $\mathcal{X}$, the domain of the value function $V^0(\cdot)$, and let $R := S \times \mathbb{U}$; R is a bounded subset of $\mathcal{Z}$. Since R is bounded, there exists a Lipschitz constant L_S such that

$$\left| V(x', u) - V(x'', u) \right| \leq L_S |x' - x''|$$

for all $x', x'' \in S$, all $u \in U$. Hence,

$$V^0(x') - V^0(x'') \leq V(x', u'') - V(x'', u'') \leq L_S |x' - x''|$$

for all $x', x'' \in S$, any $u'' \in u^0(x'')$. Interchanging x' and x'' in the above derivation yields

$$V^0(x'') - V^0(x') \leq V(x'', u') - V(x', u') \leq L_S |x'' - x'|$$

for all $x', x'' \in S$, any $u' \in u^0(x')$. Hence $V^0(\cdot)$ is Lipschitz continuous on bounded sets. ∎

We now specialize to the case where $U(x) = \{u \in \mathbb{R}^m \mid (x, u) \in \mathcal{Z}\}$ where $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{R}^m$; for each x, $U(x)$ is a polytope. This type of constraint arises in constrained optimal control problems when the system is linear and the state and control constraints are polyhedral. What we show in the sequel is that, in this special case, $U(\cdot)$ is continuous and so, therefore, is $V^0(\cdot)$. An alternative proof, which many readers may prefer, is given in Chapter 7 where we exploit the fact that if $V(\cdot)$ is strictly convex and quadratic and $\mathcal{Z}$ polyhedral, then $V^0(\cdot)$ is piecewise quadratic and continuous. Our first concern is to obtain a bound on $d(u, U(x'))$, the distance of any $u \in U(x)$ from the constraint set $U(x')$.

A bound on $d(u, U(x'))$, $u \in U(x)$. The bound we require is given by a special case of a theorem due to Clarke, Ledyaev, Stern, and Wolenski (1998, Theorem 3.1, page 126). To motivate this result, consider a differentiable convex function $f : \mathbb{R} \to \mathbb{R}$ so that $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle$ for any two points u and v in $\mathbb{R}$. Suppose also that there exists a nonempty interval $U = [a, b] \subset \mathbb{R}$ such that $f(u) \leq 0$ for

[6] A function $V(\cdot)$ is Lipschitz continuous on bounded sets if, for any bounded set S, there exists a constant $L_S \in [0, \infty)$ such that $|V(z') - V(z)| \leq L_S |z - z'|$ for all $z, z' \in S$.

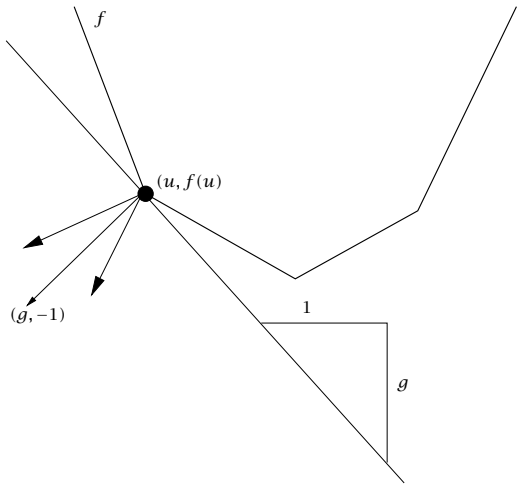


Figure C.12: Subgradient of $f(\cdot)$.

all $u \in U$ and that there exists a $\delta > 0$ such that $\Delta f(u) > \delta$ for all $u \in \mathbb{R}$. Let $u > b$ and let $v = b$ be the closest point in U to u. Then $f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle \geq \delta |v - u|$ so that $d(u, U) \leq f(u)/\delta$. The theorem of Clarke et al. (1998) extends this result to the case when $f(\cdot)$ is not necessarily differentiable but requires the concept of a subgradient of a convex function

Definition C.30 (Subgradient of convex function)**.** Suppose $f : \mathbb{R}^m \to \mathbb{R}$ is convex. Then the subgradient $\delta f(u)$ of $f(\cdot)$ at u is defined by

$$\delta f(u) := \{g \mid f(v) \geq f(u) + \langle g, v - u \rangle \ \forall v \in \mathbb{R}^m\}$$

Figure C.12 illustrates a subgradient. In the figure, g is one element of the subgradient because $f(v) \geq f(u) + \langle g, v - u \rangle$ for all v; g is the slope of the line passing through the point $(u, f(u))$. To obtain a bound on $d(u, U(x))$ we require the following result which is a special case of the much more general result of the theorem of Clarke *et al.*:

Theorem C.31 (Clarke et al. (1998))**.** *Take a nonnegative valued, convex function $\psi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Let $U(x) := \{u \in \mathbb{R}^m \mid \psi(x, u) = 0\}$ and $\mathcal{X} := \{x \in \mathbb{R}^n \mid U(x) \neq \emptyset\}$. Assume there exists a $\delta > 0$ such that*

$$u \in \mathbb{R}^m,\ x \in \mathcal{X}, \psi(x, u) > 0 \text{ and } g \in \partial_u \psi(x, u) \implies |g| > \delta$$

where $\partial_u \psi(x, u)$ denotes the convex subgradient of ψ with respect to

the variable u. Then, for each $x \in \mathcal{X}$, $d(u, U(x)) \leq \psi(x,u)/\delta$ for all $u \in \mathbb{R}^m$.

The proof of this result is given in the reference cited above. We next use this result to bound the distance of u from $U(x)$ where, for each x, $U(x)$ is polyhedral.

Corollary C.32 (A bound on $d(u, U(x'))$ for $u \in U(x)$)**.** [7] *Suppose $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{R}^m$ and let $\mathcal{X}$ denote its projection on $\mathbb{R}^n$ ($\mathcal{X} = \{x \mid \exists u \in \mathbb{R}^m$ such that $(x,u) \in \mathcal{Z}\}$). Let $\mathcal{U}(x) := \{u \mid (x, u) \in \mathcal{Z}\}$. Then there exists a $K > 0$ such that for all $x, x' \in \mathcal{X}$, $d(u, U(x')) \leq K|x' - x|$ for all $u \in U(x)$ (or, for all $x, x' \in \mathcal{X}$, all $u \in U(x)$, there exists a $u' \in U(x')$ such that $|u' - u| \leq K|x' - x|$).*

Proof. The polyhedron $\mathcal{Z}$ admits the representation $\mathcal{Z} = \{(x,u) \mid \langle m^j, u\rangle - \langle n^j, x\rangle - p^j \leq 0,\ j \in \mathcal{J}\}$ for some $m^j \in \mathbb{R}^m$, $n^j \in \mathbb{R}^n$ and $p^j \in \mathbb{R}$, $j \in \mathcal{J} := \{1, \ldots, J\}$. Define $\mathcal{D}$ to be the collection of all index sets $I \subseteq \mathcal{J}$ such that $\sum_{j\in I} \lambda^j m^j \neq 0,\ \forall \lambda \in \Lambda_I$ in which, for a particular index set I, Λ_I is defined to be $\Lambda_I := \{\lambda \mid \lambda^j \geq 0,\ \sum_{j\in I} \lambda^j = 1\}$. Because $\mathcal{D}$ is a finite set, there exists a $\delta > 0$ such that for all $I \in \mathcal{D}$, all $\lambda \in \Lambda_I$, $|\sum_{j\in I} \lambda^j m^j| > \delta$. Let $\psi(\cdot)$ be defined by $\psi(x,u) := \max\{\langle m^j, u\rangle - \langle n^j, x\rangle - p^j, 0 \mid j \in \mathcal{J}\}$ so that $(x,u) \in \mathcal{Z}$ (or $u \in \mathcal{U}(x)$) if and only if $\psi(x,u) = 0$. We now claim that, for every $(x,u) \in \mathcal{X} \times \mathbb{R}^m$ such that $\psi(x,u) > 0$ and every $g \in \partial_u \psi(x,u)$, the subgradient of ψ with respect to u at (x,u), we have $|g| > \delta$. Assuming for the moment that the claim is true, the proof of the Corollary may be completed with the aid of Theorem C.31. Assume, as stated in the Corollary, that $x, x' \in \mathcal{X}$ and $u \in \mathcal{U}(x)$; the theorem asserts

$$d(u, \mathcal{U}(x')) \leq (1/\delta)\psi(x', u),\ \forall x' \in \mathcal{X}$$

But $\psi(x,u) = 0$ (since $u \in \mathcal{U}(x)$) so that

$$d(u, \mathcal{U}(x')) \leq (1/\delta)[\psi(x', u) - \psi(x,u)] \leq (c/\delta)|x' - x|$$

where c is the Lipschitz constant for $x \mapsto \psi(x,u)$ ($\psi(\cdot)$ is piecewise affine and continuous). This proves the Corollary with $K = c/\delta$.

It remains to confirm the claim. Take any $(x,u) \in \mathcal{X} \times \mathbb{R}^m$ such that $\psi(x,u) > 0$. Then $\max_j\{\langle m^j, u\rangle - \langle n^j, x\rangle - p^j, 0 \mid j \in \mathcal{J}\} > 0$. Let

[7] The authors wish to thank Richard Vinter and Francis Clarke for providing this result.

$I^0(x,u)$ denote the active constraint set (the set of those constraints at which the maximum is achieved). Then

$$\langle m^j, u\rangle - \langle n^j, x\rangle - p^j > 0,\ \forall j \in I^0(x,u)$$

Since $x \in \mathcal{X}$, there exists a $\bar{u} \in \mathcal{U}(x)$ so that

$$\langle m^j, \bar{u}\rangle - \langle n^j, x\rangle - p^j \leq 0,\ \forall j \in I^0(x,u)$$

Subtracting these two inequalities yields

$$\langle m^j, u - \bar{u}\rangle > 0,\ \forall j \in I^0(x,u)$$

But then, for all $\lambda \in \Lambda_{I^0(x,u)}$, it follows that $|\sum_{j\in I^0(x,u)} \lambda^j m^j (u-\bar{u})| > 0$, so that

$$\sum_{j\in I^0(x,u)} \lambda^j m^j \neq 0$$

It follows that $I^0(x,u) \in \mathcal{D}$. Hence

$$\Big|\sum_{j\in I^0(x,u)} \lambda^j m^j\Big| > \delta,\ \forall \lambda \in \Lambda_{I^0(x,u)}$$

Now take any $g \in \partial_u f(x,u) = \text{co}\{m^j \mid j \in I^0(x,u)\}$ (co denotes "convex hull"). There exists a $\lambda \in \Lambda_{I^0(x,u)}$ such that $g = \sum_{j\in I^0(x,u)} \lambda^j m_j$. But then $|g| > \delta$ by the inequality above. This proves the claim and, hence, completes the proof of the Corollary. ■

Continuity of the value function when $U(x) = \{u \mid (x,u) \in \mathcal{Z}\}$. In this section we investigate continuity of the value function for the constrained linear quadratic optimal control problem $\mathbb{P}(x)$; in fact we establish continuity of the value function for the more general problem where the cost is continuous rather than quadratic. We showed in Chapter 2 that the optimal control problem of interest takes the form

$$V^0(x) = \min_u \{V(x,u) \mid (x,u) \in \mathcal{Z}\}$$

where $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{U}$ where $\mathbb{U} \subset \mathbb{R}^m$ is a polytope and, hence, is compact and convex; in MPC problems we replace the control u by the sequence of controls $\mathbf{u}$ and m by Nm. Let $u^0 : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ be defined by

$$u^0(x) := \arg\min_u \{V(x,u) \mid (x,u) \in \mathcal{Z}\}$$

and let $\mathcal{X}$ be defined by

$$\mathcal{X} := \{x \mid \exists u \text{ such that } (x,u) \in \mathcal{Z}\}$$

so that $\mathcal{X}$ is the projection of $\mathcal{Z} \subset \mathbb{R}^n \times \mathbb{R}^m$ onto $\mathbb{R}^n$. Let the set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ be defined by

$$U(x) \coloneqq \{u \in \mathbb{R}^m \mid (x, u) \in Z\}$$

The domain of $V^0(\cdot)$ and of $U(\cdot)$ is $\mathcal{X}$. The optimization problem may be expressed as $V^0(x) = \min_u \{V(x, u) \mid u \in U(x)\}$. Our first task is establish the continuity of $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$.

Theorem C.33 (Continuity of $U(\cdot)$)**.** *Suppose $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{U}$ where $\mathbb{U} \subset \mathbb{R}^m$ is a polytope. Then the set-valued function $U : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ defined above is continuous in $\mathcal{X}$.*

Proof. By Proposition C.27, the set-valued map $U(\cdot)$ is outer semicontinuous in $\mathcal{X}$ because its graph, $\mathcal{Z}$, is closed. We establish inner semicontinuity using Corollary C.32 above. Let x, x' be arbitrary points in $\mathcal{X}$ and $U(x)$ and $U(x')$ the associated control constraint sets. Let S be any open set such that $U(x) \cap S \neq \emptyset$ and let u be an arbitrary point in $U(x) \cap S$. Because S is open, there exist an $\varepsilon > 0$ such that $u \oplus \varepsilon\mathcal{B} \subset S$. Let $\varepsilon' \coloneqq \varepsilon / K$ where K is defined in Corollary 1. From Corollary C.32, there exists a $u' \in U(x')$ such that $|u' - u| \leq K|x' - x|$ which implies $|u' - u| \leq \varepsilon$ ($u' \in u \oplus \varepsilon\mathcal{B}$) for all $x' \in \mathcal{X}$ such that $|x' - x| \leq \varepsilon'$ ($x' \in (x \oplus \varepsilon'\mathcal{B}) \cap \mathcal{X}$). This implies $u \in \mathcal{U}(x') \cap S$ for all $x' \in \mathcal{X}$ such that $|x' - x| \leq \varepsilon'$ ($x' \in (x \oplus \varepsilon'\mathcal{B}) \cap \mathcal{X}$). Hence $\mathcal{U}(x') \cap S \neq \emptyset$ for all $x' \in (x \oplus \varepsilon'\mathcal{B}) \cap \mathcal{X}$, so that $\mathcal{U}(\cdot)$ is inner semicontinuous in $\mathcal{X}$. Since $U(\cdot)$ is both outer and inner semicontinuous in $\mathcal{X}$, it is continuous in $\mathcal{X}$. ∎

We can now establish continuity of the value function.

Theorem C.34 (Continuity of the value function)**.** *Suppose that $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous and that $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{U}$ where $\mathbb{U} \subset \mathbb{R}^m$ is a polytope. Then $V^0 : \mathbb{R}^n \to \mathbb{R}$ is continuous and $u^0 : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is outer semicontinuous in $\mathcal{X}$. Moreover, if $u^0(x)$ is unique (not set-valued) at each $x \in \mathcal{X}$, then $u^0 : \mathbb{R}^n \to \mathbb{R}^m$ is continuous in $\mathcal{X}$.*

Proof. Since the real-valued function $V(\cdot)$ is continuous (by assumption) and since the set-valued function $U(\cdot)$ is continuous in $\mathcal{X}$ (by Theorem C.33), it follows from Theorem C.28 that $V^0 : \mathbb{R}^n \to \mathbb{R}$ is continuous and $u^0 : \mathbb{R}^n \rightsquigarrow \mathbb{R}^m$ is outer semicontinuous in $\mathcal{X}$; it also follows that if $u^0(x)$ is unique (not set-valued) at each $x \in \mathcal{X}$, then $u^0 : \mathbb{R}^n \to \mathbb{R}^m$ is continuous in $\mathcal{X}$. ∎

Lipschitz continuity when $U(x) = \{u \mid (x,u) \in \mathcal{Z}\}$. Here we establish that $V^0(\cdot)$ is Lipschitz continuous if $V(\cdot)$ is Lipschitz continuous and $U(x) := \{u \in \mathbb{R}^m \mid (x,u) \in \mathcal{Z}\}$; this result is more general than Theorem C.29 where it is assumed that U is constant.

Theorem C.35 (Lipschitz continuity of the value function—$U(x)$)**.** *Suppose that $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous, that $\mathcal{Z}$ is a polyhedron in $\mathbb{R}^n \times \mathbb{U}$ where $\mathbb{U} \subset \mathbb{R}^m$ is a polytope. Suppose, in addition, that $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous on bounded sets.*[8] *Then $V^0(\cdot)$ is Lipschitz continuous on bounded sets.*

Proof. Let S be an arbitrary bounded set in $\mathcal{X}$, the domain of the value function $V^0(\cdot)$, and let $R := S \times \mathbb{U}$; R is a bounded subset of $\mathcal{Z}$. Let x, x' be two arbitrary points in S. Then

$$V^0(x) = V(x, \kappa(x))$$
$$V^0(x') = V(x', \kappa(x'))$$

where $V(\cdot)$ is the cost function, assumed to be Lipschitz continuous on bounded sets, and $\kappa(\cdot)$, the optimal control law, satisfies $\kappa(x) \in U(x) \subset \mathbb{U}$ and $\kappa(x') \in U(x') \subset \mathbb{U}$. It follows from Corollary C.32 that there exists a $K > 0$ such that for all $x, x' \in \mathcal{X}$, there exists a $u' \in U(x') \subset \mathbb{U}$ such that $|u' - \kappa(x)| \le K|x' - x|$. Since $\kappa(x)$ is optimal for the problem $\mathbb{P}(x)$, and since $(x, \kappa(x))$ and (x', u') both lie in $R = S \times \mathbb{U}$, there exists a constant L_R such that

$$\begin{aligned} V^0(x') - V^0(x) &\le V(x', u') - V(x, \kappa(x)) \\ &\le L_R(|(x', u') - (x, \kappa(x))|) \\ &\le L_R|x' - x| + L_R K|x' - x| \\ &\le M_S|x' - x|, \quad M_S := L_R(1 + K) \end{aligned}$$

Reversing the role of x and x' we obtain the existence of a $u \in \mathcal{U}(x)$ such that $|u - \kappa(x')| \le K|x - x'|$; it follows from the optimality of $\kappa(x')$ that

$$\begin{aligned} V^0(x) - V^0(x') &\le V(x, u) - V(x', \kappa(x')) \\ &\le M_S|x - x'| \end{aligned}$$

where, now, $u \in U(x)$ and $\kappa(x') \in U(x')$. Hence $|V^0(x') - V^0(x)| \le M_S|x - x'|$ for all x, x' in S. Since S is an arbitrary bounded set in $\mathcal{X}$, $V^0(\cdot)$ is Lipschitz continuous on bounded sets. ∎

[8] A function $V(\cdot)$ is Lipschitz continuous on bounded sets if, for any bounded set S, there exists a constant $L_S \in [0, \infty)$ such that $|V(z') - V(z)| \le L_S|z - z'|$ for all $z, z' \in S$.

C.4 Exercises

Exercise C.1: Nested optimization and switching order of optimization

Consider the optimization problem in two variables

$$\min_{(x,y)\in\mathbb{Z}} V(x,y)$$

in which $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $V : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. Assume this problem has a solution. This assumption is satisfied, for example, if V is continuous and $\mathbb{Z}$ is compact, but, in general, we do not require either of these conditions.

Define the following four sets

$$\mathbb{X}(y) = \{x \mid (x,y) \in \mathbb{Z}\} \qquad \mathbb{Y}(x) = \{y \mid (x,y) \in \mathbb{Z}\}$$
$$\mathbb{B} = \{y \mid \mathbb{X}(y) \neq \emptyset\} \qquad \mathbb{A} = \{x \mid \mathbb{Y}(x) \neq \emptyset\}$$

Note that $\mathbb{A}$ and $\mathbb{B}$ are the projections of $\mathbb{Z}$ onto $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Projection is defined in Section C.3. Show the solutions of the following two nested optimization problems exist and are equal to the solution of the original problem

$$\min_{x\in\mathbb{A}} \left(\min_{y\in\mathbb{Y}(x)} V(x,y) \right)$$
$$\min_{y\in\mathbb{B}} \left(\min_{x\in\mathbb{X}(y)} V(x,y) \right)$$

Exercise C.2: DP nesting

Prove the assertion made in Section C.1.2 that $\mathbf{u}^i = \{u, \mathbf{u}^{i+1}\} \in \mathcal{U}(x,i)$ if and only if $(x,u) \in \mathbb{Z}$, $f(x,u) \in X(i+1)$, and $\mathbf{u}^{i+1} \in \mathcal{U}(f(x,u), i+1)$.

Exercise C.3: Recursive feasibility

Prove the assertion in the proof of Theorem C.2 that $(x(j), u(j)) \in \mathbb{Z}$ and that $f(x(j), u(j)) \in X(j+1)$.

Exercise C.4: Basic minmax result

Consider the following two minmax optimization problems in two variables

$$\inf_{x\in\mathbb{X}} \sup_{y\in\mathbb{Y}} V(x,y) \qquad \sup_{y\in\mathbb{Y}} \inf_{x\in\mathbb{X}} V(x,y)$$

in which $x \in \mathbb{X} \subseteq \mathbb{R}^n$, $y \in \mathbb{Y} \subseteq \mathbb{R}^m$, and $V : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$.

(a) Show that the values are ordered as follows

$$\inf_{x\in\mathbb{X}} \sup_{y\in\mathbb{Y}} V(x,y) \geq \sup_{y\in\mathbb{Y}} \inf_{x\in\mathbb{X}} V(x,y)$$

or, if the solutions to the problems exist,

$$\min_{x\in\mathbb{X}} \max_{y\in\mathbb{Y}} V(x,y) \geq \max_{y\in\mathbb{Y}} \min_{x\in\mathbb{X}} V(x,y)$$

A handy mnemonic for this result is that the player who goes first (inner problem) has the advantage.[9]

[9] Note that different conventions are in use. Boyd and Vandenberghe (2004, p. 240) say that the player who "goes" *second* has the advantage, meaning that the inner problem is optimized *after* the outer problem has selected a value for its variable. We say that since the inner optimization is solved first, this player "goes" first.

(b) Use your results to order these three problems

$$\sup_{x\in\mathbb{X}} \inf_{y\in\mathbb{Y}} \sup_{z\in\mathbb{Z}} V(x,y,z) \qquad \inf_{y\in\mathbb{Y}} \sup_{z\in\mathbb{Z}} \sup_{x\in\mathbb{X}} V(x,y,z) \qquad \sup_{z\in\mathbb{Z}} \sup_{x\in\mathbb{X}} \inf_{y\in\mathbb{Y}} V(x,y,z)$$

Exercise C.5: Lagrange multipliers and minmax

Consider the constrained optimization problem

$$\min_{x\in\mathbb{R}^n} V(x) \qquad \text{subject to } g(x) = 0 \tag{C.30}$$

in which $V : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Introduce the Lagrange multiplier $\lambda \in \mathbb{R}^m$ and Lagrangian function $L(x,\lambda) = V(x) - \lambda' g(x)$ and consider the following minmax problem

$$\min_{x\in\mathbb{R}^n} \max_{\lambda\in\mathbb{R}^m} L(x,\lambda)$$

Show that if (x_0,λ_0) is a solution to this problem with finite $L(x_0,\lambda_0)$, then x_0 is also a solution to the original constrained optimization (C.30).

Exercise C.6: Dual problems and duality gap

Consider again the constrained optimization problem of Exercise C.5

$$\min_{x\in\mathbb{R}^n} V(x) \qquad \text{subject to } g(x) = 0$$

and its equivalent minmax formulation

$$\min_{x\in\mathbb{R}^n} \max_{\lambda\in\mathbb{R}^m} L(x,\lambda)$$

Switching the order of optimization gives the maxmin version of this problem

$$\max_{\lambda\in\mathbb{R}^m} \min_{x\in\mathbb{R}^n} L(x,\lambda)$$

Next define a new (dual) objective function $q : \mathbb{R}^m \to \mathbb{R}$ as the inner optimization

$$q(\lambda) = \min_{x\in\mathbb{R}^n} L(x,\lambda)$$

Then the maxmin problem can be stated as

$$\max_{\lambda\in\mathbb{R}^m} q(\lambda) \tag{C.31}$$

Problem (C.31) is known as the *dual* of the original problem (C.30), and the original problem (C.30) is then denoted as the *primal* problem in this context (Nocedal and Wright, 2006, p. 343–345), (Boyd and Vandenberghe, 2004, p. 223).

(a) Show that the solution to the dual problem is a lower bound for the solution to the primal problem

$$\max_{\lambda\in\mathbb{R}^m} q(\lambda) \le \min_{x\in\mathbb{R}^n} V(x) \qquad \text{subject to } g(x) = 0 \tag{C.32}$$

This property is known as *weak duality* (Nocedal and Wright, 2006, p. 345), (Boyd and Vandenberghe, 2004, p. 225).

(b) The difference between the dual and the primal solutions is known as the duality gap. *Strong duality* is defined as the property that equality is achieved in (C.32) and the duality gap is zero (Boyd and Vandenberghe, 2004, p. 225).

$$\max_{\lambda \in \mathbb{R}^m} q(\lambda) = \min_{x \in \mathbb{R}^n} V(x) \qquad \text{subject to } g(x) = 0 \tag{C.33}$$

Show that strong duality is equivalent to the existence of λ_0 such that

$$\min_{x \in \mathbb{R}^n} V(x) - \lambda_0' g(x) = \min_{x \in \mathbb{R}^n} V(x) \qquad \text{subject to } g(x) = 0 \tag{C.34}$$

Characterize the set of all λ_0 that satisfy this equation.

Exercise C.7: Example with duality gap

Consider the following function and sets (Peressini, Sullivan, and Uhl, Jr., 1988, p. 34)

$$V(x, y) = (y - x^2)(y - 2x^2) \qquad \mathbb{X} = [-1, 1] \qquad \mathbb{Y} = [-1, 1]$$

Make a contour plot of $V(\cdot)$ on $\mathbb{X} \times \mathbb{Y}$ and answer the following question. Which of the following two minmax problems has a nonzero duality gap?

$$\min_{y \in \mathbb{Y}} \max_{x \in \mathbb{X}} V(x, y)$$
$$\min_{x \in \mathbb{X}} \max_{y \in \mathbb{Y}} V(x, y)$$

Notice that the two problems are different because the first one minimizes over y and maximizes over x, and the second one does the reverse.

Exercise C.8: The Heaviside function and inner and outer semicontinuity

Consider the (set-valued) function

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x > 0 \end{cases}$$

and you are charged with deciding how to define $H(0)$.

(a) Characterize the choices of set $H(0)$ that make H outer semicontinuous. Justify your answer.

(b) Characterize the choices of set $H(0)$ that make H inner semicontinuous. Justify your answer.

(c) Can you define $H(0)$ so that H is both outer and inner semicontinuous? Explain why or why not.

Bibliography

R. E. Bellman. *Dynamic Programming.* Princeton University Press, Princeton, New Jersey, 1957.

D. P. Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, MA, second edition, 1999.

D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar. *Dynamic Programming and Optimal Control.* Athena Scientific, Belmont, MA 02478-0003, USA, 2001.

S. P. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

F. Clarke, Y. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory.* Springer-Verlag, New York, 1998.

J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, second edition, 2006.

A. L. Peressini, F. E. Sullivan, and J. J. Uhl, Jr. *The Mathematics of Nonlinear Programming.* Springer-Verlag, New York, 1988.

E. Polak. *Optimization: Algorithms and Consistent Approximations.* Springer Verlag, New York, 1997. ISBN 0-387-94971-2.

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer-Verlag, 1998.