Appendix

We gather in this appendix some background information for the analysis in the book, including definitions, proofs of results stated in the chapters, and some foundational results, such as linear programming duality and separation of convex sets.

A.1 Definitions and Basic Concepts

Sets. We assume familiarity with the ideas of *open*, *closed*, and *compact* sets. The *distance* of a point x to a set C is

$$\operatorname{dist}(x, C) = \inf_{y \in C} \|x - y\|. \tag{A.1}$$

The *closure* of a set C, denoted by $\operatorname{cl}(C)$, is the set of all points x such that $\operatorname{dist}(x, C) = 0$. The *interior* of a set C, denoted by $\operatorname{int}(C)$, is the largest open set contained in C.

A set C is *convex* if $x \in C, y \in C \implies \alpha x + (1-\alpha) y \in C$ for all $\alpha \in [0,1]$. A set C is *affine* if $x \in C, y \in C \implies \alpha x + (1-\alpha) y \in C$ for all $\alpha \in \mathbb{R}$.

The *affine hull* of a set C, denoted $\operatorname{aff}(C)$, is the smallest affine set containing C. An explicit definition is

$$\operatorname{aff}(C) := \left\{ \sum_{i=1}^m \alpha_i x^i \mid \sum_{i=1}^m \alpha_i = 1,\ x^i \in C,\ i = 1, 2, \ldots, m \right\}. \tag{A.2}$$

The *relative interior* of C, denoted $\operatorname{ri}(C)$, is the interior of C when regarded as a subset of its affine hull. Explicitly:

$$\begin{aligned} \operatorname{ri}(C) := \{x \in \operatorname{aff}(C) \mid \exists \epsilon > 0 \text{ such that } \|y - x\| < \epsilon \\ \text{and } y \in \operatorname{aff}(C) \Rightarrow y \in C\}. \end{aligned} \tag{A.3}$$

As examples, the set $C := [0,1] \times (0,1] \times \{1\} \subset \mathbb{R}^3$ has affine hull $\operatorname{aff}(C) = \mathbb{R}^2 \times \{1\}$ and relative interior $\operatorname{ri}(C) = (0,1) \times (0,1) \times \{1\}$.

When Ω is a convex set, we define multiplication by a nonnegative scalar α as follows:

$$\alpha\Omega := \{\alpha v :\ v \in \Omega\}.$$

We define set addition for convex sets Ω_i, $i = 1, 2, \ldots, m$, as follows:

$$\sum_{i=1}^{m} \Omega_i := \left\{ \sum_{i=1}^{m} v^i : \; v^i \in \Omega_i, \; i = 1, 2, \ldots, m \right\}.$$

The set $C \in \mathbb{R}^n$ is a *cone* if $x \in C \;\; \Rightarrow \;\; \alpha x \in C$ for all $\alpha > 0$. The *polar* C° of a cone C is defined by $C^\circ := \{y \mid y^T x \le 0 \;\text{ for all } x \in C\}$.

Order Notation. Given two sequences of nonnegative scalars $\{\eta_k\}$ and $\{\zeta_k\}$, with $\zeta_k \to \infty$, we write $\eta_k = O(\zeta_k)$ if there exists a constant M such that $\eta_k \le M\zeta_k$ for all k sufficiently large. The same definition holds if $\zeta_k \to 0$.

For sequences $\{\eta_k\}$ and $\{\zeta_k\}$, as before, we write $\eta_k = o(\zeta_k)$ if $\eta_k/\zeta_k \to 0$ as $k \to \infty$. We write $\eta_k = \Omega(\zeta_k)$ if both $\eta_k = O(\zeta_k)$ and $\zeta_k = O(\eta_k)$.

For a nonnegative sequence $\{\eta_k\}$, we write $\eta_k = o(1)$ if $\eta_k \to 0$.

We sometimes (as in Section 2.2) use order notation without explicitly defining sequences like $\{\eta_k\}$ and $\{\zeta_k\}$. Consider, for example, the expression (2.6), which is

$$f(x+p) = f(x) + \nabla f(x)^T p + o(\|p\|).$$

This usage can be reconciled with our previous definition by considering a sequence of vectors $\{p^k\}$ with $\|p^k\| \to 0$. We then have

$$f(x+p^k) = f(x) + \nabla f(x)^T p^k + o(\|p^k\|),$$

where the notation $o(\cdot)$ is defined as before. Even more specifically, if we define

$$r^k := f(x+p^k) - f(x) - \nabla f(x)^T p^k,$$

we have $\|r^k\| = o(\|p^k\|)$.

Convergence of Sequences. Given a sequence of points $\{x^k\}_{k=0,1,2,\ldots}$ with $x^k \in \mathbb{R}^n$ for all k, we say that $\bar{x}$ is the *limit* of this sequence if for any $\epsilon > 0$, there is k_ϵ such that $\|x^k - \bar{x}\| \le \epsilon$ for all $k > k_\epsilon$. We denote this by $\bar{x} = \lim_{k\to\infty} x^k$.

We say that $\bar{x}$ is an *accumulation point* of the sequence $\{x^k\}$ if, for any index K and any $\epsilon > 0$, there exists $k > K$ such that $\|x^k - \bar{x}\| \le \epsilon$. When this condition holds, we can define an infinite index set $\mathcal{S} \subset \{1, 2, \ldots\}$ such that $\lim_{k\to\infty, k\in\mathcal{S}} x^k = \bar{x}$.

When $\bar{x} = \lim_{k\to\infty} x^k$, we say that *the sequence* $\{x^k\}$ *converges Q-linearly to* $\bar{x}$ if there is $\rho \in (0,1)$ such that for all k sufficiently large, we have

$$\frac{\|x^{k+1} - \bar{x}\|}{\|x^k - \bar{x}\|} \le \rho.$$

We say that $\{x^k\}$ *converges R-linearly to* $\bar{x}$ if there is a sequence of positive scalars $\{\eta_k\}$ such that $\{\eta_k\}$ converges Q-linearly to zero, and $\|x^k - \bar{x}\| \le \eta_k$ for all k.

Linear Algebra. A symmetric matrix $A \in S\mathbb{R}^{n\times n}$ admits the eigenvalue decomposition $A = \sum_{i=1}^n \lambda_i u^i (u^i)^T$, where $\{u^1, u^2, \ldots, u^n\}$ is an orthonormal set of eigenvectors and $\lambda_i = \lambda_i(A)$ are the (real) eigenvalues, usually arranged in nonincreasing order.

We define $\lambda_{\max}(A) = \max_{i=1,2,\ldots,n} \lambda_i(A)$ and $\lambda_{\min}(A) = \min_{i=1,2,\ldots,n} \lambda_i(A)$. For such matrices, the trace equals the sum of eigenvalues; that is,

$$\text{trace}\,(A) = \sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i(A). \tag{A.4}$$

Jensen's Inequality and an Integral-Norm Inequality. Jensen's inequality can be stated in several forms, one of which is the following: Let $(\Omega, \mathcal{A}, \mu)$ be a probability space, so that $\mu(\Omega) = 1$. Suppose that g is a real-valued function that is μ-integrable, and that φ is a convex function on the real line. Then we have

$$\varphi\left(\int_{\Omega} g(s)\, d\mu(s)\right) \le \int_{\Omega} \varphi(g(s))\, d\mu(s). \tag{A.5}$$

Noting that the integral represents an expected value, then by relabeling the function g as a random variable X, we can rewrite this result as follows:

$$\varphi(\mathbb{E}(X)) \le \mathbb{E}(\varphi(X)). \tag{A.6}$$

A closely related result from analysis is the following: Let $(S, \mathcal{A}, \mu)$ be a measure space, $f\colon S \to X$ be integrable, where X is a Banach space equipped with norm $\|\cdot\|$. We then have

$$\left\| \int_S f(s)\, d\mu(s) \right\| \le \int_S \|f(s)\|\, d\mu(s). \tag{A.7}$$

Taylor's Theorem For Vector Functions. Taylor's theorem is a foundational result for smooth optimization, as it enables us to use derivative information about a function f at a particular point to estimate its behavior at nearby points. We included a discussion of Taylor's theorem for a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ in Chapter 2. Here we introduce a variant for vector functions $F\colon \mathbb{R}^n \to \mathbb{R}^n$, that is useful in analyzing systems of nonlinear equations.

Theorem A.1 *Let $F\colon \mathbb{R}^n \to \mathbb{R}^n$ be a system of nonlinear equations with continuously differentiable Jacobian $J(x)$. We then have for any $x, p \in \mathbb{R}^n$ that*

$$F(x+p) - F(x) = \int_0^1 J(x+tp)p\, dt.$$

Implicit Function Theorem. The implicit function theorem describes the sensitivity of a vector function $s(x) \in \mathbb{R}^p$ to its vector argument $x \in \mathbb{R}^n$, where there is an implicit relationship between function and argument that is defined in terms of another vector function $h(x, s(x)) = 0$, where $h \in \mathbb{R}^p$ has the same dimension as s.

We state the result rigorously as follows. (For a proof, see Lang, 1983, p. 131.)

Theorem A.2 *Let $h\colon \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^p$ be a function such that the following three conditions hold.*

(i) $h(x^*, s^*) = 0$ *for some* $s^* \in \mathbb{R}^p$ *and* $x^* \in \mathbb{R}^n$
(ii) $h(\cdot,\cdot)$ *is continuously differentiable in some neighborhood of* (x^*, s^*)
(iii) $\nabla_s h(x^*, s^*) \in \mathbb{R}^{p\times p}$ *is nonsingular*

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

Then there exist open sets $\mathcal{N}_s \in \mathbb{R}^p$ and $\mathcal{N}_x \in \mathbb{R}^n$ containing s^* and x^*, respectively, and a continuous function $s(\cdot): \mathbb{R}^n \to \mathbb{R}^p$, uniquely defined, such that $s(x^*) = s^*$, and $h(x, s(x)) = 0$ for all $x \in \mathcal{N}_x$. The gradient of the function s is defined by

$$\nabla s(x) = -\nabla_x h(x, s(x))[\nabla_s h(x, s(x))]^{-1}.$$

If h is $r \geq 1$ times continuously differentiable with respect to both its arguments, then $s(x)$ is also r times continuously differentiable with respect to x.

A.2 Convergence Rates and Iteration Complexity

We show here how convergence rate expressions, both linear and sublinear, can be used to obtain a lower bound on the number of iterations required to reduce the quantity of interest below a certain given threshold $\epsilon > 0$. This bound is often called the *iteration complexity* of the algorithm.

Denote by $\{\tau_k\}$ the sequence of nonnegative scalar quantities of interest, with $\tau_k \to 0$. We could have $\tau_k = f(x^k) - f^*$ (the difference between the function value at iteration k and its optimal value), or $\tau_k = \|\nabla f(x^k)\|$ (gradient norm), or $\tau_k = \text{dist}(x^k, \mathcal{S})$ (distance between current iterate x^k and the solution set), to mention three examples. We denote the target value for τ_k by $\epsilon > 0$ and obtain expressions for the number of iterations k require to guarantee $\tau_k \leq \epsilon$.

Suppose that we can prove sublinear convergence of the form

$$\tau_k \leq \frac{A}{k+B}, \quad k = 1, 2, \ldots,$$

for some scalars $A > 0$ and $B \geq 0$. Simple manipulation shows that we have $\tau_k \leq \epsilon$ whenever $k \geq (A/\epsilon) - B$.

Suppose instead that we have a slower form of sublinear convergence, namely,

$$\tau_k \leq \frac{A}{\sqrt{k+B}}, \quad k = 1, 2, \ldots .$$

In this case, we can guarantee $\tau_k \leq \epsilon$ for all $k \geq (A/\epsilon)^2 - B$.

Suppose that we are able to prove Q-linear convergence of $\{\tau_k\}$ to zero – that is,

$$\tau_{k+1} \leq (1 - \phi)\tau_k, \quad \text{for some } \phi \in (0, 1). \tag{A.8}$$

By applying the bound (A.8) recursively, we have

$$\tau_k \leq (1 - \phi)^{k-1}\tau_1, \quad k = 1, 2, \ldots .$$

Thus, we can guarantee $\tau_T \leq \epsilon$ when

$$(1 - \phi)^{T-1}\tau_1 \leq \epsilon.$$

When $\tau_1 \leq \epsilon$, we don't need to look further – $T = 1$ will suffice. Otherwise, divide both sides by τ_1 and take logs, to obtain the equivalent condition

$$(T - 1)\log(1 - \phi) \leq \log(\epsilon/\tau_1).$$

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

Now, using the fact that $\log(1+t) \leq t$ for all $t > -1$, we find that a sufficient condition for this inequality is that

$$-(T-1)\phi \leq \log(\epsilon/\tau_1),$$

or, equivalently,

$$T \geq \frac{1}{\phi}|\log(\epsilon/\tau_1)| + 1. \tag{A.9}$$

Note that the threshold ϵ enters only logarithmically into the estimate of K. The more important term involves the value ϕ, which captures the rate of linear convergence.

A.3 Algorithm 3.1 Is an Effective Line-Search Technique

We prove here that Algorithm 3.1 succeeds in identifying a value of α that satisfies the weak Wolfe conditions, unless the function f is unbounded below along the direction d. (This proof is adapted from Burke and Engle, 2018, lemma 4.2.)

Theorem A.3 *Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and that $x, d \in \mathbb{R}^n$ are such that $\nabla f(x)^T d < 0$. Then one of the following two possibilities must occur in Algorithm 3.1.*

(i) The algorithm terminates at a finite value of α for which the weak Wolfe conditions (3.26) *are satisfied.*
(ii) The algorithm does not terminate finitely, in which case U is never set to a finite value, L is set to 1 *on the first iteration and is doubled at every subsequent iteration, and $f(x+\alpha d) \to -\infty$ for the sequence of α values generated by the algorithm.*

Proof Suppose that finite termination does not occur. If, indeed, U is never finite, then L is set to 1 on the first iteration (since otherwise the algorithm would have terminated) and α is set to 2. In fact, α is doubled at every subsequent iteration, and moreover, the condition $f(x+\alpha d) \leq f(x) + c_1 \alpha \nabla f(x)^T d$ holds for all such α, which implies that $f(x+\alpha d)$ approaches $-\infty$ for some sequence of values of α approaching ∞. Hence, we are in case (ii).

Suppose now that finite termination does not occur but that U is set to a finite value at some iteration. Using l to denote the iterations of Algorithm 3.1, and L_l, α_l, and U_l denote the values of the parameters at the start of iteration l, we have initial values $L_0 = 0$, $\alpha_0 = 1$, and $U_0 = \infty$. Note too that $L_l < \alpha_l < U_l$. Since U_l is eventually finite for some l, we have that $L_l < \alpha_l < U_l$ for all l, and since the length of the interval $[L_l, U_l]$ is halved at each iteration after U_l becomes finite, there is a value $\bar{\alpha}$ such that

$$L_l \uparrow \bar{\alpha}, \quad \alpha_l \to \bar{\alpha}, \quad U_l \downarrow \bar{\alpha}. \tag{A.10}$$

If $L_l = 0$ for all l, then we have $\bar{\alpha} = 0$ and

$$\frac{f(x+\alpha_l d) - f(x)}{\alpha_l} > c_1 \nabla f(x)^T d, \quad l = 0, 1, 2, \ldots,$$

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

so by taking limits as $l \to \infty$, we have $\nabla f(x)^T d \geq c_1 \nabla f(x)^T d$, which is a contradiction since $c_1 \in (0,1)$ and $\nabla f(x)^T d < 0$. Thus, there exists an index l_0 such that $L_l > 0$ for all $l \geq l_0$.

Consider now all indices $l > l_0$. We have the following three conditions:

$$f(x + L_l d) \leq f(x) + c_1 L_l \nabla f(x)^T d, \tag{A.11a}$$

$$f(x + U_l d) > f(x) + c_1 U_l \nabla f(x)^T d, \tag{A.11b}$$

$$\nabla f(x + L_l d)^T d < c_2 \nabla f(x)^T d. \tag{A.11c}$$

Condition (A.11b) holds because each value of U_l is defined to be a value of α for which the first "if" test is satisfied – that is, $f(x + \alpha d) > f(x) + c_1 \alpha \nabla f(x)^T d$. Similarly, condition (A.11a) holds because each L_l is defined to be a value of α for which the first "if" test fails – that is, $f(x + \alpha d) \leq f(x) + c_1 \alpha \nabla f(x)^T d$. Condition (A.11c) holds because each L_l is defined to be a value of α for which the "else if" condition holds – that is, $\nabla f(x + \alpha d)^T d < c_2 \nabla f(x)^T d$.

By taking limits in (A.11c) as $l \to \infty$, we have

$$\nabla f(x + \bar{\alpha} d)^T d \leq c_2 \nabla f(x)^T d. \tag{A.12}$$

By combining (A.11a) and (A.11b) and using the mean value theorem, we have

$$c_1 (U_l - L_l) \nabla f(x)^T d \leq f(x + U_l d) - f(x + L_l d) = (U_l - L_l) \nabla f(x + \hat{\alpha}_l d)^T d,$$

for some $\hat{\alpha}_l \in (L_l, U_l)$, for all $l > l_0$. By dividing by $U_l - L_l$ and taking limits in this expression, we obtain that $c_1 \nabla f(x)^T d \leq \nabla f(x + \bar{\alpha} d)^T d$. This contradicts (A.12), since $\nabla f(x)^T d < 0$ and $0 < c_1 < c_2$. We conclude that if U is set to a finite value on some iteration, finite termination must occur. But when finite termination occurs, the final value of α satisfies the weak Wolfe conditions (3.26), so we are in case (i). □

A.4 Linear Programming Duality, Theorems of the Alternative

Linear programming duality results are important in proving optimality conditions for constrained optimization, as well as being of vital interest in their own right. We start by discussing weak and strong duality theorems, then discuss the use of these theorems in proving so-called *theorems of the alternative*. (The celebrated Farkas lemma used in constrained optimization theory is one such theorem.)

Consider the following linear program in standard form:

$$\min_x c^T x \quad \text{subject to} \quad Ax = b, \ x \geq 0, \tag{A.13}$$

where $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$, and $x \in \mathbb{R}^n$. This problem is said to be *infeasible* if there is no $x \in \mathbb{R}^n$ that satisfies the constraints $Ax = b$, $x \geq 0$, and *unbounded* if there is a sequence of vectors $\{x^k\}_{k=1,2,\ldots}$ that is feasible (that is, $Ax^k = b$, $x^k \geq 0$), with $c^T x^k \to -\infty$.

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

The dual linear program for (A.13) is

$$\max_{\lambda, s} b^T \lambda \text{ subject to } A^T \lambda + s = c, \; s \geq 0. \tag{A.14}$$

Sometimes, for compactness of expression, the "dual slack" variables s are eliminated from the dual formulation, and it is written equivalently as

$$\max_{\lambda} b^T \lambda \text{ subject to } A^T \lambda \leq c. \tag{A.15}$$

Two fundamental theorems in linear programming relate the primal and dual problems. The first, called *weak duality*, has a trivial proof.

Theorem A.4 *Suppose that x is feasible for (A.13) and (λ, s) is feasible for (A.14). Then $b^T \lambda \leq c^T x$.*

Proof

$$c^T x = (A^T \lambda + s)^T x = \lambda^T (Ax) + s^T x \geq b^T \lambda.$$

(The inequality follows from primal feasibility $Ax = b$ and the fact that $s \geq 0$ and $x \geq 0$ imply $s^T x \geq 0$.) □

The second duality result, called *strong duality*, is much more difficult to prove.

Theorem A.5 *Considering the primal-dual pair (A.13)–(A.14), exactly one of the following three statements is true.*

(i) Both (A.13) and (A.14) are feasible, both have solutions, and the objective values of the two problems are equal at the optimal points.
(ii) Exactly one of (A.13) and (A.14) is feasible, and the other is unbounded.
(iii) Both (A.13) and (A.14) are infeasible.

This result has several interesting consequences. It tells us, for example, that we cannot have a situation where one of the primal-dual pair has an optimal solution while the other is infeasible or unbounded. It also tells us that if one of the pair is unbounded, the other is infeasible.

A common proof methodology (omitted here) is via the properties of the simplex method. Using the traditional exposition of simplex via tableaus and pivot rules, it can be shown that the method terminates in one of the three states above, when appropriate anti-cycling rules are applied.

Strong duality can be used to prove theorems of the alternative, which are typically a pair of conditions, each involving linear equalities and inequalities, of which exactly one holds. One such theorem, known as the Farkas lemma, is instrumental in proving the Karush–Kuhn–Tucker (KKT) conditions, which are first-order optimality conditions for constrained optimization.

Lemma A.6 (Farkas Lemma) *Given a set of vectors $\{a^i \in \mathbb{R}^n \mid i = 1, 2, \ldots, K\}$ and a vector $b \in \mathbb{R}^n$, exactly one of the following two statements is true.*

I. There exist nonnegative coefficients $\lambda_i \geq 0$, $i = 1, 2, \ldots, K$, such that $b = \sum_{i=1}^{K} \lambda_i a^i$. That is, b is in the cone defined by $\{a^i \in \mathbb{R}^n \mid i = 1, 2, \ldots, K\}$.
II. There exists $s \in \mathbb{R}^n$ such that $b^T s < 0$ and $(a^i)^T s \geq 0$ for all $i = 1, 2, \ldots, K$.

Proof Assembling the vectors a^i into an $n \times K$ matrix $A := [a^1, a^2, \ldots, a^K]$, we consider the following linear program:

$$\min_{\lambda} 0^T \lambda \text{ subject to } A\lambda = b, \; \lambda \geq 0, \tag{A.16}$$

which has the form of (A.13) with $c = 0$. The dual is

$$\max_{t} b^T t \text{ subject to } A^T t \leq 0. \tag{A.17}$$

Since the dual is always feasible ($t = 0$ satisfies the constraints), we have from Theorem A.5 that there are only two possible outcomes: Either (A.16) is infeasible and (A.17) is unbounded (case (ii) of Theorem A.5) or both (A.16) and (A.17) both have solutions, with equal objectives. The first of these alternatives corresponds to case II: There is no vector $\lambda \geq 0$ such that $A\lambda = b$, but because of unboundedness of (A.17), we can identify t such that $b^T t > 0$ and $(a^i)^T t \leq 0$ for all $i = 1, 2, \ldots, K$. We set $s = -t$ to obtain case II. The second alternative corresponds to case I: Feasibility of (A.16) means existence of $\lambda \geq 0$ such that $b = \sum_{i=1}^{K} a^i \lambda_i$. □

A second theorem of the alternative called Gordan's theorem is useful in proving results about separating hyperplanes between convex sets. We make use of this result in Section A.6.

Theorem A.7 (Gordan's Theorem) *Given a matrix A, exactly one of the following two statements is true.*

$$A^T y > 0 \quad \textit{for some vector } y; \tag{I}$$

$$Ax = 0, \; x \geq 0, \; x \neq 0 \quad \textit{for some vector } x. \tag{II}$$

Proof Defining $\mathbf{1}$ to be the vector $(1, 1, \ldots, 1)$ with the same number of elements as there are columns in A, we note that statement (I) is equivalent to the following linear program having a solution:

$$\min_{y} 0^T y \text{ subject to } A^T y \geq \mathbf{1}. \tag{P}$$

The dual of (P) is

$$\max_{x} \mathbf{1}^T x \text{ subject to } Ax = 0, \;\; x \geq 0. \tag{D}$$

We now argue from strong duality. Suppose first that (I) is true. Then, by scaling y by a positive scalar as needed, we can say that (P) is feasible and, thus, has a solution with objective 0. Thus, from strong duality, (D) also has a solution with zero objective. But this means that (II) cannot be true, because if any x were to satisfy (II), it would be feasible for (D) with a strictly positive objective – greater than the maximum value. Hence, we have shown that if (I) is true, (II) must be false.

Suppose now that (I) is false. Then there can be no feasible point for (P) (since if there were, it would satisfy (I)). Thus, from strong duality, (D) is either infeasible or unbounded. Since it is clearly not infeasible (the vector $x = 0$ is a feasible point), it must be unbounded. In particular, there must be a vector x such that $Ax = 0$, $x \geq 0$, with $\mathbf{1}^T x > 0$, and from the latter, we can infer than $x \neq 0$. Thus, (II) holds. □

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

A.5 Limiting Feasible Directions

We now introduce the concept of limiting feasibility directions to a closed convex set Ω and derive an alternative a first-order optimality condition for $\min_{x \in \Omega} f(x)$ that will be useful in subsequent analysis. (This concept is also useful in the case of *nonconvex* feasible sets, which we do not consider in this book.)

Definition A.8 We say that $t \in \mathbb{R}^n$ is a *limiting feasible direction* for the set Ω at a point $x \in \Omega$ if there is a sequence of vectors $t^i \to t$ and a seqeunce of positive scalars $\alpha_i \to 0$ such that $x + \alpha_i t^i \in \Omega$.

Some limiting feasible directions for a set Ω with a curved boundary are shown in Figure A.1.

The following result establishes a relationship between the normal cone and limiting feasible directions for closed convex Ω.

Theorem A.9 *Given the closed convex set Ω and a point $x^* \in \Omega$, we have that $-y \in N_\Omega(x^*)$ if and only if $y^T t \geq 0$ for all limiting directions t to Ω at x^*.*

Proof Suppose first that $-y \in N_\Omega(x^*)$, so that $y^T(x - x^*) \geq 0$ for all $x \in \Omega$. Given a direction t and associated sequences t^i and $\alpha_i > 0$, we have that

$$y^T((x^* + \alpha_i t^i) - x^*) = \alpha_i y^T t^i \geq 0,$$

so, dividing by α_i, we obtain $y^T t^i \geq 0$ for all i. By taking limits as $i \to \infty$, we obtain $y^T t \geq 0$.

Suppose now that $y^T t \geq 0$ for all limiting directions t, and let x be an arbitrary element of Ω. By definining $t^i \equiv (x - x^*)$ and $\alpha_i = 1/i$ for all $i \geq 1$, we have by convexity that $x^* + \alpha_i t^i = (1 - 1/i)x^* + (1/i)x \in \Omega$, so these sequences define the limiting direction $t = (x - x^*)$. We therefore have $y^T(x - x^*) \geq 0$ for all $x \in \Omega$, so that $-y \in N_\Omega(x^*)$, completing the proof. □

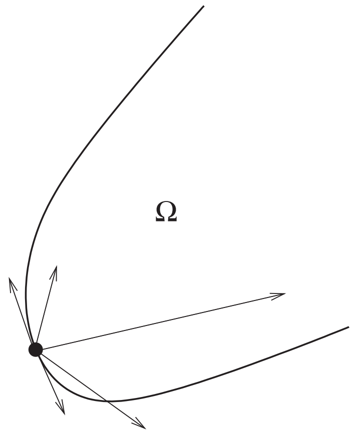

Figure A.1 Limiting feasible directions.

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

A.6 Separation Results

Here we discuss separation results, which are classical results about the existence of hyperplanes that separate two convex sets X and Y, such that X is on one side of the hyperplane and Y is on the other. These results are vital to deriving optimality conditions for convex optimization problems; we rely on them in Chapter 10.

We start with a technical result about compact sets.

Lemma A.10 *Suppose that Λ is a compact set. Let Λ_x, $x \in X$ be a collection of subsets of Λ, all closed in Λ, for some index set X. If for every finite collection of points $x^1, x^2, \dots, x^m \in X$, we have that $\cap_{i=1}^m \Lambda_{x^i} \neq \emptyset$, then $\cap_{x \in X} \Lambda_x \neq \emptyset$.*

Proof We prove the result by contradiction. Since each Λ_x is closed in Λ, its complement Λ_x^c is open in Λ. If $\cap_{x \in X} \Lambda_x = \emptyset$, then $\{\Lambda_x^c \mid x \in X\}$ is an open cover of Λ. Thus, by the Heine–Borel theorem, there is a finite subcover – that is, a set of points $x^1, x^2, \dots, x^m \in X$ such that $\cup_{i=1}^m \Lambda_{x^i}^c = \Lambda$. It follows that $\cap_{i=1}^m \Lambda_{x^i} = \emptyset$, a contradiction. □

Using this result, we show that any convex set not containing the origin can be contained in a half-space passing through the origin.

Lemma A.11 *Let X be any nonempty convex set such that $0 \notin X$. Then there is a nonzero vector $\bar{t} \in \mathbb{R}^n$ such that $\bar{t}^T x \leq 0$ for all $x \in X$.*

Proof Define $\Lambda := \{v \in \mathbb{R}^n \mid \|v\|_2 = 1\}$, and for all $x \in X$, define

$$\Lambda_x := \{v \in \Lambda \mid v^T x \leq 0\}.$$

Clearly, Λ_x is compact for all $x \in X$. Now let $x^1, x^2, \dots, x^m$ be any finite set of vectors in X. Since $0 \notin X$, then 0 is not in the convex hull of the vectors $x^1, x^2, \dots, x^m$. That is, defining A to be the matrix whose columns are $x^1, x^2, \dots, x^m$, there is no vector $p \in \mathbb{R}^m$ such that $Ap = 0$, $p \geq 0$, $\mathbf{1}^T p = 1$ (where $\mathbf{1}$ is the vector containing m elements, all of which are 1). Thus, there is no $\bar{p}$ such that

$$A\bar{p} = 0, \quad \bar{p} \geq 0, \quad \bar{p} \neq 0,$$

since if there were, then $p = \bar{p}/(\mathbf{1}^T \bar{p})$ would have the forbidden properties. It follows from Gordan's theorem (Theorem A.7) that there must be a vector t such that $A^T t > 0$, that is, $(x^i)^T t > 0$, $i = 1, 2, \dots, m$. Therefore, we have that $-t/\|t\|_2 \in \cap_{i=1,2,\dots,m} \Lambda_{x^i}$. Thus, the conditions of Lemma A.10 are satisfied, so there must exist a vector $\bar{t}$ such that $\|\bar{t}\|_2 = 1$ and $\bar{t}^T x \leq 0$ for all $x \in X$. □

The inequality $\bar{t}^T x \leq 0$ need not be strict. An example is when $X \subset \mathbb{R}^2$ is the convex set consisting of the entire left half-plane $\{(x_1, x_2)^T : \ x_1 \leq 0\}$ with the exception of the half-line $\{(0, x_2)^T : \ x_2 \leq 0\}$. The only possible choices for $\bar{t}$ here are $\bar{t} = (\beta, 0)^T$ for any $\beta > 0$, and all these choices have $\bar{t}^T x = 0$ for some $x \in X$. However, with the additional assumption of closedness of X, we can obtain strict separation.

Lemma A.12 *Let X be a nonempty, convex, and closed set with $0 \notin X$. Then there is $\bar{t} \in \mathbb{R}^n$ and $\alpha > 0$ such that $\bar{t}^T x \leq -\alpha$ for all $x \in X$.*

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

Proof Recalling the projection operator defined in (7.2), we have, by assumption, that $P_X(0) \neq 0$. (If $P_X(0)$ were zero, we would have $0 \in \mathrm{cl}(X) = X$, which is false by assumption.) We have by setting $y = 0$ in the minimum principle (7.3) that $(0 - P_X(0))^T(z - P_X(0)) \leq 0$ for all $z \in X$, which implies $P_X(0)^T z \geq \|P_X(0)\|_2^2 > 0$. We obtain the result by taking $\bar{t} = -P(0)$ and $\alpha = \|P(0)\|_2^2$. □

Having understood the issue of separation between a point and a convex set, we turn to separation between two closed convex sets. It turns out that separation is possible, but *strict* separation requires the additional condition of compactness of one of the sets. We show these facts in the next two results.

Theorem A.13 (Separation of Closed Convex Sets) *Let X and Y be two nonempty disjoint closed convex sets. Then these sets can be* separated*; that is, there is $c \in \mathbb{R}^n$ with $c \neq 0$, and $\alpha \in \mathbb{R}$ such that $c^T x - \alpha \leq 0$ for all $x \in X$ and $c^T y - \alpha \geq 0$ for all $y \in Y$.*

Proof We first define the set $X - Y$ as follows:

$$X - Y := \{x - y : \ x \in X, \ y \in Y\}. \tag{A.18}$$

An elementary argument shows that $X - Y$ is convex. Since X and Y are disjoint, we have that $0 \notin X - Y$. We can thus apply Lemma A.11 to deduce that there is $c \neq 0$ such that $c^T(x - y) \leq 0$ for all $x \in X$, $y \in Y$. By choosing an arbitrary $\hat{x} \in X$, we have that $c^T y$ is bounded below by $c^T \hat{x}$ for all $y \in Y$. Hence, the infimum of $c^T y$ over $y \in Y$ exists; we denote it by α and note that $c^T y \geq \alpha$ for all $y \in Y$. Moreover, since $c^T x \leq c^T y$ for all $x \in X$, $y \in Y$, we must have $c^T x \leq \alpha$ too. We conclude that for these definitions of c and α, the required inequalities are satisfied. □

We investigate further the properties of the set $X - Y$ defined in (A.18), where X and Y are closed convex sets. We noted above that $X - Y$ is convex, but it may not be closed. Consider the following example of two closed convex sets in $\mathbb{R}^2$:

$$X = \{(x_1, x_2) \,|\, x_1 > 0, x_2 \geq 1/x_1\}, \qquad Y = \{(y_1, y_2) \,|\, y_1 > 0, y_2 \leq -1/y_1\},$$

and define the sequences $\{x^k\}$ and $\{y^k\}$ by $x^k := (k, 1/k)^T \in X$ for all $k \geq 1$, and $y^k := (k, -1/k) \in Y$ for all $k \geq 1$. The sequence $z^k := x^k - y^k = (0, 2/k) \in X - Y$, by definition, and $z^k \to (0,0)^T$, but $(0,0)^T \notin X - Y$. Thus, $X - Y$ is not closed in this example. However, by adding a compactness assumption, we obtain closedness of $X - Y$, and thus a strict separation result.

Theorem A.14 (Strict Separation) *Let X and Y be two disjoint closed convex nonempty sets with X compact. Then these sets can be* strictly separated*, that is, there is $c \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, and $\epsilon > 0$ such that $c^T x - \alpha \leq -\epsilon$ for all $x \in X$ and $c^T y - \alpha \geq \epsilon$ for all $y \in Y$.*

Proof We first show closedness of $X - Y$. Let z^k be any sequence in $X - Y$ such that $z^k \to z$ for some z. Closedness will follow if we can show that $z \in X - Y$. By definition of $X - Y$, we can find two sequences $\{x^k\}$ in X and $\{y^k\}$ in Y such that $z^k := x^k - y^k$. Since X is compact, we have by taking a subsequence if necessary that $x^k = z^k + y^k \to x$ for some $x \in X$. Thus, we have that $y^k = x^k - z^k \to x - z$, and

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press

by closedness of Y, we have $x - z \in Y$. Thus, $z = x - (x - z) \in X - Y$, proving our claim that $X - Y$ is closed, as well as being nonempty and convex.

Since $0 \notin X - Y$, we use Lemma A.12 to choose a nonzero $\bar{t} \in \mathbb{R}^n$ and $\beta > 0$ such that $\bar{t}^T(x - y) \leq -\beta$ for all $x \in X$ and $y \in Y$. Fixing some $\bar{y} \in Y$, we have that $\bar{t}^T x \leq -\beta + \bar{t}^T \bar{y}$ for all $x \in X$. Hence, $\bar{t}^T x$ is bounded above for all $x \in X$, so there is a supremal value γ such that $\bar{t}^T x \leq \gamma$. A similar argument shows that $\bar{t}^T y$ is bounded below for all $y \in Y$, and has an infimal value δ. Moreover, we have that $\gamma + \beta \leq \delta$. Thus, for all $x \in X$ and $y \in Y$, we have that

$$\bar{t}^T x \leq \gamma < \gamma + \beta/2 < \gamma + \beta \leq \bar{t}^T y.$$

We obtain the result by setting $c = \bar{t}$, $\alpha = \gamma + \beta/2$, and $\epsilon = \beta/2$. □

Supporting Hyperplane for Convex Sets. We now prove an almost immediate consequence of the separating hyperplane theorem, a result called the *supporting hyperplane* theorem that is used in the discussion of existence of subgradients in Section 8.1. We first need the following definition. Given a set $X \subset \mathbb{R}^n$, we say that $x \in X$ is a *boundary point* of X if it is not in $\text{int}(X)$ – that is, $x \in X$ and for any $\epsilon > 0$, there exists $y \notin X$ with $\|y - x\| < \epsilon$.

Theorem A.15 (Supporting Hyperplane Theorem) *Let X be a nonempty convex set, and let x be any boundary point of X. Then there exists a nonzero $c \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that $c^T x = \alpha$ but $c^T z \leq \alpha$ for all $z \in X$. (We call the plane defined by $c^T x = \alpha$ the* supporting hyperplane.*)*

Proof If X has an interior in $\mathbb{R}^n$, then $x \notin \text{int}(X)$, and we apply Lemma A.11 to separate 0 from $\text{int}(X) - \{x\}$. This result says that there is nonzero $\bar{t} \in \mathbb{R}^n$ such that $\bar{t}^T(z - x) \leq 0$ for all $z \in \text{int}(X)$. Thus, $\bar{t}^T(z - x) \leq 0$ for all $z \in \text{cl}(X)$, and since $X \subset \text{cl}(X)$, we obtain the result by setting $c = \bar{t}$ and $\alpha = \bar{t}^T x$.

if X does not have an interior, it is contained in a hyperplane. That is, there exist nonzero $c \in \mathbb{R}^n$ and α such that $X \subset \{z \,|\, c^T z = \alpha\}$. These c and α satisfy our claim (rather trivially). □

Separating a Convex Set from a Hyperplane. Two sets C_1 and C_2 are said to be *properly separated* if there is a separating hyperplane defined by $c^T x = \alpha$ such that it is not the case that *both* C_1 and C_2 are contained in the hyperplane. Recalling the definition of relative interior of a set C from (A.3), we have the following result concerning proper separation.

Theorem A.16 *(Rockafellar, 1970, Theorem 11.3) Let C_1 and C_2 be nonempty convex sets. These sets can be properly separated if and only if their relative interiors* $\text{ri}\,(C_1)$ *and* $\text{ri}\,(C_2)$ *are disjoint.*

We refer to Rockafellar (1970) for the proof, which depends on a number of other technical results. We have the following corollary.

Corollary A.17 *Suppose that C_1 is a nonempty convex set and C_2 is a subspace, with* $\mathrm{ri}\,(C_1)$ *disjoint from C_2. Then there is a vector c such that $c^T x = 0$ for all $x \in C_2$ and $c^T x \le 0$ for all $x \in C_1$, with the inequality being strict for some $x \in C_1$.*

Proof Since C_2 is a subspace, we have $C_2 = \mathrm{aff}\,(C_2) = \mathrm{ri}\,(C_2)$. Thus $\mathrm{ri}\,(C_1)$ and $\mathrm{ri}\,(C_2)$ are disjoint, so we can apply Theorem A.16 to deduce that C_1 and C_2 are properly separable. Let (c,α) define a properly separating hyperplane, with $c^T x \le \alpha$ for all $x \in C_1$ and $c^T x \ge \alpha$ for all $x \in C_2$. Since C_2 is a subspace, we have $0 \in C_2$ and thus $\alpha \le 0$. In fact, we must have $c^T x = 0$ for all $x \in C_2$. (If this were not true – that is, $c^T x > 0$ for some $x \in C_2$ – we have from $\beta x \in C_2$ for all $\beta \in \mathbb{R}$ that $\{c^T x \,|\, x \in C_2\} = (-\infty,\infty)$, contradicting the existence of α.) If $\alpha < 0$, the claim follows immediately, from nonemptiness of C_1. If $\alpha = 0$, we have that the separating hyperplane $c^T x = 0$ contains C_2. Thus, since C_1 and C_2 are properly separated by this hyperplane, the hyperplane cannot contain C_1 as well as C_2. Thus, $c^T x < 0$ for some $x \in C_1$, as claimed. □

Normal Cone of the Intersection of an Affine Space and a Convex Set. We now restate Theorem 10.4, the critical result concerning the normal cone of the feasible set for the problem (10.1), and provide a proof.

Theorem A.18 *Suppose that $\mathcal{X} \in \mathbb{R}^n$ is a closed convex set and that $\mathcal{A} := \{x \mid Ax = b\}$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and define $\Omega := \mathcal{X} \cap \mathcal{A}$. Then for any $x \in \Omega$, we have*

$$N_\Omega(x) \supset N_{\mathcal{X}}(x) + \{A^T\lambda \,|\, \lambda \in \mathbb{R}^m\}. \tag{A.19}$$

If, in addition, the set $\mathrm{ri}\,(\mathcal{X}) \cap \mathcal{A}$ *is nonempty, then this result holds with equality; that is,*

$$N_\Omega(x) = N_{\mathcal{X}}(x) + \{A^T\lambda \,|\, \lambda \in \mathbb{R}^m\}. \tag{A.20}$$

Proof To show (A.19), take any $z \in \Omega$, and note that $z - x \in \mathrm{null}\,(A)$, so that $(z - x)^T A^T\lambda = \lambda^T A(z-x) = 0$ for all $\lambda \in \mathbb{R}^m$. For any $u \in N_{\mathcal{X}}(x)$, we have $(z-x)^T u \le 0$, by definition of $N_{\mathcal{X}}(x)$. It follows that

$$(z-x)^T(u + A^T\lambda) \le 0,$$

and so $u + A^T\lambda \in N_\Omega(x)$ for any $u \in N_{\mathcal{X}}(x)$ and any $\lambda \in \mathbb{R}^m$.

For the assertion "$\subset$" in (A.20), we choose an arbitrary $v \in N_\Omega(x)$, and aim to show that $v \in N_{\mathcal{X}}(x) + N_{\mathcal{A}}(x)$. By choice of v, we have $v^T(z - x) \le 0$ for all $z \in \Omega = \mathcal{X} \cap \mathcal{A}$. We define the following sets:

$$C_1 = \{(y,\mu) \in \mathbb{R}^{n+1} \mid y = z - x \text{ for some } z \in \mathcal{X} \text{ and } \mu \le v^T y\},$$
$$C_2 = \{(y,\mu) \in \mathbb{R}^{n+1} \mid y \in \mathrm{null}\,(A),\ \mu = 0\}.$$

Note that C_2 is a subspace and that C_1 is closed, convex, and nonempty. Note too that $\mathrm{ri}\,(C_1)$ and C_2 are disjoint, because if there were a vector $(\hat{y},\hat{\mu}) \in \mathrm{ri}\,(C_1) \cap C_2$, we would have $\hat{z} = x + \hat{y} \in \mathcal{X}$ and $A\hat{z} = Ax = b$, so that $\hat{z} \in \Omega$. Moreover, we would have $v^T\hat{y} > \hat{\mu} = 0$ and, thus, $v^T(\hat{z} - x) > 0$, contradicting $v \in N_\Omega(x)$. We can now apply Corollary A.17 to deduce the existence of a vector $(w,\gamma) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\inf_{(y,\mu)\in C_1} w^T y + \gamma\mu < \sup_{(y,\mu)\in C_1} w^T y + \gamma\mu \le 0, \tag{A.21}$$

while

$$w^T u = 0 \text{ for all } u \in \text{null}\,(A). \tag{A.22}$$

This latter equality implies that $w = A^T\lambda$ for some $\lambda \in \mathbb{R}^m$.

We note next that $\gamma \ge 0$, since otherwise we obtain $\sup_{(y,\mu)\in C_1} w^T y + \gamma\mu = \infty$ by letting μ tend to $-\infty$. We also cannot have $\gamma = 0$, as we argue now. If $\gamma = 0$, we would have from (A.21) that $\inf_{(y,\mu)\in C_1} w^T y < \sup_{(y,\mu)\in C_1} w^T y \le 0$, and so, in particular, $w^T(z-x) < 0$ for some $z \in \mathcal{X}$. For any point $\tilde{x} \in \text{ri}\,(\mathcal{X})$, we claim that $w^T(\tilde{x}-x) < 0$. If (for contradiction) we were to have $w^T(\tilde{x}-x) \ge 0$, we would find that for small positive α and the fact that $z - x \in \text{aff}\,(C_1)$ that $\tilde{x} - \alpha(z-x) \in C_1$, and hence, from (A.21), that $w^T(\tilde{x} - \alpha(z-x) - x) \le 0$. On the other hand, we have $w^T(\tilde{x} - \alpha(z-x) - x) = w^T(\tilde{x}-x) - \alpha w^T(z-x) > 0$, a contradiction. Thus, $w^T(\tilde{x}-x) < 0$ for all $\tilde{x} \in \text{ri}\,(\mathcal{X})$. It follows from (A.22) that $\tilde{x} - x \notin \text{null}\,(A)$ and, thus, $A\tilde{x} \ne Ax = b$. Thus, there exists no point $\tilde{x} \in \text{ri}\,(C) \cap \mathcal{A}$, so $\gamma = 0$ is not possible.

We thus have that γ in (A.21) is strictly positive. Taking any $z \in \mathcal{X}$, we have from (A.21), by setting $\mu = v^T y = v^T(z-x)$ in the definition of C_1, that

$$w^T(z-x) + \gamma\mu = w^T(z-x) + \gamma v^T(z-x) = (w+\gamma v)^T(z-x) \le 0.$$

Therefore, we have $w + \gamma v \in N_{\mathcal{X}}(x)$ and so $(1/\gamma)w + v = (1/\gamma)(w+\gamma v) \in N_{\mathcal{X}}(x)$. Since we already observed following (A.22) that $w = A^T\lambda$ for some $\lambda \in \mathbb{R}^m$, we have

$$v = ((1/\gamma)w + v) - (1/\gamma)w \in N_{\mathcal{X}}(x) + N_{\mathcal{A}}(x),$$

as required. □

A.7 Bounds for Degenerate Quadratic Functions

We prove here some claims concerning convex quadratic functions that may not be strongly convex. We show here that such functions satisfy the PL property (3.45). Thus, algorithms applied to these problems have similar performance as when applied to strongly convex functions. The modulus of convexity m in the standard convergence analysis can be replaced by the the minimum *nonzero* eigenvalue of the Hessian of the quadratic function.

Consider first the function $f(x) = \frac{1}{2}x^T Ax$ arising in Section 3.8, where A is positive semidefinite $n \times n$ matrix with rank $r \le n$ and eigenvalues $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_r > 0$. We claim that f satisfies (3.45) with $m = \lambda_r$. To prove the claim, we write the eigenvalue decomposition of A as follows:

$$A = \sum_{i=1}^{r} \lambda_i u^i (u^i)^T,$$

where $\{u^1, u^2, \ldots, u^r\}$ is the orthnormal set of eigenvectors. We then have that

$$\|\nabla f(x)\|^2 = \|Ax\|^2 = \left\| \sum_{i=1}^{r} u^i \lambda_i (u^i)^T x \right\|^2 = \sum_{i=1}^{r} \lambda_i^2 \left[(u^i)^T x \right]^2.$$

Meanwhile, we have

$$f(x) - f(x^*) = \frac{1}{2} x^T A x = \frac{1}{2} \sum_{i=1}^{r} \lambda_i \left[(u^i)^T x \right]^2,$$

so that

$$2\lambda_r (f(x) - f(x^*)) = \lambda_r \sum_{i=1}^{r} \lambda_i \left[(u^i)^T x \right]^2 \leq \sum_{i=1}^{r} \lambda_i^2 \left[(u^i)^T x \right]^2 = \|\nabla f(x)\|^2,$$

as required.

Next, we recall from Section 5.2.2 the Kaczmarz method, which is a type of stochastic gradient algorithm applied to the function

$$f(x) = \frac{1}{2N} \|Ax - b\|^2,$$

where $A \in \mathbb{R}^{N \times n}$, and there exists x^* (possibly not unique) such that $f(x^*) = 0$, that is, $Ax^* = b$. (Let us assume for simplicity of exposition that $N \geq n$.) We claimed in Section 5.4.2 that for any x, there exists x^* such that $Ax^* = b$ in which

$$\|Ax - b\|^2 \geq \lambda_{\text{min, nz}} \|x - x^*\|^2,$$

where $\lambda_{\text{min, nz}}$ is the smallest nonzero eigenvalue of $A^T A$. We prove this statement by writing the singular value decomposition of A as

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T,$$

where the singular values σ_i satisfy

$$\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r > \sigma_{r+1} = \cdots = \sigma_n = 0,$$

so that r is the rank of A. The left singular vectors $\{u_1, u_2, \ldots, u_n\}$ form an orthonormal set in $\mathbb{R}^N$, and the right singular vectors $\{v_1, v_2, \ldots, v_n\}$ form an orthonormal set in $\mathbb{R}^n$. The eigenvalues of $A^T A$ are σ_i^2, $i = 1, 2, \ldots, n$, so that the rank of $A^T A$ is r and the smallest nonzero eigenvalue is $\lambda_{\text{min, nz}} = \sigma_r^2$.

Solutions x^* of $Ax^* = b$ have the form

$$x^* = \sum_{i=1}^{r} \frac{u_i^T b}{\sigma_i} v_i + \sum_{i=r+1}^{n} \tau_i v_i,$$

where $\tau_{r+1}, \ldots, \tau_d$ are arbitrary coefficients. Given x, we set $\tau_i = v_i^T x$, $i = r + 1, \ldots, n$. (We leave it as an Exercise to show that this choice minimizes the distance $\|x - x^*\|$.) We then have

$$\begin{aligned}
\|Ax - b\|^2 &= \|A(x - x^*)\|^2 \\
&= \left\| \sum_{i=1}^{n} \sigma_i u_i v_i^T (x - x^*) \right\|^2 \\
&= \left\| \sum_{i=1}^{r} \sigma_i u_i v_i^T (x - x^*) \right\|^2 \\
&\geq \sigma_r^2 \sum_{i=1}^{r} [v_i^T (x - x^*)]^2 \\
&= \lambda_{\min,\,\mathrm{nz}} \sum_{i=1}^{n} [v_i^T (x - x^*)]^2 \\
&= \lambda_{\min,\,\mathrm{nz}} \|x - x^*\|^2,
\end{aligned}$$

where the last step follows from the fact that $[v_1, v_2, \ldots, v_n]$ is a $n \times n$ orthogonal matrix.

https://doi.org/10.1017/9781009004282.013 Published online by Cambridge University Press