Nonlinear Optimization

James V. Burke
University of Washington

Contents

CHAPTER 1

Introduction

In mathematical optimization we seek to either minimize or maximize a function over a set of alternatives. The function is called the *objective function*, and we allow it to be transfinite in the sense that at each point its value is either a real number or it is one of the to infinite values $\pm\infty$. The set of alternatives is called the *constraint region*. Since every maximization problem can be restated as a minimization problem by simply replacing the objective f_0 by its negative $-f_0$ (and visa versa), we choose to focus only on minimization problems. We denote such problems using the notation

$$\begin{aligned} &\underset{x\in X}{\text{minimize}} \;\; f_0(x) \\ &\text{subject to} \;\; x \in \Omega, \end{aligned} \tag{1.1}$$

where $f_0 : X \to \mathbb{R}\cup\{\pm\infty\}$ is the objective function, X is the space over which the optimization occurs, and $\Omega \subset X$ is the constraint region. Within this general framework, a taxonomy of optimization problems can be defined based on the underlying structural features that the problem possesses. For example, does the space X consist of integers, real numbers, complex numbers, matrices, or is it an infinite dimensional space of functions? Is the function f_0 discrete, continuous, or is it differentiable and how many derivatives does it posses? What is the geometry of the set Ω and how is it represented? For our purposes, we assume that Ω is a subset of $\mathbb{R}^n$ and that $f_0 : \mathbb{R}^n \to \mathbb{R}\cup\{\pm\infty\}$. Although this limits the kind of optimization problems that we study, the problem class is sufficiently broad to include a wide variety of applied problems of great practical importance and interest. For example, this framework includes *linear programming* (LP).

Linear Programming

In the case of LP, the objective function is linear, that is, there exists $c \in \mathbb{R}^n$ such that

$$f_0(x) = c^T x = \sum_{j=1}^{n} c_j x_j,$$

and the constraint region is representable as the set of solution to a finite system of linear equation and inequalities,

$$\Omega = \left\{ x \in \mathbb{R}^n \;\middle|\; \sum_{i=1}^{n} a_{ij}x_j \le b_j,\; i = 1,\dots,s,\; \sum_{i=1}^{n} a_{ij}x_j = b_j,\; i = s+1,\dots,m \right\}, \tag{1.2}$$

where $A := [a_{ij}] \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.

In this course we are primarily concerned with nonlinear problems, that is, problems that cannot be encoded using finitely many linear function alone. A natural generalization of the LP framework to the nonlinear setting is to simply replace each of the linear functions with a nonlinear function. This leads to the general *nonlinear programming* (NLP) problem which is the problem of central concern in these notes.

Nonlinear Programming

In nonlinear programming we are given nonlinear functions $f_i : \mathbb{R}^n \to \mathbb{R},\; i = 1, 2, \dots, m$, where f_0 is the objective function in (1.1) *and the functions $f_i,\; i = 1, 2, \dots, m$ are called the constraint functions which are used to define the constrain region in* (1.1) *by setting*

$$\Omega = \{ x \in \mathbb{R}^n \,|\, f_i(x) \le 0,\; i = 1,\dots,s,\; f_i(x) = 0,\; i = s+1,\dots,m \}\ . \tag{1.3}$$

If $\Omega = \mathbb{R}^n$, then we say that the problem (1.1) is an *unconstrained optimization problem*; otherwise, it called a constrained problem. We begin with unconstrained problems. They are simpler to handle since we are only concerned with minimizing the objective function and we need not concern ourselves with the constraint region. However, since we allow the objective to take infinite values, we will see that every explicitly constrained problem can be restated as an ostensibly unconstrained problem.

In the following chapter, we begin our study with what is arguably the most widely studied and used class of unconstrained unconstrained nonlinear optimization problems. This is the class of *linear least squares* problems. The theory an techniques we develop for this class of problems provides a template for how we address and exploit structure in a wide variety of other problem classes.

Linear Least Squares

A linear least squares problem is one of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}\,\|Ax-b\|_2^2\,, \tag{1.4}$$

where

$$A\in\mathbb{R}^{m\times n},\ b\in\mathbb{R}^m,\quad \text{and}\quad \|y\|_2^2 := y_1^2+y_2^2+\cdots+y_m^2\ .$$

Linear least squares problems arise in a wide range of applications. Indeed, whole books have been written about this problem, and many instances of this problem remain very active areas of research. This problem formulation is usually credited to Legendre and Gauss who made careful studies of the method around 1800. However, in the 50 years prior to their work, others applied the basic approach to the study of observational data and, in particular, to the study of planetary motion.

The second most important class of unconstrained nonlinear optimization problems is the minimization of *quadratic functions*. As we will see, the linear least squares problem is a member of this class of problems. It is important for a wide variety of reasons, not the least of which is the relationship to the second-order Taylor approximations for functions mapping $\mathbb{R}^n$ into $\mathbb{R}$.

Quadratic Functions

A function $f:\mathbb{R}^n\to\mathbb{R}$ is said to be quadratic if there exists $\alpha\in\mathbb{R}$, $g\in\mathbb{R}^n$ and $H\in\mathbb{R}^{n\times n}$ such that

$$f(x)=\alpha+g^Tx+\tfrac{1}{2}x^THx\ .$$

Notice that we may as well assume that the matrix H is symmetric ($H=H^T$) since

$$x^THx=\tfrac{1}{2}(x^THx+x^THx)=\tfrac{1}{2}((x^THx)^T+x^THx)=\tfrac{1}{2}(x^TH^Tx+x^THx)=x^T(\tfrac{1}{2}(H^T+H))x,$$

that is, we may as well replace the matrix H by its *symmetric part* $\frac{1}{2}(H^T+H)$.

Having quadratic functions in hand, one arrives at an important nonlinear generalization of linear programming where we simply replace the LP linear objective with a quadratic function. These are called *quadratic programming problems.*

The linear least squares problem and the optimization of quadratic functions are the themes for our initial forays into optimization. The theory and methods we develop for these problems, as well as certain variations on these problems, form the basis for our extensions to other problem classes. For this reason, we study these problems with great care. Notice that although these problems are nonlinear, their component pieces come from linear algebra, that is matrices. Obviously, these components play a key role in understanding the structure and behavior of these problems. For this reason, it is essential that the reader be familiar with the basic concepts and techniques of linear algebra. For this the reader should consult Appendix A.

CHAPTER 2

The Linear Least Squares Problem

In this chapter we study the linear least squares problem introduced in (1.4). Since this is such an important topic, we only briefly touch on a few aspects of this problem. We begin by introducing a few of the applications of the linear least squares from current research areas.

1. Applications

1.1. Polynomial Fitting. In many data fitting application one assumes a functional relationship between a set of "inputs" and a set of "outputs". For example, a patient is injected with a drug and the the research wishes to understand the clearance of the drug as a function of time. One way to do this is to draw blood samples over time and to measure the concentration of the drug in the drawn serum. The goal is to then provide a functional description of the concentration at any point in time.

Suppose the observed data is $y_i \in \mathbb{R}$ for each time point t_i, $i = 1, 2, \ldots, N$, respectively. The underlying assumption it that there is some function of time $f : \mathbb{R} \to \mathbb{R}$ such that $y_i = f(t_i)$, $i = 1, 2, \ldots, N$. The goal is to provide and estimate of the function f. One way to do this is to try to approximate f by a polynomial of a fixed degree, say n:

$$p(t) = x_0 + x_1 t + x_2 t^2 + \cdots + x_n t^n.$$

We now wish to determine the values of the coefficients that "best" fit the data.

If were possible to exactly fit the data, then there would exist a value for the coefficient, say $\overline{x} = (\overline{x}_0, \overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n)$ such that

$$y_i = \overline{x}_0 + \overline{x}_1 t_i + \overline{x}_2 t_i^2 + \cdots + \overline{x}_n t_i^n, \ i = 1, 2, \ldots, N.$$

But if N is larger than n, then it is unlikely that such an $\overline{x}$ exists; while if N is less than n, then there are probably many choices for $\overline{x}$ for which we can achieve a perfect fit. We discuss these two scenarios and their consequences in more depth at a future dat, but, for the moment, we assume that N is larger than n. That is, we wish to approximate f with a low degree polynomial.

When $n << N$, we cannot expect to fit the data perfectly and so there will be errors. In this case, we must come up with a notion of what it means to "best" fit the data. In the context of least squares, "best" means that we wish to minimized the sum of the squares of the errors in the fit:

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}} \ \tfrac{1}{2} \sum_{i=1}^{N} (x_0 + x_1 t_i + x_2 t_i^2 + \cdots + x_n t_i^n \ - \ y_i)^2 \ . \tag{2.1}$$

The leading one half in the objective is used to simplify certain computations that occur in the analysis to come. This minimization problem has the form

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}} \ \tfrac{1}{2} \left\| Vx - y \right\|_2^2 ,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad V = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & \ldots & t_2^n \\ \vdots & & & & \\ 1 & t_N & t_N^2 & \ldots & t_N^n \end{bmatrix},$$

since

$$Vx = \begin{pmatrix} x_0 + x_1 t_1 + x_2 t_1^2 + \cdots + x_n t_1^n \\ x_0 + x_1 t_2 + x_2 t_2^2 + \cdots + x_n t_2^n \\ \vdots \\ x_0 + x_1 t_N + x_2 t_N^2 + \cdots + x_n t_N^n \end{pmatrix}.$$

That is, the polynomial fitting problem (2.1) is an example of a linear least squares problem (1.4). The matrix V is called the *Vandermonde matrix* associated with this problem.

This is neat way to approximate functions. However, polynomials are a very poor way to approximate the clearance data discussed in our motivation to this approach. The concentration of a drug in serum typically rises quickly after injection to a maximum concentration and falls off gradually decaying exponentially. There is only one place where such a function is zero, and this occurs at time zero. On the other hand, a polynomial of degree n has n zeros (counting multiplicity). Therefore, it would seem that exponential functions would provide a better basis for estimating clearance. This motivates our next application.

1.2. Function Approximation by Bases Functions. In this application we expand on the basic ideas behind polynomial fitting to allow other kinds of approximations, such as approximation by sums of exponential functions. In general, suppose we are given data points $(z_i, y_i) \in \mathbb{R}^2,\ i = 1, 2, \ldots, N$ where it is assumed that the observation y_i is a function of an unknown function $f : \mathbb{R} \to \mathbb{R}$ evaluated at the point z_i for each $i = 1, 2, \ldots, N$. Based on other aspects of the underlying setting from which this data arises may lead us to believe that f comes from a certain space $\mathcal{F}$ of functions, such as the space of continuous or differentiable functions on an interval. This space of functions may itself be a vector space in the sense that the zero function is in the space ($0 \in \mathcal{F}$), two function in the space can be added pointwise to obtain another function in the space ($\mathcal{F}$ is closed with respect to addition), and any real multiple of a function is the space is also in the space ($\mathcal{F}$ is closed with respect to scalar multiplication). In this case, we may select from X a finite subset of functions, say $\phi_1, \phi_2, \ldots, \phi_k$, and try to approximate f as a linear combination of these functions:

$$f(x) \sim x_1\phi_1(z) + x_2\phi_2(z) + \cdots + x_n\phi_k(z).$$

This is exactly what we did in the polynomial fitting application discussed above. There $\phi_i(z) = z^i$ but we started the indexing at $i = 0$. Therefore, this idea is essentially the same as the polynomial fitting case. But the functions z^i have an additional properties. First, they are linearly independent in the sense that the only linear combination that yields the zero function is the one where all of the coefficients are zero. In addition, any continuous function on and interval can be approximated "arbitrarily well" by a polynomial assuming that we allow the polynomials to be of arbitrarily high degree (think Taylor approximations). In this sense, polynomials form a basis for the continuous function on and interval. By analogy, we would like our functions ϕ_i to be linearly independent and to come from basis of functions. There are many possible choices of bases, but a discussion of these would take us too far afield from this course.

Let now suppose that the functions $\phi_1, \phi_2, \ldots, \phi_k$ are linearly independent and arise from a set of basis function that reflect a deeper intuition about the behavior of the function f, e.g. it is well approximated as a sum of exponentials (or trig functions). Then the task to to find those coefficient $x_1, x_2, \ldots, x_n$ that best fits the data in the least squares sense:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2} \sum_{i=1}^{N} (x_1\phi_1(z_i) + x_2\phi_2(z_i) + \cdots + x_n\phi_k(z_i) - y_i)^2.$$

This can be recast as the linear least squares problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2} \left\| Ax - y \right\|_2^2,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad A = \begin{bmatrix} \phi_1(z_1) & \phi_2(z_1) & \dots & \phi_n(z_1) \\ \phi_1(z_2) & \phi_2(z_2) & \dots & \phi_n(z_2) \\ \vdots & & & \\ \phi_1(z_N) & \phi_2(z_N) & \dots & \phi_n(z_N) \end{bmatrix}.$$

May possible further generalizations of this basic idea are possible. For example, the data may be multi-dimensional: $(z_i, y_i) \in \mathbb{R}^s \times \mathbb{R}^t$. In addition, constraints may be added, e.g., the function must be monotone (either increasing of decreasing), it must be unimodal (one "bump"), etc. But the essential features are that we estimate using linear combinations and errors are measured using sums of squares. In many cases, the sum of squares error metric is not a good choice. But is can be motivated by assuming that the error are distributed using the Gaussian, or normal, distribution.

1.3. Linear Regression and Maximum Likelihood. Suppose we are considering a new drug therapy for reducing inflammation in a targeted population, and we have a relatively precise way of measuring inflammation for each member of this population. We are trying to determine the dosing to achieve a target level of inflamation. Of course, the dose needs to be adjusted for each individual due to the great amount of variability from one individual to the next. One way to model this is to assume that the resultant level of inflamation is on average a linear function of the dose and other individual specific covariates such as sex, age, weight, body surface area, gender, race, blood iron levels, desease state, etc. We then sample a collection of N individuals from the target population, registar their dose z_{i0} and the values of their individual specific covariates $z_{i1}, z_{i2}, \dots, z_{in}$, $i = 1, 2, \dots, N$. After dosing we observe that the resultant inflammation for the ith subject to be $y_i,\ i = 1, 2, \dots, N$. By saying that the "resultant level of inflamation is on average a linear function of the dose and other individual specific covariates ", we mean that there exist coefficients $x_0, x_1, x_2, \dots, x_n$ such that

$$y_i = x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} + v_i,$$

where v_i is an instance of a random variable representing the individuals deviation from the linear model. Assume that the random variables v_i are independently identically distributed $N(0, \sigma^2)$ (norm with zero mean and variance σ^2). The probability density function for the the normal distribution $N(0, \sigma^2)$ is

$$\frac{1}{\sigma\sqrt{2\pi}} \text{EXP}[-v^2/(2\sigma^2)] \ .$$

Given values for the coefficients x_i, the likelihood function for the sample $y_i,\ i = 1, 2, \dots, N$ is the joint probability density function evaluated at this observation. The independence assumption tells us that this joint pdf is given by

$$L(x; y) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \text{EXP}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} - y_i\right)^2\right] \ .$$

We now wish to choose those values of the coefficients $x_0, x_2, \dots, x_n$ that make the observation $y_1, y_2, \dots, y_n$ most probable. One way to try to do this is to maximize the likelihood function $L(x; y)$ over all possible values of x. This is called *maximum likelihood estimation*:

$$\underset{x \in \mathbb{R}^{n+1}}{\text{maximize}}\, L(x; y) \ . \tag{2.2}$$

Since the natural logarithm is nondecreasing on the range of the likelihood function, the problem (2.2) is equivalent to the problem

$$\underset{x \in \mathbb{R}^{n+1}}{\text{maximize}} \ln(L(x; y)) \ ,$$

which in turn is equivalent to the minimization problem

$$\underset{x \in \mathbb{R}^{n+1}}{\text{minimize}} -\ln(L(x; y)) \ . \tag{2.3}$$

Finally, observe that

$$-\ln(L(x;y)) = K + \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(x_0 z_{i0} + x_1 z_{i1} + x_2 z_{i2} + \cdots + x_n z_{in} - y_i\right)^2,$$

where $K = n\ln(\sigma\sqrt{2\pi})$ is constant. Hence the problem (2.3) is equivalent to the linear least squares problem

$$\underset{x\in\mathbb{R}^{n+1}}{\text{minimize}}\ \tfrac{1}{2}\|Ax-y\|_2^2,$$

where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x = \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \text{and} \quad A = \begin{bmatrix} z_{10} & z_{11} & z_{12} & \cdots & z_{1n} \\ z_{20} & z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & & & & \\ z_{N0} & z_{N1} & z_{N2} & \cdots & z_{Nn} \end{bmatrix}.$$

This is the first step in trying to select an optimal dose for each individual across a target population. What is missing from this analysis is some estimation of the variability in inflammation response due to changes in the covariates. Understanding this sensitivity to variations in the covariates is an essential part of any regression analysis. However, a discussion of this step lies beyond the scope of this brief introduction to linear regression.

1.4. System Identification in Signal Processing. We consider a standard problem in signal processing concerning the behavior of a stable, causal, linear, continuous-time, time-invariant system with input signal $u(t)$ and output signal $y(t)$. Assume that these signals can be described by the convolution integral

$$y(t) = (g * u)(t) := \int_0^{+\infty} g(\tau)u(t-\tau)d\tau\ . \tag{2.4}$$

In applications, the goal is to obtain an estimate of g by observing outputs y from a variety of known input signals u. For example, returning to our drug dosing example, the function u may represent the input of a drug into the body through a drug pump any y represent the concentration of the drug in the body at any time t. The relationship between the two is clearly causal (and can be shown to be stable). The transfer function g represents what the body is doing to the drug. In the way, the model (2.4) is a common model used in pharmaco-kinetics.

The problem of estimating g in (2.4) is an infinite dimensional problem. Below we describe a way to approximate g using the the FIR, or *finite impulse response* filter. In this model we discretize time by choosing a fixed number N of time points t_i to observe y from a known input u, and a finite time horizon $n < N$ over which to approximate the integral in (2.4). To simplify matters we index time on the integers, that is, we equate t_i with the integer i. After selecting the data points and the time horizon, we obtain the FIR model

$$y(t) = \sum_{k=1}^{n} g(k)u(t-k), \tag{2.5}$$

where we try to find the "best" values for $g(k),\ k = 0, 1, 2, \ldots, n$ to fit the system

$$y(t) = \sum_{k=0}^{n} g(k)u(t-k), \quad t = 1, 2, \ldots, N.$$

Notice that this requires knowledge of the values $u(t-k)$ for $t = t = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, n$. One often assumes a observational error in this model that is $N(0, \sigma^2)$ for a given value of σ^2. In this case,

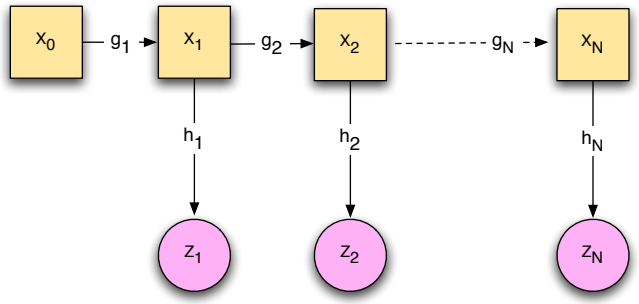


FIGURE 1. Dynamic systems amenable to Kalman smoothing methods.

the FIR model (2.5) becomes

$$y(t) = \sum_{i=1}^{n} g(k)u(t-k) + v(t), \tag{2.6}$$

where $v(t),\ t = 1, \ldots, N$ are iid $N(0, \sigma^2)$. In this case, the corresponding maximum likelihood estimation problem becomes the linear least squares problem

$$\underset{g \in \mathbb{R}^{n+1}}{\text{minimize}}\ \tfrac{1}{2} \left\| Hg - y \right\|_2^2 ,$$

where

$$y = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix}, \quad g = \begin{pmatrix} g(0) \\ g(1) \\ g(2) \\ \vdots \\ g(n) \end{pmatrix} \quad \text{and} \quad H = \begin{bmatrix} u(1) & u(0) & u(-1) & u(-2) & \ldots & u(1-n) \\ u(2) & u(1) & u(0) & u(-1) & \ldots & u(2-n) \\ u(3) & u(2) & u(1) & u(0) & \ldots & u(3-n) \\ \vdots & & & & & \\ u(N) & u(N-1) & u(N-2) & u(N-3) & \ldots & u(N-n) \end{bmatrix}.$$

Notice that the matrix H has constant "diagonals". Such matrices are called *Toeplitz matrices*.

1.5. Kalman Smoothing. Kalman smoothing is a fundamental topic in signal processing and control literature, with numerous applications in navigation, tracking, healthcare, finance, and weather. Contributions to theory and algorithms related to Kalman smoothing, and to dynamic system inference in general, have come from statistics, engineering, numerical analysis, and optimization. Here, the term 'Kalman smoother' includes any method of inference on any dynamical system fitting the graphical representation of Figure 1.

The combined mathematical, statistical, and probablistic model corresponding to Figure 1 is specified as follows:

$$\begin{array}{rcll} \mathbf{x_1} & = & g_1(x_0) + \mathbf{w_1}, & \\ \mathbf{x_k} & = & g_k(\mathbf{x_{k-1}}) + \mathbf{w_k} & k = 2, \ldots, N, \\ \mathbf{z_k} & = & h_k(\mathbf{x_k}) + \mathbf{v_k} & k = 1, \ldots, N\ , \end{array} \tag{2.7}$$

where $\mathbf{w_k}$, $\mathbf{v_k}$ are mutually independent random variables with known positive definite covariance matrices Q_k and R_k, respectively. The vectors $\{\mathbf{x_k}\}$ are called the state sequence and the vectors $\{\mathbf{z_k}\}$ the observation sequence. Here, $\mathbf{w_k}$ often, but not always, arises from a probabilistic model (discretization of an underlying stochastic differential equation in the state x, from which the names 'smoother' is derived) and $\mathbf{v_k}$ comes from a statistical model for observations. We have $\mathbf{x_k}, \mathbf{w_k} \in \mathbb{R}^n$, and $\mathbf{z_k}, \mathbf{v_k} \in \mathbb{R}^{m(k)}$, so dimensions can vary between time points. The functions g_k and h_k as well as the matrices Q_k and R_k are known and given. In addition, the observation sequence $\{\mathbf{z_k}\}$ is also known. The goal is to estimate the unobserved state sequence $\{\mathbf{x_k}\}$. For example, in our drug dosing, the amount of the drug remaining

in the body at time t is the unknown state sequence while the observation sequence is the observed concentration of the drug in each of our blood draws.

The classic case is obtained by making the following assumptions:

(1) x_0 is known, and g_k, h_k are known *linear* functions, which we denote by

$$g_k(x_{k-1}) = G_k x_{k-1} \quad h_k(x_k) \quad = H_k x_k \tag{2.8}$$

where $G_k \in \mathbb{R}^{n\times n}$ and $H_k \in \mathbb{R}^{m(k)\times n}$,

(2) $\mathbf{w_k}$, $\mathbf{v_k}$ are mutually independent *Gaussian* random variables.

In the classical setting, the connection to the linear least squares problem is obtained by formulating the maximum *a posteriori* (MAP) problem under linear and Gaussian assumptions. As in the linear regression and signal processing applications, this yields the following linear least squares problem:

$$\min_{\{x_k\}} f(\{x_k\}) := \sum_{k=1}^{N} \frac{1}{2}(z_k - H_k x_k)^T R_k^{-1}(z_k - H_k x_k) + \frac{1}{2}(x_k - G_k x_{k-1})^T Q_k^{-1}(x_k - G_k x_{k-1}) \,. \tag{2.9}$$

To simplify this expression, we introduce data structures that capture the entire state sequence, measurement sequence, covariance matrices, and initial conditions. Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

$$\mathrm{vec}(\{u_k\}) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \,,\; \mathrm{diag}(\{T_k\}) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

We now make the following definitions:

$$\begin{array}{lll} R = \mathrm{diag}(\{R_k\}) & x = \mathrm{vec}(\{x_k\}) & \\ Q = \mathrm{diag}(\{Q_k\}) & w = \mathrm{vec}(\{g_0, 0, \ldots, 0\}) & G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}, \\ H = \mathrm{diag}(\{H_k\}) & z = \mathrm{vec}(\{z_1, z_2, \ldots, z_N\}) & \end{array} \tag{2.10}$$

where $g_0 := g_1(x_0) = G_1 x_0$. With definitions in (2.10), problem (2.9) can be written

$$\min_x f(x) = \frac{1}{2}\|Hx - z\|^2_{R^{-1}} + \frac{1}{2}\|Gx - w\|^2_{Q^{-1}} \,, \tag{2.11}$$

where $\|a\|^2_M = a^\top M a$.

Since the number of time steps N can be quite large, it is essential that the underlying tri-diagonal structure is exploited in any solution procedure. This is especially true when the state-space dimension n is also large which occurs when making PET scan movies of brain metabolics or reconstructing weather patterns on a global scale.

2. Optimality in the Linear Least Squares Problem

We now turn to a discussion of optimality in the least squares problem (1.4) which we restate here for ease of reference:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \; \tfrac{1}{2}\,\|Ax - b\|_2^2 \,, \tag{2.12}$$

where

$$A \in \mathbb{R}^{m\times n},\; b \in \mathbb{R}^m, \quad \text{and} \quad \|y\|_2^2 := y_1^2 + y_2^2 + \cdots + y_m^2 \,.$$

In particular, we will address the question of when a solution to this problem exists and how they can be identified or characterized.

Suppose that $\overline{x}$ is a solution to (2.12), i.e.,

$$\|A\overline{x} - b\|_2 \leq \|Ax - b\|_2 \quad \forall\ x \in \mathbb{R}^n. \tag{2.13}$$

Using this inequality, we derive necessary and sufficient conditions for the optimality of $\overline{x}$. A useful identity for our derivation is

$$\|u+v\|_2^2 = (u+v)^T(u+v) = u^Tu + 2u^Tv + v^Tv = \|u\|_2^2 + 2u^Tv + \|v\|_2^2\,. \tag{2.14}$$

Let x be any other vector in $\mathbb{R}^n$. Then, using (2.14) with $u = A(\overline{x} - x)$ and $v = Ax - b$ we obtain

$$\begin{aligned} \|A\overline{x} - b\|_2^2 &= \|A(\overline{x} - x) + (Ax - b)\|_2^2 \\ &= \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) + \|Ax - b\|_2^2 \\ &\geq \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) + \|A\overline{x} - b\|_2^2 \qquad \text{(by (2.13))}. \end{aligned} \tag{2.15}$$

Therefore, by canceling $\|A\overline{x} - b\|_2^2$ from both sides, we know that, for all $x \in \mathbb{R}^n$,

$$0 \geq \|A(\overline{x} - x)\|_2^2 + 2(A(\overline{x} - x))^T(Ax - b) = 2(A(\overline{x} - x))^T(A\overline{x} - b) - \|A(\overline{x} - x)\|_2^2\,.$$

By setting $x = \overline{x} + tw$ for $t \in T$ and $w \in \mathbb{R}^n$, we find that

$$\frac{t^2}{2}\|Aw\|_2^2 + tw^TA^T(A\overline{x} - b) \geq 0 \quad \forall\ t \in \mathbb{R} \quad \text{and} \quad w \in \mathbb{R}^n.$$

Dividing by $t > 0$, we find that

$$\frac{t}{2}\|Aw\|_2^2 + w^TA^T(A\overline{x} - b) \geq 0 \quad \forall\ t \in \mathbb{R}_{++} \quad \text{and} \quad w \in \mathbb{R}^n,$$

Sending t to zero gives

$$w^TA^T(A\overline{x} - b) \geq 0 \quad \forall\ w \in \mathbb{R}^n,$$

which implies that $A^T(A\overline{x} - b) = 0$ (why?), or equivalently,

$$A^TA\overline{x} = A^Tb. \tag{2.16}$$

The system of equations (2.16) is called the *normal equations* associated with the linear least squares problem (2.12). This derivation leads to the following theorem.

THEOREM 2.1. *[Linear Least Squares and the Normal Equations]*
The vector $\overline{x}$ solves the problem (2.12)*, i.e.,*

$$\|A\overline{x} - b\|_2 \leq \|Ax - b\|_2 \quad \forall\ x \in \mathbb{R}^n,$$

if and only if $A^TA\overline{x} = A^Tb$.

PROOF. We have just shown that if $\overline{x}$ is a solution to (2.12), then the normal equations are satisfied, so we need only establish the reverse implication. Assume that $A^TA\overline{x} = A^Tb$ or, equivalently, $A^T(A\overline{x} - b) = 0$. Then, for all $x \in \mathbb{R}^n$,

$$\begin{aligned} \|Ax - b\|_2^2 &= \|(Ax - A\overline{x}) + (A\overline{x} - b)\|_2^2 \\ &= \|A(x - \overline{x})\|_2^2 + 2(A(x - \overline{x}))^T(A\overline{x} - b) + \|A\overline{x} - b\|_2^2 \qquad \text{(by (2.14))} \\ &\geq 2(x - \overline{x})^TA^T(A\overline{x} - b) + \|A\overline{x} - b\|_2^2 \qquad (\text{since } \|A(x - \overline{x})\|_2^2 \geq 0) \\ &= \|A\overline{x} - b\|_2^2 \qquad (\text{since } A^T(A\overline{x} - b) = 0), \end{aligned}$$

or equivalently, $\overline{x}$ solves (2.12). □

This theorem provides a nice characterization of solutions to (2.12), but it does not tell us if a solution exits. For this we use the following elementary result from linear algebra.

LEMMA 2.2. *For every matrix $A \in \mathbb{R}^{m\times n}$ we have*

$$\mathit{Null}(A^TA) = \mathit{Null}(A) \quad \text{and} \quad \mathit{Ran}(A^TA) = \mathit{Ran}(A^T)\ .$$

PROOF. Note that if $x \in \text{Null}(A)$, then $Ax = 0$ and so $A^TAx = 0$, that is, $x \in \text{Null}(A^TA)$. Therefore, $\text{Null}(A) \subset \text{Null}(A^TA)$. Conversely, if $x \in \text{Null}(A^TA)$, then

$$A^TAx = 0 \implies x^TA^TAx = 0 \implies (Ax)^T(Ax) = 0 \implies \|Ax\|_2^2 = 0 \implies Ax = 0,$$

or equivalently, $x \in \text{Null}(A)$. Therefore, $\text{Null}(A^TA) \subset \text{Null}(A)$, and so $\text{Null}(A^TA) = \text{Null}(A)$.

Since $\text{Null}(A^TA) = \text{Null}(A)$, the Fundamental Theorem of the Alternative tells us that

$$\text{Ran}(A^TA) = \text{Ran}((A^TA)^T) = \text{Null}(A^TA)^\perp = \text{Null}(A)^\perp = \text{Ran}(A^T),$$

which proves the lemma. □

This lemma immediately gives us the following existence result.

THEOREM 2.3. *[Existence and Uniqueness for the Linear Least Squares Problem] Consider the linear least squares problem* (2.12).

(1) A solution to the normal equations (2.16) *always exists.*
(2) A solution to the linear least squares problem (2.12) *always exists.*
(3) The linear least squares problem (2.12) *has a unique solution if and only if Null*$(A) = \{0\}$ *in which case* $(A^TA)^{-1}$ *exists and the unique solution is given by* $\overline{x} = (A^TA)^{-1}A^Tb$.
(4) If Ran$(A) = \mathbb{R}^m$, *then* $(AA^T)^{-1}$ *exists and* $\overline{x} = A^T(AA^T)^{-1}b$ *solves* (2.12), *indeed,* $A\overline{x} = b$.

PROOF. (1) Lemma 2.2 tells us that $\text{Ran}(A^TA) = \text{Ran}(A^T)$; hence, a solution to $A^TAx = A^Tb$ must exist.
(2) This follows from Part (1) and Theorem 2.1.
(3) By Theorem 2.1, $\overline{x}$ solves the linear least squares problem if and only if $\overline{x}$ solves the normal equations. Hence, the linear least squares problem has a uniques solution if and only if the normal equations have a unique solution. Since $A^TA \in \mathbb{R}^{n\times n}$ is a square matrix, this is equivalent to saying that A^TA is invertible, or equivalently, $\text{Null}(A^TA) = \{0\}$. However, by Lemma 2.2, $\text{Null}(A) = \text{Null}(A^TA)$. Therefore, the linear least squares problem has a uniques solution if and only if $\text{Null}(A) = \{0\}$ in which case A^TA is invertible and the unique solution is given by $\overline{x} = (A^TA)^{-1}A^Tb$.
(4) By the hypotheses, Lemma 2.2, and the Fundamental Theorem of the Alternative, $\{0\} = (\mathbb{R}^m)^\perp = (\text{Ran}(A))^\perp = \text{Null}(A^T) = \text{Null}(AA^T)$; hence, $AA^T \in \mathbb{R}^{m\times m}$ is invertible. Consequently, $\overline{x} = A^T(AA^T)^{-1}b$ is well-defined and satisfies $A\overline{x} = b$ □

Theorem 2.3 establishes the existence of solutions to the linear least squares problem as well as necessary conditions for optimality and uniqueness. When the solution is unique, it also provides a formula for this solution. However, these results do not provide a numerical mechanism for computing a solution even in the case when the solution is unique. Here the dimension of the problem, or the problem size, plays a key role. In addition, the level of accuracy in the solution as well as the greatest accuracy possible are also issues of concern. Linear least squares problems range in size from just a few variables and equations to millions. Some are so large that all of the computing resources at our disposal today are insufficient to solve them, and in many cases the matrix A is not even available in the sense that it is not stored on a computer. However, in this latter case, it is often possible to either compute or approximate the vector Ax for a given vector x. Therefore, great care and inventiveness is required in the numerical solution of these problems. Research into how to solve this class of problems remains an important area of research to this day.

In our study of numerical solution techniques we present a few classical methods. But before doing so, we study other aspects of the problem in order to gain further insight into its geometric structure.

3. Orthogonal Projection onto a Subspace

In this section we view the linear least squares problem from the perspective of a least distance problem to a subspace, or equivalently, as the problem of projecting onto a subspace. Suppose $S \subset \mathbb{R}^m$ is a given

subspace and $b \notin S$. The least distance problem for S and b is to find that element of S that is as close to b as possible. That is we wish to solve the problem

$$\min_{z \in S} \tfrac{1}{2} \|z - b\|_2^2 \ , \tag{2.17}$$

or equivalently, we wish to find the point $\overline{z} \in S$ such that

$$\|\overline{z} - b\|_2 \leq \|z - b\|_2 \qquad \forall\ z \in S.$$

If we take the subspace to be the range of A, $S = \mathrm{Ran}(A)$, then the problem (2.17) is closely related to the problem (2.12) since

(2.18)

$\overline{z} \in \mathbb{R}^m$ solves (2.17) if and only if there is an $\overline{x} \in \mathbb{R}^n$ with $\overline{z} = A\overline{x}$ such that $\overline{x}$ solves (2.12). (why?)

Below we discuss this connection and its relationship to the notion of an orthogonal projection onto a subspace.

A matrix $P \in \mathbb{R}^{m \times m}$ is said to be a *projection* if and only if $P^2 = P$. In this case we say that P is a projection onto the subspace $S = \mathrm{Ran}(P)$, the range of P. Note that if $x \in \mathrm{Ran}(P)$, then there is a $w \in \mathbb{R}^m$ such that $x = Pw$, therefore, $Px = P(Pw) = P^2 w = Pw = x$. That is, P leaves all elements of $\mathrm{Ran}(P)$ fixed. Also, note that, if P is a projection, then

$$(I - P)^2 = I - P - P + P^2 = I - P,$$

and so $(I - P)$ is also a projection. Since for all $w \in \mathbb{R}^m$,

$$w = Pw + (I - P)w,$$

we have

$$\mathbb{R}^m = \mathrm{Ran}(P) + \mathrm{Ran}(I - P).$$

In this case we say that the subspaces $\mathrm{Ran}(P)$ and $\mathrm{Ran}(I - P)$ are *complementary subspaces* since their sum is the whole space and their intersection is the origin, i.e., $\mathrm{Ran}(P) \cap \mathrm{Ran}(I - P) = \{0\}$ (why?).

Conversely, given any two subspaces S_1 and S_2 that are complementary, that is, $S_1 \cap S_2 = \{0\}$ and $S_1 + S_2 = \mathbb{R}^m$, there is a projection P such that $S_1 = \mathrm{Ran}(P)$ and $S_2 = \mathrm{Ran}(I - P)$. We do not show how to construct these projections here, but simply note that they can be constructed with the aid of bases for S_1 and S_2.

The relationship between projections and complementary subspaces allows us to define a notion of *orthogonal projection*. Recall that for every subspace $S \subset \mathbb{R}^m$, the subspace orthogonal to S is given by

$$S^\perp := \left\{ x \,\middle|\, x^T y = 0 \ \forall\ y \in S \right\}.$$

We say that S and $S^\perp$ are *orthogal subspaces*. Clearly, S and $S^\perp$ are complementary:

$$S \cap S^\perp = \{0\} \qquad \text{and} \qquad S + S^\perp = \mathbb{R}^m. \qquad \text{(why?)}$$

Therefore, there is a projection P such that $\mathrm{Ran}(P) = S$ and $\mathrm{Ran}(I - P) = S^\perp$, or equivalently,

$$((I - P)y)^T (Pw) = 0 \qquad \forall\ y, w \in \mathbb{R}^m. \tag{2.19}$$

The orthogonal projection plays a very special role among all possible projections onto a subspace. For this reason, we denote the orthogonal projection onto the subspace S by P_S.

We now use the condition (2.19) to derive a simple test of whether a linear transformation is an orthogonal projection. For brevity, we write $P := P_S$ and set $M = (I - P)^T P$. Then, by (2.19),

$$0 = e_i^T M e_j = M_{ij} \qquad \forall\ i, j = 1, \ldots, n,$$

i.e., M is the zero matrix. But then, since $0 = (I - P)^T P = P - P^T P$, we have

$$P = P^T P = (P^T P)^T = P^T.$$

Conversely, if $P = P^T$ and $P^2 = P$, then $(I - P)^T P = 0$. Therefore, a matrix P is an orthogonal projection if and only if $P^2 = P$ and $P = P^T$.

An orthogonal projection for a given subspace S can be constructed from any orthonormal basis for that subspace. Indeed, if the columns of the matrix Q form an orthonormal basis for S, then the matrix $P = QQ^T$ satisfies

$$P^2 = QQ^TQQ^T \overset{\text{why?}}{=} QI_kQ^T = QQ^T = P \qquad \text{and} \qquad P^T = (QQ^T)^T = QQ^T = P,$$

where $k = \dim(S)$, and so P is the orthogonal projection onto S since, by construction, $\text{Ran}(QQ^T) = \text{Ran}(Q) = S$. We catalogue these observations in the following lemma.

LEMMA 2.4. *[Orthogonal Projections]*

(1) The matrix $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection if and only if $P = P^2$ and $P = P^T$.

(2) If the columns of the matrix $Q \in \mathbb{R}^{n\times k}$ form an orthonormal basis for the subspace $S \subset \mathbb{R}^n$, then $P := QQ^T$ is the orthogonal projection onto S.

Let us now apply these projection ideas to the problem (2.17). Let $P := P_S$ be the orthogonal projection onto the subspace S, and let $\overline{z} = Pb$. Then, for every $z \in S$,

$$\begin{aligned}
\|z-b\|_2^2 &= \|Pz - Pb - (I-P)b\|_2^2 && \text{(since } z \in S\text{)}\\
&= \|P(z-b) + (I-P)b\|_2^2 \\
&= \|P(z-b)\|_2^2 + 2(z-b)^TP^T(I-P)b + \|(I-P)b\|_2^2 \\
&= \|P(z-b)\|_2^2 + \|(I-P)b\|_2^2 && \text{(since } P = P^T \text{ and } P = P^2\text{)}\\
&\geq \|(P-I)b\|_2^2 && \text{(since } \|P(z-b)\|_2^2 >\geq 0\text{)}\\
&= \|\overline{z}-b\|_2^2 .
\end{aligned}$$

Consequently, $\|\overline{z}-b\|_2 \leq \|z-b\|_2$ for all $z \in S$, that is, $\overline{z} = Pb$ solves (2.17). This observation yield the following theorem as an elementary consequence of the *parallelogram law*:

$$2\|u\|_2^2 + 2\|v\|_2^2 = \|u+v\|_2^2 + \|u-v\|_2^2 \qquad \forall\, u, v \in \mathbb{R}^n.$$

THEOREM 2.5. *[Subspace Projection Theorem]*
Let $S \subset \mathbb{R}^m$ be a subspace and let $b \in \mathbb{R}^m \setminus S$. Then the unique solution to the least distance problem

$$\underset{z\in S}{\text{minimize}}\, \|z-b\|_2$$

is $\overline{z} := P_Sb$, where P_S is the orthogonal projector onto S.

PROOF. Everything but the uniqueness of the solution has been established in the discussion preceeding the theorem. To show uniqueness, apply the parallelogram law to obtain

$$\|(1-t)u + tv\|_2^2 = (1-t)\|u\|_2^2 + t\|v\|_2^2 - t(1-t)\|u-v\|_2^2 \quad \forall\ 0 \leq t \leq 1 \quad \text{and} \quad u, v \in \mathbb{R}^m.$$

Let $z^1, z^2 \in \mathbb{R}^m$ be two points that solve the minimum distance problem. Then, $\|z^1-b\|_2 = \|z^2-b\|_2 =: \eta > 0$, and so by the identity given above,

$$\begin{aligned}
\|\tfrac{1}{2}(z^1+z^2)-b\|_2^2 &= \|\tfrac{1}{2}(z^1-b) + \tfrac{1}{2}(z^2-b)\|_2^2 \\
&= \tfrac{1}{2}\|z^1-b\|_2^2 + \tfrac{1}{2}\|z^2-b\|_2^2 - \frac{1}{4}\|z^1-z^2\|_2^2 \\
&= \eta^2 - \frac{1}{4}\|z^1-z^2\|_2^2 .
\end{aligned}$$

Since $\eta = \inf\{\|z-b\|_2 \,|\, z \in S\}$, we must have $z^1 = z^2$. □

Let us now reconsider the linear least-squares problem (2.12) as it relates to our new found knowledge about orthogonal projections and their relationship to least distance problems for subspaces. Consider the case where $m >> n$ and $\text{Null}(A) = \{0\}$. In this case, Theorem 2.3 tells us that $\overline{x} = (A^TA)^{-1}A^Tb$

solves (2.12), and $\overline{z} = P_S b$ solves (2.19) where P_S is the orthogonal projector onto $S = \text{Ran}(A)$. Hence, by (2.18),

$$P_S b = \overline{z} = A\overline{x} = A(A^T A)^{-1} A^T b.$$

Since this is true for all possible choices of the vector b, we have

$$P_S = P_{\text{Ran}(A)} = A(A^T A)^{-1} A^T \ ! \tag{2.20}$$

That is, the matrix $A(A^T A)^{-1} A^T$ is the orthogonal projector onto the range of A. One can also check this directly by showing that the matrix $M = A(A^T A)^{-1} A^T$ satisfies $M^2 = M$, $M^T = M$, and $\text{Ran}(M) = \text{Ran}(A)$.

PROPOSITION 2.6. *Let $A \in \mathbb{R}^{m \times n}$ with $m \leq n$ and Null$(A) = \{0\}$. Then*

$$P_{Ran(A)} = A(A^T A)^{-1} A^T.$$

4. Minimal Norm Solutions to $Ax = b$

Again let $A \in \mathbb{R}^{m \times n}$, but now we suppose that $m << n$. In this case A is short and fat so the matrix A most likely has rank m, or equivalently,

$$\text{Ran}(A) = \mathbb{R}^m \ . \tag{2.21}$$

But regardless of the range of A and the choice of the vector $b \in \mathbb{R}^m$, the set of solutions to $Ax = b$ will be infinite if a solution exists since the nullity of A is $n - m$. Indeed, if x^0 is any particular solution to $Ax = b$, then the set of solutions is given by $x^0 + \text{Null}(A) := \left\{ x^0 + z \,\middle|\, z \in \text{Null}(A) \right\}$. In this setting, one might prefer the solution to the system having least norm. This solution is found by solving the problem

$$\min_{z \in \text{Null}(A)} \tfrac{1}{2} \left\| z + x^0 \right\|_2^2 \ . \tag{2.22}$$

This problem is of the form (2.17). Consequently, the solution is given by $\overline{z} = -P_S x^0$ where P_S is the orthogonal projection onto $S := \text{Null}(A)$. In particular, this implies that the least norm solution to the system $Ax = b$ is uniquely given by the orthogonal projection of x^0 onto the range of A^T since $S^\perp = \text{Null}(A)^\perp = \text{Ran}\left(A^T\right)$ and

$$x^0 + \overline{z} = x^0 - P_{\text{Null}(A)} x^0 = (I - P_{\text{Null}(A)}) x^0 = P_{\text{Null}(A)^\perp} x^0 = P_{\text{Ran}\left(A^T\right)} x^0 \ . \tag{2.23}$$

Recall that the formula (2.20) shows that if $M \in \mathbb{R}^{k \times s}$ is such that $\text{Null}(M) = \{0\}$, then the orthogonal projector onto $\text{Ran}(M)$ is given by

$$P_{\text{Ran}(M)} = M(M^T M)^{-1} M^T \ . \tag{2.24}$$

In our case, $M = A^T$ and $M^T M = AA^T$. Thus, if we assume that (2.21) holds, then

$$\text{Null}(M) = \text{Null}(A^T) = \text{Ran}(A)^\perp = (\mathbb{R}^m)^\perp = \{0\}$$

and consequently, by (2.24), the orthogonal projector onto $\text{Ran}(A^T)$ is given by

$$P_{\text{Ran}(A^T)} = A^T (AA^T)^{-1} A \ .$$

Therefore, when (2.21) holds, the least norm solution to $Ax = b$ is uniquely given by

$$\overline{x} = A^T (AA^T)^{-1} A x^0,$$

where x^0 is any particular solution to $Ax = b$. These observations establish the following theorem.

THEOREM 2.7. *[Least Norm Solution to Linear Systems] Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and let x^0 be any solution to the system $Ax = b$. Then the least norm solution to the system $Ax = b$ is given by the orthogonal projection of x^0 onto the range of A^T. If it is further assumed that Ran$(A) = \mathbb{R}^m$, then the following hold.*

(1) The matrix AA^T is invertible.

(2) The orthogonal projection onto $\mathrm{Ran}\left(A^T\right)$ *and Null*(A) *are given by*

$$P_{\mathrm{Ran}\left(A^T\right)} = A^T(AA^T)^{-1}A \quad \text{and} \quad P_{Null(A)} = I - A^T(AA^T)^{-1}A \ .$$

(3) For every $b \in \mathbb{R}^m$*, the system* $Ax = b$ *is consistent, and the least norm solution to this system is uniquely given by*

$$\overline{x} = A^T(AA^T)^{-1}Ax^0 \ ,$$

where x^0 *is any particular solution to the system* $Ax = b$*.*

5. Gram-Schmidt Orthogonalization, the QR Factorization and Normal Equations

5.1. Gram-Schmidt Orthogonalization. In the previous sections we learned the significance of orthogonal projections for the linear least squares problem. In addition, we found that if the columns of the matrix U form an orthonormal basis for the subspace S, then the matrix UU^T is the orthogonal projection onto S. Hence, one way to obtain an orthogonal projection onto a subspace S is to compute an orthogonal basis for S. This is precisely what the *Gram-Schmidt orthogonalization* process does.

Let us recall the Gram-Schmidt orthogonalization process for a sequence of linearly independent vectors $a_1, \ldots, a_n \in \mathbb{R}^m$ (note that this implies that $n \leq m$ (why?)). In this process we define vectors $q_1, \ldots, q_n$ inductively, as follows: set

$$p_1 = a_1, \qquad q_1 = p_1/\|p_1\|,$$

$$p_j = a_j - \sum_{i=1}^{j-1} \langle a_j,\, q_i\rangle\, q_i \quad \text{and} \quad q_j = p_j/\|p_j\| \quad \text{for} \quad 2 \leq j \leq n.$$

For $1 \leq j \leq n$, $q_j \in \mathrm{Span}\{a_1, \ldots, a_j\}$, so $p_j \neq 0$ by the linear independence of $a_1, \ldots, a_j$. An elementary induction argument shows that the q_j's form an orthonormal basis for span $(a_1, \ldots, a_n)$.

If we now define

$$r_{jj} = \|p_j\| \neq 0 \quad \text{and} \quad r_{ij} = \langle q_i,\, a_j\rangle \quad \text{for} \quad 1 \leq i < j \leq n,$$

then

$$\begin{aligned} a_1 &= r_{11}\, q_1, \\ a_2 &= r_{12}\, q_1 + r_{22}\, q_2, \\ a_3 &= r_{13}\, q_1 + r_{23}\, q_2 + r_{33}\, q_3, \\ &\vdots \\ a_n &= \sum_{i=1}^{n} r_{in}\, q_i. \end{aligned}$$

Set

$$A := [a_1\ a_2\ \ldots\ a_n] \in \mathbb{R}^{m\times n}, \quad R := [r_{ij}] \in \mathbb{R}^{n\times n}, \quad \text{and} \quad Q := [q_1\ q_2\ \ldots\ q_n] \in \mathbb{R}^{m\times n} \ ,$$

where $r_{ij} = 0,\ i > j$. Then

$$A = QR \ ,$$

where Q is unitary and R is an upper triangular $n \times n$ matrix. In addition, R is invertible since the diagonal entries r_{jj} are non-zero. This is called the *QR factorization* of the matrix A.

Remark 2.8. *If the* a_j*'s for* $j = 1, \ldots, n$ *are linearly dependent, then, for at least one value of* j*,*

$$a_j \in \mathrm{Span}\{a_1, \ldots, a_{j-1}\}, \quad \textit{and so} \quad p_j = 0.$$

The process can be modified by setting $r_{jj} = 0$*, not defining a new* q_j *for this iteration, but continuing to define* $r_{ij} = \langle a_j,\, q_i\rangle \quad$ *for* $\ 1 \leq i < j$*, and proceeding. We still obtain with orthogonormal vectors* $\{q_1, q_2, \ldots, q_k\}$*, but now* $k < n$*. In general, after* n *iterations, there will be* $1 \leq k \leq n$ *vectors* $\{q_1, \ldots, q_k\}$

that form an orthonormal basis for $\mathrm{Span}\{a_1, \ldots, a_n\}$*, where* $n - k$ *is the number of diagonal entries* r_{jj} *that take the value zero. Again we obtain* $A = QR$*, but now* Q *may not be square and the matrix* R *may have zero diagonal entries in which case it is not invertible.*

REMARK 2.9. *The classical Gram-Schmidt algorithm as described above can have poor computational behavior due to the accumulation of round-off error. In particular, the computed vectors* q_j*'s are not orthogonal:* $\langle q_j, q_k \rangle$ *is small for* $j \neq k$ *with* j *near* k*, but not so small for* $j \ll k$ *or* $j \gg k$*.*

An alternate version, "Modified Gram-Schmidt," is equivalent in exact arithmetic, but behaves better numerically. In the following "pseudo-codes," p *denotes a temporary storage vector used to accumulate the sums defining the* p_j*'s.*

Classic Gram-Schmidt

For $j = 1, \cdots, n$ *do*
- $p := a_j$
- *For* $i = 1, \ldots, j-1$ *do*
  - $r_{ij} = \langle a_j, q_i \rangle$
  - $p := p - r_{ij} q_i$
- $r_{jj} := \|p\|$
- $q_j := p / r_{jj}$

Modified Gram-Schmidt

For $j = 1, \ldots, n$ *do*
- $p := a_j$
- *For* $i = 1, \ldots, j-1$ *do*
  - $r_{ij} = \langle p, q_i \rangle$
  - $p := p - r_{ij} q_i$
- $r_{jj} = \|p\|$
- $q_j := p / r_{jj}$

The only difference is in the computation of r_{ij}*: in Modified Gram-Schmidt, we orthogonalize the accumulated partial sum for* p_j *against each* q_i *successively.*

THEOREM 2.10. *[The Full QR Factorization] Suppose* $A \in \mathbb{R}^{m \times n}$ *with* $m \geq n$*. Then there exists a permutation matrix* $P \in \mathbb{R}^{n \times n}$*, a unitary matrix* $Q \in \mathbb{R}^{m \times m}$*, and an upper triangular matrix* $R \in \mathbb{R}^{m \times n}$ *such that* $AP = QR$*. Let* $Q_1 \in \mathbb{R}^{m \times n}$ *denote the first* n *columns of* Q*,* Q_2 *the remaining* $(m - n)$ *columns of* Q*, and* $R_1 \in \mathbb{R}^{n \times n}$ *the first* n *rows of* R*, then*

$$AP = QR = [Q_1 \; Q_2] \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1. \tag{2.25}$$

Moreover, we have the following:

(a) We may choose R *to have nonnegative diagonal entries.*

(b) If A *is of full rank, then we can choose* R *with positive diagonal entries, in which case we obtain the condensed factorization* $A = Q_1 R_1$*, where* $R_1 \in \mathbb{R}^{n \times n}$ *invertible and the columns of* Q_1 *forming an orthonormal basis for the range of* A*.*

(c) If $\mathrm{rank}\,(A) = k < n$*, then*

$$R_1 = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

where R_{11} *is a* $k \times k$ *invertible upper triangular matrix and* $R_{12} \in \mathbb{R}^{k \times (n-k)}$*. In particular, this implies that* $AP = Q_{11}[R_{11} \; R_{12}]$*, where* Q_{11} *are the first* k *columns of* Q*. In this case, the columns of* Q_{11} *form an orthonormal basis for the range of* A*.*

REMARK 2.11. *We call the factorization* $AP = Q_{11}[R_{11} \; R_{12}]$ *in Part (c) above the condensed QR Factorization. Note that if* P *is a permutation matrix, then so is* P^T *with* $P^{-1} = P^T$ *(i.e. permutation matrices are unitary). The role of the permutation matrix is to make the first* $k = \mathrm{rank}\,(A)$ *columns of* AP *linearly independent.*

To distinguish the condensed QR Factorization from the factorization in (2.25) *with* Q *an* $m \times m$ *unitary matrix, we will refer the factorization where* Q *is unitary as the full QR factorization.*

PROOF. If necessary, permute the columns of A so that the first $k = \operatorname{rank}(A)$ columns of A are linearly independent and let P denote the permutation matrix that accomplishes this task so the the first k columns of AP are linearly independent. Apply the Gram-Schmidt orthogonalization process to obtain the matrix

$$Q_1 = [q_1, \ldots, q_k] \in \mathbb{R}^{m\times k} \quad \text{and the upper triangular matrix} \quad \widetilde{R}_{11} = [r_{ij}] \in \mathbb{R}^{k\times k}$$

so that Q_1R_1 gives the first k columns of A. The write the remaining columns of A as linear combinations of the columns of Q_1 to obtain the coefficient matrix $R_{12} \in \mathbb{R}^{k\times(n-k)}$ yielding $AP = Q_1[R_{11}\ R_{12}]$. Finally, extend $\{q_1, \ldots, q_k\}$ to an orthonormal basis $\{q_1, \ldots, q_m\}$ of $\mathbb{R}^m$, and set

$$Q = [q_1, \ldots, q_m] \quad \text{and} \quad R = \begin{bmatrix} R_1 \\ 0 \end{bmatrix} \in \mathbb{R}^{m\times n}, \text{ so } AP = QR.$$

As $r_{jj} > 0$ in the Gram-Schmidt process, we have (b). □

REMARK 2.12. *There are more efficient and better computationally behaved ways of calculating the Q and R factors. The idea is to create zeros below the diagonal (successively in columns $1, 2, \ldots$) as in Gaussian Elimination, except instead of doing this by successive left multiplication by Gaussian elimination matrices, we left multiply by unitary matrices. Below, we show how this can be done with Householder transformations. But another popular approach is to use Givens rotations.*

In practice, every $A \in \mathbb{R}^{m\times n}$ has a QR-factorization, even when $m < n$. This follows immediately from Part (c) Theorem 2.10.

COROLLARY 2.12.1. *[The General Condensed QR Factorization] Let $A \in \mathbb{R}^{m\times n}$ have rank $k \leq \min\{m, n\}$. Then there exist*

$$\begin{aligned} Q &\in \mathbb{R}^{m\times k} \quad \textit{with orthonormal columns,} \\ R &\in \mathbb{R}^{k\times n} \quad \textit{full rank upper triangular, and} \\ P &\in \mathbb{R}^{n\times n} \quad \textit{a permutation matrix} \end{aligned}$$

such that

$$AP = QR.$$

In particular, the columns of the matrix Q form a basis for the range of A. Moreover, the matrix R can be written in the form

$$R = [R_1\ R_2],$$

where $R_1 \in \mathbb{R}^{k\times k}$ is nonsingular.

REMARK 2.13. *The permutation P in the corollary above can be taken to be any permutation that re-orders the columns of A so that the first k columns of A are linearly independent, where k is the rank of A (similarly for $\widetilde{P}$ in permuting the columns of A^T).*

COROLLARY 2.13.1. *[Orthogonal Projections onto the Four Fundamental Subspaces] Let $A \in \mathbb{R}^{m\times n}$ have rank $k \leq \min\{m, n\}$. Let A and A^T have generalized QR factorizations*

$$AP = Q[R_1\ R_2] \quad \text{and} \quad A^T\widetilde{P} = \widetilde{Q}[\widetilde{R}_1\ \widetilde{R}_2].$$

Since row rank equals column rank, $P \in \mathbb{R}^{n\times n}$ is a permutation matrix, $\widetilde{P} \in \mathbb{R}^{m\times m}$ is a permutation matrix, $Q \in \mathbb{R}^{m\times k}$ and $\widetilde{Q} \in \mathbb{R}^{n\times k}$ have orthonormal columns, $R_1, \widetilde{R}_1 \in \mathbb{R}^{k\times k}$ are both upper triangular nonsingular matrices, $R_2 \in \mathbb{R}^{k\times(n-k)}$, and $\widetilde{R}_2 \in \mathbb{R}^{k\times(m-k)}$. Moreover,

$$\begin{aligned} QQ^T &\quad \textit{is the orthogonal projection onto } \textit{Ran}(A), \\ I - QQ^T &\quad \textit{is the orthogonal projection onto } \textit{Null}(A^T), \\ \widetilde{Q}\widetilde{Q}^T &\quad \textit{is the orthogonal projection onto } \textit{Ran}(A^T), \textit{ and} \\ I - \widetilde{Q}\widetilde{Q}^T &\quad \textit{is the orthogonal projection onto } \textit{Null}(A)^\perp. \end{aligned}$$

PROOF. The result follows immediately from Corollary 2.12.1 and the Fundamental Theorem of the Alternative. □

EXERCISE 2.1. *Verify the representations of the orthogonal projections onto Ran(A) and Null(A) given in Corollary 2.13.1 correspond to those given in Proposition 2.6 and Theorem 2.7.*

5.2. Solving the Normal Equations with the QR Factorization. Let's now reconsider the linear least squares problem (2.12) and how the QR factorization can be used in its solution. Specifically, we examine how is can be used to solve the normal equations $A^TAx = A^Tb$. Let A and b be as in (2.12), and let

$$AP = Q[R_1 \ R_2]$$

be the general condensedQR factorization of A, where $P \in \mathbb{R}^{n\times n}$ is a permutation matrix, $Q \in \mathbb{R}^{m\times k}$ has orthonormal columns, $R_1 \in \mathbb{R}^{k\times k}$ is nonsingular and upper triangular, and $R_2 \in \mathbb{R}^{k\times(n-k)}$ with $k = \operatorname{rank}(A) \leq \min\{n, m\}$. Replacing A by $A = Q[R_1 \ R_2]P^T$ in the normal equations gives the following equivalent system:

$$P\begin{bmatrix}R_1^T\\R_2^T\end{bmatrix}Q^TQ\begin{bmatrix}R_1 & R_2\end{bmatrix}P^Tx = P\begin{bmatrix}R_1^T\\R_2^T\end{bmatrix}\begin{bmatrix}R_1 & R_2\end{bmatrix}P^Tx = A^Tb = P\begin{bmatrix}R_1^T\\R_2^T\end{bmatrix}Q^Tb,$$

since $Q^TQ = I_k$ the $k\times k$ identity matrix. By multiplying on the left by P^T, replacing b by $\hat{b} := Q^Tb \in \mathbb{R}^k$ and $\begin{bmatrix}R_1 & R_2\end{bmatrix}P^Tx$ by z, we obtain

$$\begin{bmatrix}R_1^T\\R_2^T\end{bmatrix}z = \begin{bmatrix}R_1^T\\R_2^T\end{bmatrix}\hat{b}.$$

Let us see if we can reconstruct a solution to the normal equations by choosing the most obvious solution to the this system, namely, $\overline{z} := \hat{b}$. If this is to yield a solution to the normal equations, then, by the equivalence $z = \begin{bmatrix}R_1 & R_2\end{bmatrix}P^Tx$, we obtain $\overline{x}$ by solving the system

$$\overline{z} = \hat{b} = \begin{bmatrix}R_1 & R_2\end{bmatrix}P^Tx \ .$$

Set

$$\begin{bmatrix}w_1\\w_2\end{bmatrix} := P^Tx \ ,$$

where $w_1 \in \mathbb{R}^k$ and $w_2 \in \mathbb{R}^{(n-k)}$, and consider the system

$$R_1w_1 = \hat{b} \ \in \mathbb{R}^k.$$

Since $R_1 \in \mathbb{R}^{k\times k}$ is invertible, this system has a unique solution $\overline{w}_1 := R_1^{-1}\hat{b}$. Indeed, this system is very easy to solve using *back substitution* since R_1 is upper triangular. Next set $\overline{w}_2 = 0 \in \mathbb{R}^{(n-k)}$ and

$$\overline{x} := P\overline{w} = P\begin{bmatrix}R_1^{-1}\hat{b}\\0\end{bmatrix}.$$

Then

$$\begin{aligned}
A^TA\overline{x} &= A^TAP\begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix} \\
&= A^TQ\begin{bmatrix} R_1 & R_2 \end{bmatrix}P^TP\begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix} \\
&= A^TQR_1R_1^{-1}\hat{b} && \text{(since } P^TP = I\text{)} \\
&= A^TQ\hat{b} \\
&= A^TQQ^Tb \\
&= P\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix}Q^TQQ^Tb && \text{(since } A^T = P\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix}Q^T\text{)} \\
&= P\begin{bmatrix} R_1^T \\ R_2^T \end{bmatrix}Q^Tb && \text{(since } Q^TQ = I\text{)} \\
&= A^Tb,
\end{aligned}$$

that is, $\overline{x}$ solves the normal equations!

Let us now consider the computational cost of obtaining the solution to the linear least squares problem in this way. The key steps is this computation are as follows:

$AP = Q[R_1\ R_2]$	the general condensed QR factorization	$o(m^2n)$
$\hat{b} = Q^Tb$	a matrix-vector product	$o(km)$
$\overline{w}_1 = R_1^{-1}\hat{b}$	a back solve	$o(k^2)$
$\overline{x} = P\begin{bmatrix} R_1^{-1}\hat{b} \\ 0 \end{bmatrix}$	a matrix-vector product	$o(kn)$.

Therefore, the majority of the numerical effort is in the computation of the QR factorization.

5.3. Computing the Full QR Factorization using Householder Reflections. In subsection 5.1 we showed how to compute the QR factorization using the Gram-Schmidt orthogonalization procedure. We also indicated that due to numerical round-off error this procedure has difficulty in preserving the orthogonality of the columns of the matrix Q. To address this problem we presented the mathematically equivalent *modified* Gram-Schmidt process which has improved performance. We now present a very different method for obtaining the full QR factorization. The approach we describe is very much like Gauss-Jordan Elimination to obtain reduced echelon form. However, now we successively multiply A on the left by unitary matrices, rather than Gauss-Jordan elimination matrices, which eventually put A into upper triangular form. The matrices we multiply by are the *Householder reflection matrices*.

Given $w \in \mathbb{R}^n$ we can associate the matrix

$$U = I - 2\frac{ww^T}{w^Tw}$$

which reflects $\mathbb{R}^n$ across the hyperplane $\text{Span}\{w\}^\perp$. The matrix U is call the Householder reflection across this hyperplane.

Given a pair of vectors x and y with

$$\|x\|_2 = \|y\|_2, \quad \text{and} \quad x \neq y,$$

the Householder reflection

$$U = I - 2\frac{(x-y)(x-y)^T}{(x-y)^T(x-y)}$$

is such that $y = Ux$, since

$$\begin{aligned} Ux &= x - 2(x-y)\frac{\|x\|^2 - y^T x}{\|x\|^2 - 2y^T x + \|y\|^2} \\ &= x - 2(x-y)\frac{\|x\|^2 - y^T x}{2(\|x\|^2 - y^T x)} \qquad (\text{since } \|x\| = \|y\|) \\ &= y\ . \end{aligned}$$

We now show how Householder reflections can be used to obtain the QR factorization. Let $\mu := \min\{n, m\}$. The procedure described below terminates in at most $\kappa \leq \tau$ steps. The approach is based on a numerical linear algebra procedure call *deflation* where the dimension of the problem is reduced at each iteration. Here we describe the basic idea of a deflation step in the QR-factorization of the matrix $A_0 \in \mathbb{R}^{m\times n} \setminus \{0\}$. Begin by block decomposing A_0 as

$$A_0 = \begin{bmatrix} \alpha_0 & a_0^T \\ b_0 & \tilde{A}_0 \end{bmatrix}, \text{ with } \tilde{A}_0 \in \mathbb{R}^{(m-1)\times(n-1)},$$

and set

$$\nu_0 = \left\| \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \right\|_2 .$$

If $\nu_0 = 0$, then multiply A_0 on the left by a permutation matrix P_0 to bring a non-zero (largest magnitude) column in A_0 into the first column and the zero column to the last column. Then block decompose $A_0 P_0$ as above with

$$A_0 P_0 = \begin{bmatrix} \alpha_0 & a_0^T \\ b_0 & \tilde{A}_0 \end{bmatrix}, \text{ with } \tilde{A}_0 \in \mathbb{R}^{(m-1)\times(n-1)},$$

and set

$$\nu_0 = \left\| \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \right\|_2 \neq 0.$$

Let H_0 be the Householder transformation that maps

$$\begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} \mapsto \nu_0\, e_1 \quad :$$

$$H_0 = I - 2\frac{ww^T}{w^T w} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_0 \\ b_0 \end{pmatrix} - \nu_0 e_1 = \begin{pmatrix} \alpha_0 - \nu_0 \\ b_0 \end{pmatrix}.$$

Then, there is a matrix $A_1 \in \mathbb{R}^{(m-1)\times(n-1)}$ and a vector $a_1 \in \mathbb{R}^{n-1}$ such that

$$H_0 A_0 P_0 = \begin{bmatrix} \nu_0 & a_1^T \\ 0 & A_1 \end{bmatrix}$$

If $\tau = 1$ or $A_1 = 0$, we are done; otherwise, repeat the process on the matrix A_1. Decompose A_1 as

$$A_1 = \begin{bmatrix} \alpha_1 & a_1^T \\ b_1 & \tilde{A}_1 \end{bmatrix}, \text{ with } \tilde{A}_1 \in \mathbb{R}^{(m-2)\times(n-2)},$$

and set

$$\nu_1 = \left\| \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \right\|_2 .$$

If $\nu_1 = 0$, then multiply A_1 on the left by a permutation matrix P_1 to bring a non-zero (largest magnitude) column in A_1 into the first column and the zero column to the last column. Then block decompose A_1P_1 as above with

$$A_1P_1 = \begin{bmatrix} \alpha_1 & a_1^T \\ b_1 & \tilde{A}_1 \end{bmatrix}, \text{ with } \tilde{A}_1 \in \mathbb{R}^{(m-2)\times(n-2)},$$

and set

$$\nu_1 = \left\| \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \right\|_2 \neq 0.$$

Let H_1 be the Householder transformation that maps

$$\begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} \mapsto \nu_1 \, e_1 \quad :$$

$$H_1 = I - 2\frac{ww^T}{w^Tw} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_1 \\ b_1 \end{pmatrix} - \nu_1 e_1 = \begin{pmatrix} \alpha_1 - \nu_1 \\ b_1 \end{pmatrix}.$$

Then, there is a matrix $A_2 \in \mathbb{R}^{(m-2)\times(n-2)}$ and a vector $a_2 \in \mathbb{R}^{n-1}$ such that

$$H_1A_1P_1 = \begin{bmatrix} \nu_2 & a_2^T \\ 0 & A_2 \end{bmatrix}.$$

Consequently,

$$\begin{aligned} \begin{bmatrix} 1 & 0 \\ 0 & H_1 \end{bmatrix} H_0A_0P_0 \begin{bmatrix} 1 & 0 \\ 0 & P_1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & H_1 \end{bmatrix} \begin{bmatrix} \nu_0 & a_1^T \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & P_1 \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_1^T \\ 0 & H_1A_1P_1 \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_1^T \\ 0 & \begin{bmatrix} \nu_2 & a_2^T \\ 0 & A_2 \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} \nu_0 & a_{12} & \tilde{a}_1^T \\ 0 & \nu_2 & a_2^T \\ 0 & 0 & A_2 \end{bmatrix}. \end{aligned}$$

If $\tau = 2$ or $A_2 = 0$, we are done; otherwise repeat as above on the matrix A_2. This process terminates after $\kappa \leq \tau$ iterations with an upper triangular factorization of the form

$$\widetilde{H}_\kappa\widetilde{H}_{\kappa-1}\dots\widetilde{H}_0A_0P_0\widetilde{P}_1\dots\widetilde{P}_\kappa = \begin{bmatrix} \nu_1 & a_{12} & a_{13} & a_{14} & \cdots & a_{1\kappa} & \cdots & a_{1n} \\ 0 & \nu_2 & a_{23} & a_{24} & \cdots & a_{2\kappa} & \cdots & a_{2n} \\ 0 & 0 & \nu_3 & a_{34} & \cdots & a_{3\kappa} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \nu_\kappa & \cdots & a_{\kappa n} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} = R,$$

where the zeros below the κ row are absent if $m = \kappa$. Then $P := P_0\widetilde{P}_1\dots\widetilde{P}_\kappa$ is a permutation matrix and $Q := \widetilde{H}_\kappa\widetilde{H}_{\kappa-1}\dots\widetilde{H}_0$ is unitary with $A_0P = QR$. The matrix A_0 is surjective if and only if $\kappa = m$ in which case $P = I$. On the other hand, A_0 is injective if and only if $\kappa = n$ and again $P = I$.

If the above method is implemented by always permuting the column of greatest magnitude into the current pivot column, then

$$AP = QR$$

gives a QR-factorization with the diagonal entries of R nonnegative and listed in the order of descending magnitude, i.e. $\nu_1 \geq \nu_2 \cdots \geq \nu_\kappa > 0$. Since Q is unitary, this is the full QR factorization in (2.25).

The numerical stability of the procedure can be improved with a slight change to the Householder transformations at each step. Let H_s be the Householder transformation used at iteration s. Redefine H_s so that it maps

$$\begin{pmatrix} \alpha_s \\ b_s \end{pmatrix} \mapsto -\text{sign}(\alpha_s)\nu_s \; e_1 \quad :$$

$$H_s = I - 2\frac{ww^T}{w^T w} \qquad \text{where} \qquad w = \begin{pmatrix} \alpha_s \\ b_s \end{pmatrix} + \text{sign}(\alpha_s)\nu_1 e_1 = \begin{pmatrix} \alpha_s + \text{sign}(\alpha_s)\nu_s \\ b_s \end{pmatrix},$$

where

$$\text{sign}(\alpha) := \begin{cases} 1 & , \; \alpha \geq 0 \\ -1 & , \; \alpha < 0 \\ 0. & \end{cases}$$

Note that this avoids the possibility of subtraction by nearly like terms and increases the magnitude of the vector w. The monotonicity and non-negativity of the ν_s's can be recovered on termination by redefining $Q := QD$ and $R := DR$, where $D := \text{diag}(\text{sign}(\nu_1), \text{sign}(\nu_2), \cdots, \text{sign}(\nu_\kappa), 1, \cdots, 1) \in \mathbb{R}^{m \times m}$ is a diagonal unitary matrix since $D^T D = I$.

CHAPTER 3

Optimization of Quadratic Functions

In this chapter we study the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}x^THx + g^Tx + \beta, \tag{3.1}$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric, $g \in \mathbb{R}^n$, and $\beta \in \mathbb{R}$. It has already been observed that we may as well assume that H is symmetric since

$$x^THx = \tfrac{1}{2}x^THx + \tfrac{1}{2}(x^THx)^T = x^T\left[\tfrac{1}{2}(H + H^T)\right]x,$$

where $\frac{1}{2}(H + H^T)$ is called the *symmetric part* of H. Therefore, in this chapter we assume that H is symmetric. In addition, we have also noted that an objective function can always be shifted by a constant value without changing the solution set to the optimization problem. Therefore, we assume that $\beta = 0$ for most of our discussion. However, just as in the case of integration theory where it is often helpful to choose a particular constant of integration, in many applications there is a "natural" choice for β that helps one interpret the problem as well as its solution.

The class of problems (3.1) is important for many reasons. Perhaps the most common instance of this problem is the linear least squares problem:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}\left\|Ax - b\right\|_2^2, \tag{3.2}$$

where $A \in \mathbb{R}^{m\times n},$ and $b \in \mathbb{R}^m$. By expanding the objective function in (3.2), we see that

$$\tfrac{1}{2}\left\|Ax - b\right\|_2^2 = \tfrac{1}{2}x^T(A^TA)x - (A^Tb)^Tx + \tfrac{1}{2}\left\|b\right\|_2^2 = \tfrac{1}{2}x^THx + g^Tx + \beta, \tag{3.3}$$

where $H = A^TA$, $g = -A^Tb$, and $\beta = \frac{1}{2}\left\|b\right\|_2^2$. This connection to the linear least squares problem will be explored in detail later in this chapter. For the moment, we continue to exam the general problem (3.1). As in the case of the linear least squares problem, we begin by discussing characterizations of the solutions as well as their existence and uniqueness. In this discussion we try to follow the approach taken for the the linear least squares problem. However, in the case of (3.2), the matrix $H := A^TA$ and the vector $g = -A^Tb$ possess special features that allowed us to establish very strong results on optimality conditions as well as on the existence and uniqueness of solutions. In the case of a general symmetric matrix H and vector g it is possible to obtain similar results, but there are some twists. Symmetric matrices have many special properties that can be exploited to help us achieve our goal. Therefore, we begin by recalling a few of these properties, specifically those related to eigenvalue decomposition.

1. Optimality Properties of Quadratic Functions

Recall that for the linear least squares problem, we were able to establish a necessary and sufficient condition for optimality, namely the normal equations, by working backward from a known solution. We now try to apply this same approach to quadratic functions, in particular, we try to extend the derivation in (2.15) to the objective function in (3.4). Suppose $\overline{x}$ is a local solution to the quadratic optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \tfrac{1}{2}x^THx + g^Tx, \tag{3.4}$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric and $g \in \mathbb{R}^n$, i.e., there is an $\epsilon > 0$ such that

$$\tfrac{1}{2}\overline{x}^T H\overline{x} + g^T\overline{x} \le \tfrac{1}{2}x^T Hx + g^T x \quad \forall\, x \in \overline{x} + \epsilon\mathbb{B}_2, \tag{3.5}$$

where $\overline{x} + \epsilon\mathbb{B}_2 := \{\overline{x} + \epsilon u \,|\, u \in \mathbb{B}_2\}$ and $\mathbb{B}_2 := \{u \,|\, \|u\|_2 \le 1\}$ (hence, $\overline{x} + \epsilon\mathbb{B}_2 = \{x \,|\, \|\overline{x} - x\|_2 \le \epsilon\}$). Note that, for all $x \in \mathbb{R}^n$,

$$\begin{aligned}
\overline{x}^T H\overline{x} &= (x + (\overline{x} - x))^T H(x + (\overline{x} - x)) \\
&= x^T Hx + 2x^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2(\overline{x} + (x - \overline{x}))^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2\overline{x}^T H(\overline{x} - x) + 2(x - \overline{x})^T H(\overline{x} - x) + (\overline{x} - x)^T H(\overline{x} - x) \\
&= x^T Hx + 2\overline{x}^T H(\overline{x} - x) - (\overline{x} - x)^T H(\overline{x} - x).
\end{aligned} \tag{3.6}$$

Therefore, for all $x \in \overline{x} + \epsilon\mathbb{B}_2$,

$$\begin{aligned}
\tfrac{1}{2}\overline{x}^T H\overline{x} + g^T\overline{x} &= (\tfrac{1}{2}x^T Hx + g^T x) + (H\overline{x} + g)^T(\overline{x} - x) - \tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x) \\
&\ge (\tfrac{1}{2}\overline{x}^T H\overline{x} + g^T\overline{x}) + (H\overline{x} + g)^T(\overline{x} - x) - \tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x)\ , \qquad \text{(since } \overline{x} \text{ is a local solution)}
\end{aligned}$$

and so

$$\tfrac{1}{2}(\overline{x} - x)^T H(\overline{x} - x) \ge (H\overline{x} + g)^T(\overline{x} - x) \qquad \forall\, x \in \overline{x} + \epsilon\mathbb{B}_2. \tag{3.7}$$

Let $0 \le t \le \epsilon$ and $v \in \mathbb{B}_2$ and define $x = \overline{x} + tv \in \overline{x} + \epsilon\mathbb{B}_2$. If we plug $x = \overline{x} + tv$ into (3.7), then

$$\frac{t^2}{2}v^T Hv \ge -t(H\overline{x} + g)^T v. \tag{3.8}$$

Dividing this expression by $t > 0$ and taking the limit as $t \downarrow 0$ tells us that

$$0 \le (H\overline{x} + g)^T v \quad \forall\ v \in \mathbb{B}_2\ ,$$

which implies that $H\overline{x} + g = 0$. Plugging this information back into (3.8) gives

$$\frac{t^2}{2}v^T Hv \ge 0 \quad \forall\, v \in \mathbb{B}_2\ .$$

Dividing by $t^2/2$ for $t \ne 0$ tells us that

$$v^T Hv \ge 0 \quad \forall\, v \in \mathbb{B}_2$$

or equivalently, that $v^T Hv \ge 0 \quad \forall\, v \in \mathbb{R}^n$. This latter condition on the matrix H plays an important role in optimization theory and practice. We codify this property in the following definition.

DEFINITION 3.1. *Let $H \in \mathbb{R}^{n\times n}$.*

(1) H is said to be positive definite if $x^T Hx > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
(2) H is said to be positive semi-definite if $x^T Hx \ge 0$ for all $x \in \mathbb{R}^n$.
(3) H is said to be negative definite if $x^T Hx < 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.
(4) H is said to be positive semi-definite if $x^T Hx \le 0$ for all $x \in \mathbb{R}^n$.
(5) H is said to be indefinite if H is none of the above.

In addition, if $\mathcal{S}^n$ is the linear space of real $n \times n$ symmetric matrices, we denote the set of real symmetric positive semi-definite and positive definite matrices by $\mathcal{S}^n_+$ and $\mathcal{S}^n_{++}$, respectively. Similarly, we denote by $\mathcal{S}^n_-$ and $\mathcal{S}^n_{--}$, the sets of real symmetric negative semi-definite and negative definite matrices, respectively.

These observations motivate the following theorem.

THEOREM 3.2. *[Existence and Uniqueness in Quadratic Optimization] Let $H \in \mathbb{R}^{n\times n}$ and $g \in \mathbb{R}^n$ be as in* (3.4).

(1) A local solution to the problem (3.4) *exists if and only if $H \in \mathcal{S}^n_+$ and there exists a solution $\overline{x}$ to the equation $Hx + g = 0$ in which case $\overline{x}$ is a local solution to* (3.4).

(2) If $\overline{x}$ is a local solution to (3.4)*, then it is a global solution to* (3.4)*.*
(3) The problem (3.4) *has a unique global solution if and only if H is positive definite in which case this solution is given by $\overline{x} = -H^{-1}g$.*
(4) If either H is not positive semi-definite or there is no solution to the equation $Hx + g = 0$ (or both), then

$$-\infty = \inf_{x\in\mathbb{R}^n} \tfrac{1}{2}x^THx + g^Tx \ .$$

PROOF. (1) We have already shown that if a local solution $\overline{x}$ to (3.4) exists, then $H\overline{x} + g = 0$ and H is positive semi-definite. On the other hand, suppose that H is positive semi-definite and $\overline{x}$ is a solution to $Hx + g = 0$. Then, for all $x \in \mathbb{R}^n$, we can interchange the roles of x and $\overline{x}$ in the second line of (3.6) to obtain

$$x^THx = \overline{x}^TH\overline{x} + 2\overline{x}^TH(x - \overline{x}) + (x - \overline{x})^TH(x - \overline{x}). \tag{3.9}$$

Hence, for all $x \in \mathbb{R}^n$,

$$\tfrac{1}{2}x^THx + g^Tx = \tfrac{1}{2}\overline{x}^TH\overline{x} + g^T\overline{x} + (H\overline{x} + g)^T(x - \overline{x}) + \tfrac{1}{2}(x - \overline{x})^TH(x - \overline{x}) \geq \tfrac{1}{2}\overline{x}^TH\overline{x} + g^T\overline{x} \ ,$$

since $H\overline{x}+g = 0$ and H is positive semi-definite. That is, $\overline{x}$ is a global solution to (3.4) and hence a local solution.

(2) The proof of part (1) shows that if H is positive semi-definite and $H\overline{x} + g = 0$, then $\overline{x}$ is a global solution to (3.4). Hence the result follows from part (1).

(3) If (3.4) has a unique global solution $\overline{x}$, then $\overline{x}$ must be the unique solution to the equation $Hx+g = 0$. This can only happen if H is invertible. Hence, H is invertible and positive semi-definite which implies that H is positive definite. On the other hand, if H is positive definite, then, in particular, it is positive semi-definite and there is a unique solution to the equation $Hx + g = 0$, i.e., (3.6) has a unique global solution.

(4) The result follows if we can show that $f(x) := \frac{1}{2}x^THx+g^Tx$ is unbounded below when either H is not positive semi-definite or there is no solution to the equation $Hx + g = 0$ (or both). Let us first suppose that H is not positive semi-definite, or equivalently, there is a $\hat{x} \in \mathbb{R}^n$ such that $\hat{x}^TH\hat{x} < 0$. Then, for every $t \in \mathbb{R}$, $f(t\hat{x}) = \frac{t^2}{2}\hat{x}^TH\hat{x} + tg^t\hat{x}$. Therefore, $f(t\hat{x}) \downarrow -\infty$ as $t \uparrow +\infty$, and so f is unbounded below.

Next assume that H is positive semi-definite but $g \notin \mathrm{Ran}\,(H)$, i.e. there is no solution to $Hx + g = 0$. Let P be the orthogonal projection onto $\mathrm{Ran}\,(H)$ so that $I - P$ is the orthogonal projection onto $\mathrm{Ran}\,(H)^\perp = \mathrm{Nul}\left(H^T\right) = \mathrm{Nul}\,(H)$. Set $g_1 := Pg$ and $g_2 := (I - P)g$ so that $g = g_1 + g_2$. Since $g \notin \mathrm{Ran}\,(H)$, $g_2 \neq 0$. Then, for every $t \in \mathbb{R}$, $f(tg_2) = \frac{t^2}{2}g_2^THg_2 + tg^Tg_2 = t\,\|g_2\|_2^2$. Therefore, $f(tg_2) \downarrow -\infty$ as $t \downarrow -\infty$, and so again f is unbounded below. □

The identity (3.9) is a very powerful tool in the analysis of quadratic functions. It was the key tool in showing that every local solution to (3.4) is necessarily a global solution. We now show how these results can be extended to problems with linear equality constraints.

2. Minimization of a Quadratic Function on an Affine Set

In this section we consider the problem

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}x^THx + g^Tx \\ &\text{subject to } Ax = b, \end{aligned} \tag{3.10}$$

where $H \in \mathbb{R}^{n\times n}$ is symmetric, $g \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. We assume that the system $Ax = b$ is consistent. That is, there exists $\hat{x} \in \mathbb{R}^n$ such that $A\hat{x} = b$ in which case

$$\{x \,|\, Ax = b\} = \hat{x} + \mathrm{Null}(A).$$

Consequently, the problem (3.10) is of the form

$$\underset{x\in\hat{x}+S}{\text{minimize}}\ \tfrac{1}{2}x^THx + g^Tx \ , \tag{3.11}$$

where S is a subspace of $\mathbb{R}^n$. This representation of the problem shows that the problem (3.10) is trivial if $\text{Null}(A) = \{0\}$ since then the unique solution $\hat{x}$ to $Ax = b$ is also the unique solution to (3.10) since it is the only feasible point. Hence, when considering the problem (3.10) we assumed that $\text{Null}(A) \neq \{0\}$, and, more specifically, that $m < n$.

DEFINITION 3.3. *[Affine Sets] A subset K of $\mathbb{R}^n$ is said to be affine if there exists a point $\hat{x} \in \mathbb{R}^n$ and a subspace $S \subset \mathbb{R}^n$ such that $K = \hat{x} + S = \{\hat{x} + u \,|\, u \in S\}$.*

We now develop necessary and sufficient optimality conditions for the problem (3.11), that is, for the minimization of a quadratic function over an affine set. Assume that $v^1, v^2, \ldots, v^k$ is a basis for S where $k = \dim(S)$. Let $V \in \mathbb{R}^{n\times k}$ be the matrix whose columns are the vectors $v^1, v^2, \ldots, v^k$ giving $S = \text{Ran}(V)$. Then $\hat{x} + S = \left\{\hat{x} + Vz \,\middle|\, z \in \mathbb{R}^k\right\}$. This allows us to rewrite the problem (3.11) as

$$\underset{z\in\mathbb{R}^k}{\text{minimize}}\ \tfrac{1}{2}(\hat{x} + Vz)^T H(\hat{x} + Vz) + g^T(\hat{x} + Vz)\ . \tag{3.12}$$

By expanding the objective function in (3.12), we obtain

$$\tfrac{1}{2}(\hat{x} + Vz)^T H(\hat{x} + Vz) + g^T(\hat{x} + Vz) = \tfrac{1}{2}z^T V^T HVz + (V^T(H\hat{x} + g))^T z + f(\hat{x}),$$

where $f(x) := \frac{1}{2}x^T Hx + g^T x$. If we now set $\widehat{H} := V^T HV$, $\hat{g} := V^T(H\hat{x} + g)$, and $\beta := f(\hat{x})$, then problem (3.12) has the form of (3.1):

$$\underset{z\in\mathbb{R}^k}{\text{minimize}}\ \tfrac{1}{2}z^T\widehat{H}z + \hat{g}^T z\ , \tag{3.13}$$

where, as usual, we have dropped the constant term $\beta = f(\hat{x})$.

PROPOSITION 3.4. *Consider the two problems* (3.11) *and* (3.12)*, where the columns of the matrix V form a basis for the subspace S. The set of optimal solution to these problems are related as follows:*

$$\{\overline{x} \,|\, \overline{x} \text{ solves } (3.11)\} = \{\hat{x} + V\overline{z} \,|\, \overline{z} \text{ solves } (3.13)\}\,.$$

This relationship allows us to obtain necessary and sufficient conditions for optimality for problem for the problem (3.11) by applying Theorem 3.2 to (3.13).

THEOREM 3.5. *[Optimization of Quadratics on Affine Sets]*

1. *The point $\overline{x} \in \hat{x} + S$ is a local solution to* (3.11) *if and only if $u^T Hu \geq 0$ for all $u \in S$ and $H\overline{x} + g \in S^\perp$.*
2. *A point $\overline{x}$ is a local solution to* (3.11) *if and only if it is a global solution to* (3.11)*.*
3. *The problem* (3.11) *has a unique global solution if and only if $u^T Hu > 0$ for all $u \in S \setminus \{0\}$.*
4. *If $V \in \mathbb{R}^{n\times k}$ is any matrix such that Ran$(V) = S$ where $k = \dim(S)$, then* (3.11) *has a unique global solution if and only if the matrix $V^T HV$ is positive definite in which case the unique global solution $\overline{x}$ is given by*

$$\overline{x} = [I - V(V^T HV)^{-1}V^T H]\hat{x} - V(V^T HV)^{-1}V^T g\ .$$

5. *If either there exists $\overline{u} \in S$ such that $\overline{u}^T H\overline{u} < 0$ or there does not exist $x \in \hat{x} + S$ such that $Hx + g \in S^\perp$ (or both), then*

$$-\infty = \inf_{x\in\hat{x}+S} \tfrac{1}{2}x^T Hx + g^T x\ .$$

PROOF. (1) By Proposition 3.4, a solution to (3.11) exists if and only if a solution to (3.12) exists. By Theorem 3.2, a solution to (3.12) exists if and only if $V^T HV$ is positive semi-definite and there is a solution $\overline{z}$ to the equation

$$0 = \widehat{H}z + \hat{g} = V^T HVz + V^T(H\hat{x} + g) = V^T(H(\hat{x} + Vz) + g) = V^T(Hx + g),$$

where $x := \hat{x} + Vz$, in which case $\overline{z}$ solves (3.12) and, by Proposition 3.4, $\overline{x} := \hat{x} + V\overline{z}$ solves (3.11). The condition that $V^T HV$ is positive semi-definite is equivalent to the statement that $z^T V^T HVz \geq 0$

for all $z \in \mathbb{R}^k$, or equivalently, $u^T H u \geq 0$ for all $u \in S$. The condition, $V^T(H\overline{x} + g) = 0$ is equivalent to $H\overline{x} + g \in \text{Null}(V^T) = \text{Ran}(V)^\perp = S^\perp$.

(2) This is an immediate consequence of Proposition 3.4 and Part (2) of Theorem 3.2.

(3) By Theorem 3.2, the problem (3.12) has a unique solution if and only if $V^T H V$ is positive definite. The result follows by observing that $V^T H V$ is positive definite if and only if $u^T H u > 0$ for all $u \in S \setminus \{0\}$.

(4) Again, by Theorem 3.2, the problem (3.12) has a unique solution if and only if $V^T H V$ is positive definite with the unique solution given by $\overline{z} = -(V^T H V)^{-1} V^T (H\hat{x} + g)$. Consequently, by Proposition 3.4, the unique solution to (3.11) is

$$\overline{x} = \hat{x} + V\overline{z} = \hat{x} - V(V^T H V)^{-1} V^T (H\hat{x} + g) = [I - V(V^T H V)^{-1} V^T H]\hat{x} - V(V^T H V)^{-1} V^T g.$$

(4) This is established in Problem Set 4 where you will show that the result follows by following the pattern of proof used in Part (4) of Theorem 3.2. □

Theorem 3.5 and the equivalence between problems (3.10) and (3.11) yield the main result of this section.

THEOREM 3.6. *[Optimization of Quadratics Subject to Linear Equality Constraints]*

(1) A local solution to the problem (3.10) *exists if and only if* $u^T H u \geq 0$ *for all* $u \in$ *Null*(A) *and there exists a vector pair* $(\overline{x}, \overline{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ *such that* $H\overline{x} + A^T\overline{y} + g = 0$*, in which case* $\overline{x}$ *is a local solution to* (3.11)*.*

(2) If $\overline{x}$ *is a local solution to* (3.11)*, then it is a global solution.*

(3) The problem (3.11) *has a unique global solution if and only if* $u^T H u > 0$ *for all* $u \in$ *Null*$(A) \setminus \{0\}$*.*

(4) If $u^T H u > 0$ *for all* $u \in$ *Null*$(A) \setminus \{0\}$ *and* $\text{rank}\,(A) = m$*, the matrix*

$$M := \begin{bmatrix} H & A^T \\ A & 0 \end{bmatrix} \quad \textit{is invertible, and the vector} \quad \begin{bmatrix} \overline{x} \\ \overline{y} \end{bmatrix} = M^{-1} \begin{bmatrix} -g \\ b \end{bmatrix}$$

has $\overline{x}$ *as the unique global solution to* (3.11)*.*

(5) If either there exists $\overline{u} \in$ *Null*(A) *such that* $\overline{u}^T H \overline{u} < 0$ *or there does not exist a vector pair* $(\overline{x}, \overline{y}) \in \mathbb{R}^n \times \mathbb{R}^m$ *such that* $H\overline{x} + A^T\overline{y} + g = 0$ *(or both), then*

$$-\infty = \inf_{x \in \hat{x} + S} \tfrac{1}{2} x^T H x + g^T x \ .$$

REMARK 3.7. *The condition that* $\text{rank}\,(A) = m$ *in Part (4) of the theorem can always be satisfied by replacing* A *by first row reducing* A *to echelon form.*

PROOF. (1) Recall that $\text{Null}(A)^\perp = \text{Ran}(A^T)$. Hence, $w \in \text{Null}(A)$ if and only if there exists $y \in \mathbb{R}^m$ such that $w = A^T y$. By setting $w = H\overline{x} + g$ the result follows from Part (1) of Theorem 3.5.

(2) Again, this is an immediate consequence of Proposition 3.4 and Part (2) of Theorem 3.2.

(3) This is just Part (3) of Theorem 3.5.

(4) Suppose $M \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, then $Hx + A^T y = 0$ and $Ax = 0$. If we multiply $Hx + A^T y$ on the left by x^T, we obtain $0 = x^T H x + x^T A^T y = x^T H x$ which implies that $x = 0$ since $x \in \text{Null}(A)$. But then $A^T y = 0$, so that $y = 0$ since $\text{rank}\,(A) = m$. Consequently, $\text{Null}(M) = \{0\}$, i.e., M is invertible. The result now follows from Part (1).

(5) By Part (1), this is just a restatement of Theorem 3.5 Part (4). □

The vector $\overline{y}$ appearing in this Theorem is call a *Lagrange multiplier* vector. Lagrange multiplier vectors play an essential role in constrained optimization and lie at the heart of what is called *duality theory*. This theory is more fully developed in Chapter **??**.

We now study how one might check when H is positive semi-definite as well as solving the equation $Hx + g = 0$ when H is positive semi-definite.

3. The Principal Minor Test for Positive Definiteness

Let $H \in \mathcal{S}^n$. We wish to obtain a test of when H is positive definite. First note that $H_{ii} = e_i^T H e_i$, so that H can be positive definite only if $H_{ii} > 0,\ i = 1, \ldots, n$. This is only a "sanity check" for whether a matrix is positive definite. That is, if any diagonal element of H is not positive, then H cannot be positive definite. In this section we develop a necessary and sufficient condition for H to be positive definite based on the determinant. We begin with the following lemma.

LEMMA 3.8. *Let $H \in \mathcal{S}^n$, $u \in \mathbb{R}^n$, and $\alpha \in \mathbb{R}$, and consider the block matrix*

$$\hat{H} := \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} \in \mathcal{S}^{(n+1)}.$$

(1) The matrix $\hat{H}$ is positive semi-definite if and only if H is positive semi-definite and there exists a vector $z \in \mathbb{R}^n$ such that $u = Hz$ and $\alpha \geq z^T H z$.
(2) The matrix $\hat{H}$ is positive definite if and only if H is positive definite and $\alpha > u^T H^{-1} u$.

PROOF. (1) Suppose H is positive semi-definite, and there exists z such that $u = Hz$ and $\alpha \geq z^T H z$. Then for any $\hat{x} = \begin{bmatrix} x \\ x_n \end{bmatrix}$ where $x_n \in \mathbb{R}$ and $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \hat{x}^T \hat{H} \hat{x} &= x^T H x + 2x^T H x_n z + x_n^2 \alpha \\ &= (x + x_n z)^T H (x + x_n z) + x_n^2 (\alpha - z^T H z) \geq 0. \end{aligned}$$

Hence, $\hat{H}$ is positive semi-definite.

Conversely, suppose that $\hat{H}$ is positive semi-definite. Write $u = u_1 + u_2$ where $u_1 \in \mathrm{Ran}(H)$ and $u_2 \in \mathrm{Ran}(H)^\perp = \mathrm{Null}(H)$, so that there is a $z \in \mathbb{R}^n$ such that $u_1 = Hz$. Then, for all $\hat{x} = \begin{pmatrix} x \\ x_n \end{pmatrix} \in \mathbb{R}^{(n+1)}$,

$$\begin{aligned} 0 \leq \hat{x}^T \hat{H} \hat{x} &= x^T H x + 2x_n u^T x + \alpha x_n^2 \\ &= x^T H x + 2x_n (u_1 + u_2)^T x + \alpha x_n^2 \\ &= x^T H x + 2x_n z^T H x + x_n^2 z^T H z + x_n^2 (\alpha - z^T H z) + 2x_n u_2^T x \\ &= (x + x_n z)^T H (x + x_n z) + x_n^2 (\alpha - z^T H z) + 2x_n u_2^T x. \end{aligned}$$

Taking $x_n = 0$ tells us that H is positive semi-definite, and taking $\hat{x} = \begin{pmatrix} -tu_2 \\ 1 \end{pmatrix}$ for $t \in \mathbb{R}$ gives

$$\alpha - 2t \|u_2\|_2^2 \geq 0 \quad \text{for all } t \in \mathbb{R},$$

which implies that $u_2 = 0$. Finally, taking $\hat{x} = \begin{pmatrix} -z \\ 1 \end{pmatrix}$, tells us that $z^T H z \leq \alpha$ which proves the result.

(2) The proof follows the pattern of Part (1) but now we can take $z = H^{-1} u$. □

If the matrix H is invertible, we can apply a block Gaussian elimination to the matrix $\hat{H}$ in the lemma to obtain a matrix with block upper triangular structure:

$$\begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix} \begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} = \begin{bmatrix} H & u \\ 0 & (\alpha - u^T H^{-1} u) \end{bmatrix}.$$

This factorization tells us that

$$\begin{aligned}\det\begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} &= \det\begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix}\det\begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix} \\ &= \det\left(\begin{bmatrix} I & 0 \\ (-H^{-1}u)^T & 1 \end{bmatrix}\begin{bmatrix} H & u \\ u^T & \alpha \end{bmatrix}\right) \\ &= \det\begin{bmatrix} H & u \\ 0 & (\alpha - u^T H^{-1}u) \end{bmatrix} \\ &= \det(H)(\alpha - u^T H^{-1}u).\end{aligned} \tag{3.14}$$

We use this determinant identity in conjunction with the previous lemma to establish a test for whether a matrix is positive definite based on determinants. The test requires us to introduce the following elementary definition.

DEFINITION 3.9. *[Principal Minors] The kth principal minor of a matrix $B \in \mathbb{R}^{n\times n}$ is the determinant of the upper left–hand corner $k\times k$–submatrix of B for $1 \le k \le n$.*

PROPOSITION 3.10. *[The Principal Minor Test] Let $H \in \mathcal{S}^n$. Then H is positive definite if and only if each of its principal minors is positive.*

PROOF. The proof proceeds by induction on the dimension n of H. The result is clearly true for $n = 1$. We now assume the result is true for $1 \le k \le n$ and show it is true for dimension $n + 1$. Write

$$H := \begin{bmatrix} \hat{H} & u \\ u^T & \alpha \end{bmatrix}.$$

Then Lemma 3.8 tells us that H is positive definite if and only if $\hat{H}$ is positive definite and $\alpha > u^T\hat{H}^{-1}u$. By the induction hypothesis, $\hat{H}$ is positive definite if and only if all of its principal minors are positive. If we now combine this with the expression (3.14), we get that H is positive definite if and only if all principal minors of $\hat{H}$ are positive and, by (3.14), $\det(H) = \det(\hat{H})(\alpha - u^T\hat{H}^{-1}u) > 0$, or equivalently, all principal minors of H are positive. □

This result only applies to positive definite matrices, and does not provide insight into how to solve linear equations involving H such as $Hx+g = 0$. These two issues can be addressed through the Cholesky factorization.

4. The Cholesky Factorizations

We now consider how one might solve a quadratic optimization problem. Recall that a solution only exists when H is positive semi-definite and there is a solution to the equation $Hx + g = 0$. Let us first consider solving the equation when H is positive definite. We use a procedure similar to the LU factorization but which also takes advantage of symmetry.

Suppose

$$H = \begin{bmatrix} \alpha_1 & h_1^T \\ h_1 & \widetilde{H}_1 \end{bmatrix}, \quad \text{where} \quad \widetilde{H}_1 \in \mathcal{S}^n.$$

Note that $\alpha_1 = e_1^T H e_1 > 0$ since H is positive definite (if $\alpha_1 \le 0$, then H cannot be positive definite), so there is no need to apply a permutation. Multiply H on the left by the Gaussian elimination matrix for the first column, we obtain

$$L_1^{-1}H = \begin{bmatrix} 1 & 0 \\ -\frac{h_1}{\alpha_1} & I \end{bmatrix}\begin{bmatrix} \alpha_1 & h_1^T \\ h_1 & \widetilde{H}_1 \end{bmatrix} = \begin{bmatrix} \alpha_1 & h_1^T \\ 0 & \widetilde{H}_1 - \alpha_1^{-1}h_1h_1^T \end{bmatrix}.$$

By symmetry, we have

$$L_1^{-1}HL_1^{-T} = \begin{bmatrix} \alpha_1 & h_1^T \\ 0 & \widetilde{H}_1 - \alpha_1^{-1}h_1h_1^T \end{bmatrix}\begin{bmatrix} 1 & -\frac{h_1^T}{\alpha_1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \widetilde{H}_1 - \alpha_1^{-1}h_1h_1^T \end{bmatrix}.$$

Set $H_1 = \widetilde{H}_1 - \alpha^{-1} h_1 h_1^T$. Observe that for every non-zero vector $v \in \mathbb{R}^{(n-1)}$,

$$v^T H_1 v = \begin{pmatrix} 0 \\ v \end{pmatrix}^T \begin{bmatrix} \alpha_1 & 0 \\ 0 & H_1 \end{bmatrix} \begin{pmatrix} 0 \\ v \end{pmatrix} = \left(L_1^{-T} \begin{pmatrix} 0 \\ v \end{pmatrix} \right)^T H \left(L_1^{-T} \begin{pmatrix} 0 \\ v \end{pmatrix} \right) > 0,$$

which shows that H_1 is positive definite. Decomposing H_1 as we did H gives

$$H_1 = \begin{bmatrix} \alpha_2 & h_2^T \\ h_2 & \widetilde{H}_2 \end{bmatrix}, \quad \text{where} \quad \widetilde{H}_2 \in \mathcal{S}^{(n-1)}.$$

Again, $\alpha_2 > 0$ since H_1 is positive definite (if $\alpha_2 \leq 0$, then H cannot be positive definite). Hence, we can repeat the reduction process for H_1. Continuing in this way, if at any stage we discover and $\alpha_i \leq 0$, then we terminate, since H cannot be positive definite.

If we can continue this process n times, we will have constructed a lower triangular matrix

$$L := L_1 L_2 \cdots L_n \quad \text{such that} \quad L^{-1} H L^{-T} = D, \quad \text{where} \quad D := \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$$

is a diagonal matrix with strictly positive diagonal entries. On the other hand, if at some point in the process we discover an α_i that is not positive, then H cannot be positive definite and the process terminates. That is, this computational procedure simultaneously tests whether H is positive definite as it tries to diagonalize H. We will call this process the *Cholesky diagonalization procedure*. It is used to establish the following factorization theorem.

THEOREM 3.11. *[The Cholesky Factorization] Let $H \in \mathcal{S}_+^n$ have rank k. Then there is a lower triangular matrix $L \in \mathbb{R}^{n \times k}$ such that $H = LL^T$. Moreover, if the rank of H is n, then there is a positive diagonal matrix D and a lower triangular matrix $\widetilde{L}$ with ones on its diagonal such that $H = \widetilde{L} D \widetilde{L}^T$.*

PROOF. Let the columns of the matrix $V_1 \in \mathbb{R}^{n \times k}$ be an orthonormal basis for $\operatorname{Ran}(H)$ and the columns of $V_2 \in \mathbb{R}^{n \times (n-k)}$ be an orthonormal basis for $\operatorname{Null}(H)$ and set $V = [V_1 \ V_2] \in \mathbb{R}^{n \times n}$. Then

$$\begin{aligned} V^T H V &= \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} H [V_1 \ V_2] \\ &= \begin{bmatrix} V_1^T H V_1 & V_1^T H V_2 \\ V_2^T H V_1 & V_2^T H V_2 \end{bmatrix} \\ &= \begin{bmatrix} V_1^T H V_1 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Since $\operatorname{Ran}(H) = \operatorname{Null}(H^T)^\perp = \operatorname{Null}(H)^\perp$, $V_1 H V_1^T \in \mathbb{R}^{k \times k}$ is symmetric and positive definite. By applying the procedure described prior to the statement of the theorem, we construct a nonsingular lower triangular matrix $\widetilde{L} \in \mathbb{R}^{k \times k}$ and a diagonal matrix $D = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_k)$, with $\alpha_i > 0$, $i = 1, \ldots, k$, such that $V_1 H V_1^T = \widetilde{L} D \widetilde{L}^T$. Set $\widehat{L} = \widetilde{L} D^{1/2}$ so that $V_1 H V_1^T = \widehat{L} \widehat{L}^T$. If $k = n$, taking $V = I$ proves the theorem by setting $L = \widehat{L}$. If $k < n$,

$$H = [V_1 \ V_2] \begin{bmatrix} \widehat{L}\widehat{L}^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = (V_1 \widehat{L})(V_1 \widehat{L})^T.$$

Let $(V_1 \widehat{L})^T \in \mathbb{R}^{k \times n}$ have reduced QR factorization $(V_1 \widehat{L})^T = QR$ (see Theorem 2.10). Since $\widehat{L}^T$ has rank k, $Q \in \mathbb{R}^{k \times k}$ is unitary and $R = [R_1 \ R_2]$ with $R_1 \in \mathbb{R}^{k \times k}$ nonsingular and $R_2 \in \mathbb{R}^{k \times (n-k)}$. Therefore,

$$H = (V_1 \widehat{L})(V_1 \widehat{L})^T = R^T Q^T Q R = R^T R.$$

The theorem follows by setting $L = R^T$. □

When H is positive definite, the factorization $H = LL^T$ is called the Cholesky factorization of H, and when $\operatorname{rank}(H) < n$ it is called the *generalized Cholesky factorization* of H. In the positive definite case, the Cholesky diagonalization procedure computes the Cholesky factorization of H. On the other hand, when H is only positive semi-definite, the proof of the theorem provides a guide for obtaining the

generalized Cholesky factorization, but it requires a QR-factorization of H to obtain an orthonormal basis of $\mathrm{Ran}\,(H)$.

When H is positive semidefinite, there is an alternative but equally useful factorization of H which does not require a QR-factorization. We call this an *echelon* Cholesky factorization. It is also rank revealing. The factorization is based on Part (1) of Lemma 3.8 as we now illustrate.

Let $L \in \mathbb{R}^{n\times k}$ be such that L^T is in upper echelon form and there is a permutation matrix $P \in \mathbb{R}^{n\times n}$ for which

$$PL = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix}$$

with $L_1 \in \mathbb{R}^{k\times k}$ a nonsingular lower triangular matrix with positive diagonal entries. Consider the matrix

$$H = \begin{bmatrix} LL^T & v \\ v^T & \alpha \end{bmatrix},$$

where $v \in \mathbb{R}^n$ and $\alpha \geq 0$. By Lemma 3.8, H is positive semidefinite if and only if there exists $u \in \mathbb{R}^k$ such that $v = Lu$ and $\alpha \geq u^T u$ (Why?). Set $\hat{v} = Pv$ and define $\hat{v}_1 \in \mathbb{R}^k$ and $\hat{v}_2 \in \mathbb{R}^{n-k}$ so that $\hat{v} = (\hat{v}_1^T,\ \hat{v}_2^T)^T$. Set $\overline{u} = L_1^{-1}\hat{v}_1$ (this is a forward solve on the equation $\hat{v}_1 = L_1 u$). If either $\hat{v}_2 \neq L_2 u$ or $\alpha < u^T u$, then H is not positive definite and the process terminates since either there is no solution to the equation $v = Lu$ or the unique solution satisfies $\alpha < u^T u$. If H is positive semidefinite, then

$$H = \begin{bmatrix} LL^T & v \\ v^T & \alpha \end{bmatrix} = \begin{bmatrix} LL^T & L\overline{u} \\ \overline{u}^T L^T & \alpha \end{bmatrix} = \begin{bmatrix} L & 0 \\ \overline{u}^T & \sqrt{\alpha - \overline{u}^T\overline{u}} \end{bmatrix} \begin{bmatrix} L & 0 \\ \overline{u}^T & \sqrt{\alpha - \overline{u}^T\overline{u}} \end{bmatrix}^T = \widehat{L}\widehat{L}^T,$$

where $\widehat{L}$ is lower triangular with full rank $k+1$ if $\alpha > u^T u$. If $\alpha = \overline{u}^T\overline{u}$, reset

$$\widehat{L} = \begin{bmatrix} L \\ \overline{u}^T \end{bmatrix},$$

so that again $H = \widehat{L}\widehat{L}^T$ with $\widehat{L}$ of full rank k. In this way we obtain a factorization of H of the form $H = \widehat{L}\widehat{L}$, where $\widehat{L} \in \mathbb{R}^{(n+1)\times\hat{k}}$ with $\hat{k} = \mathrm{rank}\left(\widehat{L}\right) = \mathrm{rank}\,(H)$ has the same properties as L, i.e. $\widehat{L}^T$ is in upper echelon form and there is a permutation matrix $\widehat{P} \in \mathbb{R}^{n\times n}$ for which

$$\widehat{P}\widehat{L} = \begin{bmatrix} \widehat{L}_1 \\ \widehat{L}_2 \end{bmatrix}$$

with $\widehat{L}_1 \in \mathbb{R}^{k\times k}$ a nonsingular lower triangular matrix with positive diagonal entries.

This discussion describes the inductive step of an algorithm for computing the generalized Cholesky factorization of any $H \in \mathcal{S}^n$ if the first diagonal entry of H is nonzero and H is positive semidefinite. If H is not positive semidefinite, this is discovered in the inductive step by either observing that $v^2 \neq L_2 u$ is not solvable or observing that $\alpha < \overline{u}^T\overline{u}$, in which case the procedure is terminated.

The wrinkle in this induction occurs if the first diagonal entry in H, namely $(H)_{11}$, is zero. In this case the induction cannot be initialized since the initial L satisfies $LL^T = (H)_{11}$ must have full rank. This is a some what unusual circumstance and it has a significant implication for the structure of H as is seen in the following proposition.

PROPOSITION 3.12 (Zero structure in Positive Semidefinite Matrices). *Let $H \in \mathcal{S}^n$ and $V \in \mathbb{R}^{m\times n}$. Then the block matrix*

$$\widehat{H} := \begin{bmatrix} 0_{m\times m} & V \\ V^T & H \end{bmatrix}$$

is positive semidefinite if and only if $H \in \mathcal{S}^n_+$ and $V = 0$. In particular, if $W \in \mathcal{S}^n$ has zero diagonal, then $W \in \mathcal{S}^n_+$ if and only if $H = 0$.

PROOF. Clearly if $H \in \mathcal{S}^n_+$ and $V = 0$, then $\widehat{H}$ is positive semidefinte, so we only prove the forward implication. If $\widehat{H}$ is positive semidefinite, then, clearly $H \in \mathcal{S}^n_+$; moreover,

$$0 \leq \begin{pmatrix} -tVy \\ y \end{pmatrix}^T \widehat{H} \begin{pmatrix} -tVy \\ y \end{pmatrix} = -2t \|Vy\|_2 + y^T H y \quad \forall\, t \geq 0 \ \text{ and } \ y \in \mathbb{R}^n.$$

Consequently, $Vy = 0$ for all $y \in \mathbb{R}^n$ which proves the result. □

If $(H)_{11} = 1$, a modified Cholesky factorization can be obtained by applying a permutation P that brings any positive diagonal entry of H into the one-one position in PHP^T. By Proposition 3.12, if no such positive diagonal entry exits, then H must be the zero matrix if it to be positive semidefinite in which case the Cholesky factorization is not required.

EXAMPLE 3.13. *Compute the Cholesky factorization of*

$$H = \begin{bmatrix} 4 & 2 & 2 & 2 \\ 2 & 17 & 5 & 9 \\ 2 & 5 & 2 & 3 \\ 2 & 9 & 3 & 6 \end{bmatrix}.$$

Step 1: *Let* $L_0 := 2$ *so that* $L_0 L_0^T = 4$.
Next set $H_0 := \begin{bmatrix} 4 & 2 \\ 2 & 17 \end{bmatrix} = \begin{bmatrix} L_0 L_0^T & 2 \\ 2 & 17 \end{bmatrix}$. *Then* $v_0 = 2$ *and* $\alpha_0 = 17$. *Solve* $2 = v_0 = L_0 u_0 = 2u_0$ *to get* $u_0 = 1$. *Then* $\sqrt{\alpha_0 - u_0^T u_0} = 4$ *which yields* $L_1 = \begin{bmatrix} L_0 & 0 \\ u_0^T & \sqrt{\alpha_0 - u_0^T u_0} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix}$. *Observe that* $L_1 L_1^T = H_0 = \begin{bmatrix} 4 & 2 \\ 2 & 17 \end{bmatrix}$.

Step 2: *Set* $H_1 := \begin{bmatrix} L_1 L_1^T & v_1 \\ v_1^T & \alpha_1 \end{bmatrix}$ *where* L_1 *is given above,* $v_1 = (2,\ 5)^T$, *and* $\alpha_1 = 2$. *Then forward solve* $v_1 = L_1 u_1$ *for* u_1 *to get* $u_1 = (1,\ 1)^T$ *and* $\sqrt{\alpha_1 - u_1^T u_1} = 0$. *Hence*

$$L_2 = \begin{bmatrix} L_1 & 0 \\ u_1^T & \sqrt{\alpha_1 - u_1^T u_1} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Since $\sqrt{\alpha_1 - u_1^T u_1} = 0$, *we replace* L_2 *by*

$$L_2 = \begin{bmatrix} 2 & 0 \\ 1 & 4 \\ 1 & 1 \end{bmatrix} = \left[\begin{array}{cc} 2 & 0 \\ 1 & 4 \\ \hline 1 & 1 \end{array}\right] = \begin{bmatrix} L_{21} \\ L_{22} \end{bmatrix} \quad \text{so that} \quad H_1 = L_2 L_2^T = \begin{bmatrix} 2 & 0 \\ 1 & 4 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 1 & 4 \\ 1 & 1 \end{bmatrix}^T = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 17 & 5 \\ 2 & 5 & 2 \end{bmatrix}$$

is the generalized Cholesky factorization of H_1.

Step 3: *Set* $H_2 := H = \begin{bmatrix} L_2 L_2^T & v_2 \\ v_2^T & \alpha_2 \end{bmatrix}$ *where* L_2 *is given above,* $v_2 = (2,\ 9,\ 3)^T$, *and* $\alpha_2 = 6$. *Set* $v_{21} := (2,\ 9)^T$ *and* $v_{22} = 3$ *so that* $v_2 := \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix}$. *Then forward solve* $v_{21} = L_{21} u_2$ *for* u_2 *to get* $u_2 = (1,\ 2)^T$. *We now verify that* $3 = v_{22} = L_{22} u_2 = 3$ *to establish the consistency of the system* $v_2 = L_2 u_2$ *as required in order to proceed. Next we verify that* $6 = \alpha_2 \geq u_2^T u_2 = 5$, *which holds, and set*

$$L_3 := \begin{bmatrix} L_2 & 0 \\ u_2^T & \sqrt{\alpha_2 - u_2^T u_2} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix}.$$

Then

$$\begin{bmatrix} 4 & 2 & 2 & 2 \\ 2 & 17 & 5 & 9 \\ 2 & 5 & 2 & 3 \\ 2 & 9 & 3 & 6 \end{bmatrix} = H = L_3 L_3^T = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 1 \\ 0 & 4 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Finally, observe that L_3^T is in echelon form and the permutation matrix associated with L_3 is

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

4.1. Computing Solutions to the Quadratic Optimization Problem via Cholesky Factorizations.

Step 1: Apply the procedure described in the previous section for computing the generalized Cholesky factorization of H. If it is determined that H is not positive definite, then proceed no further since the problem (3.1) has no solution and the optimal value is $-\infty$.

Step 2: Step 1 provides us with the generalized Cholesky factorization for $H = LL^T$ with $L^T = [L_1^T \; L_2^T]$, where $L_1 \in \mathbb{R}^{k\times k}$ and $L_2 \in \mathbb{R}^{(n-k)\times k}$ with $k = \operatorname{rank}(H)$. Write

$$g = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix},$$

where $g_1 \in \mathbb{R}^k$ and $g_2 \in \mathbb{R}^{(n-k)}$. Since $\operatorname{Ran}(H) = \operatorname{Ran}(L)$, the system $Hx + g = 0$ is solvable if and only if $-g \in \operatorname{Ran}(L)$. That is, there exists $w \in \mathbb{R}^k$ such that $Lw = -g$, or equivalently,

$$L_1 w = -g_1 \quad \text{and} \quad L_2 w = -g_2.$$

Since L_1 is invertible, the system $L_1 w = -g_1$ has as its unique solution $\overline{w} = L_1^{-1} g_1$. Note that $\overline{w}$ is easy to compute by forward substitution since L_1 is lower triangular. Having $\overline{w}$ check to see if $L_2\overline{w} = -g_2$. If this is not the case, then proceed no further, since the system $Hx + g = 0$ has no solution and so the optimal value in (3.1) is $-\infty$. Otherwise, proceed to Step 3.

Step 3: Use back substitution to solve the equation $L_1^T y = \overline{w}$ for $\overline{y} := L_1^{-T}\overline{w}$ and set

$$\overline{x} = \begin{pmatrix} \overline{y} \\ 0 \end{pmatrix}.$$

Then

$$H\overline{x} = LL^T\overline{x} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} [L_1^T \; L_2^T] \begin{pmatrix} \overline{y} \\ 0 \end{pmatrix} = \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} \overline{w} = -g \; .$$

Hence, $\overline{x}$ solves the equation $Hx + g = 0$ and so is an optimal solution to the quadratic optimization problem (3.1).

5. Linear Least Squares Revisited

We have already seen that the least squares problem is a special case of the problem of minimizing a quadratic function. But what about the reverse? Part (4) of Theorem 3.2 tells us that, in general, the reverse cannot be true since the linear least squares problem always has a solution. But what about the case when the quadratic optimization problem has a solution? In this case the matrix H is necessarily positive semi-definite and a solution to the system $Hx + g = 0$ exists. By Theorem 3.11, there is a lower triangular matrix $L \in \mathbb{R}^{n\times k}$, where $k = \operatorname{rank}(H)$, such that $H = LL^T$. Set $A := L^T$. In particular, this implies that $\operatorname{Ran}(H) = \operatorname{Ran}(L) = \operatorname{Ran}(A^T)$. Since $Hx + g = 0$, we know that $-g \in \operatorname{Ran}(H) = \operatorname{Ran}(A^T)$, and so there is a vector $b \in \mathbb{R}^k$ such that $-g = A^T b$. Consider the linear least squares problem

$$\min_{x\in\mathbb{R}^n} \tfrac{1}{2} \|Ax - b\|_2^2 .$$

As in (3.3), expand the objective in this problem to obtain

$$\begin{aligned}\tfrac{1}{2}\|Ax-b\|_2^2 &= \tfrac{1}{2}x^T(A^TA)x-(A^Tb)^Tx+\tfrac{1}{2}\|b\|_2^2\\ &= \tfrac{1}{2}x^TLL^Tx+g^Tx+\beta\\ &= \tfrac{1}{2}x^THx+g^Tx+\beta,\end{aligned}$$

where $\beta=\frac{1}{2}\|b\|_2^2$. We have just proved the following result.

PROPOSITION 3.14. *A quadratic optimization problem of the form* (3.1) *has an optimal solution if and only if it is equivalent to a linear least squares problem.*

6. The Conjugate Gradient Algorithm

The Cholesky factorization is an important and useful tool for computing solutions to the quadratic optimization problem, but it is too costly to be employed in many very large scale applications. In some applications, the matrix H is too large to be stored or it is not available as a data structure. However, in these problems it is often the case that the matrix vector product Hx can be obtained for a given vector $x\in\mathbb{R}^n$. This occurs, for example, in signal processing applications. In this section, we develop an algorithm for solving the quadratic optimization problem (3.4) that only requires access to the matrix vector products Hx. Such an algorithm is called a *matrix free* method since knowledge the whole matrix H is not required. In such cases the Cholesky factorization is inefficient to compute. The focus of this section is the study of the matrix free method known as the *conjugate gradient algorithm*. Throughout this section we assume that the matrix H is positive definite.

6.1. Conjugate Direction Methods. Consider the problem (3.4) where it is known that H is symmetric and positive definite. In this case it is possible to define a notion of *orthogonality* or *congugacy* with respect to H.

DEFINITION 3.15 (Conjugacy). *Let $H\in\mathcal{S}^n_{++}$. We say that the vectors $x,y\in\mathbb{R}^n\backslash\{0\}$ are H-conjugate (or H-orthogonal) if $x^THy=0$.*

PROPOSITION 3.16. *[Conjugacy implies Linear Independence]*
If $H\in\mathcal{S}^n_{++}$ and the set of nonzero vectors d^0, $d^1,\ldots,d^k$ are (pairwise) H-conjugate, then these vectors are linearly independent.

PROOF. If $0=\sum_{i=0}^{k}\mu_id^i$, then for $\bar{i}\in\{0,1,\ldots,k\}$

$$0=(d^{\bar{i}})^TH[\sum_{i=0}^{k}\mu_id^i]=\mu_{\bar{i}}(d^{\bar{i}})^THd^{\bar{i}},$$

Hence $\mu_i=0$ for each $i=0,\ldots,k$. □

Let $x^0\in\mathbb{R}^n$ and suppose that the vectors $d^0,d^1,\ldots,d^{k-1}\in\mathbb{R}^n$ are H-conjugate. Set $S=\mathrm{Span}(d^0,d^1,\ldots,d^{k-1})$. Theorem 3.5 tells us that there is a unique optimal solution $\overline{x}$ to the problem $\min\left\{\frac{1}{2}x^THx+g^Tx\,\middle|\,x\in x^0+S\right\}$, and that $\overline{x}$ is uniquely identified by the condition $H\overline{x}+g\in S^\perp$, or equivalently, $0=(d^j)^T(H\overline{x}+g),\ j=0,1,\ldots,k-1$. Since $\overline{x}\in x_0+S$, there are scalars $\mu_0,\ldots,\mu_{n-1}$ such that

$$\overline{x}=x^0+\mu_0d^0+\ldots+\mu_{k-1}d^{k-1}, \tag{3.15}$$

and so, for each $j = 0, 1, \ldots, k-1$,

$$\begin{aligned} 0 &= (d^j)^T(H\overline{x} + g) \\ &= (d^j)^T\left(H(x^0 + \mu_0 d^0 + \ldots + \mu_{k-1}d^{k-1}) + g\right) \\ &= (d^j)^T(Hx^0 + g) + \mu_0(d^j)^T Hd^0 + \ldots + \mu_{k-1}(d^j)^T Hd^{k-1} \\ &= (d^j)^T(Hx^0 + g) + \mu_j(d^j)^T Hd^j \ . \end{aligned}$$

Therefore,

$$\mu_j = \frac{-(Hx^0 + g)^T(d^j)}{(d^j)^T Hd^j} \quad j = 0, 1 \ldots, k-1 \ . \tag{3.16}$$

This observation motivates the following theorem.

THEOREM 3.17. *[Expanding Subspace Theorem]*
Consider the problem (3.4) *with* $H \in \mathcal{S}^n_{++}$, *and set* $f(x) = \frac{1}{2}x^T Hx + g^T x$. *Let* $\{d^i\}_{i=0}^{n-1}$ *be a sequence of nonzero* H*-conjugate vectors in* $\mathbb{R}^n$. *Then, for any* $x^0 \in \mathbb{R}^n$ *the sequence* $\{x^k\}$ *generated according to*

$$x^{k+1} := x^k + t_k d^k,$$

with

$$t_k := \arg\min\{f(x^k + td^k) : t \in \mathbb{R}\},$$

has the property that $f(x) = \frac{1}{2}x^T Hx + g^T x$ *attains its minimum value on the affine set* $x^0 +$ *Span* $\{d^0, \ldots, d^{k-1}\}$ *at the point* x^k. *In particular, if* $k = n$, *then* x^n *is the unique global solution to the problem* (3.4).

PROOF. Let us first compute the value of the t_k's. For $j = 0, \ldots, k-1$, define $\varphi_j : \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} \varphi_j(t) &= f(x^j + td^j) \\ &= \tfrac{t^2}{2}(d^j)^T Hd^j + t(g^j)^T d^j + f(x^j), \end{aligned}$$

where $g^j = Hx^j + g$. Then, for $j = 0, \ldots, k-1$, $\varphi_j'(t) = t(d^j)^T Hd^j + (g^j)^T d^j$ and $\varphi_j''(t) = (d^j)^T Hd^j > 0$. Since $\varphi_j''(t) > 0$, our one dimensional calculus tells us that φ_j attains its global minmimum value at the unique solution t_j to the equation $\varphi_j'(t) = 0$, i.e.,

$$t_j = -\frac{(g^j)^T d^j}{(d^j)^T Hd^j}.$$

Therefore,

$$x^k = x^0 + t_0 d^0 + t_1 d^1 + \cdots + t_k d^k$$

with

$$t_j = -\frac{(g^j)^T d^j}{(d^j)^T Hd^j}, \ j = 0, 1, \ldots, k.$$

In the discussion preceding the theorem it was shown that if $\overline{x}$ is the solution to the problem

$$\min\left\{f(x) \,\middle|\, x \in x^0 + \text{Span}(d^0, d^1, \ldots, d^k)\right\} \ ,$$

then $\overline{x}$ is given by (3.15) and (3.16). Therefore, if we can now show that $\mu_j = t_j, \ j = 0, 1, \ldots, k$, then $\overline{x} = x_k$ proving the result. For each $j \in \{0, 1, \ldots, k\}$ we have

$$\begin{aligned} (g^j)^T d^j &= (Hx^j + g)^T d^j \\ &= \left(H(x^0 + t_0 d^0 + t_1 d^1 + \cdots + t_{j-1}d^{j-1}) + g\right)^T d^j \\ &= (Hx^0 + g)^T d^j + t_0(d^0)^T Hd^j + t_1(d^1)^T Hd^j + \cdots + t_{j-1}(d^{j-1})^T Hd^j \\ &= (Hx^0 + g)^T d^j \\ &= (g^0)^T d^j \ . \end{aligned}$$

Therefore, for each $j \in \{0, 1, \ldots, k\}$,

$$t_j = \frac{-(g^j)^T d^j}{(d^j)^T H d^j} = \frac{-(g^0)^T d^j}{(d^j)^T H d^j} = \mu_j,$$

which proves the result. □

6.2. The Conjugate Gradient Algorithm. The major drawback of the Conjugate Direction Algorithm of the previous section is that it seems to require that a set of H-conjugate directions must be obtained before the algorithm can be implemented. This is in opposition to our working assumption that H is so large that it cannot be kept in storage since any set of H-conjugate directions requires the same amount of storage as H. However, it is possible to generate the directions d^j one at a time and then discard them after each iteration of the algorithm. One example of such an algorithm is the Conjugate Gradient Algorithm.

The C-G Algorithm:

Initialization: $x^0 \in \mathbb{R}^n$, $d^0 = -g^0 = -(Hx^0 + g)$.

For $k = 0, 1, 2, \ldots$

$$\begin{array}{lll} t_k & := -(g^k)^T d^k / (d^k)^T H d^k & \\ x^{k+1} & := x^k + t_k d^k & \\ g^{k+1} & := Hx^{k+1} + g & \text{(STOP if } g^{k+1} = 0\text{)} \\ \beta_k & := (g^{k+1})^T H d^k / (d^k)^T H d^k & \\ d^{k+1} & := -g^{k+1} + \beta_k d^k & \\ k & := k + 1. & \end{array}$$

THEOREM 3.18. *[CONJUGATE GRADIENT THEOREM]*
The C-G algorithm is a conjugate direction method. If it does not terminate at x^k (i.e. $g^k \neq 0$), then

(1) Span $[g^0, g^1, \ldots, g^k]$ = span $[g^0, Hg^0, \ldots, H^k g^0]$
(2) Span $[d^0, d^1, \ldots, d^k]$ = span $[g^0, Hg^0, \ldots, H^k g^0]$
(3) $(d^k)^T H d^i = 0$ for $i \leq k - 1$
(4) $t_k = (g^k)^T g^k / (d^k)^T H d^k$
(5) $\beta_k = (g^{k+1})^T g^{k+1} / (g^k)^T g^k$.

PROOF. We first prove (1)-(3) by induction. The results are clearly true for $k = 0$. Now suppose they are true for k, we show they are true for $k + 1$. First observe that

$$g^{k+1} = g^k + t_k H d^k$$

so that $g^{k+1} \in$ Span$[g^0, \ldots, H^{k+1} g^0]$ by the induction hypothesis on (1) and (2). Also $g^{k+1} \notin$ Span $[d^0, \ldots, d^k]$, otherwise, by Theorem 3.5 Part (1), $g^{k+1} = Hx^{k+1} + g = 0$ since the method is a conjugate direction method up to step k by the induction hypothesis. Hence $g^{k+1} \notin$ Span $[g^0, \ldots, H^k g^0]$ and so Span $[g^0, g^1, \ldots, g^{k+1}] =$ Span $[g^0, \ldots, H^{k+1} g^0]$, which proves (1).

To prove (2) write

$$d^{k+1} = -g^{k+1} + \beta_k d^k$$

so that (2) follows from (1) and the induction hypothesis on (2).

To see (3) observe that

$$(d^{k+1})^T H d^i = -(g^{k+1})^T H d^i + \beta_k (d^k)^T H d^i.$$

For $i = k$ the right hand side is zero by the definition of β_k. For $i < k$ both terms vanish. The term $(g^{k+1})^T H d^i = 0$ by Theorem 3.17 since $Hd^i \in$ Span$[d^0, \ldots, d^k]$ by (1) and (2). The term $(d^k)^T H d^i$ vanishes by the induction hypothesis on (3).

To prove (4) write

$$-(g^k)^T d^k = (g^k)^T g^k - \beta_{k-1} (g^k)^T d^{k-1}$$

where $(g^k)^T d^{k-1} = 0$ by Theorem 3.17.

To prove (5) note that $(g^{k+1})^T g^k = 0$ by Theorem 3.17 because $g^k \in \text{Span}[d^0, \ldots, d^k]$. Hence

$$(g^{k+1})^T H d^k = \frac{1}{t_k}(g^{k+1})^T [g^{k+1} - g^k] = \frac{1}{t_k}(g^{k+1})^T g^{k+1}.$$

Therefore,

$$\beta_k = \frac{1}{t_k}\frac{(g^{k+1})^T g^{k+1}}{(d^k)^T H d^k} = \frac{(g^{k+1})^T g^{k+1}}{(g^k)^T g^k}.$$

□

Remarks:

(1) The C–G method is an example of a *descent method* since the values

$$f(x^0),\ f(x^1),\ \ldots, f(x^n)$$

form a decreasing sequence.

(2) It should be observed that due to the occurrence of round-off error the C-G algorithm is best implemented as an iterative method. That is, at the end of n steps, x^n may not be the global optimal solution and the intervening directions d^k may not be H-conjugate. Consequently, the algorithm is usually iterated until $\left\|g^k\right\|_2$ is sufficiently small. Due to the observations in the previous remark, this approach is guarenteed to continue to reduce the function value if possible since the overall method is a descent method. In this sense the C–G algorithm is self correcting.

CHAPTER 4

Optimality Conditions

1. Existence of Optimal Solutions

Consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ where f is continuous on all of $\mathbb{R}^n$ over a set $\Omega \subset \mathbb{R}^n$:

$$\mathcal{P} \qquad \min_{x \in \Omega} f(x).$$

As we have seen, there is no guarantee that f has a minimum value over Ω, or if it does, it may not be attained. To clarify this situation, we examine conditions under which a solution is guaranteed to exist. Recall that we already have at our disposal a rudimentary existence result for constrained problems. This is the Weierstrass Extreme Value Theorem.

THEOREM 4.1. *(*WEIERSTRASS EXTREME VALUE THEOREM*) Every continuous function on a compact set attains its extreme values on that set. In particular, this implies that continuous functions are bounded on compact sets.*

PROOF. We establish the result for the infimum since the proof is essentially identical for the supremum. Letv $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on the compact set $C \subset \mathbb{R}^n$ and set $\mu := \inf_{s \in C} f(x)$. Then there is a sequence $\{x^k\} \subset C$ such that $f(x^k) \downarrow \mu$. Since C is compact there is a subsequence $\mathbb{J} \subset \mathbb{N}$ and point $\overline{x} \in C$ such that $x^k \overset{\mathbb{J}}{\to} \overline{x}$. By continuity $f(\overline{x}) = \mu$, hence f attains its infimal value on C. □

The necessity of compactness in this result is illustrated by the following examples.

EXAMPLE 4.2. *(1) A continuous function can be unbounded on a bounded set if that set is not closed. For example, the function $f(x) = 1/x$ is continuous on the half open interval $(0, 1]$, but f is not bounded on this bounded set. The same is true for the function $f(x) = \tan x$ on the open interval $(-\pi/2,\ \pi/2)$.*

(2) A bounded function may not attain its extremal values. For example, the function $f(x) = (e^x - e^{-x})/(e^x + e^{-x})$ only takes values on the open interval $(-1,\ 1)$. Its infimum is -1 and its supremum is 1. But no value of x attains these values.

We build on Theorem 4.1 by introducing the notion of a coercive function.

DEFINITION 4.3. *Let $\Omega \subset \mathbb{R}^n$. A function $f : \mathbb{R}^n \to \Omega$ is said to be coercive over Ω if for every sequence $\{x^\nu\} \subset \Omega$ for which $\|x^\nu\| \to \infty$ it must be the case that $f(x^\nu) \to +\infty$ as well.*

Continuous coercive functions can be characterized by an underlying compactness property on their lower level sets.

THEOREM 4.4. *(Coercivity and Compactness) Let $\Omega \subset \mathbb{R}^n$ be closed, and let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous on Ω. The function f is coercive over Ω if and only if for every $\alpha \in \mathbb{R}$ the set $\{x \in \Omega \,|\, f(x) \leq \alpha\}$ is compact.*

PROOF. We first show that the coercivity of f over Ω implies the compactness of the sets $\{x \in \Omega \,|\, f(x) \leq \alpha\}$. We begin by noting that the continuity of f implies the closedness of the sets $\{x \,|\, f(x) \leq \alpha\}$, and so the sets $\{x \,|\, f(x) \leq \alpha\} \cap \Omega = \{x \in \Omega \,|\, f(x) \leq \alpha\}$ are closed as well. Thus, it remains only to show that any set of the form $\{x \in \Omega \,|\, f(x) \leq \alpha\}$ is bounded. We show this by contradiction. Suppose to the contrary that there is an $\alpha \in \mathbb{R}^n$ such that the set $S = \{x \in \Omega \,|\, f(x) \leq \alpha\}$ is unbounded. Then there must exist

a sequence $\{x^\nu\} \subset S$ with $\|x^\nu\| \to \infty$. But then, by the coercivity of f over Ω, we must also have $f(x^\nu) \to \infty$. This contradicts the fact that $f(x^\nu) \leq \alpha$ for all $\nu = 1, 2, \ldots$. Therefore the set S must be bounded.

Let us now assume that each of the sets $\{x \in \Omega \,|\, f(x) \leq \alpha\}$ is bounded and let $\{x^\nu\} \subset \Omega$ be such that $\|x^\nu\| \to \infty$. Let us suppose that there exists a subsequence of the integers $J \subset \mathbb{N}$ such that the set $\{f(x^\nu)\}_J$ is bounded above. Then there exists $\alpha \in \mathbb{R}^n$ such that $\{x^\nu\}_J \subset \{x \in \Omega \,|\, f(x) \leq \alpha\}$. But this cannot be the case since each of the sets $\{x \in \Omega \,|\, f(x) \leq \alpha\}$ is bounded while every subsequence of the sequence $\{x^\nu\}$ is unbounded by definition. Therefore, the set $\{f(x^\nu)\}_J$ cannot be bounded, and so the sequence $\{f(x^\nu)\}$ contains no bounded subsequence, i.e. $f(x^\nu) \to \infty$. □

This result in conjunction with Weierstrass's Theorem immediately yields the following existence result for the problem $\mathcal{P}$.

THEOREM 4.5. *(Coercivity implies existence) Let $\Omega \subset \mathbb{R}^n$ be closed, and let $f : \mathbb{R}^n \to \Omega$ be continuous on Ω. If f is coercive over Ω, then f has at least one global minimizer over Ω.*

PROOF. Let $\alpha \in \mathbb{R}$ be chosen so that the set $S = \{x \in \Omega \,|\, f(x) \leq \alpha\}$ is non-empty. By coercivity, this set is compact. By Weierstrass's Theorem, the problem $\min\{f(x) \,|\, x \in S\}$ has at least one global solution. Obviously, the set of global solutions to the problem $\min\{f(x) \,|\, x \in S\}$ is a global solution to $\mathcal{P}$ which proves the result. □

REMARK 4.6. *It should be noted that the coercivity hypothesis is stronger than is strictly required in order to establish the existence of a solution. Indeed, a global minimizer must exist if there exist one non-empty compact lower level set. We do not need all of them to be compact. However, in practice, coercivity is easy to check.*

This existence result is quite useful, but it does not give us a test for optimality. That is, knowing that a solution exists does not directly assist in determining whether any given point is or is not an optimal solution. The goal of the next section is to develops such tests or certificates of optimality. These tests are based on variational properties of the function f.

2. First-Order Optimality Conditions without Constraints

When walking in the mountains, you know that you are at a local minima when every direction you look in takes you up. One way to find a local minima is to choose a direction that points downward and walk in that direction until you are no longer going down, and then choose another downward direction and repeat. This process is essentially what an important class of optimization algorithms called *descent methods*. Given a point, your current location, you find a direction d that goes downward. We call such a diretion a *descent direction*. We then proceed in the direction d for a positive distance called the stepsize $t > 0$ to arrive at the new point $x + td$ such that the value of the function at the new point is less than that at the old point, i.e., $f(x + td) < f(x)$. If no such downward direction exists, then you are at a local minimum.

We now formalize these notions mathematically. The first step is to formalize the notion of a descent direction. This is done using the *directional derivative*. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x, d \in \mathbb{R}$. The directional derivative of f at x in the direction d defined as

$$f'(x; d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

It is important to observe that this is a one sided derivative since $t \downarrow 0$.

EXAMPLE 4.7 (The Directional Derivative). *The directional derivative at a point can exist even if the function is not differentiable at that point. In addition, we allow the directional derivative to take the values $\pm\infty$. We give two examples to illustrate this behaviour.*

(1) *Consider the absolute value function* $f(x) = |x|$ *on* $\mathbb{R}$. *Clearly,* f *is not differentiable at the origin, however, the directional derivative exists for all directions there. Indeed, one can easily show that* $f'(0;d) = |d|$.
(2) *Consider the function* $f(x) = \sqrt{|x|}$ *on* $\mathbb{R}$. *Again,* f *is not differentiable at the origin. By direct computation, we have* $f'(0;d) = +\infty$ *for all* $d \neq 0$ *with* $f'(0;0) = 0$.

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then

$$f(x + td) = f(x) + t\nabla f(x)^T d + o(t),$$

and so

$$f'(x;d) = \lim_{t \downarrow 0} \frac{f(x) + t\nabla f(x)^T d + o(t) - f(x)}{t} = \nabla f(x)^T d + \lim_{t \downarrow 0} \frac{o(t)}{t} = \nabla f(x)^T d.$$

We call d a *direction of strict descent* for f at x if $f'(x;d) < 0$. Note that if $f'(x;d) < 0$, then there must be a $\bar{t} > 0$ such that

$$\frac{f(x + td) - f(x)}{t} < 0 \quad \text{whenever} \quad 0 < t < \bar{t}\ .$$

In this case, we must have

$$f(x + td) < f(x) \quad \text{whenever} \quad 0 < t < \bar{t}\ .$$

That is, we can always reduce the function value at x by moving in the direction d an arbitrarily small amount. In particular, if there is a direction d such that $f'(x;d)$ exists with $f'(x;d) < 0$, then x cannot be a local solution to the problem $\min_{x \in \mathbb{R}^n} f(x)$. Equivalently, if x is a local to the problem $\min_{x \in \mathbb{R}^n} f(x)$, then $f'(x;d) \geq 0$ whenever $f'(x;d)$ exists. We state this elementary result in the following lemma.

LEMMA 4.8 (Basic First-Order Optimality Result using the Directional Derivative). *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *and let* $\overline{x} \in \mathbb{R}^n$ *be a local solution to the problem* $\min_{x \in \mathbb{R}^n} f(x)$. *Then*

$$f'(x;d) \geq 0$$

for every direction $d \in \mathbb{R}^n$ *for which* $f'(x;d)$ *exists.*

However, the directional derivative does not always exist. To cover such cases, we generalize the notion of the directional derivative and define the *subderivative* of f at x in a direction $d \in \mathbb{R}^n$ by

$$df(x)(d) := \liminf_{\substack{\tau \searrow 0 \\ d' \to d}} \frac{f(x + \tau d') - f(x)}{\tau}\ .$$

The subderivative always exists as a number in $\mathbb{R} \cup \{\pm\infty\}$ even when the directional derivative does not. Again, when $f$is differentiable at x, then $df(x)(d) = \nabla f(x)^T d$. In a manner parallel to Lemma 4.8, we obtain the following theorem.

THEOREM 4.9 (Basic First-Order Optimality Result using the Subderivative). *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *and let* $\overline{x} \in \mathbb{R}^n$ *be a local solution to the problem* $\min_{x \in \mathbb{R}^n} f(x)$. *Then*

$$df(\overline{x})(d) \geq 0 \qquad \forall\ d \in \mathbb{R}^n.$$

When $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, this lemma gives the following well-known result.

THEOREM 4.10 (First-Order Optimality for Differentiable Functions). *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable at a point* $\overline{x} \in \mathbb{R}^n$. *If* $\overline{x}$ *is a local minimum of* f, *then* $\nabla f(\overline{x}) = 0$.

PROOF. Since f is differentiable at $\overline{x}$, we have $f'(\overline{x};d) = df(x)(d) = \nabla f(\overline{x})^T d$ for all $d \in \mathbb{R}^n$. By Lemma 4.8, we have

$$0 \leq f'(\overline{x};d) = \nabla f(\overline{x})^T d \quad \text{for all} \quad d \in \mathbb{R}^n\ .$$

Taking $d = -\nabla f(\overline{x})$ we get

$$0 \leq -\nabla f(\overline{x})^T \nabla f(\overline{x}) = -\left\|\nabla f(\overline{x})\right\|^2 \leq 0.$$

Therefore, $\nabla f(\overline{x}) = 0$. □

EXAMPLE 4.11 (First-Order Optimality and the Linear Least Squares Problem). *The objective function for the linear least squares problem is* $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ *whose gradient is* $\nabla f(x) = A^T(Ax-b)$. *The condition* $0 = \nabla f(x)$ *is therefore equivalent to the normal equations. However, in the case of LLS, the normal equations are both necessary and sufficient for optimality. Moreover, a solution always exists. In this section, we examine when these properties hold for other types of optimization problems.*

EXAMPLE 4.12 (First-Order Optimality for Quadratic Functions). *The quadratic objective function takes the form* $f(x) = \frac{1}{2}x^THx + g^Tx$, *where* $H \in \mathcal{S}^n$. *The gradient is* $\nabla f(x) = Hx + g$. *The condition* $0 = \nabla f(x)$ *only specifies one of the two conditions required for a global (equivalently, local) minima. The missing condition is that* H *is positive semidefinite. This is a condition on the Hessian* $\nabla^2 f(x) = H$. *In this section we examine how this Hessian condition translates to other settings.*

Points at which $\nabla f(x) = 0$ are said to be a stationary, or critical, points of f. In our next result we link the notions of coercivity and stationarity.

THEOREM 4.13. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be differentiable on all of* $\mathbb{R}^n$. *If* f *is coercive, then* f *has at least one global minimizer and these global minimizers can be found from among the set of critical points of* f.

PROOF. Since differentiability implies continuity, we already know that f has at least one global minimizer. Differentiabilty implies that this global minimizer is critical. □

Thus one way to find a global minimizer of a coercive differentiable function is to first find all critical points and then from among these determine those yielding the smallest function value.

The condition $\nabla f(x) = 0$ is called a first-order optimality condition because it uses only first derivative information. This information is sufficient in the case of the linear least squares problem, but is insufficient in the case of general quadratic functions where information about the Hessian is required. Conditions involving the Hessian matrix are called second-order conditions, and are examine in the following section.

3. Second-Order Optimality Conditions without Constraints

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. The theory of second-order optimality conditions is motivated by the optimality conditions we established for quadratic functions and are based on the second-order Taylor expansion of f at a given point $\overline{x}$:

$$f(x) = f(\overline{x}) + \nabla f(\overline{x})^T(x-\overline{x}) + \frac{1}{2}(x-\overline{x})^T\nabla^2 f(\overline{x})(x-\overline{x}) + o(\|x-\overline{x}\|^2 \tag{4.1}$$

where

$$\lim_{x\to\overline{x}} \frac{o(\|x-\overline{x}\|^2)}{\|x-\overline{x}\|^2} = 0.$$

THEOREM 4.14. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be twice continuously differentiable at the point* $\overline{x} \in \mathbb{R}^n$.

1. *(Necessity) If* $\overline{x}$ *is a local minimum of* f, *then* $\nabla f(\overline{x}) = 0$ *and* $\nabla^2 f(\overline{x})$ *is positive semidefinite.*
2. *(Sufficiency) If* $\nabla f(\overline{x}) = 0$ *and* $\nabla^2 f(\overline{x})$ *is positive definite, then there is an* $\alpha > 0$ *such that* $f(x) \geq f(\overline{x}) + \alpha\|x-\overline{x}\|^2$ *for all* x *near* $\overline{x}$.

PROOF. (1) We make use of the second-order Taylor series expansion (4.1) and the fact that $\nabla f(\overline{x}) = 0$ by Theorem 4.10. Given $d \in \mathbb{R}^n$ and $t > 0$ set $x := \overline{x} + td$, plugging this into (4.1) we find that

$$0 \leq \frac{f(\overline{x}+td) - f(\overline{x})}{t^2} = \frac{1}{2}d^T\nabla^2 f(\overline{x})d + \frac{o(t^2)}{t^2}$$

since $\nabla f(\overline{x}) = 0$ by Theorem 4.10. Taking the limit as $t \to 0$ we get that

$$0 \leq d^T\nabla^2 f(\overline{x})d.$$

Since d was chosen arbitrarily, $\nabla^2 f(\overline{x})$ is positive semi-definite.

(2) Suppose the result is false. Then there are sequences $\alpha_k \downarrow 0$ and $x^k \to \overline{x}$ such that

$$\alpha_k \left\|x^k - \overline{x}\right\| \geq f(x_k) - f(\overline{x}) = \nabla f(\overline{x})^T(x^k - \overline{x}) + \frac{1}{2}(x^k - \overline{x})^T\nabla^2 f(\overline{x})(x^k - \overline{x}) + o(\|x^k - \overline{x}\|^2$$
$$= \frac{1}{2}(x^k - \overline{x})^T\nabla^2 f(\overline{x})(x^k - \overline{x}) + o(\|x^k - \overline{x}\|^2, \quad \forall\ k = 1, 2, \ldots.$$

By compactness, there is a $\overline{u} \in \mathbb{R}^n$ with $\|\overline{u}\| = 1$ and a subsequence $\mathbb{J} \subset \mathbb{N}$ such that $\frac{x^k - \overline{x}}{\|x^k - \overline{x}\|} \to \overline{u}$. Hence is we divide this inequality through by $\left\|x^k - \overline{x}\right\|$ and take the limit over $\mathbb{J}$ we find that $\overline{u}^T\nabla^2 f(\overline{x})\overline{u} \leq 0$ which contradicts the hypothesis that $\nabla^2 f(\overline{x})$ is positive definite. This contradiction proves the result.

□

A symmetric matrix $H \in \mathcal{S}^n$ is said to be an *indefinite symmetric matrix* if there exist $\overline{u}, \overline{v} \in \mathbb{R}^n$ such that $\overline{u}^T H\overline{u} > 0$ and $\overline{v}^T H\overline{v} < 0$ Critical points $\overline{x}$ for a function $f : \mathbb{R}^n \to \mathbb{R}$ at which the Hessian $\nabla^2 f(x)$ is indefinite are called *saddle points*. Saddle points are neither local minima or local maxima. From our previous study of symmetric matrices, we know that any symmtric matrix having at least one positive and one negative entry must be indefinite.

EXAMPLE 4.15. *Let $\Delta_i(H)$ denote the i^{th} principal minor of the symmetric matrix*

$$H = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 5 & 1 \\ -1 & 1 & 4 \end{bmatrix}.$$

Then

$$\Delta_1(H) = 1, \quad \Delta_2(H) = \begin{vmatrix} 1 & 1 \\ 1 & 5 \end{vmatrix} = 4, \quad \textit{and}$$
$$\Delta_3(H) = \det(H) = 1 \cdot \begin{vmatrix} 5 & 1 \\ 1 & 4 \end{vmatrix} - 1 \cdot \begin{vmatrix} 1 & -1 \\ 1 & 4 \end{vmatrix} - 1 \cdot \begin{vmatrix} 1 & -1 \\ 5 & 1 \end{vmatrix} = 19 - 5 - 6 = 8.$$

Therefore, H is positive definite.

4. First-Order Optimality Conditions with Constraints

We study first–order optimality conditions for the constrained problem

$$\mathcal{P}: \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \Omega, \end{array}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and $\Omega \subset \mathbb{R}^n$ is closed and non-empty. The first step in the analysis of the problem $\mathcal{P}$ is to derive conditions that allow us to recognize when a particular vector $\overline{x}$ is a solution, or local solution, to the problem. In the unconstrained case, the first-order necessary condition for optimality is that thje gradient is zero, and the second-order necessary condition is that the Hessian is positive semidefinite. If the Hessian is positive definite, then we know that the critcal point is a local minima. When constraints are present, the situation is much more complicated. Nonetheless, out goal is to somehow preserve that natural feel of our results for the unconstrained case.

To begin, let us first suppose that we are at a feasible point $x \in \Omega$ and we wish to find a better point $\tilde{x}$ using the basic structure of a descent algorithm. For this we must find a direction of descent for f that also maintains feasibility. That is, we wish to find a direction d such that $f'(x; d) < 0$ and $x + td \in \Omega$ for all $t \in (0, \overline{t}]$ for some $\overline{t} > 0$. Search directions that can maintain feasibility in this way are given a special name.

DEFINITION 4.16 (Feasible Directions). *Given a subset Ω of $\mathbb{R}^n$ and a point $x \in \Omega$, we say that a direction $d \in \mathbb{R}^n$ is a feasible direction for Ω at x if there is a $\overline{t} > 0$ such that $x + td \in \Omega$ for all $t \in [0, \overline{t}]$.*

THEOREM 4.17 (First-Order Optimality with Feasible Directions). *If $\overline{x}$ is a local solution to the problem $\mathcal{P}$, then $f'(\overline{x};d) \geq 0$ for all feasible directions d for Ω at $\overline{x}$ for which $f'(\overline{x};d)$ exists.*

PROOF. The proof is a straightforward application of the definitions. If the result were false, then there would be a direction of descent for f at $\overline{x}$ that is also a feasible direction for Ω at $\overline{x}$. But then moving a little bit in this direction both keeps us in Ω and strictly reduces the value of f. This contradicts the assumption that $\overline{x}$ is a local solution. Therefore, the result must be true. □

Unfortunately, this result is really only useful when the set Ω is a convex polyhedron. In this instance, it will be shown that the set of feasible directions suffices. However, in general nonlinear settings, this set is woefully inadequate in describing the first-order properties of constraint regions. To illustrate this, consider the set

$$\Omega = \{(x_1, x_2)^T \,:\, x_1^2 + x_2^2 = 1\}.$$

Here the only feasible direction at any point of Ω is the zero direction. Hence, regardless of the objective function f and the point $\overline{x} \in \Omega$, we have that $f'(\overline{x};d) \geq 0$ for every feasible direction to Ω at $\overline{x}$. In this case, Theorem 4.17 has no content.

To overcome this deficiency we introduce a general notion of *tangency* that considers all directions d *pointing into* Ω at $x \in \Omega$ in a limiting sense. A set $K \subset \mathbb{R}^n$ is called a *cone* if $\lambda x \in K$ for all $\lambda \geq 0$ whenever $x \in K$. Define the *tangent cone* to Ω at a point $x \in \Omega$ to be the set of limiting directions obtained from sequences in Ω that converge to x. Specifically, the tangent cone is given by

$$T(x \,|\, \Omega) := \{d \,:\, \exists\, \tau_i \searrow 0 \text{ and } \{x_i\} \subset \Omega, \text{ with } x_i \to x, \text{ such that } \tau_i^{-1}(x_i - x) \to d\}.$$

It is easily seen that $T(x \,|\, \Omega)$ is a closed cone.

EXAMPLE 4.18 (Tangent Cone Examples).
(1) If $x \in \operatorname{intr}(\Omega) \subset \mathbb{R}^n$, then $T(x \,|\, \Omega) = \mathbb{R}^n$. Here, $\operatorname{intr}(\Omega)$ denotes the interior of the set Ω. In particular, this implies that if Ω is open, then $T(x \,|\, \Omega) = \mathbb{R}^n$ for all $x \in \Omega$.
(2) If $\overline{x} \in \mathbb{R}^n$ and $\Omega = \{\overline{x}\}$, then $T(x \,|\, \Omega) = \{0\}$.
(3) If $\Omega = \{x \,:\, Ax = b\}$, where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, then $T(x \,|\, \Omega) = Nul(A)$ for every $x \in \Omega$.
(4) If Ω is a convex polyhedron, then the set the feasible directions at $x \in \Omega$ coincides with the tangent cone $T(x \,|\, \Omega) = \bigcup_{t \geq 0}(\Omega - x)$.
(5) If $\Omega = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$, then $T(x \,|\, \Omega) = \{v \in \mathbb{R}^n \mid \langle v,\, x\rangle = 0\}$. In particular, if $n = 2$, then $\Omega = \left\{(x_1,\ x_2)^T \,\middle|\, x_1^2 + x_2^2 = 1\right\}$ and $T(x \,|\, \Omega) = \left\{v \in \mathbb{R}^2 \,\middle|\, v_1 x_1 + v_2 x_2 = 0\right\}$ is the line tangent to Ω at $(x_1,\ x_2)^T$ translated to pass through the origin.
(6) If Ω is a smooth manifold, then $T(x \,|\, \Omega)$ is the usual tangent space to the manifold at x. That is, if $F : \mathbb{R}^n \to \mathbb{R}^m$ is smooth and $\Omega = \{x \,|\, F(x) = 0\}$, then $T(x \,|\, \Omega) = \{d \,|\, \nabla F(x)d = 0\}$ for every $x \in \Omega$.

We have the following analogue of Theorem 4.9 in the constrained case.

THEOREM 4.19 (The Basic First-Order Optimality Conditions: Constrained Case). *Let Ω and f be as in $\mathcal{P}$. If $\overline{x}$ is a local solution to $\mathcal{P}$, then*

$$\nabla f(\overline{x})^T d \geq 0 \quad \forall\, d \in T(\overline{x} \,|\, \Omega).$$

PROOF. Let $d \in T(\overline{x} \,|\, \Omega)$ so that there exists $\{x^k\} \in \Omega$ and $\{t_k\} \in \mathbb{R}_{++}$ such that $x^k \to \overline{x}$, $t_k \downarrow 0$ and $d^k := t_k^{-1}(x^k - \overline{x}) \to d$. Then $x^k = \overline{x} + t_k d^k$ for all $k = 1, 2, \ldots$ so that

$$\nabla f(\overline{x})^T d = \lim_{k\to\infty} \nabla f(\overline{x})^T d^k = \lim_{k\to\infty} \frac{(f(\overline{x}) + t_k \nabla f(\overline{x})^T d^k + o(t_k)) - f(\overline{x})}{t_k} = \lim_{k\to\infty} \frac{f(x^k) - f(\overline{x})}{t_k} \geq 0,$$

where the final inequality follows from the fact that $\overline{x}$ is a local solution to $\mathcal{P}$. □

In particular, this result yields Theorem 4.10 as a special case. However, from a practical point of view, the application of this theorem is limited to those cases where the tangent cone to Ω has a convenient representation. We now show how this can be done in an important special case.

5. First-Order Conditions for Nonlinear Programming: KKT Points

When Ω has the form

$$\Omega := \{x : c_i(x) \leq 0, i = 1, \ldots, s, c_i(x) = 0, i = s+1, \ldots, m\}, \tag{4.2}$$

where each $c_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^n$, the optimization problem $\mathcal{P}$ ic called a *nonlinear program*:

$$\begin{array}{lll} \text{(NLP)} & \text{minimize} & f(x) \\ & \text{subjectto} & c_i(x) \leq 0,\ i = 1, \ldots, s, \\ & & c_i(x) = 0,\ i = s+1, \ldots, m. \end{array}$$

Observe that if $x \in \Omega$ and $d \in T(x \,|\, \Omega)$ then there are sequences $\{x_k\} \subset \Omega$ and $\tau_k \searrow 0$ with $x_k \to x$ such that $\tau_k^{-1}(x_k - x) \to d$. Setting $d_k = \tau_k^{-1}(x_k - x)$ for all k we have that

$$c_i'(x; d) = \lim_{k \to \infty} \frac{c_i(x + \tau_k d_k) - c_i(x)}{\tau_k}$$

equals 0 for $i \in \{s+1, \ldots, m\}$ and is less than or equal to 0 for $i \in I(x)$ where

$$I(x) := \{i : i \in \{1, \ldots, s\},\ c_i(x) = 0\}\ .$$

Consequently,

$$T(x \,|\, \Omega) \subset \{d : \nabla c_i(x)^T d \leq 0,\ i \in I(x),\ \nabla c_i(x)^T d = 0, i = s+1, \ldots, m\}\ . \tag{4.3}$$

The set on the right hand side of this inclusion,

$$L_\Omega(x) := \{d : \nabla c_i(x)^T d \leq 0,\ i \in I(x),\ \nabla c_i(x)^T d = 0, i = s+1, \ldots, m\},$$

is a convex polyhedral cone and as such is computationally tractable. However, there is a subtlety here that we must address. Specifically, the set Ω typically has many possible representations. For some of these equivalence in (4.3) may not hold while for others is does hold. Therefore, whether or not the equivalence in (4.3) holds is representation dependent.

Example 4.20 (Regularity Counterexamples).

(1) Let $\Omega := \left\{x \in \mathbb{R}^2 \,\middle|\, (x_1^2 + x_2^2) - 1 \leq 0,\ ((x_1 - 2)^2 + x_2^2) - 1 \leq 0\right\}$ *so that* Ω *is the single point* $\overline{x} = (1, 0)^T$. *In this case,* $T(\overline{x} \,|\, \Omega) = \left\{(0, 0)^T\right\}$ *while* $L_\Omega(\overline{x}) = \left\{(0, \lambda)^T \,\middle|\, \lambda \in \mathbb{R}\right\}$. *However, if we simply represent* Ω *by* $\Omega := \{x \,|\, x_1 = 2,\ x_2 = 0\}$, *then equality occurs in* (4.3).

(2) The representation of the set

$$\Omega := \{x \in \mathbb{R}^2 | -x_1^3 \leq x_2 \leq x_1^3\}$$

is not regular at the origin. To see this, one first verifies that

$$T_\Omega(0) = \left\{(d_1, d_2)^T \,\middle|\, d_1 \geq 0, d_2 = 0\right\}.$$

Then set

$$c_1(x_1, x_2) = -x_1^3 - x_2 \qquad \text{and} \qquad c_2(x_1, x_2) = -x_1^3 + x_2,$$

so that $\Omega = \left\{(x_1, x_2)^T \,\middle|\, c_1(x_1, x_2) \leq 0, c_2(x_1, x_2) \leq 0\right\}$. *Finally, it can be shown that*

$$\left\{d \,\middle|\, \nabla c_1(0, 0)^T d \leq 0, \nabla c_2(0, 0)^T d \leq 0\right\} = \left\{(d_1, d_2)^T \,\middle|\, d_2 = 0\right\} \neq T_\Omega(0).$$

Can you show that this set does not have any representation that is regular at the origin?

Nonetheless, in a certain sense, the cases where these two sets do not coincide are exceptional. Conditions under which they do coincide are called *constraint qualifications* which are explored in in the following section. Since having equality in (4.3) is central to our discussion of first-oerder optimality conditions, we make the following definition.

DEFINITION 4.21 (Regularity). *We say that the representation of the set Ω defined in* (4.2) *is regular at $x \in \Omega$ if*

$$T(x \,|\, \Omega) = \{d \in \mathbb{R}^n : c_i'(x;d) \le 0, i \in I(x), c_i'(x;d) = 0 \ i = s+1, \dots, m\}.$$

With this representation of the tangent cone in hand, we return to explore its implications in the context of Theorem 4.19. The theorem tells us that if $\overline{x}$ is a local solution to (NLP) at which the representation of Ω is regular, then $\nabla f(x)^T d \ge 0$ for all $d \in \mathbb{R}^n$ such that $\nabla c_i(\bar{x})^T d \le 0$ for all $i \in I(\bar{x})$ and $\nabla c_i(\bar{x})^T d = 0$ for all $i = s+1, \dots, m$. Consequently, the following linear program has an optimal value of zero:

$$\begin{array}{lll} \max & (-\nabla f(\bar{x}))^T d & \\ \text{subject to} & \nabla c_i(\bar{x})^T d \le 0 & i \in I(\bar{x}) \\ & \nabla c_i(\bar{x})^T d = 0 & i = s+1, \dots, m. \end{array} \tag{4.4}$$

Hence, the Strong Duality Theorem of linear programming tells us that the dual linear program is feasible and has zero optimal value:

$$\begin{array}{ll} \min & 0 \\ \text{subject to} & \sum_{i \in I(\bar{x})} u_i \nabla c_i(\bar{x}) + \sum_{i=s+1}^m u_i \nabla c_i(\bar{x}) = -\nabla f(\bar{x}) \\ & 0 \le u_i, \ i \in I(\bar{x}). \end{array} \tag{4.5}$$

The feasibility of the dual implies that there exist scalars $u_i, \ i \in I(\overline{x}) \cup \{s+1, \dots, m\}$ with $u_i \ge 0$ for $i \in I(\overline{x})$ such that

$$0 = \nabla f(\overline{x}) + \sum_{i \in I(\overline{x})} u_i \nabla c_i(\overline{x}) + \sum_{i=s+1}^m u_i \nabla c_i(\overline{x}). \tag{4.6}$$

These observations yield our main theorem in this chapter on first-order optimality for (NLP). The central novel ingredient in this result is the *Lagrangian function* $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$L(x,u) := f(x) + \sum_{i=1}^m u_i c_i(x),$$

where the variables u_i are referred to as the dual variables, or *Lagrange multipliers*.

THEOREM 4.22 (Constrained First-Order Optimality Conditions). *Let $\overline{x} \in \Omega$ be a local solution to (NLP) at which the representation of Ω is regular. Then there exist $u \in \mathbb{R}^m$ such that*

(1) $0 = \nabla_x L(\overline{x}, u)$,
(2) $0 = u_i c_i(\overline{x})$ for $i = 1, \dots, s$, and
(3) $0 \le u_i$, $i = 1, \dots, s$.

PROOF. For $i \in I(\overline{x}) \cup \{s+1, \dots, m\}$ let u_i be as given in (4.6) and for $i \in \{1, \dots, s\} \setminus I(\overline{x})$ set $u_i = 0$. Then this choice of $u \in \mathbb{R}^m$ satisfies (1)–(3) above. □

DEFINITION 4.23 (KKT Conditions). *Let $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. We say that (x,u) is a Karush-Kuhn-Tucker (KKT) pair for (NLP) if*

(1) $c_i(x) \le 0$ $i = 1, \dots, s$, $c_i(x) = 0$ $i = s+1, \dots, m$ (Primal Feasibility),
(2) $u_i \ge 0$ for $i = 1, \dots, s$ (Dual Feasibility),
(3) $0 = u_i c_i(x)$ for $i = 1, \dots, s$ (Complementarity), and
(4) $0 = \nabla_x L(x,u)$ (stationarity of the Lagrangian).

Given $x \in \mathbb{R}^n$, if there is a $u \in \mathbb{R}^m$ such that (x,u) is a Karush-Kuhn-Tucker pair for (NLP), then we say that x is a KKT point for (NLP) (we also refer to such an x as a stationary point for (NLP)). □

5.1. Example: Computing KKT Points. In general, computing KKT points by hand is difficult since the problem is nonlinear and so, in general, there is no closed form solution. But in some cases it is possible. A simple strategy is to make reasonable guesses about the nature of the solution and then apply elimination techniques. Consider the following problem:

$$\begin{array}{ll} \text{minimize} & x_1^2 + x_2^2 - 4x_1 - 4x_2 \\ \text{subject to} & x_1^2 \leq x_2 \\ & x_1 + x_2 \leq 2 \ . \end{array}$$

First write the problem in the standard (NLP) form required for the application of the KKT theory: In this example there are no equality constraints, so $s = 0$ and $m = 2$. We have

$$\begin{aligned} f(x_1, x_2) &= x_1^2 + x_2^2 - 4x_1 - 4x_2 = (x_1 - 2)^2 + (x_2 - 2)^2 - 8 \\ c_1(x_1, x_2) &= x_1^2 - x_2 \\ c_2(x_1, x_2) &= x_1 + x_2 - 2 \ . \end{aligned}$$

Note that we can ignore the constant term in the objective function since it does not effect the optimal solution, so henceforth $f(x_1, x_2) = x_1 - 2)^2 + (x_2 - 2)^2$. At this point it is often helpful to graph the solution set if possible. The graph shows that the constraint region is a slice of a parabola.

Since all of these functions are convex, this is an example of a convex programming problem and so the KKT conditions are both necessary and sufficient for global optimality (see Chapter 5 Section 6). Hence, if we locate a KKT point, we know that it is necessarily a globally optimal solution.

The Lagrangian for this problem is

$$L((x_1, x_2), (u_1, u_2)) = (x_1 - 2)^2 + (x_2 - 2)^2 + u_1(x_1^2 - x_2) + u_2(x_1 + x_2 - 2) \ .$$

The KKT conditions are

(1) (Primal Feasibility) $x_1^2 \leq x_2$ and $x_1 + x_2 \leq 2$
(2) (Dual Feasibility) $0 \leq u_1$ and $0 \leq u_2$
(3) (Complementarity) $u_1(x_1^2 - x_2) = 0$ and $u_2(x_1 + x_2 - 2) = 0$
(4) (Stationarity of the Lagrangian)

$$0 = \nabla_x L((x_1, x_2), (u_1, u_2)) = \begin{pmatrix} 2(x_1 - 2) + 2u_1 x_1 + u_2 \\ 2(x_2 - 2) - u_1 + u_2 \end{pmatrix},$$

or equivalently

$$\begin{aligned} 4 &= 2x_1 + 2u_1 x_1 + u_2 \\ 4 &= 2x_2 - u_1 + u_2. \end{aligned}$$

Observe that the global minimizer for the objective function is $(x_1, x_2) = (2, 2)$. Thus, if this point is feasible, it would be the global solution and the multipliers would both be zero. But it is not feasible. Indeed, both constraints are violated by this point. Hence, we conjecture that both constraints are active at the solution. In this case, the KKT pair $((x_1, x_2), (u_1, u_2))$ must satisfy the following 4 equations

$$\begin{aligned} x_2 &= x_2^2 \\ 2 &= x_1 + x_2 \\ 4 &= 2x_1 + 2u_1 x_1 + u_2 \\ 4 &= 2x_2 - u_1 + u_2. \end{aligned} \tag{4.7}$$

We now have 4 equations in 4 unknowns that we can try to solve by elimination. Using the first equation to eliminate x_2 from the second equation, we see that x_1 must satisfy

$$0 = x_1^2 + x_1 - 2 = (x_1 + 2)(x_1 - 1),$$

so $x_1 = -2$ or $x_1 = 1$. Thus, either $(x_1, x_2) = (-2, 4)$ or $(x_1, x_2) = (1, 1)$. Since $(1, 1)$ is closer to the global minimizer of the objective f, we first investigate $(x_1, x_2) = (1, 1)$ to see if it is a KKT point. For this we must find the KKT multipliers (u_1, u_2).

By plugging $(x_1, x_2) = (1, 1)$ into the second of the equations (4.7), we get

$$2 = 2u_1 + u_2 \text{ and } 2 = -u_1 + u_2 .$$

By subtracting these two equations, we get $0 = 3u_1$ so $u_1 = 0$ and $u_2 = 2$. Since both of these values are non-negative, we have found a KKT pair for the original problem. Hence, by convexity, we know that $(x_1, x_2) = (1, 1)$ is the global solution to the problem.

6. Regularity and Constraint Qualifications for NLP

Regularity in nonlinear programming is a powerful tool in the derivation of optimality conditions for (NLP). In this section we examine mild testable conditions under which regularity is guaranteed to hold.

Definition 4.24 (The LICQ and MFCQ Constraint Qualifications). *Let $\bar{x}$ be feasible for (NLP) and put $I(\bar{x}) := \{i \mid c_i(\bar{x}) = 0, i = 1, 2, \ldots, s\}$. We say that*

a) *the linear independence constraint qualification (LICQ) holds at $\bar{x}$ (and write* LICQ($\bar{x}$)*) if the gradients*

$$\nabla c_i(\bar{x}) \ (i \in I(\bar{x})), \quad \nabla c_i(\bar{x}) \ (i = s + 1, \ldots, m)$$

are linearly independent.

b) *the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at $\bar{x}$ (and write* MFCQ($\bar{x}$)*) if the gradients*

$$\nabla c_i(\bar{x}) \ (i = s + 1, \ldots, m)$$

are linearly independent and there exists a vector $d \in \mathbb{R}^n$ such that

$$\nabla c_i(\bar{x})^T d < 0 \ (i \in I(\bar{x})), \quad \nabla c_i(\bar{x})^T d = 0 \ (j = s + 1, \ldots, m). \tag{4.8}$$

The key property of the MFCQ is given in the following technical lemma which relies in the Implicit Function Theorem (see Appendix B, Section 6).

Lemma 4.25 (The Access Lemma for (NLP)). *Let $\bar{x} \in \Omega$ be such that MFCQ is satsified at $\bar{x}$. Then for every direction d satisfying* (4.8) *there exists $\varepsilon > 0$ and a C^1-curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $x(t) \in \Omega$ for all $t \in [0, \varepsilon)$, $x(0) = \bar{x}$ and $x'(0) = d$.*

Proof. Define $C_e : \mathbb{R}^n \to \mathbb{R}^{(m-s)}$ by $C_e(x) := (c_{s+1}, \ldots, c_m(x))^T$ and define $\widehat{C} : \mathbb{R}^{(m-s)+1} \to \mathbb{R}^{(m-s)}$ by

$$\widehat{C}_i(y, t) = c_i(\bar{x} + td + \nabla C_e(\bar{x})^T y) \quad i = s + 1, \ldots, m,$$

where $\nabla C_e(\bar{x})$ denotes the Jacobian of C_e at $\bar{x}$. The nonlinear equation $\widehat{C}(y, t) = 0$ has the solution$(\bar{y}, \bar{t}) = (0, 0)$ with

$$\nabla_y \widehat{C}(0, 0) = \nabla C_e(\bar{x}) \nabla C_e(\bar{x})^T$$

and the latter matrix is non-singular (even positive definite) due to the linear independence of the vectors $\nabla c_i(\bar{x}) \, (i = s+1, \ldots, m)$. The *implicit function theorem* (see Appendix B, Section 6) yields a C^1-function $y : (-\varepsilon, \varepsilon) \to \mathbb{R}^{(m-s)}$ such that $y(0) = 0$, $\widehat{C}(y(t), t) = 0$ and

$$y'(t) = -\nabla_y \widehat{C}(y(t), t)^{-1} \nabla_t \widehat{C}(y(t), t)$$

for all $t \in (-\varepsilon, \varepsilon)$. Hence, we have

$$y'(0) = -\nabla_y \widehat{C}(0, 0)^{-1} \nabla_t \widehat{C}(0, 0) = -\nabla_y \widehat{C}(0, 0)^{-1} \nabla C_e(\bar{x}) d = 0.$$

Now, put $x(t) = \bar{x} + td + \nabla C_e(\bar{x})^T y(t)$ for all $t \in (-\varepsilon, \varepsilon)$. Reducing ε if necessary, $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ has all desired properties. Obviously, $x \in C^1$, $x(0) = \bar{x}$, $x'(0) = d$ and $c_i(x(t)) = 0$ for all $t \in (-\varepsilon, \varepsilon)$ and

$i = s+1, \ldots, m$. Moreover, by continuity, $g_i(x(t)) < 0$ for all $i \notin I(\bar{x})$ and $|t|$ sufficiently small. For $i \in I(\bar{x})$ we have $c_i(x(0)) = c_i(\bar{x}) = 0$ and

$$\frac{d}{dt} c_i(x(0)) = \nabla c_i(\bar{x})^T d < 0$$

and hence $c_i(x(t)) < 0$ for all $t > 0$ sufficiently small. □

THEOREM 4.26 (LICQ and MFCQ imply Regularity). *Let $\bar{x}$ be feasible for (NLP). Then the following implications hold:*

$$\text{LICQ}(\bar{x}) \implies \text{MFCQ}(\bar{x}) \implies [\textit{the representation of } \Omega \textit{ at } \overline{x} \textit{ is regular}]. \tag{4.9}$$

PROOF. LICQ($\bar{x}$) $\Longrightarrow$ MFCQ($\bar{x}$): The vectors $\nabla c_i(\bar{x})$ $(i = s+1, \ldots, m)$ are linear independent by assumption. It remains to find a suitable vector d. For this purpose, consider the matrix

$$\begin{pmatrix} \nabla c_i(\bar{x})^T & (i \in I(\bar{x})) \\ \nabla c_i(\bar{x})^T & (i = s+1, \ldots, m) \end{pmatrix} \in \mathbb{R}^{(|I(\bar{x})|+(m-s))\times n},$$

which has full rank by LICQ($\bar{x}$). Hence we can add rows to obtain non-singular matrix $A(\bar{x}) \in \mathbb{R}^{n\times n}$. Then the linear equation

$$A(\bar{x})d = \begin{pmatrix} -e \\ 0 \end{pmatrix},$$

with $e \in \mathbb{R}^{|I(\bar{x})|}$ being the vector of all ones, has a solution $\hat{d}$, which fulfills the requirements for MFCQ($\bar{x}$).

MFCQ($\bar{x}$) $\Longrightarrow$ [the representation of Ω at $\bar{x}$ is regular]: In view of (4.3), it suffices to show that

$$L(\bar{x}) := \{d \,:\, \nabla c_i(\overline{x})^T d \le 0,\, i \in I(\overline{x}),\ \nabla c_i(\overline{x})^T d = 0, i = s+1, \ldots, m\} \subset T\left(\overline{x} \,|\, \Omega\right).$$

Let $d \in L(\bar{x})$ and $\hat{d}$ given by MFCQ($\bar{x}$) be such that

$$\nabla c_i(\bar{x})^T \hat{d} < 0 \ \forall\, i \in I(\bar{x}), \quad \nabla c_i(\bar{x})^T \hat{d} = 0 \text{ for } i = s+1, \ldots, m.$$

Put $d(t) := d + t\hat{d}$ for $t > 0$. Then for all $t > 0$ we have

$$\nabla c_i(\bar{x})^T d(t) < 0 \ \forall i \in I(\bar{x}), \quad \nabla c_i(\bar{x})^T d(t) = 0 \text{ for } i = s+1, \ldots, m.$$

We claim that this implies $d(t) \in T\left(\overline{x} \,|\, \Omega\right)$ for all $t > 0$. To see this, note that Lemma 4.25 implies the existence a C^1-curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $x(t) \in \Omega$ for all $t \in [0, \varepsilon)$, $x(0) = \bar{x}$ and $x'(0) = d(t)$. For an arbitrary sequence $t_k \downarrow 0$ and $x^k := x(t_k)$ we infer that $x^k \overset{\Omega}{\to} \bar{x}$ and thus

$$d(t) = x'(0) = \lim_{k\to\infty} \frac{x(t_k) - \bar{x}}{t_k - 0} = \lim_{k\to\infty} \frac{x^k - \bar{x}}{t_k} \in T\left(\overline{x} \,|\, \Omega\right).$$

And since $T\left(\overline{x} \,|\, \Omega\right)$ is closed, this implies $d = \lim_{t\downarrow 0} d(t) \in T\left(\overline{x} \,|\, \Omega\right)$. □

The following dual formulation of the MFCQ (obtained through linear programming duality) is often very useful in practice due to its relationship to the KKT conditions given in Definition 4.23.

PROPOSITION 4.27 (Dual Formulation of the MFCQ). *The MFCQ is satisfied at a point $\overline{x} \in \Omega$ if and only if the only solution to the system*

$$\begin{aligned} 0 &= \sum_{i=1}^{m} u_i \nabla c_i(\overline{x}), \\ 0 &= u_i c_i(\overline{x}) && i = 1, 2, \cdots, s, \ \textit{and} \\ 0 &\le u_i && i = 1, 2, \cdots, s, \end{aligned} \tag{4.10}$$

is $u_i = 0,\ i = 1, 2, \cdots, m$.

PROOF. First note that the MFCQ is satisfied at $\overline{x}$ if and only if the gradients $\{\nabla c_i(\overline{x}) \mid i = s+1, \cdots, m\}$ are linearly independent and the following LP is feasible:

$$\begin{array}{lll} \min & 0 & \\ \text{subject to} & \nabla c_i(\overline{x})^T d \leq -1 & i \in I(\overline{x}) \\ & \nabla c_i(\overline{x})^T d = 0 & i = s+1, \cdots, m. \end{array} \tag{4.11}$$

The dual to this LP (4.11) is

$$\begin{array}{ll} \min & \sum_{i \in I(\overline{x})} u_i \\ \text{subject to} & \sum_{i \in I(\overline{x})} u_i \nabla c_i(\overline{x}) + \sum_{i=s+1}^{m} u_i \nabla c_i(\overline{x}) = 0 \\ & 0 \leq u_i, \ i \in I(\overline{x}). \end{array} \tag{4.12}$$

This LP is easily seen to be feasible by taking all u_i's equal to zero. Hence, by the Strong Duality Theorem of Linear Programming, the LP (4.11) is feasible if and only if the LP (4.12) is finite valued in which case the optimal value in both the primal and dual is zero. That is, the MFCQ holds at $\overline{x}$ if and only if the optimal value in (4.12) is zero and the gradients $\{\nabla f_i(\overline{x}) \mid i = s+1, \cdots, m\}$ are linearly independent. The latter statement is equivalent to the statement that the only solution to the system

$$\begin{aligned} \sum_{i=1}^{m} u_i \nabla c_i(\overline{x}) &= 0, \\ u_i c_i(\overline{x}) &= 0 \quad i = 1, 2, \cdots, s, \text{ and} \\ u_i &\geq 0 \qquad i = 1, 2, \cdots, s, \end{aligned}$$

is $u_i = 0, \ i = 1, 2, \cdots, m$. □

The formulation of the MFCQ given in Proposition 4.27 allows us to easily characterize the compacness of the KKT multiplier set in NLP.

THEOREM 4.28 (MFCQ $\iff$ Compact Multiplier Set). *Let $\overline{x} \in \Omega$ be a local solution to (NLP) at which the set of Karush-Kuhn-Tucker multiplier set*

$$KKT(\overline{x}) := \left\{ u \in \mathbb{R}^m \;\middle|\; \begin{array}{c} \nabla_x L(\overline{x}, u) = 0 \\ u_i f_i(\overline{x}) = 0, \ i = 1, 2, \cdots, s, \\ 0 \leq u_i, \ i = 1, 2, \cdots, s \end{array} \right\} \tag{4.13}$$

is non-empty. Then $KKT(\overline{x})$ is a compact set if and only if the MFCQ is satisfied at $\overline{x}$.

PROOF. ($\Rightarrow$) If MFCQ is not satisfied at $\overline{x}$, then from the Strong Duality Theorem for linear programming, Lemma 4.27, and the LP (4.12) guarentees the existence of a non-zero vector $\bar{u} \in \mathbb{R}^m$ satisfying

$$\sum_{i=1}^{m} u_i \nabla f_i(\overline{x}) = 0 \text{ and } 0 \leq u_i \text{ with } 0 = u_i f_i(\overline{x}) \text{ for } i = 1, 2, \cdots, s.$$

Then for each $u \in KKT(\overline{x})$ we have that $u + t\bar{u} \in KKT(\overline{x})$ for all $t > 0$. Consequently, $KKT(\overline{x})$ cannot be compact.
($\Leftarrow$) If $KKT(\overline{x})$ is not compact, there is a sequence $\{u^j\} \subset KKT(\overline{x})$ with $\|u^j\| \uparrow +\infty$. With no loss is generality, we may assume that

$$\frac{u^j}{\|u^j\|} \to u.$$

But then

$$\begin{array}{rll} u_i \geq 0, & i = 1, 2, \cdots, s, & \\ u_i f_i(\overline{x}) = & \lim_{i \to \infty} \frac{u^j}{\|u^j\|} f_i(\overline{x}) = & 0, \ i = 1, 2, \cdots, s, \text{ and} \\ \sum_{i=1}^{m} u_i f_i(\overline{x}) = & \lim_{i \to \infty} \frac{\nabla_x L(\overline{x}, u^j)}{\|u^j\|} = & 0. \end{array}$$

Hence, by Lemma 4.27, the MFCQ cannot be satisfied at $\overline{x}$. □

7. Second-Order Optimality Conditions for Nonlinear Programming

Second–order conditions are introduced by way of the Lagrangian. As is illustrated in the following result, the multipliers provide a natural way to incorporate the curvature of the constraints.

THEOREM 4.29 (Constrained Second-Order Sufficiency). *Let Ω have representation (4.2) and suppose that each of the functions c_i, $i = 0, 1, 2, \dots m$ are twice continuously differentiable. If $\overline{x} \in \Omega$ is such that for every $d \in T(\overline{x} \,|\, \Omega) \setminus \{0\}$ with $\nabla f(\overline{x})^T d = 0$ there exists $\overline{u} \in KKT(\overline{x})$ such that*

$$d^T \nabla_x^2 L(\overline{x}, \overline{u}) d > 0,$$

then there is an $\epsilon > 0$ and $\nu > 0$ such that

$$f(x) \geq f(\overline{x}) + \nu \|x - \overline{x}\|^2$$

for every $x \in \Omega$ with $\|x - \overline{x}\| \leq \epsilon$. In particular, $\overline{x}$ is a strict local solution to (NLP).

PROOF. Suppose to the contrary that no such $\epsilon > 0$ and $\nu > 0$ exist, then there exist sequences $\{x^k\} \subset \Omega$, $\{\nu_k\} \subset \mathbb{R}_+$ such that $x^k \to \overline{x}$, $\nu_k \downarrow 0$, and

$$f(x^k) \leq f(\overline{x}) + \nu_k \|x^k - \overline{x}\|^2 \quad \forall\, k = 1, 2, \dots . \tag{4.14}$$

With no loss of generality, we can assume that

$$d_k := \frac{x^k - \overline{x}}{\|x^k - \overline{x}\|} \to \overline{d} \in T(\overline{x} \,|\, \Omega).$$

By (4.14),

$$f(\overline{x}) + \nabla f(\overline{x})^T (x^k - \overline{x}) + o(\left\|x^k - \overline{x}\right\|) \leq f(\overline{x}) + \nu_k \left\|x^k - \overline{x}\right\|^2 \quad \forall\, k = 1, 2, \dots .$$

Consequently,

$$\nabla f(\overline{x})^T \frac{(x^k - \overline{x})}{\|x^k - \overline{x}\|} + \frac{o(\left\|x^k - \overline{x}\right\|)}{\|x^k - \overline{x}\|} \leq \nu_k \left\|x^k - \overline{x}\right\| \quad \forall\, k = 1, 2, \dots .$$

Taking the limit in k tells us that $\nabla f(\overline{x})^T d \leq 0$. On the other hand,

$$\nabla f(\overline{x})^T d = -\sum_{i \in I(\overline{x})} u_i \nabla c_i(\overline{x})^T d \geq 0 \qquad \forall\ u \in KKT(\overline{x})$$

since

$$T(\overline{x} \,|\, \Omega) \subset \{d : \nabla c_i(\overline{x})^T d \leq 0,\ i \in I(\overline{x}),\ \nabla c_i(\overline{x})^T d = 0,\ i = s+1, \dots, m\}.$$

Consequently, $\nabla f(\overline{x})^T d = 0$, and so, by hypothesis, there is a $\overline{u} \in KKT(\overline{x})$ such that $d^T \nabla_x^2 L(\overline{x}, \overline{u}) d > 0$.

Note that the feasibility of the sequence $\{x^k\}$ implies that

$$\begin{aligned} L(x^k, \overline{u}) \leq f(x^k) \quad &\leq f(\overline{x}) + \nu_k \|x^k - \overline{x}\|^2 \\ &= L(\overline{x}, \overline{u}) + \nu_k \|x^k - \overline{x}\|^2, \end{aligned}$$

and so

$$\begin{aligned} L(\overline{x}, \overline{u}) \quad & + \nabla_x L(\overline{x}, \overline{u})^T (x^k - \overline{x}) \\ & + \tfrac{1}{2} (x^k - \overline{x})^T \nabla_x^2 L(\overline{x}, \overline{u}) (x^k - \overline{x}) + o(\|x^k - \overline{x}\|^2) \\ & \leq L(\overline{x}, \overline{u}) + \nu_k \|x^k - \underline{\mathrm{x}}\|^2 , \end{aligned} \tag{4.15}$$

or equivalently,

$$\tfrac{1}{2} (x^k - \overline{x})^T \nabla_x^2 L(\overline{x}, \overline{u}) (x^k - \overline{x}) + o(\|x^k - \overline{x}\|^2) \leq \nu_k \|x^k - \underline{\mathrm{x}}\|^2.$$

Dividing by $\left\|x^k - \overline{x}\right\|^2$ and taking the limit in k yields $d^T \nabla_x^2 L(\overline{x}, \overline{u}) d \leq 0$. This contradicts our hypothesis, so no such sequences $\{x^k\}$ and $\{\nu_k\}$ can exist and the result is established. □

CHAPTER 5

Convexity

1. Introduction

In the previous chapter we established first- and second-order optimality conditions for the problem

$$\mathcal{P} \qquad \min_{x\in\Omega} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous on an open set containing $\Omega \subset \mathbb{R}^n$. These conditions are based on local information only and so refer to properties of local extrema. In this chapter we study the notion of convexity which allows us to provide global optimality conditions. Convexity is a watershed concept in optimization where the principal devision between problem types in not between linear and nonlinear, rather it is between convex and nonconvex.

Definition 5.1 (Convex Sets and Functions: The Epi-graphical Perspective).

(1) A set $C \subset \mathbb{R}^n$ is said to be a convex set if for every $x, y \in C$ and $\lambda \in [0,1]$ one has

$$(1-\lambda)x + \lambda y \in C\ .$$

(2) Given an extended real-valued function $f : \mathbb{R}^n \to \mathbb{R}_e := \mathbb{R} \cup \{\pm\infty\}$, the epi-graph and domain of f are given by

$$\operatorname{epi}(f) =: \{(x,\mu) \in \mathbb{R}^n \times \mathbb{R} \,|\, f(x) \le \mu\} \quad \text{and} \quad \operatorname{dom} f := \{x \in \mathbb{R}^n \,|\, f(x) \le +\infty\},$$

respectively.

(3) The function $f : \mathbb{R}^n \to \mathbb{R}_e$ is said to be convex if $\operatorname{epi}(f)$ is convex.

(4) The function $f : \mathbb{R}^n \to \mathbb{R}_e$ is said to be closed, or lower semi-continuous (lsc), if $\operatorname{epi}(f)$ is closed.

Example 5.2 (Convex Sets).

(1) Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$ and let $\mathbb{B} := \{x \,|\, \|x\| \le 1\}$ be its unit ball. Then $\mathbb{B}$ is convex since for every $x^1, x^2 \in \mathbb{B}$ and $\lambda \in [0,1]$ we have

$$\left\|(1-\lambda)x^1 + \lambda x^2\right\| \le (1-\lambda)\left\|x^1\right\| + \lambda\left\|x^2\right\| \le 1.$$

(2) Every affine set, i.e. a translate of a subspace, is a convex set. Indeed, $S \subset \mathbb{R}^n$ is affine if and only if $(1-\lambda)x^1 + \lambda x^2 \in S$ for all $x^1, x^2 \in S$ and $\lambda \in \mathbb{R}$.

(3) The positive orthant $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n \,|\, x_1 \ge 0 \ \forall\ i = 1, 2, \ldots, n\}$ is convex.

(4) Given points $x^1, x^2, \ldots, x^k \in \mathbb{R}^n$ the convex hull of these points, defined by

$$\operatorname{conv}\{x^1, x^2, \ldots, x^k\} := \left\{ \sum_{j=1}^k \lambda_j x^j \;\middle|\; \sum_{j=1}^k \lambda_j = 1 \text{ and } 0 \le \lambda_j \ j = 1, \ldots, k \right\},$$

is a convex set since for any two such points $\sum_{j=1}^k \bar\lambda_j x^j$ and $\sum_{j=1}^k \hat\lambda_j x^j$ we have, for all $\lambda \in [0,1]$,

$$(1-\lambda)\left(\sum_{j=1}^k \bar\lambda_j x^j\right) + \lambda\left(\sum_{j=1}^k \hat\lambda_j x^j\right) = \sum_{j=1}^k ((1-\lambda)\bar\lambda_j + \lambda\hat\lambda_j)x^j,$$

where $\sum_{j=1}^k ((1-\lambda)\bar\lambda_j + \lambda\hat\lambda_j) = (1-\lambda)\sum_{j=1}^k \bar\lambda_j + \lambda\sum_{j=1}^k \hat\lambda_j = (1-\lambda) + \lambda = 1$. As an application the reader should show that $\mathbb{B}_1 = \operatorname{conv}\{\pm e_i,\ i = 1, \ldots, n\}$.

Definition 5.1 differs from the typically definition of a convex function since it emphasizes the epigraphical perspective. The great advantage of the epigraphical approach is that it allows us to directly translate facts about convex sets to facts about convex functions. In addition, it allows a useful trick in handling constraints. Observe that if the set $\Omega \subset \mathbb{R}^n$ is a convex set, then the function

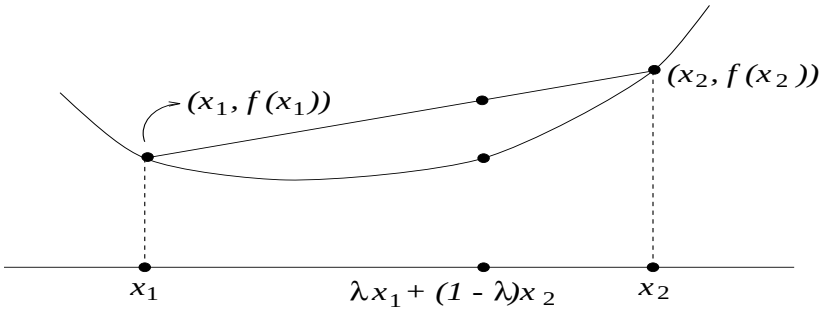
$$\delta(x \,|\, \Omega) := \begin{cases} 0, & \text{if } x \in \Omega, \text{ and} \\ +\infty, & \text{if } x \notin \Omega \end{cases}$$

is also convex. This function is called the convex *indicator function* for Ω. Using the indicator function we can write the optimization problem $\mathcal{P}$ as $\min_{x\in\mathbb{R}^n} \hat{f}(x) := f(x) + \delta(x \,|\, \Omega)$, where the function $\hat{f}$ defined in this way is a convex function when f and Ω are convex since the sum of any two convex functions is convex.

The convexity of $\operatorname{epi}(f)$ implies that for every pair of points $x_1, x_2 \in \operatorname{dom} f$ the secant line connecting $(x_1, f(x_1))$ and $(x_2, f(x_2))$ lies above the graph of f on the line segment $\lambda x_1 + (1-\lambda)x_2,\ \lambda \in [0,1]$ as illustrated in the graph below. Consequently,

$$f((1-\lambda)x^1 + \lambda x^2) \le (1-\lambda)f(x^1) + \lambda f(x^2) \qquad \forall\ x_1, x_2 \in \operatorname{dom} f. \tag{5.1}$$

The next lemma shows that the converse is also true, that is, if (5.2) holds, then f is convex.



Lemma 5.3 (Convexity and Secant Lines). *The function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ is convex if and only if, for every $x^1, x^2 \in \operatorname{dom} f$ and $\lambda \in [0,1]$, we have*

$$f((1-\lambda)x^1 + \lambda x^2) \le (1-\lambda)f(x^1) + \lambda f(x^2). \tag{5.2}$$

That is, the secant line connecting $(x^1, f(x^1))$ and $(x^2, f(x^2))$ lies above the graph of f.

Proof. Clearly we need only show that (5.2) implies convexity since the reverse implication follows from the definition of a convex set. To this end let $(x_i, \mu_i) \in \operatorname{epi}(f)$, $i = 1, 2$, and $\lambda \in [0,1]$. Then

$$f((1-\lambda)x^1 + \lambda x^2) \le (1-\lambda)f(x^1) + \lambda f(x^2) \le (1-\lambda)\mu_1 + \lambda\mu_2.$$

Therefore, $(1-\lambda)(x_1,\ \mu_1) + \lambda(x_2,\ \mu_2) = ((1-\lambda)x^1 + \lambda x^2,\ (1-\lambda)\mu_1 + \lambda\mu_2) \in \operatorname{epi}(f)$. Hence $\operatorname{epi}(f)$ is convex. □

Definition 5.4 (Strict Convexity). *The function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be strictly convex if*

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda)f(x_2) \quad \forall\ x^1, x^2 \in \operatorname{dom} f \text{ and } 0 < \lambda < 1.$$

Example 5.5 (Convex Functions).

(1) **Norms:** *Every norm is a convex function since it satisfies the triangle inequality and is positive homogeneous.*

(2) **Support Functions:** *Given a nonempty set $S \subset \mathbb{R}^n$, the support function associated with S is defined to be the function*

$$\sigma_S(x) := \sup\{\langle x, v\rangle \,|\, v \in S\}.$$

The convexity of a support function follows from the fact that they are sublinear, that is, they are both positively homogeneous and subadditive:

$$\begin{aligned}\sigma_S(\lambda x) &= \sup\{\langle \lambda x,\, v\rangle \mid v \in S\}\\ &= \lambda \sup\{\langle x,\, v\rangle \mid v \in S\}\\ &= \lambda\sigma_S(x) \quad \forall\, \lambda \geq 0. && \textit{(positive homogeneity)}\end{aligned}$$

$$\begin{aligned}\sigma_S\left(x^1+x^2\right) &= \sup\left\{\left\langle x^1+x^2,\, v\right\rangle \mid v \in S\right\}\\ &= \sup\left\{\left\langle x^1,\, v^1\right\rangle + \left\langle x^2,\, v^2\right\rangle \mid v^1 = v^2 \in S\right\}\\ &\leq \sup\left\{\left\langle x^1,\, v^1\right\rangle + \left\langle x^2,\, v^2\right\rangle \mid v^1, v^2 \in S\right\}\\ &\leq \sup\left\{\left\langle x^1,\, v^1\right\rangle \mid v^1 \in S\right\} + \sup\left\{\left\langle x^2,\, v^2\right\rangle \mid v^2 \in S\right\}\\ &= \sigma_S\left(x^1\right) + \sigma_S\left(x^2\right) \quad \forall\, x^1, x^2 \in \mathbb{R}^n. && (\textit{subadditivity})\end{aligned}$$

A function that is both sublinear and subadditive is called sublinear. Observe that if S is a convex polyhedron, then evaluating $\sigma_S(x)$ is equivalent to solving a linear program.

(3) **Quadratic Functions:** *The quadratic function $f(x) := \frac{1}{2}x^TQx$ is convex if and only if $Q \in \mathcal{S}^n_+$. Moreover, f is strictly convex if and only if $Q \in \mathcal{S}^n_{++}$. These facts follow immediately from the equivalence*

$$\begin{aligned}((1-\lambda)x+\lambda y)^TQ((1-\lambda)x+\lambda y) &- (1-\lambda)x^TQx - \lambda y^TQy\\ &= (1-\lambda)^2x^TQx + 2\lambda(1-\lambda)x^TQy + \lambda^2y^TQy - (1-\lambda)x^TQx - \lambda y^TQy\\ &= (1-\lambda)(1-\lambda-1)x^TQx + 2\lambda(1-\lambda)x^TQy + \lambda(\lambda-1)y^TQy\\ &= -\lambda(1-\lambda)(x^TQx - 2x^TQy + y^TQy)\\ &= -\lambda(1-\lambda)(x-y)^TQ(x-y).\end{aligned}$$

The prominent role of convexity in optimization theory is illustrated by the following result.

THEOREM 5.6 (Convexity ⇒ Local = Global for $\mathcal{P}$). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be convex. If $\overline{x} \in \operatorname{dom} f$ is a local solution to the problem $\mathcal{P}$, then $\overline{x}$ is a global solution to the problem $\mathcal{P}$. Moreover, if f is strictly convex, then $\mathcal{P}$ has at most one global solution.*

PROOF. If $f(\bar{x}) = -\infty$ we are done, so let us assume that $-\infty < f(\bar{x})$. Suppose there is a $\widehat{x} \in \mathbb{R}^n$ with $f(\widehat{x}) < f(\overline{x})$. Let $\epsilon > 0$ be such that $f(\overline{x}) \leq f(x)$ whenever $\|x - \overline{x}\| \leq \epsilon$. Consequently, $\|\overline{x} - \hat{x}\| > \epsilon$ and $\lambda := \epsilon(2\|\overline{x} - \widehat{x}\|)^{-1} < 1$. Set $x_\lambda := \overline{x} + \lambda(\widehat{x} - \overline{x})$. Then $\|x_\lambda - \overline{x}\| \leq \epsilon/2$ and $f(x_\lambda) \leq (1-\lambda)f(\overline{x}) + \lambda f(\widehat{x}) < f(\overline{x})$. This contradicts the choice of ϵ, hence no such $\widehat{x}$ exists.

To see the second statement in the theorem, let x^1 and x^2 be distinct global minimizers of f. Then, for $\lambda \in (0,1)$,

$$f((1-\lambda)x^1 + \lambda x^2) < (1-\lambda)f(x^1) + \lambda f(x^2) = f(x^1)\ ,$$

which contradicts the assumption that x^1 is a global minimizer. □

2. Properties of the Directional Derivative

Convex functions possess a very rich variational structure. This structure is extremely useful in the construction of optimization algorithms.

LEMMA 5.7 (Sublinearity of $f'(x;d)$). *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex.*

(1) Given $x \in \operatorname{dom} f$ and $d \in \mathbb{R}^n$ the difference quotient

$$\frac{f(x+td) - f(x)}{t} \tag{5.3}$$

is a non-decreasing function of t on $(0, +\infty)$.

(2) For every $x \in \operatorname{dom} f$ and $d \in \mathbb{R}^n$ the directional derivative $f'(x;d)$ always exists and is given by

$$f'(x;d) := \inf_{t>0} \frac{f(x+td)-f(x)}{t}. \tag{5.4}$$

(3) For every $x \in \operatorname{dom} f$, the function $f'(x;\cdot)$ is sublinear, i.e. $f'(x;\cdot)$ is positively homogeneous,

$$f'(x;\alpha d) = \alpha f'(x;d) \quad \forall\ d \in \mathbb{R}^n,\ 0 \le \alpha,$$

and subadditive,

$$f'(x;u+v) \le f'(x;u) + f'(x;v).$$

In particular, for all $x \in \operatorname{dom} f$, $f'(x;d)$ is a convex function of d.

REMARK 5.8. *It is important to note that the infimum on the right-hand side of* (5.4) *may be $\pm\infty$. For example, consider the convex function*

$$f(x) := \begin{cases} -\sqrt{1-x^2}, & \text{if } |x| \le 1, \\ +\infty, & \text{otherwise.} \end{cases}$$

It is easily shown that $f'(1;1) = +\infty$ and $f'(1;-1) = -\infty$.

PROOF. (1) Let $x \in \operatorname{dom} f$ and $d \in \mathbb{R}^n$. If $x + td \notin \operatorname{dom} f$ for all $t > 0$, then the result is obviously true. Thus, we may assume that

$$0 < \bar{t} = \sup\{t \,:\, x + td \in \operatorname{dom} f\}.$$

Let $0 < t_1 < t_2 < \bar{t}$ (we allow the possibility that $t_2 = \bar{t}$ if $\bar{t} < +\infty$). Then

$$\begin{aligned} f(x+t_1 d) &= f\left(x + \left(\tfrac{t_1}{t_2}\right) t_2 d\right) \\ &= f\left[\left(1 - \left(\tfrac{t_1}{t_2}\right)\right) x + \left(\tfrac{t_1}{t_2}\right)(x + t_2 d)\right] \\ &\le \left(1 - \tfrac{t_1}{t_2}\right) f(x) + \left(\tfrac{t_1}{t_2}\right) f(x + t_2 d). \end{aligned}$$

Hence

$$\frac{f(x+t_1 d) - f(x)}{t_1} \le \frac{f(x+t_2 d) - f(x)}{t_2}.$$

(2) We use (5.3) to show (5.4). If $x + td \notin \operatorname{dom} f$ for all $t > 0$, then the result is obviously true. Therefore, we may as well assume that there is a $\bar{t} > 0$ such that $x + td \in \operatorname{dom} f$ for all $t \in (0, \bar{t}]$. Recall that

$$f'(x;d) := \lim_{t \downarrow 0} \frac{f(x+td) - f(x)}{t}. \tag{5.5}$$

Now if the difference quotient (5.3) is non-decreasing in t on $(0, +\infty)$, then the limit in (5.5) is necessarily given by the infimum in (5.4). This infimum always exists and so $f'(x;d)$ always exists and is given by (5.4).

(3) To see that $f'(x;\cdot)$ is positively homogeneous let $d \in \mathbb{R}^n$ and $\alpha > 0$ and note that

$$f'(x;\alpha d) = \alpha \lim_{t \downarrow 0} \frac{f(x + (t\alpha)d) - f(x)}{(t\alpha)} = \alpha f'(x;d).$$

To see that $f'(x;\cdot)$ is subadditive let $u, v \in \mathbb{R}^n$, then

$$\begin{aligned} f'(x;u+v) &= \lim_{t\downarrow 0} \frac{f(x+t(u+v))-f(x)}{t} \\ &= \lim_{t\downarrow 0} \frac{f(x+\frac{t}{2}(u+v))-f(x)}{t/2} \\ &= \lim_{t\downarrow 0} 2\frac{f(\frac{1}{2}(x+tu)+\frac{1}{2}(x+tv))-f(x)}{t} \\ &\leq \lim_{t\downarrow 0} 2\frac{\frac{1}{2}f(x+tu)+\frac{1}{2}f(x+tv)-f(x)}{t} \\ &= \lim_{t\downarrow 0} \frac{f(x+tu)-f(x)}{t} + \frac{f(x+tv)-f(x)}{t} \\ &= f'(x;u)+f(x;v)\ . \end{aligned}$$

□

The representation of the directional derivative in (5.4) is extremely useful. In particular, if we set $t = 1$ and $d = y - x$ into the right hand side of (5.4), where y is *any* other point in $\mathbb{R}^n$, we obtain the celebrated *subdifferential inequality*

$$f(y) \geq f(x) + f'(x; y-x) \quad \text{for all} \quad y \in \mathbb{R}^n \text{ and } x \in \operatorname{dom} f\ . \tag{5.6}$$

The subdifferential inequality immediately yields the following powerful result.

THEOREM 5.9 (Convexity and Optimality). *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex, $\Omega \subset \mathbb{R}^n$ convex, $\overline{x} \in \operatorname{dom} f \cap \Omega$. Then $\overline{x}$ solves $\min_{x\in\Omega} f(x)$ if and only if $f'(\overline{x}; y-\overline{x}) \geq 0$ for all $y \in \Omega$.*

PROOF. Let $y \in \Omega$ so that $\overline{x} + t(y-\overline{x}) \in \Omega$ for all $t \in [0,1]$. Then $f(\overline{x}) \leq f(\overline{x}+t(y-\overline{x}))$ for all $t \in [0,1]$. Therefore, $f'(\overline{x}; y-\overline{x}) = \lim_{t\downarrow 0} t^{-1}(f(\overline{x}+t(y-\overline{x})) - f(\overline{x})) \geq 0$. The reverse implication is an immediate consequence of the subdifferential inequality (5.6). □

In order to establish the full implication for the differentiable case it is helpful to have a more refined understanding of the tangent cone to a convex set.

LEMMA 5.10 (The Tangent Cone to a Convex Set). *Let $\Omega \subset \mathbb{R}^n$ be convex with $\overline{x} \in \Omega$. Then*

$$T(\overline{x}\,|\,\Omega) = \operatorname{cl}\left(\bigcup_{\lambda\geq 0} \lambda(\Omega - \overline{x})\right) := \operatorname{cl}\left(\{\lambda(x-\overline{x})\,|\,\lambda \geq 0 \text{ and } x \in \Omega\}\right). \tag{5.7}$$

PROOF. Let $x \in \Omega$. Then $x_\mu := \overline{x} + \mu(x-\overline{x}) \in \Omega$ for all $\mu \in [0,1]$. Let $\lambda > 0$ and $\{t_k\} \subset [0, \min\{1, \lambda^{-1}]$ be any sequence converging To zero and set $x^k := x_{\lambda t_k}$. Then $\lambda(x-\overline{x}) = lim_k t_k^{-1}(x^k - \overline{x}) \in T(\overline{x}\,|\,\Omega)$. Since $T(\overline{x}\,|\,\Omega)$ is closed, then set on the right hand side of (5.7) is contained in the left hand side.

On the other hand, if $d \in T(\overline{x}\,|\,\Omega)$, then there is a sequence $\{x^k\} \subset \Omega$ and a sequence $\{t_k\} \subset \mathbb{R}_+$ converging to zero such that $d^k := t_k^{-1}(x^k - \overline{x}) \to d$. But $d^k \in \bigcup_{\lambda\geq 0} \lambda(\Omega - \overline{x})$ for all $k = 1, 2, \ldots$. Hence d is in the set on the right hand side of (5.7). □

COROLLARY 5.10.1. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, Ω, and $\overline{x}$ be as in Theorem 5.9, and, in addition, assume that f is differentiable at $\overline{x}$. Then the following are equivalent:*

(1) $\overline{x}$ solves $\min_{x\in\Omega} f(x)$.
(2) $\nabla f(\overline{x})^T(x-\overline{x}) \geq 0$ for all $x \in \Omega$.
(3) $\nabla f(\overline{x})^T d \geq 0$ for all $d \in T(\overline{x}\,|\,\Omega)$.

In particular, if $\overline{x} \in \operatorname{intr}(\Omega)$, then $\nabla f(\overline{x}) = 0$.

PROOF. This is an immediate consequence of Lemma 5.10, Theorem 5.9 and the continuity of the mapping $d \mapsto \nabla f(\overline{x})^T d = f'(\overline{x}; d)$. □

This corollary provides the basis for a dual interpretation of the optimality condition. Given a set $C \subset \mathbb{R}^n$, define the *polar* of C to be the set

$$C^\circ := \{z \mid \langle z, x\rangle \leq 1 \ \forall\, x \in C\}.$$

The polar is necessarily a closed convex set since it is the intersection of a collection of closed half space indexed by the elements of C (see Theorem 5.15). Polar convex sets are discussed in more depth later. If K is a non-empty cone ($\lambda K \subset K \ \forall\, \lambda > 0$), then

$$K^\circ = \{z \mid \langle z, x\rangle \leq 0 \ \forall\, x \in K\}$$

is a closed convex cone. Finally, if K is also a closed convex cone, then $K^{\circ\circ} := (K^\circ)^\circ = K$.

We define the *normal cone* to Ω at x to be the polar of the tangent cone $T(x \mid \Omega)$:

$$N(x \mid \Omega) := T(x \mid \Omega)^\circ.$$

We immediately have the following corollary to Corollary 5.10.2.

COROLLARY 5.10.2. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, Ω, and $\overline{x}$ be as in Corollary 5.10.2. Then $\overline{x}$ solves $\min_{x\in\Omega} f(x)$ if and only if $-\nabla f(\overline{x}) \in N(\overline{x} \mid \Omega)$.*

REMARK 5.11. *Let $\Omega \subset \mathbb{R}^n$ be non-empty closed and convex with $x \in \Omega$.*

(1) If $x \in \operatorname{intr}(\Omega)$, then $T(x \mid \Omega) = \mathbb{R}^n$ and $N(x \mid \Omega) = \{0\}$.
(2) If $\Omega := \{x\}$, then $T(x \mid \Omega) = \{0\}$ and $N(x \mid \Omega) = \mathbb{R}^n$.
(3) If $x = (x^1, x^2, \ldots, x^k)$ and $\Omega := \Omega_1 \times \Omega_2 \times \cdots \times \Omega_k$ where $x^j \in \Omega_j \subset \mathbb{R}^{n_j}$, $j = 1, \ldots, k$ with $n = \sum_{j=1}^k n_j$, then

$$T(x \mid \Omega) = T\left(x^1 \,\middle|\, \Omega_1\right) \times \cdots \times T\left(x^k \,\middle|\, \Omega_k\right) \quad \text{and} \quad N(x \mid \Omega) = N\left(x^1 \,\middle|\, \Omega_1\right) \times \cdots \times N\left(x^k \,\middle|\, \Omega_k\right).$$

3. Local Lipschitz Continuity of Convex Functions

Recall that a function $F : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *Lipschitz continuous* on a set $S \subset \mathbb{R}^n$ if there is a constants $L > 0$ such that

$$\|F(x) - F(y)\| \leq L\,\|x - y\| \quad \forall\, x, y \in S.$$

The function F is said to be *locally Lipschitz* on an open set $V \subset \mathbb{R}^n$ if for every $\overline{x} \in V$ there is an $\epsilon > 0$ and $L > 0$ such that

$$\|F(x) - F(y)\| \leq L\,\|x - y\| \quad \forall\, x, y \in \overline{x} + \epsilon\mathbb{B} \subset V.$$

In this section we establish the remarkable fact that a convex function is locally Lipschitz continuous on the interior of its domain. It is possible to generalize this results to convex functions whose domains have no interior. But this requires an understanding of the relative topology of convex sets. We begin by establishing the local boundedness of a convex function on the interior of its domain.

PROPOSITION 5.12 (Local Boundedness of Convex Functions). *Let $\mathbb{R}^n : \overline{\mathbb{R}} \to$ be convex. Then for every $\overline{x} \in \operatorname{intr}(\operatorname{dom}) f$ and $\epsilon > 0$ such that $\overline{x} + \epsilon\mathbb{B}_1 \subset \operatorname{dom} f$ there is an $M > 0$ such that $f(x) \leq M$ for all $x \in \overline{x} + \epsilon\mathbb{B}_1$.*

PROOF. Let $\overline{x} \in \operatorname{intr}(\operatorname{dom}) f$ and $\epsilon > 0$ such that $\overline{x} + \epsilon\mathbb{B}_1 \subset \operatorname{dom} f$. Since $\mathbb{B}_1 = \operatorname{conv}\{\pm e_i,\ i = 1, \ldots, n\}$, it is easily seen that $\overline{x} + \epsilon\mathbb{B}_1 = \operatorname{conv}\{\overline{x} \pm \epsilon e_i,\ i = 1, \ldots, n\}$. Consequently, for every $x \in \overline{x} + \epsilon\mathbb{B}_1$ there exit $0 \leq \lambda_i, \mu_i$, $i = 1, \ldots, n$ with $\sum_{j=1}^n (\lambda_i + \mu_i) = 1$ such that $x = \sum_{j=1}^n \lambda_i(\overline{x} + \epsilon e_i) + \sum_{j=1}^n \mu_i(\overline{x} - \epsilon e_i)$. Therefore,

$$f(x) \leq \sum_{j=1}^n \lambda_i f(\overline{x} + \epsilon e_i) + \sum_{j=1}^n \mu_i f(\overline{x} - \epsilon e_i) \leq \max\{f(\overline{x} + \epsilon e_i), f(\overline{x} - \epsilon e_i) \mid i = 1, \ldots, n\} =: M.$$

□

Lipschitz continuity now follows.

THEOREM 5.13 (Lipschitz Continuity of Convex Function). *Let $\mathbb{R}^n : \overline{\mathbb{R}} \to$ be convex. Then for every $\overline{x} \in \operatorname{intr}(\operatorname{dom}) f$ there is an $\epsilon > 0$ such that f is Lipschitz continuous on $\overline{x} + \epsilon\mathbb{B}$.*

PROOF. By Proposition 5.12, let $\epsilon > 0$ and $M > 0$ be such that $\overline{x}+2\epsilon\mathbb{B} \subset \operatorname{intr}(\operatorname{dom}) f$ with $f(x) \leq M$ for all $x \in \overline{x}+2\epsilon\mathbb{B}_1$. Set $h(x) := (2M)^{-1}[f(x+\overline{x})-f(\overline{x})]$. We need only show that h is Lipschitz continuous near 0 to show that f is Lipschitz continuous near $\overline{x}$. Observe that $h(0) = 0$ and $h(x) \leq 1$ for all $x \in 2\epsilon\mathbb{B}_1$. Moreover, for every $x \in 2\epsilon\mathbb{B}_1$, $0 = h(0) = h(\frac{1}{2}x - \frac{1}{2}x) \leq \frac{1}{2}h(x) + \frac{1}{2}h(-x)$ so that $-1 \leq -h(x) \leq h(-x)$. That is, $-1 \leq h(x) \leq 1$ for all $x \in 2\epsilon\mathbb{B}_1$. For $x, y \in \epsilon\mathbb{B}_1$ with $x \neq y$ set $\alpha := \|x - y\|_1$ and $\beta := \epsilon/\alpha$. Define $w := y + \beta(y - x) \in 2\epsilon\mathbb{B}_1$. Then

$$y = (1 + \beta)^{-1}[w + \beta x] = \frac{1}{1+\beta}w + \frac{\beta}{1+\beta}x.$$

The convexity of h implies that

$$\begin{aligned} h(y) - h(x) &\leq \frac{1}{1+\beta}h(w) + \frac{\beta}{1+\beta}h(x) - h(x) \\ &= \frac{1}{1+\beta}[h(w) - h(x)] \leq \frac{2}{1+\beta} = \frac{2}{\alpha+\epsilon}\|x - y\|_1 \leq 2\epsilon^{-1}\|x - y\|_1 . \end{aligned}$$

Since the argument is symmetric in x and y, we have established the local Lipshitz continuity of h. □

4. Tests for Convexity

As Theorems 5.6 and 5.9 as well as Corollary 5.10.2 demonstrate, convex functions are well suited to optimization theory. For this reason, it is important that we are able to recognize when a function is convex.

THEOREM 5.14 (Convexity Tests). *Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$.*

(1) If f is differentiable on $\mathbb{R}^n$, then the following statements are equivalent:
   (a) f is convex,
   (b) $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in \mathbb{R}^n$
   (c) $(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0$ for all $x, y \in \mathbb{R}^n$.

(2) If f is twice differentiable then f is convex if and only if $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^n$.

PROOF. (a) ⇒ (b) If f is convex, then 5.14 holds. By setting $t := 1$ and $d := y - x$ we obtain (b).

(b) ⇒ (c) Let $x, y \in \mathbb{R}^n$. From (b) we have

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

and

$$f(x) \geq f(y) + \nabla f(y)^T(x - y).$$

By adding these two inequalities we obtain (c).

(c) ⇒ (b) Let $x, y \in \mathbb{R}^n$. By the Mean Value Theorem there exists $0 < \lambda < 1$ such that

$$f(y) - f(x) = \nabla f(x_\lambda)^T(y - x)$$

where $x_\lambda := \lambda y + (1 - \lambda)x$. By hypothesis,

$$\begin{aligned} 0 &\leq [\nabla f(x_\lambda) - \nabla f(x)]^T(x_\lambda - x) \\ &= \lambda[\nabla f(x_\lambda) - \nabla f(x)]^T(y - x) \\ &= \lambda[f(y) - f(x) - \nabla f(x)^T(y - x)]. \end{aligned}$$

Hence $f(y) \geq f(x) + \nabla f(x)^T(y - x)$.

(b) $\Rightarrow$ (a) Let $x, y \in \mathbb{R}^n$ and set

$$\alpha := \max_{\lambda \in [0,1]} \varphi(\lambda) := [f(\lambda y + (1-\lambda)x) - (\lambda f(y) + (1-\lambda)f(x))].$$

We need to show that $\alpha \leq 0$. Since $[0,1]$ is compact and φ is continuous, there is a $\lambda \in [0,1]$ such that $\varphi(\lambda) = \alpha$. If λ equals zero or one, we are done. Hence we may as well assume that $0 < \lambda < 1$ in which case

$$0 = \varphi'(\lambda) = \nabla f(x_\lambda)^T (y - x) + f(x) - f(y)$$

where $x_\lambda = x + \lambda(y - x)$, or equivalently

$$\lambda f(y) = \lambda f(x) - \nabla f(x_\lambda)^T (x - x_\lambda).$$

But then

$$\begin{aligned} \alpha &= f(x_\lambda) - (f(x) + \lambda(f(y) - f(x))) \\ &= f(x_\lambda) + \nabla f(x_\lambda)^T (x - x_\lambda) - f(x) \\ &\leq 0 \end{aligned}$$

by (b).

2) Suppose f is convex and let $x, d \in \mathbb{R}^n$, then by (b) of Part (1),

$$f(x + td) \geq f(x) + t \nabla f(x)^T d$$

for all $t \in \mathbb{R}$. Replacing the left hand side of this inequality with its second-order Taylor expansion yields the inequality

$$f(x) + t \nabla f(x)^T d + \frac{t^2}{2} d^T \nabla^2 f(x) d + o(t^2) \geq f(x) + t \nabla f(x)^T d,$$

or equivalently,

$$\frac{1}{2} d^t \nabla^2 f(x) d + \frac{o(t^2)}{t^2} \geq 0.$$

Letting $t \to 0$ yields the inequality

$$d^T \nabla^2 f(x) d \geq 0.$$

Since d was arbitrary, $\nabla^2 f(x)$ is positive semi-definite.

Conversely, if $x, y \in \mathbb{R}^n$, then by the Mean Value Theorem there is a $\lambda \in (0,1)$ such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x_\lambda)(y - x)$$

where $x_\lambda = \lambda y + (1 - \lambda)_x$. Hence

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

since $\nabla^2 f(x_\lambda)$ is positive semi-definite. Therefore, f is convex by (b) of Part (1).

□

5. Building New from Old

This section is a brief introduction to the calculus of convex sets and functions. The calculus shows how to construct new convex sets and functions from known convex sets and functions.

THEOREM 5.15 (Convex Set Calculus). *Let C and C_ν ($\nu \in N$) be convex sets where N is an arbitrary index set, and let $\nu_i \in N$ and $\lambda_i \in \mathbb{R}$ for $i = 1, \ldots, m$. Also, let $A, B^T \in \mathbb{R}^{k \times n}$. Then the following sets are also convex.*

(1) $AC := \{Ax \,|\, x \in C\}$ *(linear image)*
(2) $B^{-1}C := \{z \,|\, Bz \in C\}$ *(inverse linear image)*
(3) $\sum_{i=1}^m \lambda_i C_{\nu_i} := \left\{ \sum_{i-1}^m \lambda_i x^i \,\middle|\, x^i \in C_{\nu_i} \right\}$ *(linear composition)*
(4) $\bigcap_{\nu \in I} C_\nu$ *(intersection)*

(5) $C_{\nu_1} \times C_{\nu_2} \times \cdots \times C_{\nu_m}$ *(product sets)*

PROOF. These are all straightforward and left to the reader as exercises. Note that some of them follow from the others. □

THEOREM 5.16 (Convex Function Calculus). *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $h : \mathbb{R}^s \times \mathbb{R}^k \to \overline{\mathbb{R}}$ and $f_\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex functions for $\nu \in N$ where N is an arbitrary index set, and let $\nu_i \in N$ and $\lambda_i \geq 0,\ i = 1, \ldots, m$. Also, let $A \in \mathbb{R}^{n\times k}$ and $b \in \mathbb{R}^n$. Then the following functions are also convex.*

(1) $\psi := \phi \circ f$, *where* $\phi : \mathbb{R} \to \mathbb{R}$ *is non-decreasing and convex.* *(composition)*
(2) $\psi(z) := f(Az + b)$ *(affine composition)*
(3) $\psi(z) := \inf\left\{ f(x) \,\middle|\, A^T x = z \right\}$ *(inverse linear composition)*
(4) $\psi(x) := \sum_{i=1}^m \lambda_i f_{\nu_i}(x)$ *(Non-negative linear combinations)*
(5) $\psi(x) := \sup_{\nu\in N} f_\nu(x)$ *(pointwise max)*
(6) $\psi(x) := \inf\left\{ \sum_{i=1}^m f_{\nu_i}(x^i) \,\middle|\, x = \sum_{i=1}^m x^i \right\}$ *(infimal convolution)*
(7) $f^*(y) := \sup_{x\in\mathbb{R}^n}[y^T x - f(x)]$ *(convex conjugation)*
(8) $\psi(y) = \inf_{x\in\mathbb{R}^s} h(x, y)$ *(infimal projection)*

PROOF. All of these facts are either straightforward of follow directly from item (8) and Theorem 5.15. We only prove (8). Let $y^1, y^2 \in \operatorname{dom}\psi$ and $\epsilon > 0$. Then there exit $x^i \in \mathbb{R}^s$ such that $h(x^i, y^i) < \psi(y^i) + \epsilon$ so that $((x^i, y^i), \psi(y^i) + \epsilon) \in \operatorname{epi}(h)$, $i = 1, 2$. Consequently, for $\lambda \in [0, 1]$,

$$\left(((1-\lambda)x^1 + \lambda x^2,\ (1-\lambda)y^1 + \lambda y^2),\ (1-\lambda)(\psi(y^1) + \epsilon) + \lambda(\psi(y^2) + \epsilon)\right) \in \operatorname{epi}(h),$$

which implies that

$$\begin{aligned}\psi((1-\lambda)y^1 + \lambda y^2 &\leq h((1-\lambda)x^1 + \lambda x^2,\ (1-\lambda)y^1 + \lambda y^2)\\ &\leq (1-\lambda)(\psi(y^1) + \epsilon) + \lambda(\psi(y^2) + \epsilon) = (1-\lambda)\psi(y^1) + \lambda\psi(y^2) + \epsilon.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we find that ψ is convex. □

6. Convex Nonlinear Programming

We now return to the nonlinear programming problem (NLP) studied in Chapter 4 Sections 5-7 where the constraint region takes the form given in (4.2):

$$\Omega := \{x : c_i(x) \leq 0, i = 1, \ldots, s, c_i(x) = 0, i = s+1, \ldots, m\}, \tag{5.8}$$

with each $c_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on $\mathbb{R}^n$. However, now we specify that

$$\begin{aligned}&c_i,\ i = 1, \ldots, s \ \text{ are convex and differentiable, and}\\ &c_i,\ i = s+1, \ldots, m \ \text{ are affine.}\end{aligned} \tag{5.9}$$

These additional hypotheses guarantee that Ω is a closed convex set. Since $c_i,\ i = s+1, \ldots, m$ are affine functions, there exists $A \in \mathbb{R}^{(m-s)\times n}$ and $b \in \mathbb{R}^{(m-s)}$ such that

$$\begin{pmatrix} c_{s+1}(x) \\ c_{s+2}(x) \\ \vdots \\ c_m(x) \end{pmatrix} = Ax - b. \tag{5.10}$$

With no loss in generality we may assume that A is surjective ($\operatorname{rank}(A) = m - s$) since otherwise we can row reduce A to obtain an equivalent linear system that is surjective.

For ease of reference, we provide a notation for specifying the version of the nonlinear programming problem discussed in this section:

$$\begin{aligned}\text{(C-NLP)}\quad &\text{minimize} && f(x)\\ &\text{subject to} && c_i(x) \leq 0,\ i = 1, \ldots, s,\\ &&& c_i(x) = 0,\ i = s+1, \ldots, m,\end{aligned}$$

where the functions c_i satisfy (5.9) and the matrix A in (5.10) is surjective.

Our goal is to obtain optimality conditions based on Lagrange multipliers as in Theorem 4.22. As we will see the Lagrange multiplier theory reaches its full potential in the convex case leading to a powerful *convex duality theory*. Recall from Section 6 in Chapter 4 that multipliers are only guaranteed to exist if a constraint qualification is satisfied. In the convex setting the standard constraint qualification is called the *Slater constraint qualification*

DEFINITION 5.17 (Slater Constraint Qualification). *The set Ω in* (5.8) *under the convexity hypotheses* (5.9) *is said to satisfy the Slater constraint qualification if there is an $\hat{x} \in \Omega$ such that*

$$c_i(\hat{x}) < 0,\ i = 1, \ldots, s \quad \textit{and} \quad c_i(\hat{x}) = 0,\ i = s+1, \ldots, m. \tag{5.11}$$

THEOREM 5.18 (Slater CQ = MFCQ). *Let Ω be as given in* (5.8) *under the convexity hypotheses* (5.9) *where it is assumed that the matrix A in* (5.10) *is surjective. If the Slater CQ is satisfied, then the MFCQ is satisfied at every point of Ω. Conversely, if the MFCQ is satisfied at a single point of Ω, then the Slater CQ is satisfied.*

PROOF. First let us suppose that the Slater CQ is satisfied and let $\overline{x} \in \Omega$. We show that the dual form of the MFCQ described in Proposition 4.27 holds at $\overline{x}$. To this end, let the multipliers $u \in \mathbb{R}^m$ satisfy (4.10), and $\hat{x} \in \Omega$ satisfy (5.11). Then $0 > c_i(\hat{x}) \geq c_i(\overline{x}) + \nabla c_i(\overline{x})^T(\hat{x} - \overline{x}),\ i = 1, \ldots, s$. Multiplying by the u_i's, summing over all i and applying (5.11) tells us that

$$0 \geq \sum_{i=1}^m u_i c_i(\hat{x}) \geq \sum_{i=1}^m u_i(c_i(\overline{x}) + \nabla c_i(\overline{x})^T(\hat{x} - \overline{x})) = \sum_{i=1}^m u_i c_i(\overline{x}) + \left(\sum_{i=1}^m u_i \nabla c_i(\overline{x})\right)^T (\hat{x} - \overline{x}) = 0,$$

so that $0 = \sum_{i=1}^m u_i c_i(\hat{x})$ which in turn implies that $u_i c_i(\hat{x}) = 0,\ i = 1, \ldots, s$ since $c_i(\hat{x}) = 0,\ i = s+1, \ldots, m$. Hence, $u_i = 0\ i = 1, \ldots, s$. But then, by (4.10), $\sum_{i=s+1}^m u_i \nabla c_i(\overline{x}) = 0$ whereby the surjectivity hypothesis implies that $u_i = 0\ i = s+1, \ldots, m$. Consequently, the dual form of the MFCQ holds.

For the reverse implication, suppose that $\overline{x}$ is any point in Ω satisfying the MFCQ and let $d \in \mathbb{R}^n$ be the vector given in (4.8). Then, for all $t > 0$ sufficiently small

$$\begin{aligned} &c_i(\overline{x} + td) = c(\overline{x}) + t\nabla c_i(\overline{x})^T d + o(t) < 0 \quad &&\text{for } \ i = 1, \ldots, s, \text{ and} \\ &c_i(\overline{x} + td) = 0 &&\text{for } \ i = s+1, \ldots, m, \end{aligned}$$

which establishes that the Slater CQ is satisfied. □

The main first-order optimality theorem for (C-NLP) now follows.

THEOREM 5.19 (First-Order Optimality in Convex NLP). *Consider the problem (C-NLP) where the the convexity hypotheses* (5.8) *are satisfied. If $\overline{x}$ is a solution to (C-NLP) at which the Slater CQ is satisfied, then $\overline{x}$ is a KKT point for (C-NLP). Conversely, if $\overline{x}$ is a KKT point for (C-NLP), then $\overline{x}$ solves (C-NLP).*

PROOF. If $\overline{x}$ is a solution to (C-NLP) at which the Slater CQ is satisfied, then Theorem 5.18 and Theorem 4.22 tells us that $\overline{x}$ is a KKT point for (C-NLP). Conversely, suppose that $\overline{x}$ is a KKT point for (C-NLP) and let $\overline{u}$ be the associated Lagrange multipliers. Since $L(x, \overline{u})$ is a convex function of x, the KKT conditions imply that $\overline{x}$ solves $\min_{x \in \mathbb{R}^n} L(x, \overline{u})$. Therefore, for all $x \in \Omega$, the KKT conditions imply that

$$f(\overline{x}) = L(\overline{x}, \overline{u}) \leq L(x, \overline{u}) \leq f(x).$$

Hence $\overline{x}$ solves (C-NLP). □

7. Saddle Point Theory and Lagrangian Duality

The Lagrangian function for (C-NLP) is

$$L(x,y) = f(x) + y_1c_1(x) + y_2c_2(x) + \cdots + y_mc_m(x) - \delta\left(y \,|\, K^\circ\right),$$

where $K := \mathbb{R}^s_- \times \{0\}^{m-s}$ and $K^\circ = \mathbb{R}^s_+ \times \mathbb{R}^{m-s}$. Define

$$\operatorname{dom} L := \{(x,y) \,|\, -\infty < L(x,y) < +\infty\}.$$

Correspondingly,

$$\operatorname{dom}_x(L) := \{x \,|\, \exists\, y \in \mathbb{R}^m \text{ s.t. } \ (x,y) \in \operatorname{dom} L\}$$

and

$$\operatorname{dom}_y(L) := \{y \,|\, \exists\, x \in \mathbb{R}^n \text{ s.t. } \ (x,y) \in \operatorname{dom} L\} \ .$$

A pair $(\overline{x},\, \overline{y}) \in \mathbb{R}^n \times K^\circ$ is said to be a saddle point for L if

$$L(\overline{x},\, y) \le L(\overline{x},\overline{y}) \le L(x,\overline{y}) \qquad \forall\ (x,y) \in \mathbb{R}^n \times K^\circ. \tag{5.12}$$

THEOREM 5.20 (Saddle Point Theorem). *Let $\overline{x} \in \mathbb{R}^n$. If there exists $\overline{y} \in K^\circ$ such that $(\overline{x},\overline{y})$ is a saddle point for the Lagrangian L, then $\overline{x}$ solves (C-NLP). Conversely, if $\overline{x}$ is a solution to (C-NLP) at which the Slater CQ (see Definition 5.17) is satisfied, then there is a $\overline{y} \in K^\circ$ such that $(\overline{x},\overline{y})$ is a saddle point for L.*

PROOF. If $(\overline{x},\overline{y}) \in \mathbb{R}^n \times K^\circ$ is a saddle point for $\mathcal{P}$ then

$$\sup_{y\in K^\circ} L(\overline{x},y) = \sup_{y\in K^\circ} f(\overline{x}) + y_1c_1(\overline{x}) + y_2c_2(\overline{x}) + \cdots + y_mc_m(\overline{x}) \le L(\overline{x},\overline{y}).$$

If $c_i(\overline{x}) > 0$ for some $i \in \{1,\dots,s\}$, then the supremum on the left is $+\infty$ since we can send $y_i \uparrow +\infty$. Hence, we must have $c_i(\overline{x}) \le 0,\ i = 1,\dots,s$. Similarly, if $c_i(\overline{x}) \ne 0$ for some $i \in \{s+1,\dots,m\}$, then the supremum on the left is again $+\infty$ since we could send $y_i \to -\operatorname{sign}(c_i(\overline{x}))\infty$. Hence, we must also have $c_i(\overline{x}) = 0,\ i = s+1,\dots,m$. That is, It must be the case that $\overline{x} \in \Omega$. Since $L(\overline{x},\overline{y}) = \sup_{y\in K^\circ} L(\overline{x},y)$, we also know that $\sum_{i=1}^s \overline{y}_i c_i(\overline{x}) = \sum_{i=1}^m \overline{y}_i c_i(\overline{x}) = 0$. Therefore, the right half of the saddle point condition (5.12) implies that

$$f(\overline{x}) = L(\overline{x},\overline{y}) \le \inf_x L(x,\overline{y}) \le \inf_{x\in\Omega} L(x,\overline{y}) \le \inf_{x\in\Omega} f(x) \le f(\overline{x}),$$

and so $\overline{x}$ solves (C-NLP).

Conversely, if $\overline{x}$ is a solution to (C-NLP) at which the Slater CQ is satisfied, then, by Theorem 5.19, there is a vector $\overline{y}$ such that $(\overline{x},\overline{y})$ is a KKT pair for (C-NLP). Primal feasibility ($\overline{x} \in \Omega$), dual feasibility ($\overline{y} \in K^\circ$), and complementarity ($\overline{y}_i c_i(\overline{x}) = 0,\ i = 1,\dots,s$) imply that

$$L(\overline{x},y) \le f(\overline{x}) = L(\overline{x},\overline{y}) \quad \forall\ y \in K^\circ.$$

On the other hand, dual feasibility and convexity imply the convexity of the function $L(x,\overline{y})$ in x. Hence the condition $0 = \nabla_x L(\overline{x},\overline{y})$ implies that $\overline{x}$ is a global minimizer for the function $x \to L(x,\overline{y})$, that is

$$L(\overline{x},\overline{y}) \le L(x,\overline{y}) \quad \forall\ x \in \mathbb{R}^n.$$

Therefore, $(\overline{x},\overline{y})$ is a saddle point for L. □

It is always the case that

$$\sup_{y\in K^\circ} \inf_{x\in\mathbb{R}^n} L(x,y) \le \inf_{x\in\mathbb{R}^n} \sup_{y\in K^\circ} L(x,y) \tag{5.13}$$

since the largest minimum is always smaller that the smallest maximum. On the other hand, if $(\overline{x},\overline{y})$ is a saddle point for L, then

$$\inf_{x\in\mathbb{R}^n} \sup_{y\in K^\circ} L(x,y) \le \sup_{y\in K^\circ} L(\overline{x},y) \le L(\overline{x},\overline{y}) \le \inf_{x\in\mathbb{R}^n} L(x,\overline{y}) \le \sup_{y\in K^\circ} \inf_{x\in\mathbb{R}^n} L(x,y).$$

Hence, if a saddle point for L exists on $\mathbb{R}^n \times K^\circ$, then

$$\sup_{y\in K^\circ} \inf_{x\in\mathbb{R}^n} L(x,y) = \inf_{x\in\mathbb{R}^n} \sup_{y\in K^\circ} L(x,y).$$

A theorems that establishes the existence of a saddle point are called mini-max theorems. These theorems provide conditions under which one can exchange an inf-sup for a sup-inf. Mini-max theorems can be used as a basis for convex duality theory.

In the proof of Theorem 5.20 it was shown that

$$\psi(x) := \sup_{y\in K^\circ} L(x,y) = \begin{cases} +\infty & \text{if } x \notin \Omega, \\ f(x) & \text{if } x \in \Omega. \end{cases}$$

For this reason, we call ψ the *primal objective* and $\inf_x \psi(x)$ the *primal problem* since

$$\inf_x \psi(x) = \inf_{x\in\mathbb{R}^n} \sup_{y\in K^\circ} L(x,y) = \inf_{x\in\Omega} f(x) \ .$$

This is the inf-sup side of the saddle point problem. The other side, the sup-inf problem, is called the *dual problem* with *dual objective* function

$$\phi(y) := \inf_{x\in\mathbb{R}^n} L(x,y) \ .$$

From (5.13), we obtain the *Weak Duality Theorem*

$$\sup_{y\in K^\circ} \phi(y) \le \inf_{x\in\mathbb{R}^n} \psi(x) \ . \tag{5.14}$$

The Saddle Point Theorem says that if $(\overline{x},\overline{y})$ is a saddle point for L, then $\overline{x}$ solves the primal problem, $\overline{y}$ solves the dual problem, and the optimal values in the primal and dual problems coincide. Illustrations of how the Lagrangian Duality Theory can be used to generate dual problems are given in the following subsections.

7.1. Linear Programming Duality. Consider the following LP:

$$\begin{array}{ll} \mathcal{P} & \text{minimize} \quad b^T x \\ & \text{subject to} \quad A^T x \ge c,\ 0 \le x \ . \end{array}$$

The Lagrangian is

$$L(x,y,v) = b^T x + y^T(c - A^T x) - v^T x, \quad \text{where } 0 \le y,\ 0 \le v \ .$$

The dual objective function is

$$g(y,u) = \min_{x\in\mathbb{R}^n} L(x,y,v) = \min_{x\in\mathbb{R}^n} b^T x + y^T(c - A^T x) - v^T x \ .$$

Our first goal is to obtain a closed form expression for $g(y,u)$. This is accomplished by using the optimality conditions for minimizing $L(x,y,u)$ to eliminate x from the definition of L. Since $L(x,y,v)$ is a convex function in x, the global solution to $\min_{x\in\mathbb{R}^n} L(x,y,v)$ is obtained by solving the equation $0 = \nabla_x L(x,y,u) = b - Ay - v$ with $0 \le y,\ 0 \le v$. Using this condition in the definition of L we get

$$L(x,y,u) = b^T x + y^T(c - A^T x) - v^T x = (b - Ay - v)^T x + c^T y = c^T y,$$

subject to $b - A^T y = v$ and $0 \le y,\ 0 \le v$. Hence the Lagrangian dual problem

$$\begin{array}{ll} & \text{maximize} \quad g(y,v) \\ & \text{subject to} \quad 0 \le y,\ 0 \le v \end{array}$$

can be written as

$$\begin{array}{ll} \mathcal{D} & \text{maximize} \quad c^T y \\ & \text{subject to} \quad b - Ay = v,\ 0 \le y,\ 0 \le v \ . \end{array}$$

Note that we can treat the variable v as a slack variable in this LP and write

$$\begin{array}{ll} \mathcal{D} & \text{maximize} \quad c^T y \\ & \text{subject to} \quad Ay \le b,\ 0 \le y \ . \end{array}$$

The linear program $\mathcal{D}$ is the dual to the linear program $\mathcal{P}$.

7.2. Convex Quadratic Programming Duality. Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and positive semi-definite, and let $c \in \mathbb{R}^n$. Consider the convex quadratic program

$$\begin{array}{lll} \mathcal{P} & \text{minimize} & \frac{1}{2}x^TQx + c^Tx \\ & \text{subject to} & Ax \leq b,\ 0 \leq x\ . \end{array}$$

Let $Q = LL^T$ is the Cholesky factorization of Q, where $L \in \mathbb{R}^{k\times n}$ with $k = \operatorname{rank}(Q)$. We can then rewrite the quadratic program $\mathcal{P}$ as

$$\begin{array}{lll} \mathcal{P}' & \text{minimize} & \frac{1}{2}\left\|z\right\|_2^2 + c^Tx \\ & \text{subject to} & Ax \leq b,\ L^Tx = z,\ 0 \leq x\ . \end{array}$$

The Lagrangian is given by

$$\mathsf{L}((x,z),(y,u,v)) = \tfrac{1}{2}\left\|z\right\|_2^2 + c^Tx + y^T(Ax-b) + u^T(L^Tx - z) - v^Tx \quad \text{where } 0 \leq y,\ 0 \leq v.$$

Here (x,z) are the primal variables and (y,u,v) are the dual variables. The dual objective function is

$$g(y,u,v) = \min_{(x,z)\in\mathbb{R}^n} \mathsf{L}((x,z),(y,u,v))\ .$$

The goal is to obtain a closed form expression for g with the variables (x,z) removed by using the first-order optimality condition $0 = \nabla_{(x,z)}\mathsf{L}((x,z),(y,u,v))$. This optimality condition completely identifies the solution since L is convex in (x,z). We have

$$\begin{aligned} 0 &= \nabla_x\mathsf{L}((x,z),(y,u,v)) = c + A^Ty + Lu - v, \\ 0 &= \nabla_z\mathsf{L}((x,z),(y,u,v)) = z - u. \end{aligned} \tag{5.15}$$

Rewriting the Lagrangian to better expose these optimality conditions and then plugging them in yields

$$\begin{aligned} \mathsf{L}((x,z),(y,u,v)) &= (c + A^Ty + Lu - v)^Tx - u^Tz + \tfrac{1}{2}\left\|z\right\|_2^2 - b^Ty \\ &= -[b^Ty + \tfrac{1}{2}\left\|u\right\|_2^T] \end{aligned}$$

where $0 \leq y,\ 0 \leq v$. Hence the dual problem is

$$\begin{array}{ll} \text{maximize} & -[b^Ty + \frac{1}{2}\left\|u\right\|_2^2] \\ \text{subject to} & c + A^Ty + Lu = v,\ 0 \leq y,\ 0 \leq v. \end{array}$$

By treating v as slack variable, we obtain the dual problem

$$\begin{array}{lll} \text{maximize} & -[b^Ty + \frac{1}{2}\left\|u\right\|_2^2] & \qquad = \qquad - \quad \text{minimize} \quad b^Ty + \frac{1}{2}\left\|u\right\|_2^2 \\ \text{subject to} & 0 \leq c + A^Ty + Lu,\ 0 \leq y & \qquad\qquad\quad \text{subject to} \quad 0 \leq c + A^Ty + Lu\ ,\ 0 \leq y. \end{array}$$

8. The Projection Theorem for Convex Sets

An important application our results on optimality conditions is the nearest point problem for a closed convex set. That is, given a closed convex set $C \subset \mathbb{R}^n$ and a point $x \notin C$, find a point $\overline{z} \in C$ such that

$$\left\|x - \overline{z}\right\|_2 \leq \left\|x - z\right\|_2 \quad \forall\, z \in C.$$

That is, we wish to solve the optimization problem

$$\begin{array}{ll} \mathcal{P}\mathit{roj}_C: & \min f(z) := \frac{1}{2}\left\|x - z\right\|_2^2 \\ & \text{subject to } z \in C\ . \end{array}$$

Here the function f is continuously differentiable with $\nabla f(z) := z - x$. The Corollary 5.10.2 says that $\overline{x}$ is a local solution to $\mathcal{P}\mathit{roj}_C$ if and only if

$$\nabla f(\overline{x})^T(y - \overline{x}) \geq 0 \quad \forall\, y \in C \ \text{ or equivalently } \quad -\nabla f(\overline{x}) \in N(\overline{x}\,|\,C)\,. \tag{5.16}$$

Moreover, since f is coercive and strictly convex, then a unique solution exists. We have just proved the celebrated *projection theorem* for convex sets.

THEOREM 5.21 (The Projection Theorem for Convex Sets). *Let $y \in \mathbb{R}^n$ and let $C \subset \mathbb{R}^n$ be nonempty, closed and convex set. Then $\overline{x} \in C$ solves the problem*

$$\min\{\frac{1}{2}\|y-x\|_2^2 : x \in C\}$$

if and only if

$$(y-\overline{x})^T(x-\overline{x}) \leq 0 \quad \forall\, x \in C, \quad \text{or equivalently,} \quad (y-\overline{x}) \in N(\overline{x} \,|\, C)\,. \tag{5.17}$$

Moreover, the solution $\overline{x}$ always exists and is unique.

DEFINITION 5.22 (The Projection Mapping). *Let $C \subset \mathbb{R}^n$ be nonempty, closed, and convex. Define the projection onto C to be the mapping $P_C : \mathbb{R}^n \to C$ given by*

$$\frac{1}{2}\|y - P_\Omega(y)\|_2^2 = \min\{\frac{1}{2}\|y-x\|_2^2 : x \in C\}.$$

Observe that P_C is well-defined by Theorem 5.21.

If $C = S$ where S is a subspace, then $P_S(y)$ is the orthogonal projection onto S. In this case we have

$$\|P_S y - P_S w\|_2^2 + \|(I-P_S)y - (I-P_S)w\|_2^2 = \|y-w\|_2^2\ .$$

A version of this result extends to arbitrary closed convex sets.

THEOREM 5.23 (The Projection is a Contraction). *Given a non-empty closed convex set $C \subset \mathbb{R}^n$, we have*

$$\|P_C y - P_C w\|_2^2 + \|(I-P_C)y - (I-P_C)w\|_2^2 \leq \|y-w\|_2^2\ .$$

In particular, this implies that the projection onto a convex set is non-expansive, i.e.

$$\|P_C y - P_C w\|_2 \leq \|y-w\|_2\ .$$

PROOF. Set $P_C = P$, then

$$\begin{aligned}\|y-w\|_2^2 &= \|(Py-Pw) + ((I-P)y - (I-P)w)\|_2^2 \\ &= \|Py-Pw\|_2^2 + \|(I-P)y-(I-P)w\|_2^2 + 2\,\langle Py-Pw,\,(I-P)y-(I-P)w\rangle\,,\end{aligned}$$

and

$$\langle Py-Pw,\,(I-P)y-(I-P)w\rangle = \langle Py-Pw,\,y-Py\rangle + \langle Pw-Py,\,w-Pw\rangle\,,$$

where, by the Projection Theorem 5.21,

$$0 \leq \langle Py-Pw,\,y-Py\rangle \quad \text{and} \quad 0 \leq \langle Pw-Py,\,w-Pw\rangle\ .$$

□

We have the following useful corollaries to Theorem 5.21.

COROLLARY 5.23.1. *Given $x \in C$, $z \in N(x \,|\, C)$ if and only if, for any $\lambda > 0$, $x = P_C(x+\lambda z)$.*

COROLLARY 5.23.2 (Gradient Projections and Optimality). *Let $f : \mathbb{R}^n \to \mathbb{R}$. Then $\overline{x}$ solves $\min_{x\in C} f(x)$ if and only if, for any $\lambda > 0$,*

$$P_C(\overline{x} - \lambda \nabla f(\overline{x})) = \overline{x}\ . \tag{5.18}$$

8.1. Examples: Computing Projections. We now address the question of implementation. Specifically, how does one compute the projection onto the convex set Ω. In general this is not a finite process. Nonetheless, for certain important convex sets Ω it can be done quite efficiently.

Projection onto box constraints

Let us suppose that Ω is given by $\Omega := \{x \in \mathbb{R}^n : \ell \leq x \leq u\}$, where $\ell, u \in \overline{\mathbb{R}}^n$ with $\overline{\mathbb{R}} = \Omega \cup \{+\infty, -\infty\}$ and $\ell :\leq u$, $i = 1, \ldots, n, \ell_i \neq +\infty$ $i = 1, \ldots, n$ and $u_i \neq -\infty$ $i = 1, \ldots, n$. Then P_Ω can be expressed componentwise as

$$[P_\Omega(x)]_i := \begin{cases} \ell_i & \text{if } x_i \leq \ell_i \\ x_i & \text{if } \ell_i < x_i < u_i \\ u_i & \text{if } u_i \leq x_i \end{cases}$$

Thus, for example, if $\Omega = \mathbb{R}^n_+$, then

$$P_\Omega(x) = x_+.$$

Projection onto a Polyhedron

Let Ω be the polyhedron given by

$$\Omega := \{x \in \mathbb{R}^n : a_i^T x \leq \alpha_i, i = 1, \ldots, 3, a_i^T x = \alpha_i, i = s+1, \ldots, m\}.$$

Then P_Ω is determined by solving the quadratic program

$$\begin{array}{ll} \min \frac{1}{2}\|x - y\|_2^2 & \\ \text{subject to} & a_i^T x \leq \alpha_i \quad i = 1, \ldots, s \\ & a_i^T x = \alpha_i \quad i = s+1, \ldots, m. \end{array}$$

CHAPTER 6

Line Search Methods

Let $f : \mathbb{R}^n \to \mathbb{R}$ be given and suppose that x_{old} is our current best estimate of a solution to

$$\mathcal{P} \quad \min_{x \in \mathbb{R}^n} f(x) \ .$$

A standard method for improving the estimate x_{old} is to choose a direction of search $d \in \mathbb{R}^n$ and compute a step length $t^* \in \mathbb{R}$ so that $x_{old} + t^* d$ reduces the value of f along the line $\{x_{old} + td \,|\, t \in \mathbb{R}\}$. The new estimate for the solution to $\mathcal{P}$ is then $x_{new} = x_{old} + t^* d$. The procedure for choosing t^* is called a *line search method*. If t^* is taken to be the global solution to the problem

$$\min_{t \in \mathbb{R}} f(x_{old} + td) \ ,$$

then t^* is called the *Curry step length*. However, except in certain very special cases, the Curry step length is far too costly to compute. For this reason we focus on easily computed step lengths. We begin the simplest and the most commonly used line search method called backtracking.

1. The Basic Backtracking Algorithm

In the *backtracking line search* we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and that we are given a direction d of strict descent at the current point x_{old}, that is $f'(x_{old}; d) < 0$.

INITIALIZATION: Choose $\gamma \in (0, 1)$ and $c \in (0, 1)$.

Having x_{old} obtain x_{new} as follows:

STEP 1: Compute the backtracking stepsize

$$\begin{array}{rcl} t^* & := & \max \gamma^\nu \\ & & \text{s.t.} \nu \in \{0, 1, 2, \ldots\} \text{ and} \\ & & f(x_{old} + \gamma^\nu d) \leq f(x_{old}) + c\gamma^\nu f'(x_{old}; d). \end{array}$$

STEP 2: Set $x_{new} = x_{old} + t^* d$.

The backtracking line search forms the basic structure upon which most line search methods are built. Due to the importance of this method, we take a moment to emphasize its key features.

(1) The update to x_{old} has the form

$$x_{new} = x_{old} + t^* d \ . \tag{6.1}$$

Here d is called the *search direction* while t^* is called the *step length* or *stepsize*.

(2) The search direction d must satisfy

$$f'(x_{old}; d) < 0.$$

Any direction satisfying this strict inequality is called a *direction of strict descent* for f at x_{old}. If $\nabla f(x_{old}) \neq 0$, then a direction of strict descent always exists since we can use the direction of *steepest descent* $d = -\nabla f'(x_{old})$ which gives

$$f'(x_{old}; -\nabla f'(x_{old})) = -\nabla f(x_{old})^T \nabla f(x_{old}) = -\left\|\nabla f'(x_{old})\right\|^2 .$$

It is important to note that if d is a direction of strict descent for f at x_{old}, then there is a $\bar{t} > 0$ such that

$$f(x_{old} + td) < f(x_{old}) \quad \forall\ t \in (0, \bar{t})$$

since

$$f'(x_{old}; d) = \lim_{t \downarrow 0} \frac{f(x_{old} + td) - f(x_{old})}{t}.$$

Hence, if $f'(x_{old}; d) < 0$, there is a $\bar{t} > 0$ such that

$$\frac{f(x_{old} + td) - f(x_{old})}{t} < 0 \quad \forall\ t \in (0, \bar{t}),$$

that is

$$f(x_{old} + td) < f(x_{old}) \quad \forall\ t \in (0, \bar{t}).$$

(3) In Step 1 of the algorithm, we require that the step length t^* be chosen so that

$$f(x_{old} + t^* d) \leq f(x_{old}) + ct^* f'(x_{old}; d). \tag{6.2}$$

This inequality is called the *Armijo-Goldstein inequality*. It is named after the first two researchers topropose it in the design of line search routines (Allen Goldstein is a Professor Emeritus here at the University of Washington). Observe that this inequality guarantees that

$$f(x_{old} + t^* d) < f(x_{old}).$$

For this reason, the algorithm described above is called a *descent algorithm*. It was observed in point (2) above that it is always possible to choose t^* so that $f(x_{old} + t^* d) < f(x_{old})$. But the Armijo-Goldstein inequality is a somewhat stronger statement. To see that it too can be satisfied observe that since $f'(x_{old}; d) < 0$,

$$\lim_{t \downarrow 0} \frac{f(x_{old} + td) - f(x_{old})}{t} = f'(x_{old}; d) < cf'(x_{old}; d) < 0.$$

Hence, there is a $\bar{t} > 0$ such that

$$\frac{f(x_{old} + td) - f(x_{old})}{t} \leq cf'(x_{old}; d) \quad \forall\ t \in (0, \bar{t}),$$

that is

$$f(x_{old} + td) \leq f(x_{old}) + tcf'(x_{old}; d) \quad \forall\ t \in (0, \bar{t}).$$

(4) The Armijo-Goldstein inequality is known as a *sufficient decrease condition*. It is essential that we do not choose t^* too small. This is the reason for setting t^* equal to the first (largest) member of the geometric sequence $\{\gamma^\nu\}$ for which the Armijo-Goldstein inequality is satisfied. In general, we always wish to choose t^* as large as possible since it is often the case that some effort was put into the selection of the search direction d. Indeed, as we will see, for Newton's method we must take $t^* = 1$ in order to achieve rapid local convergence.

(5) There is a balance that must be struck between taking t^* as large as possible and not having to evaluating the function at many points. Such a balance is obtained with an appropriate selection of the parameters γ and c. Typically one takes $\gamma \in [.5, .8]$ while $c \in [.001, .1]$ with adjustments depending on the cost of function evaluation and degree of nonlinearity.

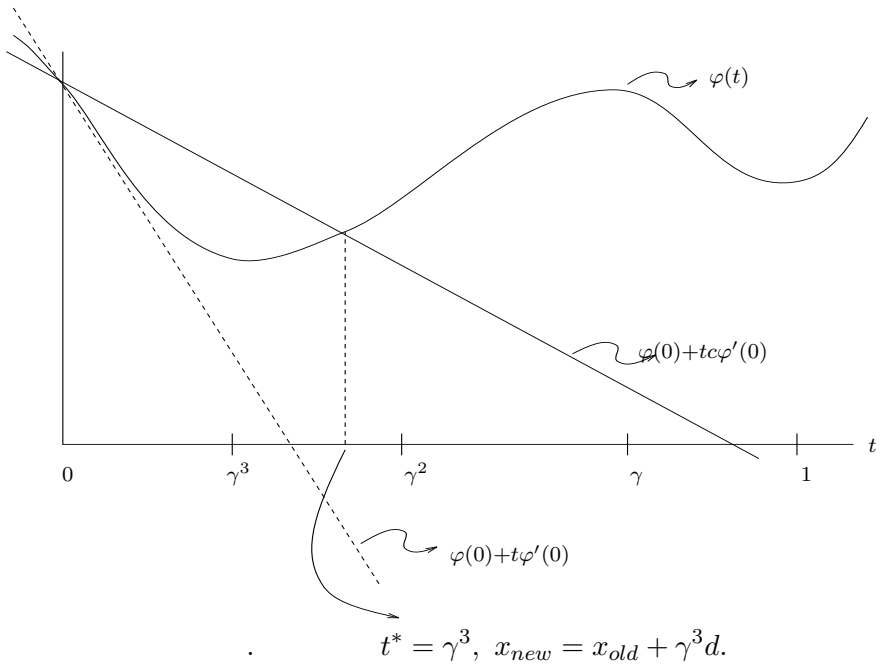
(6) The backtracking procedure of Step 1 is easy to program. A pseudo-code follows:

$$
\begin{array}{lrcl}
 & f_c & = & f(x_{old}) \\
 & \Delta f & = & cf'(x_{old}; d) \\
\texttt{if} & \Delta f & \geq & 0, \quad \texttt{print "line search failure, nondescentdirection"}, \texttt{break} \\
\texttt{else} & \text{new}f & = & f(x_{old} + d) \\
 & t & = & 1 \\
\quad \texttt{while} & \text{new}f & > & f_c + t\Delta f \\
 & t & = & \gamma t \\
 & \text{new}f & = & f(x_{old} + td) \\
\quad \texttt{endwhile} & & & \\
\texttt{endif} & & &
\end{array}
$$

Point (3) above guarantees that this procedure is finitely terminating in infinite precision arithmatic.

(7) The most prevalent cause of termination in a numerical optimization routine is a line search failure. The line search can fail for many reasons, but the most common is an issue with the choice of search direction.

(8) The backtracking procedure has a nice graphical illustration. Set $\varphi(t) = f(x_{old} + td)$ so that $\varphi'(0) = f'(x_{old}; d)$.



. $t^* = \gamma^3,\ x_{new} = x_{old} + \gamma^3 d.$

The three most studied search directions are

(1) Steepest Descent (or Cauchy Direction):

$$d = -\nabla f(x_{old}) \ .$$

(2) Newton Direction: when $\nabla^2 f(x_{old}) \in \mathcal{S}^n_{++}$

$$d = -\nabla^2 f(x_{old})^{-1} \nabla f(x_{old}) \ .$$

(3) Newton-Like Direction: for $H \in \mathcal{S}^n_{++}$,

$$d = -H \nabla f(x_{old}).$$

The advantage of the Cauchy direction is that it always provides a direction of strict descent. However, once the iterates get "close" to a stationary point, the procedure takes a very long time to obtain a moderately accurate estimate of the stationary point. Close to a solution numerical error dominates and the iterates behave chaotically.

On the other hand, Newton's method may not define directions of strict descent until one is very close to a stationary point satisfying the second-order sufficiency condition $\nabla^2 f(x_{old}) \in \mathcal{S}^n_{++}$. However, once one is near such a stationary point, then Newton's method zooms in on the stationary point very rapidly. This behavior will be made precise when we establish our convergence result for Newton's method.

The basic convergence result for the backtracking line search now follows.

THEOREM 6.1. (CONVERGENCE FOR BACKTRACKING) *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $x_0 \in \mathbb{R}$ be such that f is differentiable on $\mathbb{R}^n$ with ∇f Lipschitz continuous on an open convex set containing the set $\{x : f(x) \leq f(x_0)\}$. Let $d : \mathbb{R}^n \to \mathbb{R}^n$ be such that $f'(x : d(x)) \leq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $d(x) = 0$. Consider the following iteration:*

Initialization: $x^0 \in \mathbb{R}^n$, $0 < \gamma < 1$, $0 < c < 1$, $k = 0$.
Iteration: *For $k = 0, 1, \ldots$*
Direction: $d^k := d(x^k)$. *If $d^k = 0$, terminate iteration.*
Stepsize: $t_k := \max\left\{\gamma^s \,\middle|\, f(x^k + \gamma^s d^k) \leq f(x^k) + c\gamma^s f'(x^k; d^k),\ s = 1, 2, \ldots\right\}$.
Update: $x^{k+1} := x^k + t_k d^k,\ k = k + 1$.

Let $\{x^k\}$ be a sequence generated by this iteration. Then one of the following statements must be true:

(i) There is a k_0 such that $d^{k_0} = 0$ and the iteration terminates.
(ii) $f(x^k) \searrow -\infty$
(iii) The sequence $\{\|d^k\|\}$ diverges ($\|d^k\| \to \infty$).
(iv) For every subsequence $J \subset \mathbb{N}$ for which $\{d^k : k \in J\}$ is bounded, we have

$$\lim_{k \in J} f'(x^k; d^k) = 0.$$

REMARK 6.2. *It is important to note that this theorem says nothing about the convergence of the sequence $\{x^k\}$. Indeed, this sequence may diverge. Moreover, due to the definition of the function $d : \mathbb{R}^n \to \mathbb{R}^n$ and our discussion of the backtracking procedure, we know that the iteration is well defined and terminates if and only if $d^k = 0$, and $f(x^{k+1}) < f(x^k)$ whenever $d^k \neq 0$.*

Before proving the theorem, we consider some important corollaries concerning the Cauchy and Newton search directions. Each corollary assumes that the hypotheses of Theorem 6.1 hold.

COROLLARY 6.2.1. *If the sequences $\{d^k\}$ and $\{f(x^k)\}$ are bounded and $d^k = 0$ only if $\nabla f(x^k) = 0$, then*

$$\lim_{k \to \infty} f'(x^k; d^k) = 0.$$

PROOF. We may assume that the sequence $\{x^k\}$ is infinite otherwise it terminates with $\nabla f(x^k) = 0$. The hypotheses imply that either (i) or (iv) with $J = \mathbb{N}$ occurs in Theorem 6.1. Hence, $\lim_{k \to \infty} f'(x^k; d^k) = 0$. □

COROLLARY 6.2.2. *Suppose $d^k = 0$ when $\nabla f(x^k) = 0$; otherwise, d^k is the normalized Cauchy direction $-\nabla f'(x^k)/\left\|\nabla f(x^k)\right\|$. Then every accumulation point, $\overline{x}$, of the sequence $\{x^k\}$ satisfies $\nabla f(\overline{x}) = 0$.*

PROOF. We may assume that the sequence $\{x^k\}$ is infinite otherwise it terminates with $\nabla f(x^k) = 0$. Consequently, the sequence $\{f(x^k)\}$ is decreasing. If $\overline{x}$ is any accumulation point of the sequence $\{x^k\}$, then we must have $f(x^k) \downarrow f(\overline{x})$ since otherwise for some k_0 and $\epsilon > 0$

$$f(x^k) + \epsilon < f(\overline{x})$$

for all $k > k_0$ since $\{f(x^k)\}$ is decreasing. But $\overline{x}$ is a cluster point of $\{x^k\}$ and f is continuous. Hence, there is a $\widehat{k} > k_0$ such that

$$|f(\overline{x}) - f(x^{\widehat{k}})| < \epsilon/2$$

which contradicts the previous inequality. Therefore, the sequence $\{f(x^k)\}$ is bounded so Corollary 6.2.1 applies, that is,

$$0 = \lim_{k\to\infty} f'\left(x^k; \frac{-\nabla f(x^k)}{\|\nabla f(x^k)\|}\right) = \lim_{k\to\infty} -\left\|\nabla f(x^k)\right\|.$$

Since ∇f is continuous, $\nabla f(\overline{x}) = 0$. □

COROLLARY 6.2.3. *Let us further assume that f is twice continuously differentiable and that there is a $\beta > 0$ such that, for all $u \in \mathbb{R}^n$, $\beta \|u\|^2 < u^T\nabla^2 f(x)u$ on $\{x : f(x) \leq f(x^0)\}$. If the iteration is implemented using the Newton search directions,*

$$d^k = -\nabla^2 f(x^k)^{-1}\nabla f(x^k),$$

then every accumulation point, $\overline{x}$, of the sequence $\{x^k\}$ satisfies $\nabla f(\overline{x}) = 0$.

PROOF. Again, we may assume that the sequence $\{x^k\}$ is infinite otherwise it terminates with $\nabla f(x^k) = 0$. Let $\overline{x}$ be an accumulation point of the sequence $\{x^k\}$ and let $J \subset \mathbb{N}$ be such that $x^k \xrightarrow{J} \overline{x}$. Clearly, $\{x^k : k \in J\}$ is bounded. Hence, the continuity of ∇f and $\nabla^2 f$, along with the Weierstrass Compactness Theorem, imply that the sets $\{\left\|\nabla f(x^k)\right\| : k \in J\}$ and $\{\left\|\nabla^2 f(x^k)\right\| : k \in J\}$ are also bounded. Let M_1 be a bound on the values $\{\left\|\nabla f(x^k)\right\| : k \in J\}$ and let M_2 be an upper bound on the values $\{\left\|\nabla^2 f(x^k)\right\| : k \in J\}$. Recall that by hypotheses $\beta \|u\|^2$ is a uniform lower bound on the values $\{u^T\nabla^2 f(x^k)u\}$ for every $u \in \mathbb{R}^n$. Take $u = d^k$ to obtain the bound

$$\beta \left\|d^k\right\|^2 \leq \nabla f(x^k)^T\nabla^2 f(x^k)^{-1}\nabla f(x^k) \leq \left\|d^k\right\| \left\|\nabla f(x^k)\right\|,$$

and so

$$\left\|d^k\right\| \leq \beta^{-1} M_1 \ \forall\, k \in J.$$

Therefore, the sequence $\{d^k : k \in J\}$ is bounded. Moreover, as in the proof of Corollary 6.2.2, the sequence $\{f(x^k)\}$ is also bounded. On the other hand,

$$\left\|\nabla f(x^k)\right\| = \left\|\nabla^2 f(x^k)d^k\right\| \leq M_2 \left\|d^k\right\| \ \forall\, k \in J.$$

Therefore,

$$M_2^{-1} \left\|\nabla f(x^k)\right\| \leq \left\|d^k\right\| \ \forall\, k \in J.$$

Consequently, Theorem 6.1 Part (iv) implies that

$$\begin{aligned} 0 &= \lim_{k\in J} |f'(x^k;d^k)| \\ &= \lim_{k\in J} |\nabla f(x^k)^T \nabla^2 f(x^k)^{-1} \nabla f(x^k)| \\ &\geq \lim_{k\in J} \beta \left\|d^k\right\|^2 \\ &\geq \lim_{k\in J} \beta M_2^{-2} \left\|\nabla f(x^k)\right\|^2 \\ &= \beta M_2^{-2} \|\nabla f(\overline{x})\|^2, \end{aligned}$$

and so $\nabla f(\overline{x}) = 0$. □

Proof of Theorem 6.1: We assume that none of (i), (ii), (iii), and (iv) hold and establish a contradiction.

Since (i) does not occur, $d^k \neq 0$ for all $k = 1, 2, \ldots$. Since (ii) does not occur, the sequence $\{f(x^k)\}$ is bounded below. Since $\{f(x^k)\}$ is a bounded decreasing sequence in $\mathbb{R}$, we have $f(x^k) \searrow \overline{f}$ for some $\overline{f}$. In particular, $(f(x^{k+1}) - f(x^k)) \to 0$. Next, since (iii) and (iv) do not occur, there is a subsequence $J \subset \mathbb{N}$ and a vector $\overline{d}$ such that $d^k \xrightarrow{J} \overline{d}$ and

$$\sup_{k\in J} f'(x^k;d^k) =: \beta < 0.$$

The Armijo-Goldstein inequality combined with the fact that $(f(x^{k+1}) - f(x^k)) \to 0$, imply that

$$t_k f'(x^k;d^k) \to 0.$$

Since $f'(x^k;d^k) \leq \beta < 0$ for $k \in J$, we must have $t_k \xrightarrow{J} 0$. With no loss in generality, we assume that $t_k < 1$ for all $k \in J$. Hence,

$$c\gamma^{-1} t_k f'(x^k;d^k) < f(x^k + t_k\gamma^{-1}d^k) - f(x^k) \tag{6.3}$$

for all $k \in J$ due to the stepsize choice and our assumption that $\tau_k < 1$ for all $k \in J$. By the Mean Value Theorem, there exists for each $k \in J$ a $\theta_k \in (0,1)$ such that

$$f(x^k + t_k\gamma^{-1}d^k) - f(x^k) = t_k\gamma^{-1} f'(\widehat{x}^k;d^k)$$

where

$$\begin{aligned} \widehat{x}^n &:= (1-\theta_k)x^k + \theta_k(x^k + t_k\gamma^{-1}d^k) \\ &= x^k + \theta_k t_k \gamma^{-1} d^k. \end{aligned}$$

Now, since ∇f is Lipschitz continuous, we have

$$\begin{aligned} f(x^k + t_k\gamma^{-1}d^k) - f(x^k) &= t_k\gamma^{-1} f'(\widehat{x}^k;d^k) \\ &= t_k\gamma^{-1} f'(x^k;d^k) + t_k\gamma^{-1}[f'(\widehat{x}^k;d^k) - f'(x^k;d^k)] \\ &= t_k\gamma^{-1} f'(x^k;d^k) + t_k\gamma^{-1}[\nabla f(\widehat{x}^k) - \nabla f(x^k)]^T d^k \\ &\leq t_k\gamma^{-1} f'(x^k;d^k) + t_k\gamma^{-1} L \left\|\widehat{x}^k - x^k\right\| \left\|d^k\right\| \\ &= t_k\gamma^{-1} f'(x^k;d^k) + L(t_k\gamma^{-1})^2 \theta_k \left\|d^k\right\|^2. \end{aligned}$$

Combining this inequality with inequality (6.3) yields the inequality

$$ct_k\gamma^{-1} f'(x^k;d^k) < t_k\gamma^{-1} f'(x^k;d^k) + L(t_k\gamma^{-1})^2 \theta_k \left\|d^k\right\|^2.$$

By rearranging and then substituting β for $f'(x^k;d^k)$ we obtain

$$0 < (1-c)\beta + (t_k\gamma^{-1})L \left\|d^k\right\|^k \xrightarrow{J} (1-c)\beta < 0.$$

This contradiction establishes the result. □

2. The Wolfe Conditions

We now consider a couple of modifications to the basic backtracking line search that attempt to better approximate an exact line-search (Curry line search), i.e. the stepsize t_k is chosen to satisfy

$$f(x^k + t_k d^k) = \min_{t \in \mathbb{R}} f(x^k + t d^k).$$

In this case, the first-order optimality conditions tell us that $0 = \nabla f(x^k + t_k d^k)^T d^k$. The Wolfe conditions try to combine the Armijo-Goldstein sufficient decrease condition with a condition that tries to push $\nabla f(x^k + t_k d^k)^T d^k$ either toward zero, or at least to a point where the search direction d^k is less of a direction of descent. To describe these line search conditions, we take parameters $0 < c_1 < c_2 < 1$.

Weak Wolfe Conditions

$$f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k f'(x^k; d^k) \tag{6.4}$$

$$c_2 f'(x^k; d^k) \leq f'(x^k + t_k d^k; d^k)\ . \tag{6.5}$$

Strong Wolfe Conditions

$$f(x^k + t_k d^k) \leq f(x^k) + c_1 t_k f'(x^k; d^k) \tag{6.6}$$

$$|f'(x^k + t_k d^k; d^k)| \leq c_2 |f'(x^k; d^k)|\ . \tag{6.7}$$

The weak Wolfe condition (6.5) tries to make d^k less of a direction of descent (and possibly a direction of ascent) at the new point, while the strong Wolfe condition tries to push the directional derivative in the direction d^k closer to zero at the new point. Imposing one or the other of the Wolfe conditions on a line search procedure has become standard practice for optimization software based on line search methods.

We now give a result showing that there exists stepsizes satisfying the weak Wolfe conditions. A similar result (with a similar proof) holds for the strong Wolfe conditions.

LEMMA 6.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and suppose that $x, d \in \mathbb{R}^n$ are such that the set $\{f(x + td) : t \geq 0\}$ is bounded below and $f'(x; d) < 0$, then for each $0 < c_1 < c_2 < 1$ the set*

$$\left\{ t \;\middle|\; \begin{array}{c} t > 0, f'(x + td; d) \geq c_2 f'(x; d), \text{ and} \\ f(x + td) \leq f(x) + c_1 t f'(x; d) \end{array} \right\}$$

has non–empty interior.

PROOF. Set $\phi(t) = f(x + td) - (f(x) + c_1 t f'(x; d))$. Then $\phi(0) = 0$ and $\phi'(0) = (1 - c_1) f'(x; d) < 0$. So there is a $\bar{t} > 0$ such that $\phi(t) < 0$ for $t \in (0, \bar{t})$. Moreover, since $f'(x; d) < 0$ and $\{f(x + td) : t \geq 0\}$ is bounded below, we have $\phi(t) \to +\infty$ as $t \uparrow \infty$. Hence, by the continuity of f, there exists $\hat{t} > 0$ such that $\phi(\hat{t}) = 0$. Let $t^* = \inf \left\{ \hat{t} \,\middle|\, 0 \leq t,\ \phi(\hat{t}) = 0 \right\}$. Since $\phi(t) < 0$ for $t \in (0, \bar{t})$, $t^* > 0$ and by continuity $\phi(t^*) = 0$. By Rolle's theorem (or the mean value theorem) there must exist $\tilde{t} \in (0, t^*)$ with $\phi'(\tilde{t}) = 0$. That is,

$$\nabla f(x + \tilde{t} d)^T d = c_1 \nabla f(x)^T d > c_2 \nabla f(x)^T d.$$

From the definition of t^* and the fact that $\tilde{t} \in (0, t^*)$, we also have

$$f(x + td) - (f(x) + c_1 \tilde{t} \nabla f(x)^T d) < 0\ .$$

The result now follows from the continuity of f and ∇f. □

We now describe a bisection method that either computes a stepsize satisfying the weak Wolfe conditions or sends the function values to $-\infty$. Let x and d in $\mathbb{R}^n$ be such that $f'(x; d) < 0$.

A Bisection Method for the Weak Wolfe Conditions

INITIALIZATION: Choose $0 < c_1 < c_2 < 1$, and set $\alpha = 0$, $t = 1$, and $\beta = +\infty$.

REPEAT
  If $f(x+td) > f(x) + c_1 t f'(x;d)$,
    set $\beta = t$ and reset $t = \frac{1}{2}(\alpha+\beta)$.
  Else if $f'(x+td;d) < c_2 f'(x;d)$,
    set $\alpha = t$ and reset

$$t = \begin{cases} 2\alpha, & \text{if } \beta = +\infty \\ \frac{1}{2}(\alpha+\beta), & \text{otherwise.} \end{cases}$$

  Else, STOP.
END REPEAT

LEMMA 6.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and suppose that $x, d \in \mathbb{R}^n$ are such that $f'(x;d) < 0$. Then one of the following two possibilities must occur in the Bisection Method for the Weak Wolfe Condition described above.*

(i) The procedure terminates finitely at a value of t for which the weal Wolfe conditions are satisfied.
(ii) The procedure does not terminate finitely, the parameter β is never set to a finite value, the parameter α becomes positive on the first iteration and is doubled in magnitude at every iteration thereafter, and $f(x+td) \downarrow -\infty$.

PROOF. Let us suppose that the procedure does not terminate finitely. If the parameter β is never set to a finite value, then it must be the case that that α becomes positive on the first iteration (since we did not terminate) and is doubled on each subsequent iteration with

$$f(x+\alpha d) \le f(x) + c_1 \alpha f'(x;d).$$

But then $f(x+td) \downarrow -\infty$ since $f'(x;d) < 0$. That is, option (ii) above occurs. Hence, we may as well assume that β is eventually finite and the procedure is not finitely terminating. For the sake of clarity, let us index the bounds and trial steps by iteration as follows: $\alpha_k < t_k < \beta_k,\ k = 1, 2, \ldots$. Since β is eventually finite, the bisection procedure guarantees that there is a $\bar{t} > 0$ such that

$$\alpha_k \uparrow \bar{t}, \quad t_k \to \bar{t}, \quad \text{and} \quad \beta_k \downarrow \bar{t}\ . \tag{6.8}$$

If $\alpha_k = 0$ for all k, then $\bar{t} = 0$ and

$$\frac{f(x+t_k d) - f(x)}{t_k} - c_1 f'(x;d) > 0 \quad \forall k.$$

But then, taking the limit in k, we obtain $f'(x;d) \ge c_1 f'(x;d)$, or equivalently, $0 > (1-c_1) f'(x;d) \ge 0$ which is a contradiction. Hence, we can assume that eventually $\alpha_k > 0$.

We now have that the sequences $\{\alpha_k\}$, $\{t_k\}$, and $\{\beta_k\}$ are infinite with (6.8) satisfied, and there is a k_0 such that $0 < \alpha_k < t_k < \beta_k < \infty$ for all $k \ge k_0$. By construction, we know that for all $k > k_0$

$$f(x+\alpha_k d) \le f(x) + c_1 \alpha_k f'(x;d) \tag{6.9}$$

$$f(x) + c_1 \beta_k f'(x;d) < f(x+\beta_k d) \tag{6.10}$$

$$f'(x+\alpha_k d;d) < c_2 f'(x;d)\ . \tag{6.11}$$

Taking the limit in k in (6.11) tells us that

$$f'(x+\bar{t}d;d) \le c_2 f'(x;d)\ . \tag{6.12}$$

Adding (6.9) and (6.10) together and using the Mean Value Theorem gives

$$c_1(\beta_k - \alpha_k) f'(x;d) \le f(x+\beta_k d) - f(x+\alpha_k d) = (\beta_k - \alpha_k) f'(x+\hat{t}_k d;d) \quad \forall\ k > k_0,$$

where $\alpha_k \le \hat{t}_k \le \beta_k$. Dividing by $(\beta_k - \alpha_k) > 0$ and taking the limit in k gives $c_1 f'(x;d) \le f'(x+\bar{t}d;d)$ which combined with (6.12) yields the contradiction $f'(x+\bar{t}d;d) \le c_2 f'(x;d) < c_1 f'(x;d) \le f'(x+\bar{t}d;d)$. Consequently, option (i) above must occur if (ii) does not. □

A global convergence result for a line search routine based on the Weak Wolfe conditions now follows.

THEOREM 6.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$, $x^0 \in \mathbb{R}^n$, and $0 < c_1 < c_2 < 1$. Assume that $\nabla f(x)$ exists and is Lipschitz continuous on an open set containing the set $\{x \mid f(x) \le f(x^0)\}$. Let $\{x^\nu\}$ be a sequence initiated at x^0 and generated by the following algorithm:*

Step 0: *Set $k = 0$.*
Step 1: *Choose $d^k \in \mathbb{R}^n$ such that $f'(x^k; d^k) < 0$.*
If no such d^k exists, then STOP.
First-order necessary conditions for optimality are satisfied at x^k.
Step 2: *Let t^k be a stepsize satisfying the Weak Wolfe conditions (6.4) and (6.5).*
If no such t^k exists, then STOP.
The function f is unbounded below.
Step 3: *Set $x^{k+1} = x^k + t_k d^k$, reset $k = k + 1$, and return to Step 1.*

One of the following must occur:

(i) The algorithm terminates finitely at a first-order stationary point for f.
(ii) For some k the stepsize selection procedure generates a sequence of trial stepsizes $t_{k\nu} \uparrow +\infty$ such that $f(x^k + t_{k\nu} d^k) \to -\infty$.
(iii) $f(x^k) \downarrow -\infty$.
(iv) $\displaystyle\sum_{k=0}^{\infty} \left\|\nabla f(x^k)\right\|^2 \cos^2 \theta_k < +\infty$, where $\displaystyle\cos\theta_k = \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\| \, \|d^k\|}$ for all $k = 1, 2, \ldots$.

PROOF. We assume that (i), (ii), and (iii) do not occur and show that (iv) occurs. Since (i) and (ii) do not occur the sequence $\{x^\nu\}$ is infinite and $f'(x^k; d^k) < 0$ for all $k = 1, 2, \ldots$. Since (ii) does not occur, the weak Wolfe conditions are satisfied at every iteration. The condition (6.4) implies that the sequence $\{f(x^k)\}$ is strictly decreasing. In particular, this implies that $\{x^\nu\} \subset \{x \mid f(x) \le f(x^0)\}$. The condition (6.5) implies that

$$(c_2 - 1)\nabla f(x^k)^T d^k \le (\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k$$

for all k. Combining this with the Lipschitz continuity of ∇f on an open neighborhood of $\{x \mid f(x) \le f(x^0)\}$, gives

$$(c_2 - 1)\nabla f(x^k)^T d^k \le (\nabla f(x^{k+1}) - \nabla f(x^k))^T d^k \le L t_k \left\|d^k\right\|^2 .$$

Hence

$$t_k \ge \frac{c_2 - 1}{L} \frac{\nabla f(x^k)^T d^k}{\|d^k\|^2} > 0.$$

Plugging this into (6.4) give the inequality

$$f(x^{k+1}) \le f(x^k) - c_1 \frac{1 - c_2}{L} \frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} = f(x^k) - c_1 \frac{1 - c_2}{L} \left\|\nabla f(x^k)\right\|^2 \cos^2 \theta_k.$$

Setting $c = c_1 \frac{1 - c_2}{L}$ and summing over k gives

$$f(x^{k+1}) \le f(x^0) - c \sum_{\nu=0}^{k} \|\nabla f(x^\nu)\|^2 \cos^2 \theta_\nu \, .$$

Since (iii) does not occur, we can take the limit in k and obtain

$$\sum_{\nu=0}^{\infty} \|\nabla f(x^\nu)\|^2 \cos^2 \theta_\nu < +\infty \, .$$

□

If the function f is bounded below and the algorithm does not terminate finitely, then Part (iv) of this theorem states that

$$\left\|\nabla f(x^k)\right\| \cos^2\theta_k \to 0 \ .$$

Hence, if the search directions d^k are chosen so that there is a $\delta > 0$, independent of the iteration k, such that $\cos\theta_k < -\delta$ for all k, then it must be the case that $\left\|\nabla f(x^k)\right\| \to 0$ so that every cluster point of the sequence $\{x^k\}$ is a first-order stationary point for f. For example, we have the following corollary to the theorem.

COROLLARY 6.5.1. *Let f and $\{x^k\}$ be as in the theorem, and let $\{B_k\}$ be a sequence of symmetric positive definite matrices for which there exists $\bar{\lambda} > \underline{\lambda} > 0$ such that*

$$\underline{\lambda}\,\|u\|^2 \le u^T B_k u \le \bar{\lambda}\,\|u\|^2 \ \ \forall\, u \in \mathbb{R}^n \text{ and } k = 1, 2, \dots \ . \tag{6.13}$$

Let us further assume that f is bounded below. If the search directions d^k are given by

$$d^k = -B_k \nabla f(x^k) \ \forall\, k = 1, 2, \dots \ ,$$

then $\nabla f(x^k) \to 0$.

PROOF. It is easily shown (see exercises) that the condition (6.13) implies that the eigenvalues of the sequence $\{B_k\}$ are uniformly lower bounded by $\underline{\lambda}$ and uniformly upper bounded by $\bar{\lambda}$. In particular, this implies that

$$\underline{\lambda}\,\|u\| \le \|B_k u\| \le \bar{\lambda}\,\|u\| \ \ \forall\, u \in \mathbb{R}^n \text{ and } k = 1, 2, \dots$$

(see exercises). Hence for all k

$$\begin{aligned}
\cos\theta_k &= \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|\,\|d^k\|} \\
&= -\frac{\nabla f(x^k)^T B_k \nabla f(x^k)}{\|\nabla f(x^k)\|\,\|B_k \nabla f(x^k)\|} \\
&\le -\frac{\underline{\lambda}\left\|\nabla f(x^k)\right\|^2}{\|\nabla f(x^k)\|\,\|B_k \nabla f(x^k)\|} \\
&\le -\frac{\underline{\lambda}\left\|\nabla f(x^k)\right\|^2}{\|\nabla f(x^k)\|\,\bar{\lambda}\,\|\nabla f(x^k)\|} \\
&= -\underline{\lambda}/\bar{\lambda} \\
&< 0 \ .
\end{aligned}$$

Therefore $\nabla f(x^k) \to 0$. □

A possible choice for the matrices B_k in the above result is $B_k = I$ for all k. This essentially gives the method of steepest descent.

CHAPTER 7

Search Directions for Unconstrained Optimization

In this chapter we study the choice of search directions used in our basic updating scheme

$$x^{k+1} = x^k + t_k d^k \ .$$

for solving

$$\mathcal{P} \qquad \min_{x \in \mathbb{R}^n} f(x).$$

All of the search directions considered can be classified as *Newton-like* since they are all of the form

$$d^k = -H_k \nabla f(x^k),$$

where H_k is a symmetric $n \times n$ matrix. If $H_k = \mu_k I$ for all k, the resulting search directions a a scaled steepest descent direction with scale factors μ_k. More generally, we choose H_k to approximate $\nabla^2 f(x^k)^{-1}$ in order to approximate Newton's method for optimization. The Newton is important since it possesses rapid local convergence properties, and can be shown to be *scale independent*. We precede our discussion of search directions by making precise a useful notion of speed or *rate of convergence*.

1. Rate of Convergence

We focus on notions of quotient rates convergence, or Q-convergence rates. Let $\{x^\nu\} \subset \mathbb{R}^n$ and $\overline{x} \in \mathbb{R}^n$ be such that $\overline{x}^\nu \to \overline{x}$. We say that $\overline{x}^\nu \to \overline{x}$ at a *linear* rate if

$$\limsup_{\nu \to \infty} \frac{\left\| x^{\nu+1} - \overline{x} \right\|}{\| x^\nu - \overline{x} \|} < 1 \ .$$

The convergence is said to be *superlinear* if this limsup is 0. The convergence is said to be *quadratic* if

$$\limsup_{\nu \to \infty} \frac{\left\| x^{\nu+1} - \overline{x} \right\|}{\| x^\nu - \overline{x} \|^2} < \infty \ .$$

For example, given $\gamma \in (0,1)$ the sequence $\{\gamma^\nu\}$ converges linearly to zero, but not superlinearly. The sequence $\{\gamma^{\nu^2}\}$ converges superlinearly to 0, but not quadratically. Finally, the sequence $\{\gamma^{2^\nu}\}$ converges quadratically to zero. Superlinear convergence is much faster than linear convergences, but quadratic convergence is much, much faster than superlinear convergence.

2. Newton's Method for Solving Equations

Newton's method is an iterative scheme designed to solve nonlinear equations of the form

$$g(x) = 0, \tag{7.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be continuously differentiable. Many problems of importance can be posed in this way. In the context of the optimization problem $\mathcal{P}$, we wish to locate critical points, that is, points at which $\nabla f(x) = 0$. We begin our discussion of Newton's method in the usual context of equation solvng.

Assume that the function g in (7.1) is continuously differentiable and that we have an approximate solution $x^0 \in \mathbb{R}^n$. We now wish to improve on this approximation. If $\overline{x}$ is a solution to (7.1), then

$$0 = g(\overline{x}) = g(x^0) + g'(x^0)(\overline{x} - x^0) + o\|\overline{x} - x^0\|.$$

Thus, if x^0 is "close" to $\overline{x}$, it is reasonable to suppose that the solution to the linearized system

$$0 = g(x^0) + g'(x^0)(x - x^0) \tag{7.2}$$

is even closer. This is Newton's method for finding the roots of the equation $g(x) = 0$. It has one obvious pitfall. Equation (7.2) may not be consistent. That is, there may not exist a solution to (7.2).

For the sake of the present argument, we assume that (3) holds, i.e. $g'(x^0)^{-1}$ exists. Under this assumption (7.2) defines the iteration scheme,

$$x^{k+1} := x^k - [g'(x^k)]^{-1} g(x^k), \tag{7.3}$$

called the Newton iteration. The associated direction

$$d^k := -[g'(x^k)]^{-1} g(x^k). \tag{7.4}$$

is called the Newton direction. We analyze the convergence behavior of this scheme under the additional assumption that only an approximation to $g'(x^k)^{-1}$ is available. We denote this approximation by J_k. The resulting iteration scheme is

$$x^{k+1} := x^k - J_k g(x^k). \tag{7.5}$$

Methods of this type are called *Newton-Like methods*.

THEOREM 7.1. *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable, $x^0 \in \mathbb{R}^n$, and $J_0 \in \mathbb{R}^{n\times n}$. Suppose that there exists $\overline{x}$, $x^0 \in \mathbb{R}^n$, and $\epsilon > 0$ with $\left\|x^0 - \overline{x}\right\| < \epsilon$ such that*

(1) $g(\overline{x}) = 0$,
(2) $g'(x)^{-1}$ exists for $x \in B(\overline{x};\epsilon) := \{x \in \mathbb{R}^n : \|x - \overline{x}\| < \epsilon\}$ with

$$\sup\{\|g'(x)^{-1}\| : x \in B(\overline{x};\epsilon)] \le M_1$$

(3) g' is Lipschitz continuous on $c\ell B(\overline{x};\epsilon)$ with Lipschitz constant L, and
(4) $\theta_0 := \frac{LM_1}{2}\|x^0 - \overline{x}\| + M_0 K < 1$ where $K \ge \|(g'(x^0)^{-1} - J_0)y^0\|$, $y^0 := g(x^0)/\|g(x^0)\|$, and $M_0 = \max\{\|g'(x)\| : x \in B(\overline{x};\epsilon)\}$.

Further suppose that iteration (7.5) is initiated at x^0 where the J_k's are chosen to satisfy one of the following conditions;

(i) $\|(g'(x^k)^{-1} - J_k)y^k\| \le K$,
(ii) $\|(g'(x^k)^{-1} - J_k)y^k\| \le \theta_1^k K$ for some $\theta_1 \in (0,1)$,
(iii) $\|(g'(x^k)^{-1} - J_k)y^k\| \le \min\{M_3\|x^k - x^{k-1}\|, K\}$, for some $M_2 > 0$, or
(iv) $\|(g'(x^k)^{-1} - J_k)y^k\| \le \min\{M_2\|g(x^k)\|, K\}$, for some $M_3 > 0$,

where for each $k = 1, 2, \ldots$, $y^k := g(x^k)/\left\|g(x^k)\right\|$.

These hypotheses on the accuracy of the approximations J_k yield the following conclusions about the rate of convergence of the iterates x^k.

(a) If (i) holds, then $x^k \to \overline{x}$ linearly.
(b) If (ii) holds, then $x^k \to \overline{x}$ superlinearly.
(c) If (iii) holds, then $x^k \to \overline{x}$ two step quadratically.
(d) If (iv) holds, then $x^k \to \overline{x}$ quadratically.

PROOF. We begin by inductively establishing the basic inequalities

$$\|x^{k+1} - \overline{x}\| \le \frac{LM_1}{2}\|x^k - \overline{x}\|^2 + \|(g'(x^k)^{-1} - J_k)g(x^k)\|, \tag{7.6}$$

and

$$\|x^{k+1} - \overline{x}\| \le \theta_0\|x^k - \overline{x}\| \tag{7.7}$$

as well as the inclusion

$$x^{k+1} \in B(\overline{x};\epsilon) \tag{7.8}$$

for $k = 0, 1, 2, \ldots$. For $k = 0$ we have

$$\begin{aligned} x^1 - \overline{x} &= x^0 - \overline{x} - g'(x^0)^{-1} g(x^0) + \left[g'(x^0)^{-1} - J_0\right] g(x^0) \\ &= g'(x^0)^{-1} \left[g(\overline{x}) - (g(x^0) + g'(x^0)(\overline{x} - x^0))\right] \\ &\quad + \left[g'(x^0)^{-1} - J_0\right] g(x^0), \end{aligned}$$

since $g'(x^0)^{-1}$ exists by the hypotheses. Consequently, the hypothese (1)–(4) plus the quadratic bound lemma imply that

$$\begin{aligned} \left\|x^{k+1} - \overline{x}\right\| &\leq \left\|g'(x^0)^{-1}\right\| \left\|g(\overline{x}) - \left(g(x^0) + g'(x^0)(\overline{x} - x^0)\right)\right\| \\ &\quad + \left\|\left(g'(x^0)^{-1} - J_0\right) g(x^0)\right\| \\ &\leq \frac{M_1 L}{2} \left\|x^0 - \overline{x}\right\|^2 + K \left\|g(x^0) - g(\overline{x})\right\| \\ &\leq \frac{M_1 L}{2} \left\|x^0 - \overline{x}\right\|^2 + M_0 K \left\|x^0 - \overline{x}\right\| \\ &\leq \theta_0 \left\|x^0 - \overline{x}\right\| < \epsilon, \end{aligned}$$

whereby (7.6) – (7.8) are established for $k = 0$.

Next suppose that (7.6) – (7.8) hold for $k = 0, 1, \ldots, s-1$. We show that (7.6) – (7.8) hold at $k = s$. Since $x^s \in B(\overline{x}, \epsilon)$, hypotheses (2)–(4) hold at x^s, one can proceed exactly as in the case $k = 0$ to obtain (7.6). Now if any one of (i)–(iv) holds, then (i) holds. Thus, by (7.6), we find that

$$\begin{aligned} \left\|x^{s+1} - \overline{x}\right\| &\leq \tfrac{M_1 L}{2} \left\|x^s - \overline{x}\right\|^2 + \left\|(g'(x^s)^{-1} - J_s) g(x^s)\right\| \\ &\leq \left[\tfrac{M_1 L}{2} \theta_0^s \left\|x^0 - \overline{x}\right\| + M_0 K\right] \left\|x^s - \overline{x}\right\| \\ &\leq \left[\tfrac{M_1 L}{2} \left\|x^0 - \overline{x}\right\| + M_0 K\right] \left\|x^s - \overline{x}\right\| \\ &= \theta_0 \left\|x^s - \overline{x}\right\|. \end{aligned}$$

Hence $\left\|x^{s+1} - \overline{x}\right\| \leq \theta_0 \left\|x^s - \overline{x}\right\| \leq \theta_0 \epsilon < \epsilon$ and so $x^{s+1} \in B(\overline{x}, \epsilon)$. We now proceed to establish (a)–(d).
(a) This clearly holds since the induction above established that

$$\left\|x^{k+1} - \overline{x}\right\| \leq \theta_0 \left\|x^k - \overline{x}\right\|.$$

(b) From (7.6), we have

$$\begin{aligned} \left\|x^{k+1} - \overline{x}\right\| &\leq \frac{LM_1}{2} \left\|x^k - \overline{x}\right\|^2 + \left\|(g'(x^k)^{-1} - J_k) g(x^k)\right\| \\ &\leq \frac{LM_1}{2} \left\|x^k - \overline{x}\right\|^2 + \theta_1^k K \left\|g(x^k)\right\| \\ &\leq \left[\frac{LM_1}{2} \theta_0^k \left\|x^0 - \overline{x}\right\| + \theta_1^k M_0 K\right] \left\|x^k - \overline{x}\right\| \end{aligned}$$

Hence $x^k \to \overline{x}$ superlinearly.

(c) From (7.6) and the fact that $x^k \to \overline{x}$, we eventually have

$$\begin{aligned}
\left\|x^{k+1}-\overline{x}\right\| &\leq \frac{LM_1}{2}\left\|x^k-\overline{x}\right\|^2+\left\|(g'(x^k)^{-1}-J_k)g(x^k)\right\| \\
&\leq \frac{LM_1}{2}\left\|x^k-\overline{x}\right\|^2+M_2\left\|x^k-x^{k-1}\right\|\left\|g(x^k)\right\| \\
&\leq \left[\frac{LM_1}{2}\left\|x^k-\overline{x}\right\|+M_0M_2\left[\left\|x^{k-1}-\overline{x}\right\|+\left\|x^k-\overline{x}\right\|\right]\right]\left\|x^k-\overline{x}\right\| \\
&\leq \left[\frac{LM_1}{2}\theta_0\left\|x^{k-1}-\overline{x}\right\|+M_0M_2(1+\theta_0)\left\|x^{k-1}-\overline{x}\right\|\right] \\
&\qquad \times\theta_0\left\|x^{k-1}-\overline{x}\right\| \\
&= \left[\frac{LM_1}{2}\theta_0+M_0M_2(1+\theta_0)\right]\theta_0\left\|x^{k-1}-\overline{x}\right\|^2.
\end{aligned}$$

Hence $x^k \to \overline{x}$ two step quadratically.

(d) Again by (7.6) and the fact that $x^k \to \overline{x}$, we eventually have

$$\begin{aligned}
\left\|x^{k+1}-\overline{x}\right\| &\leq \frac{LM_1}{2}\left\|x^k-\overline{x}\right\|^2+\left\|(g'(x^k)^{-1}-J_k)g(x^k)\right\| \\
&\leq \frac{LM_1}{2}\left\|x^k-\overline{x}\right\|^2+M_2\left\|g(x^k)\right\|^2 \\
&\leq \left[\frac{LM_1}{2}+M_2M_0^2\right]\left\|x^k-\overline{x}\right\|^2.
\end{aligned}$$

□

Note that the conditions required for the approximations to the Jacobian matrices $g'(x^k)^{-1}$ given in (i)–(ii) do not imply that $J_k \to g'(\overline{x})^{-1}$. The stronger conditions

$(i)'$ $\left\|g'(x^k)^{-1}-J_k\right\| \leq \left\|g'(x^0)^{-1}-J_0\right\|$,
$(ii)'$ $\left\|g'(x^{k+1})^{-1}-J_{k+1}\right\| \leq \theta_1\left\|g'(x^k)^{-1}-J_k\right\|$ for some $\theta_1 \in (0,1)$,
$(iii)'$ $\left\|g'(x^k)^{-1}-J_k\right\| \leq \min\{M_2\left\|x^{k+1}-x^k\right\|, \left\|g'(x^0)^{-1}-J_0\right\|\}$ for some $M_2>0$, or
$(iv)'$ $g'(x^k)^{-1}=J_k$,

which imply the conditions (i) through (iv) of Theorem 7.1 respectively, all imply the convergence of the inverse Jacobian approximates to $g'(\overline{x})^{-1}$. The conditions $(i)'$–$(iv)'$ are less desirable since they require greater expense and care in the construction of the inverse Jacobian approximates.

3. Newton's Method for Minimization

We now translate the results of previous section to the optimization setting. The underlying problem is

$$\mathcal{P} \qquad \min_{x\in\mathbb{R}^n} f(x)\ .$$

The Newton-like iterations take the form

$$x^{k+1} = x^k - H_k\nabla f(x^k),$$

where H_k is an approximation to the inverse of the Hessian matrix $\nabla^2 f(x^k)$.

THEOREM 7.2. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be twice continuously differentiable, $x^0\in\mathbb{R}^n$, and $H_0\in\mathbb{R}^{n\times n}$. Suppose that*

(1) there exists $\overline{x}\in\mathbb{R}^n$ and $\epsilon > \left\|x^0-\overline{x}\right\|$ such that $f(\overline{x})\leq f(x)$ whenever $\|x-\overline{x}\|\leq\epsilon$,

(2) there is a $\delta > 0$ such that $\delta \|z\|_2^2 \leq z^T \nabla^2 f(x) z$ for all $x \in B(\overline{x}, \epsilon)$,
(3) $\nabla^2 f$ is Lipschitz continuous on $\mathrm{cl}\,(B)\,(\overline{x}; \epsilon)$ with Lipschitz constant L, and
(4) $\theta_0 := \frac{L}{2\delta} \left\|x^0 - \overline{x}\right\| + M_0 K < 1$ where $M_0 > 0$ satisfies $z^T \nabla^2 f(x) z \leq M_0 \|z\|_2^2$ for all $x \in B(\overline{x}, \epsilon)$ and $K \geq \|(\nabla^2 f(x^0)^{-1} - H_0) y^0\|$ with $y^0 = \nabla f(x^0) / \left\|\nabla f(x^0)\right\|$.

Further, suppose that the iteration

$$x^{k+1} := x^k - H_k \nabla f(x^k) \tag{7.9}$$

is initiated at x^0 where the H_k's are chosen to satisfy one of the following conditions:

(i) $\left\|(\nabla^2 f(x^k)^{-1} - H_k) y^k\right\| \leq K$,
(ii) $\left\|(\nabla^2 f(x^k)^{-1} - H_k) y^k\right\| \leq \theta_1^k K$ for some $\theta_1 \in (0, 1)$,
(iii) $\left\|(\nabla^2 f(x^k)^{-1} - H_k) y^k\right\| \leq \min\{M_2 \left\|x^k - x^{k-1}\right\|, K\}$, for some $M_2 > 0$, or
(iv) $\left\|(\nabla^2 f(x^k)^{-1} - H_k) y^k\right\| \leq \min\{M_3 \left\|\nabla f(x^k)\right\|, K\}$, for some $M_3 > 0$,

where for each $k = 1, 2, \ldots$ $y^k := \nabla f(x^k) / \left\|\nabla f(x^k)\right\|$.

These hypotheses on the accuracy of the approximations H_k yield the following conclusions about the rate of convergence of the iterates x^k.

(a) If (i) holds, then $x^k \to \overline{x}$ linearly.
(b) If (ii) holds, then $x^k \to \overline{x}$ superlinearly.
(c) If (iii) holds, then $x^k \to \overline{x}$ two step quadratically.
(d) If (iv) holds, then $x^k \to \overline{x}$ quadradically.

To more fully understand the convergence behavior described in this theorem, let us examine the nature of the controling parameters L, M_0, and M_1. Since L is a Lipschitz constant for $\nabla^2 f$ it loosely corresponds to a bound on the third–order behavior of f. Thus the assumptions for convergence make implicit demands on the third derivative. The constant δ is a local lower bound on the eigenvalues of $\nabla^2 f$ near $\overline{x}$. That is, f behaves locally as if it were a *strongly convex function* (see exercises) with modulus δ. Finally, M_0 can be interpreted as a local Lipschitz constant for ∇f and only plays a role when $\nabla^2 f$ is approximated inexactly by H_k's.

We now illustrate the performance differences between the method of steepest descent and Newton's method on a simple one dimensional problem. Let $f(x) = x^2 + e^x$. Clearly, f is a strongly convex function with

$$\begin{aligned} f(x) &= x^2 + e^x \\ f'(x) &= 2x + e^x \\ f''(x) &= 2 + e^x \ > \ 2 \\ f'''(x) &= e^x. \end{aligned}$$

If we apply the steepest descent algorithm with backtracking ($\gamma = 1/2,\ c = 0.01$) initiated at $x^0 = 1$, we get the following table

k	x^k	$f(x^k)$	$f'(x^k)$	s
0	1	.37182818	4.7182818	0
1	0	1	1	0
2	$-.5$	.8565307	-0.3934693	1
3	$-.25$	.8413008	0.2788008	2
4	$-.375$	.8279143	$-.0627107$	3
5	$-.34075$	.8273473	.0297367	5
6	$-.356375$	.8272131	$-.01254$	6
7	$-.3485625$	.8271976	.0085768	7
8	$-.3524688$	.8271848	$-.001987$	8
9	$-.3514922$	.8271841	.0006528	10
10	$-.3517364$	.827184	$-.0000072$	12

If we apply Newton's method from the same starting point and take a unit step at each iteration, we obtain a dramatically different table.

x	$f'(x)$
1	4.7182818
0	1
$-1/3$	.0498646
$-.3516893$	.00012
$-.3517337$	.00000000064

In addition, one more iteration gives $|f'(x^5)| \leq 10^{-20}$. This is a stunning improvement in performance and shows why one always uses Newton's method (or an approximation to it) whenever possible.

Our next objective is to develop numerically viable methods for approximating Jacobians and Hessians in Newton-like methods.

4. Matrix Secant Methods

Let us return to the problem of finding $\overline{x} \in \mathbb{R}^n$ such that $g(\overline{x}) = 0$ where $g : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. In this section we consider Newton-Like methods of a special type. Recall that in a Newton-Like method the iteration scheme takes the form

$$x^{k+1} := x^k - J_k g(x^k), \tag{7.10}$$

where J_k is meant to approximate the inverse of $g'(x^k)$. In the one dimensional case, a method proposed by the Babylonians 3700 years ago is of particular significance. Today we call it the *secant method*:

$$J_k = \frac{x^k - x^{k-1}}{g(x^k) - g(x^{k-1})}. \tag{7.11}$$

With this approximation one has

$$g'(x^k)^{-1} - J_k = \frac{g(x^{k-1}) - [g(x^k) + g'(x^k)(x^{k-1} - x^k)]}{g'(x^k)[g(x^{k-1}) - g(x^k)]}.$$

Near a point x^* at which $g'(x^*) \neq 0$ one can use the MVT to show there exists an $\alpha > 0$ such that

$$\alpha \left\| x - y \right\| \leq \left\| g(x) - g(y) \right\|.$$

Consequently, by the Quadratic Bound Lemma,

$$\left\| g'(x^k)^{-1} - J_k \right\| \leq \frac{\frac{L}{2} \left\| x^{k-1} - x^k \right\|^2}{\alpha \left\| g'(x^k) \right\| \left\| x^{k-1} - x^k \right\|} \leq K \left\| x^{k-1} - x^k \right\|$$

for some constant $K > 0$ whenever x^k and x^{k-1} are sufficiently close to x^*. Therefore, by our convergence Theorem for Newton Like methods, the secant method is locally two step quadratically convergent to a non–singular solution of the equation $g(x) = 0$. An additional advantage of this approach is that no extra function evaluations are required to obtain the approximation J_k.

4.0.1. *Matrix Secant Methods for Equations.* Unfortunately, the secant approximation (7.11) is meaningless if the dimension n is greater than 1 since division by vectors is undefined. But this can be rectified by multiplying (7.11) on the right by $(g(x^{k-1}) - g(x^k))$ and writing

$$J_k(g(x^k) - g(x^{k-1})) = x^k - x^{k-1}. \tag{7.12}$$

Equation (7.12) is called the *Quasi-Newton equation* (QNE), or *matrix secant equation* (MSE), at x^k. Here the matrix J_k is unknown, but is required to satisfy the n linear equations of the MSE. These equations determine an n dimensional affine manifold in $\mathbb{R}^{n \times n}$. Since J_k contains n^2 unknowns, the n linear equations in (7.12) are not sufficient to uniquely determine J_k. To nail down a specific J_k further conditions on the update J_k must be given. What conditions should these be?

To develop sensible conditions on J_k, let us consider an overall iteration scheme based on (7.10). For convenience, let us denote J_k^{-1} by B_k (i.e. $B_k = J_k^{-1}$). Using the B_k's, the MSE (7.12) becomes

$$B_k(x^k - x^{k-1}) = g(x^k) - g(x^{k-1}). \tag{7.13}$$

At every iteration we have (x^k, B_k) and compute x^{k+1} by (7.10). Then B_{k+1} is constructed to satisfy (7.13). If B_k is close to $g'(x^k)$ and x^{k+1} is close to x^k, then B_{k+1} should be chosen not only to satisfy (7.13) but also to be as "close" to B_k as possible. With this in mind, we must now decide what we mean by "close". From a computational perspective, we prefer "close" to mean *easy to compute*. That is, B_{k+1} should be *algebraically close* to B_k in the sense that B_{k+1} is only a rank 1 modification of B_k. Since we are assuming that B_{k+1} is a rank 1 modification to B_k, there are vectors $u, v \in \mathbb{R}^n$ such that

$$B_{k+1} = B_k + uv^T. \tag{7.14}$$

We now use the matrix secant equation (7.13) to derive conditions on the choice of u and v. In this setting, the MSE becomes

$$B_{k+1}s^k = y^k,$$

where

$$s^k := x^{k+1} - x^k \qquad \text{and} \qquad y^k := g(x^{k+1}) - g(x^k)\ .$$

Multiplying (**??**) by s^k gives

$$y^k = B_{k+1}s^k = B_k s^k + uv^T s^k\ .$$

Hence, if $v^T s^k \neq 0$, we obtain

$$u = \frac{y^k - B_k s^k}{v^T s^k}$$

and

$$B_{k+1} = B_k + \frac{\left(y^k - B_k s^k\right) v^T}{v^T s^k}. \tag{7.15}$$

Equation (7.15) determines a whole class of rank one updates that satisfy the MSE where one is allowed to choose $v \in \mathbb{R}^n$ as long as $v^T s^k \neq 0$. If $s^k \neq 0$, then an obvious choice for v is s^k yielding the update

$$B_{k+1} = B_k = \frac{\left(y^k - B_k s^k\right) {s^k}^T}{{s^k}^T s^k}. \tag{7.16}$$

This is known as Broyden's update. It turns out that the Broyden update is also analytically close.

THEOREM 7.3. *Let $A \in \mathbb{R}^{n\times n}$, s, $y \in \mathbb{R}^n$, $s \neq 0$. Then for any matrix norms $\|\cdot\|$ and $\|\cdot\|_2$ such that*

$$\|AB\| \leq \|A\|\ \|B\|_2$$

and

$$\left\| \frac{vv^T}{v^T v} \right\|_2 \leq 1,$$

the solution to

$$\min\{\|B - A\| : Bs = y\} \tag{7.17}$$

is

$$A_+ = A + \frac{(y - As)s^T}{s^T s}. \tag{7.18}$$

In particular, (7.18) *solves* (7.17) *when $\|\cdot\|$ is the ℓ_2 matrix norm, and* (7.18) *solves* (7.17) *uniquely when $\|\cdot\|$ is the Frobenius norm.*

PROOF. Let $B \in \{B \in \mathbb{R}^{n\times n} : Bs = y\}$, then

$$\begin{aligned} \|A_+ - A\| &= \left\| \frac{(y - As)s^T}{s^T s} \right\| = \left\| (B - A)\frac{ss^T}{s^T s} \right\| \\ &\leq \|B - A\| \left\| \frac{ss^T}{s^T s} \right\|_2 \leq \|B - A\| . \end{aligned}$$

Note that if $\|\cdot\|_2 = \|\cdot\|_2$, then

$$\begin{aligned} \left\| \frac{vv^T}{v^T v} \right\|_2 &= \sup \left\{ \left\| \frac{vv^T}{v^T v} x \right\|_2 \middle| \|x\|_2 = 1 \right\} \\ &= \sup \left\{ \sqrt{\frac{(v^T x)^2}{\|v\|^2}} \middle| \|x\|_2 = 1 \right\} \\ &= 1, \end{aligned}$$

so that the conclusion of the result is not vacuous. For uniqueness observe that the Frobenius norm is strictly convex and $\|A \cdot B\|_F \leq \|A\|_F \|B\|_2$. □

Therefore, the Broyden update (7.16) is both algebraically and analytically close to B_k. These properties indicate that it should perform well in practice and indeed it does.

Algorithm: Broyden's Method

Initialization: $x^0 \in \mathbb{R}^n$, $B_0 \in \mathbb{R}^{n\times n}$

Having (x^k, B_k) compute (x^{k+1}, B_{x+1}) as follows:

Solve $B_k s^k = -g(x^k)$ for s^k and set

$$\begin{aligned} x^{k+1} : &= x^k + s^k \\ y^k : &= g(x^{k+1}) - g(x^k) \\ B_{k+1} : &= B_k + \frac{(y^k - B_k s^k){s^k}^T}{{s^k}^T s^k}. \end{aligned}$$

We would prefer to write the Broyden update in terms of the matrices $J_k = B_k^{-1}$ so that we can write the step computation as $s^k = -J_k g(x^k)$ avoiding the need to solve an equation. To obtain the formula for J_k we use the the following important lemma for matrix inversion.

LEMMA 7.4. (Sherman-Morrison-Woodbury) *Suppose $A \in \mathbb{R}^{n\times n}$, $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{n\times k}$ are such that both A^{-1} and $(I + V^T A^{-1} U)^{-1}$ exist, then*

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}U(I + V^T A^{-1}U)^{-1}V^T A^{-1}$$

The above lemma verifies that if $B_k^{-1} = J_k$ exists and ${s^k}^T J_k y^k = {s^k}^T B_k^{-1} y^k \neq 0$, then

$$J_{k+1} = \left[B_k + \frac{(y^k - B_k s^k){s^k}^T}{{s^k}^T s^k} \right]^{-1} = B_k^{-1} + \frac{(s^k - B_k^{-1} y^k){s^k}^T B_k^{-1}}{{s^k}^T B_k^{-1} y} = J_k + \frac{(s^k - J_k y^k){s^k}^T J_k}{{s^k}^T J_k y}. \tag{7.19}$$

In this case, it is possible to directly update the inverses J_k. It should be cautioned though that this process can become numerically unstable if $|{s^k}^T J_k y^k|$ is small. Therefore, in practise, the value $|{s^k}^T J_k y^k|$ must be monitored to avoid numerical instability.

Although we do not pause to establish the convergence rates here, we do give the following result due to Dennis and Moré (1974).

THEOREM 7.5. *Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Assume that there exists $x^* \in \mathbb{R}^n$ and $r, \beta > 0$ such that $x^* + r\mathbb{B} \subset D$, $g(x^*) = 0$, $g'(x^*)^{-1}$ exists with $\|g'(x^*)^{-1}\| \leq \beta$, and g' is Lipschitz continuous on $x^* + r\mathbb{B}$ with Lipschitz constant $\gamma > 0$. Then there exist positive constants ϵ and δ such that if $\|x^0 - x^*\|_2 \leq \epsilon$ and $\|B_0 - g'(x^0)\| \leq \delta$, then the sequence $\{x^k\}$ generated by the iteration*

$$\left[\begin{array}{rcl} x^{k+1} & := & x^k + s^k \text{ where } s^k \text{ solves } 0 = g(x^k) + B_k s \\ B_{k+1} & := & B_k + \frac{(y^k - B_k s^k)s_k^T}{s_k^T s^k} \text{ where } y^k = g(x^{k+1}) - g(x^k) \end{array}\right.$$

is well-defined with $x^k \to x^*$ superlinearly.

4.0.2. *Matrix Secant Methods for Minimization.* We now extend these matrix secant ideas to optimization, specifically minimization. The underlying problem we consider is

$$\mathcal{P} : \underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(x) \ ,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is assumed to be twice continuously differentiable. In this setting, we wish to solve the equation $\nabla f(x) = 0$ and the MSE (7.13) becomes

$$H_{k+1} y^k = s^k \ , \tag{7.20}$$

where $s^k := x^{k+1} - x^k$ and

$$y^k := \nabla f(x^{k+1}) - \nabla f(x^k).$$

Here the matrix H_k is intended to be an approximation to the inverse of the hessian matrix $\nabla^2 f(x^k)$. Writing $M_k = H_k^{-1}$, a straightforward application of Broyden's method gives the update

$$M_{k+1} = M_k + \frac{(y^k - M_k s^k){s^k}^T}{{s^k}^T s^k}.$$

However, this is unsatisfactory for two reasons:

(1) Since M_k approximates $\nabla^2 f(x^k)$ it must be symmetric.
(2) Since we are minimizing, then M_k must be positive definite to insure that $s^k = -M_k^{-1}\nabla f(x^k)$ is a direction of descent for f at x^k.

To address problem 1 above, one could return to equation (7.15) an find an update that preserves symmetry. Such an update is uniquely obtained by setting

$$v = (y^k - M_k s^k).$$

This is called the symmetric rank 1 update or SR1. Although this update can on occasion exhibit problems with numerical stability, it has recently received a great deal of renewed interest. The stability problems occur whenever

$$v^T s^k = (y^k - M_k s^k)^T s^s \tag{7.21}$$

has small magnitude. The inverse SR1 update is given by

$$H_{k+1} = H_k + \frac{(s^k - H_k y^k)(s^k - H_k y^k)^T}{(s^k - H_k y^k)^T y^k}$$

which exists whenever $(s^k - H_k y^k)^T y^k \neq 0$.

We now approach the question of how to update M_k in a way that addresses both the issue of symmetry and positive definiteness while still using the Broyden updating ideas. Given a symmetric positive definite matrix M and two vectors s and y, our goal is to find a symmetric positive definite matrix $\bar{M}$ such that $\bar{M}s = y$. Since M is symmertic and positive definite, there is a non-singular $n \times n$ matrix L such that $M = LL^T$. Indeed, L can be chosen to be the lower triangular Cholesky factor of M. If $\overline{M}$ is also

symmetric and positive definite then there is a matrix $J \in \mathbb{R}^{n\times n}$ such that $\overline{M} = JJ^T$. The MSE (**??**) implies that if

$$J^T s = v \tag{7.22}$$

then

$$Jv = y. \tag{7.23}$$

Let us apply the Broyden update technique to (7.23), J, and L. That is, suppose that

$$J = L + \frac{(y - Lv)v^T}{v^T v}. \tag{7.24}$$

Then by (7.22)

$$v = J^T s = L^T s + \frac{v(y - Lv)^T s}{v^T v}. \tag{7.25}$$

This expression implies that v must have the form

$$v = \alpha L^T s$$

for some $\alpha \in \mathbb{R}$. Substituting this back into (7.25) we get

$$\alpha L^T s = L^T s + \frac{\alpha L^T s(y - \alpha LL^T s)^T s}{\alpha^2 s^T LL^T s}.$$

Hence

$$\alpha^2 = \left[\frac{s^T y}{s^T M s}\right]. \tag{7.26}$$

Consequently, such a matrix J satisfying (7.25) exists only if $s^T y > 0$ in which case

$$J = L + \frac{(y - \alpha Ms)s^T L}{\alpha s^T M s},$$

with

$$\alpha = \left[\frac{s^T y}{s^T M s}\right]^{1/2},$$

yielding

$$\overline{M} = M + \frac{yy^T}{y^T s} - \frac{Mss^T M}{s^T M s}. \tag{7.27}$$

Moreover, the Cholesky factorization for $\overline{M}$ can be obtained directly from the matrices J. Specifically, if the QR factorization of J^T is $J^T = QR$, we can set $\overline{L} = R$ yielding

$$\overline{M} = JJ^T = R^T Q^T QR = \overline{L}\,\overline{L}^T.$$

The formula for updating the inverses is again given by applying the Sherman-Morrison-Woodbury formula to obtain

$$\overline{H} = H + \frac{(s + Hy)^T yss^T}{(s^T y)^2} - \frac{Hys^T + sy^T H}{s^T y}, \tag{7.28}$$

where $H = M^{-1}$. The update (7.27) is called the BFGS update and (7.28) the inverse BFGS update. The letter BFGS stand for Broyden, Flethcher, Goldfarb, and Shanno.

We have shown that beginning with a symmetric positive definite matrix M_k we can obtain a symmetric and positive definite update M_{k+1} that satisfies the MSE $M_{k+1}s_k = y_k$ by applying the formula (7.27) whenever ${s^k}^T y^k > 0$. We must now address the question of how to choose x^{k+1} so that ${s^k}^T y^k > 0$. Recall that

$$y = y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$$

and

$$s^k = x^{k+1} - x^k = t_k d^k \ ,$$

where

$$d^k = -t_k H_k \nabla f(x^k)$$

is the matrix secant search direction and t_k is the stepsize. Hence

$$\begin{aligned} {y^k}^T s^k &= \nabla f(x^{k+1})^T s^k - \nabla f(x^k)^T s^k \\ &= t_k(\nabla f(x^k + t_k d_k)^T d^k - \nabla f(x^k)^T d^k) \ , \end{aligned}$$

where $d^k := -H_k \nabla f(x^k)$. Since H_k is positive definite the direction d^k is a descent direction for f at x^k and so $t_k > 0$. Therefore, to insure that ${s^k}^T y^k > 0$ we need only show that $t_k > 0$ can be choosen so that

$$\nabla f(x^k + t_k d^k)^T d^k \geq \beta \nabla f(x^k)^T d^k \tag{7.29}$$

for some $\beta \in (0, 1)$ since in this case

$$\nabla f(x^k + t_k d_k)^T d^k - \nabla f(x^k)^T d^k \geq (\beta - 1)\nabla f(x^k)^T d^k > 0.$$

But this precisely the second condition in the weak Wolfe conditions with $\beta = c_2$. Hence a successful BFGS update can always be obtained. The BFGS update and is currently considered the best matrix secant update for minimization.

BFGS Updating

$$\begin{aligned} \sigma &:= \sqrt{{s^k}^T y^k} \\ \hat{s}^k &:= s^k/\sigma \\ \hat{y}^k &:= y^k/\sigma \\ H_{k+1} &:= H_k + (\hat{s}^k - H_k\hat{y}^k)(\hat{s}^k)^T + \hat{s}^k(\hat{s}^k - H_k\hat{y}^k)^T - (\hat{s}^k - H_k\hat{y}^k)^T\hat{y}^k\hat{s}^k(\hat{s}^k)^T \end{aligned}$$

APPENDIX A

Review of Matrices and Block Structures

Numerical linear algebra lies at the heart of modern scientific computing and computational science. Today it is not uncommon to perform numerical computations with matrices having millions of components. The key to understanding how to implement such algorithms is to exploit underlying structure within the matrices. In these notes we touch on a few ideas and tools for dissecting matrix structure. Specifically we are concerned with the *block structure* matrices.

1. Rows and Columns

Let $A \in \mathbb{R}^{m\times n}$ so that A has m rows and n columns. Denote the element of A in the ith row and jth column as A_{ij}. Denote the m rows of A by $A_{1\cdot},\ A_{2\cdot},\ A_{3\cdot},\ \ldots, A_{m\cdot}$ and the n columns of A by $A_{\cdot 1},\ A_{\cdot 2},\ A_{\cdot 3},\ \ldots, A_{\cdot n}$. For example, if

$$A = \begin{bmatrix} 3 & 2 & -1 & 5 & 7 & 3 \\ -2 & 27 & 32 & -100 & 0 & 0 \\ -89 & 0 & 47 & 22 & -21 & 33 \end{bmatrix},$$

then $A_{2,4} = -100$,

$$A_{1\cdot} = \begin{bmatrix} 3 & 2 & -1 & 5 & 7 & 3 \end{bmatrix},\ A_{2\cdot} = \begin{bmatrix} -2 & 27 & 32 & -100 & 0 & 0 \end{bmatrix},\ A_{3\cdot} = \begin{bmatrix} -89 & 0 & 47 & 22 & -21 & 33 \end{bmatrix}$$

and

$$A_{\cdot 1} = \begin{bmatrix} 3 \\ -2 \\ -89 \end{bmatrix},\ A_{\cdot 2} = \begin{bmatrix} 2 \\ 27 \\ 0 \end{bmatrix},\ A_{\cdot 3} = \begin{bmatrix} -1 \\ 32 \\ 47 \end{bmatrix},\ A_{\cdot 4} = \begin{bmatrix} 5 \\ -100 \\ 22 \end{bmatrix},\ A_{\cdot 5} = \begin{bmatrix} 7 \\ 0 \\ -21 \end{bmatrix},\ A_{\cdot 6} = \begin{bmatrix} 3 \\ 0 \\ 33 \end{bmatrix}.$$

Exercise A.1. *If*

$$C = \begin{bmatrix} 3 & -4 & 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 3 \end{bmatrix},$$

what are $C_{4,4}$, $C_{\cdot 4}$ and $C_{4\cdot}$? For example, $C_{2\cdot} = \begin{bmatrix} 2 & 2 & 0 & 0 & 1 & 0 \end{bmatrix}$ and $C_{\cdot 2} = \begin{bmatrix} -4 \\ 2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$.

The block structuring of a matrix into its rows and columns is of fundamental importance and is extremely useful in understanding the properties of a matrix. In particular, for $A \in \mathbb{R}^{m\times n}$ it allows us to write

$$A = \begin{bmatrix} A_{1\cdot} \\ A_{2\cdot} \\ A_{3\cdot} \\ \vdots \\ A_{m\cdot} \end{bmatrix} \text{ and } A = \begin{bmatrix} A_{\cdot 1} & A_{\cdot 2} & A_{\cdot 3} & \ldots & A_{\cdot n} \end{bmatrix}.$$

These are called the row and column block representations of A, respectively

1.1. *Matrix vector Multiplication.* Let $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$. In terms of its coordinates (or components), we can also write $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ with each $x_j \in \mathbb{R}$. The term x_j is called the jth component of x. For example if

$$x = \begin{bmatrix} 5 \\ -100 \\ 22 \end{bmatrix},$$

then $n = 3,\ x_1 = 5,\ x_2 = -100,\ x_3 = 22$. We define the matrix-vector product Ax by

$$Ax = \begin{bmatrix} A_{1\cdot} \bullet x \\ A_{2\cdot} \bullet x \\ A_{3\cdot} \bullet x \\ \vdots \\ A_{m\cdot} \bullet x \end{bmatrix},$$

where for each $i = 1, 2, \ldots, m$, $A_{i\cdot} \bullet x$ is the dot product of the ith row of A with x and is given by

$$A_{i\cdot} \bullet x = \sum_{j=1}^{n} A_{ij}x_j \ .$$

For example, if

$$A = \begin{bmatrix} 3 & 2 & -1 & 5 & 7 & 3 \\ -2 & 27 & 32 & -100 & 0 & 0 \\ -89 & 0 & 47 & 22 & -21 & 33 \end{bmatrix} \text{ and } x = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 2 \\ 3 \end{bmatrix},$$

then

$$Ax = \begin{bmatrix} 24 \\ -29 \\ -32 \end{bmatrix}.$$

Exercise A.2. *If*

$$C = \begin{bmatrix} 3 & -4 & 1 & 1 & 0 & 0 \\ 2 & 2 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 3 \end{bmatrix} \text{ and } x = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 2 \\ 3 \end{bmatrix},$$

what is Cx?

Note that if $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$, then Ax is always well defined with $Ax \in \mathbb{R}^m$. In terms of components, the ith component of Ax is given by the dot product of the ith row of A (i.e. $A_{i\cdot}$) and x (i.e. $A_{i\cdot} \bullet x$).

The view of the matrix-vector product described above is the *row-space* perspective, where the term *row-space* will be given a more rigorous definition at a later time. But there is a very different way of viewing the matrix-vector product based on a *column-space* perspective. This view uses the notion of the linear combination of a collection of vectors.

Given k vectors $v^1, v^2, \ldots, v^k \in \mathbb{R}^n$ and k scalars $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$, we can form the vector

$$\alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_k v^k \in \mathbb{R}^n \ .$$

Any vector of this kind is said to be a *linear combination* of the vectors $v^1, v^2, \ldots, v^k$ where the $\alpha_1, \alpha_2, \ldots, \alpha_k$ are called the coefficients in the linear combination. The set of all such vectors formed as linear combinations of $v^1, v^2, \ldots, v^k$ is said to be the *linear span* of $v^1, v^2, \ldots, v^k$ and is denoted

$$\operatorname{span}\left(v^1, v^2, \ldots, v^k\right) := \left\{ \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_k v^k \,\middle|\, \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R} \right\}.$$

Returning to the matrix-vector product, one has that

$$Ax = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + A_{13}x_3 + \cdots + A_{1n}x_n \\ A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + \cdots + A_{2n}x_n \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ A_{m1}x_1 + A_{m2}x_2 + A_{m3}x_3 + \cdots + A_{mn}x_n \end{bmatrix} = x_1 A_{\cdot 1} + x_2 A_{\cdot 2} + x_3 A_{\cdot 3} + \cdots + x_n A_{\cdot n},$$

which is a linear combination of the columns of A. That is, we can view the matrix-vector product Ax as taking a linear combination of the columns of A where the coefficients in the linear combination are the coordinates of the vector x.

We now have two fundamentally different ways of viewing the matrix-vector product Ax.

Row-Space view of Ax:

$$Ax = \begin{bmatrix} A_{1\cdot} \bullet x \\ A_{2\cdot} \bullet x \\ A_{3\cdot} \bullet x \\ \vdots \\ A_{m\cdot} \bullet x \end{bmatrix}$$

Column-Space view of Ax:

$$Ax = x_1 A_{\cdot 1} + x_2 A_{\cdot 2} + x_3 A_{\cdot 3} + \cdots + x_n A_{\cdot n} \ .$$

2. Matrix Multiplication

We now build on our notion of a matrix-vector product to define a notion of a matrix-matrix product which we call *matrix multiplication*. Given two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$ note that each of the columns of B resides in $\mathbb{R}^n$, i.e. $B_{\cdot j} \in \mathbb{R}^n$ $i = 1, 2, \ldots, k$. Therefore, each of the matrix-vector products $AB_{\cdot j}$ is well defined for $j = 1, 2, \ldots, k$. This allows us to define a matrix-matrix product that exploits the block column structure of B by setting

$$AB := \begin{bmatrix} AB_{\cdot 1} & AB_{\cdot 2} & AB_{\cdot 3} & \cdots & AB_{\cdot k} \end{bmatrix}. \tag{A.1}$$

Note that the jth column of AB is $(AB)_{\cdot j} = AB_{\cdot j} \in \mathbb{R}^m$ and that $AB \in \mathbb{R}^{m\times k}$, i.e.

if $H \in \mathbb{R}^{m\times n}$ and $L \in \mathbb{R}^{n\times k}$, then $HL \in \mathbb{R}^{m\times k}$.

Also note that

if $T \in \mathbb{R}^{s\times t}$ and $M \in \mathbb{R}^{r\times \ell}$, then the matrix product TM is only defined when $t = r$.

For example, if

$$A = \begin{bmatrix} 3 & 2 & -1 & 5 & 7 & 3 \\ -2 & 27 & 32 & -100 & 0 & 0 \\ -89 & 0 & 47 & 22 & -21 & 33 \end{bmatrix} \text{ and } B = \begin{bmatrix} 2 & 0 \\ -2 & 2 \\ 0 & 3 \\ 0 & 0 \\ 1 & 1 \\ 2 & -1 \end{bmatrix},$$

then

$$AB = \begin{bmatrix} A\begin{bmatrix} 2 \\ -2 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix} & A\begin{bmatrix} 0 \\ -2 \\ 3 \\ 0 \\ 1 \\ -1 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 15 & 5 \\ -58 & 150 \\ -133 & 87 \end{bmatrix}.$$

EXERCISE A.3. *if*

$$C = \begin{bmatrix} 3 & -4 & 1 & 1 \\ 2 & 2 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 1 \end{bmatrix} \text{ and } D = \begin{bmatrix} -1 & 0 & 2 & 4 & 3 \\ 0 & -2 & -1 & 4 & 5 \\ 5 & 2 & -4 & 1 & 1 \\ 3 & 0 & 1 & 0 & 0 \end{bmatrix},$$

is CD well defined and if so what is it?

The formula (A.1) can be used to give further insight into the individual components of the matrix product AB. By the definition of the matrix-vector product we have for each $j = 1, 2, \ldots, k$

$$AB_{\cdot j} = \begin{bmatrix} A_{1\cdot} \bullet B_{\cdot j} \\ A_{2\cdot} \bullet B_{\cdot j} \\ A_{m\cdot} \bullet B_{\cdot j} \end{bmatrix}.$$

Consequently,

$$(AB)_{ij} = A_{i\cdot} \bullet B_{\cdot j} \quad \forall\, i = 1, 2, \ldots m,\ j = 1, 2, \ldots, k.$$

That is, the element of AB in the ith row and jth column, $(AB)_{ij}$, is the dot product of the ith row of A with the jth column of B.

2.1. Elementary Matrices. We define the *elementary unit coordinate matrices* in $\mathbb{R}^{m\times n}$ in much the same way as we define the elementary unit coordinate vectors. Given $i \in \{1, 2, \ldots, m\}$ and $j \in \{1, 2, \ldots, n\}$, the elementary unit coordinate matrix $E_{ij} \in \mathbb{R}^{m\times n}$ is the matrix whose ij entry is 1 with all other entries taking the value zero. This is a slight abuse of notation since the notation E_{ij} is supposed to represent the ijth entry in the matrix E. To avoid confusion, we reserve the use of the letter E when speaking of matrices to the elementary matrices.

EXERCISE A.4. *(Multiplication of square elementary matrices)Let $i, k \in \{1, 2, \ldots, m\}$ and $j, \ell \in \{1, 2, \ldots, m\}$. Show the following for elementary matrices in $\mathbb{R}^{m\times m}$ first for $m = 3$ and then in general.*

(1) $E_{ij}E_{k\ell} = \begin{cases} E_{i\ell} & \text{, if } j = k, \\ 0 & \text{, otherwise.} \end{cases}$

(2) For any $\alpha \in \mathbb{R}$, if $i \neq j$, then $(I_{m\times m} - \alpha E_{ij})(I_{m\times m} + \alpha E_{ij}) = I_{m\times m}$ so that

$$(I_{m\times m} + \alpha E_{ij})^{-1} = (I_{m\times m} - \alpha E_{ij}).$$

(3) For any $\alpha \in \mathbb{R}$ with $\alpha \neq 0$, $(I + (\alpha^{-1} - 1)E_{ii})(I + (\alpha - 1)E_{ii}) = I$ so that

$$(I + (\alpha - 1)E_{ii})^{-1} = (I + (\alpha^{-1} - 1)E_{ii}).$$

EXERCISE A.5. *(Elementary permutation matrices)Let $i, \ell \in \{1, 2, \ldots, m\}$ and consider the matrix $P_{ij} \in \mathbb{R}^{m\times m}$ obtained from the identity matrix by interchanging its i and ℓth rows. We call such a matrix an elementary permutation matrix. Again we are abusing notation, but again we reserve the letter P for permutation matrices (and, later, for projection matrices). Show the following are true first for $m = 3$ and then in general.*

(1) $P_{i\ell}P_{i\ell} = I_{m\times m}$ *so that* $P_{i\ell}^{-1} = P_{i\ell}$.
(2) $P_{i\ell}^T = P_{i\ell}$.
(3) $P_{i\ell} = I - E_{ii} - E_{\ell\ell} + E_{i\ell} + E_{\ell i}$.

EXERCISE A.6. *(Three elementary row operations as matrix multiplication)In this exercise we show that the three elementary row operations can be performed by left multiplication by an invertible matrix. Let* $A \in \mathbb{R}^{m\times n}$, $\alpha \in \mathbb{R}$ *and let* $i, \ell \in \{1, 2, \ldots, m\}$ *and* $j \in \{1, 2, \ldots, n\}$. *Show that the following results hold first for* $m = n = 3$ *and then in general.*

(1) *(row interchanges) Given* $A \in \mathbb{R}^{m\times n}$, *the matrix* $P_{ij}A$ *is the same as the matrix* A *except with the* i *and* j*th rows interchanged.*
(2) *(row multiplication) Given* $\alpha \in \mathbb{R}$ *with* $\alpha \neq 0$, *show that the matrix* $(I + (\alpha - 1)E_{ii})A$ *is the same as the matrix* A *except with the* i*th row replaced by* α *times the* i*th row of* A.
(3) *Show that matrix* $E_{ij}A$ *is the matrix that contains the* j*th row of* A *in its* i*th row with all other entries equal to zero.*
(4) *(replace a row by itself plus a multiple of another row) Given* $\alpha \in \mathbb{R}$ *and* $i \neq j$, *show that the matrix* $(I + \alpha E_{ij})A$ *is the same as the matrix* A *except with the* i*th row replaced by itself plus* α *times the* j*th row of* A.

2.2. Associativity of matrix multiplication. Note that the definition of matrix multiplication tells us that this operation is associative. That is, if $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times k}$, and $C \in \mathbb{R}^{k\times s}$, then $AB \in \mathbb{R}^{m\times k}$ so that $(AB)C$ is well defined and $BC \in \mathbb{R}^{n\times s}$ so that $A(BC)$ is well defined, and, moreover,

$$(AB)C = \begin{bmatrix}(AB)C_{\cdot 1} & (AB)C_{\cdot 2} & \cdots & (AB)C_{\cdot s}\end{bmatrix} \tag{A.2}$$

where for each $\ell = 1, 2, \ldots, s$

$$\begin{aligned}(AB)C_{\cdot\ell} &= \begin{bmatrix}AB_{\cdot 1} & AB_{\cdot 2} & AB_{\cdot 3} & \cdots & AB_{\cdot k}\end{bmatrix} C_{\cdot\ell} \\ &= C_{1\ell}AB_{\cdot 1} + C_{2\ell}AB_{\cdot 2} + \cdots + C_{k\ell}AB_{\cdot k} \\ &= A\begin{bmatrix}C_{1\ell}B_{\cdot 1} + C_{2\ell}B_{\cdot 2} + \cdots + C_{k\ell}B_{\cdot k}\end{bmatrix} \\ &= A(BC_{\cdot\ell})\ .\end{aligned}$$

Therefore, we may write (A.2) as

$$\begin{aligned}(AB)C &= \begin{bmatrix}(AB)C_{\cdot 1} & (AB)C_{\cdot 2} & \cdots & (AB)C_{\cdot s}\end{bmatrix} \\ &= \begin{bmatrix}A(BC_{\cdot 1}) & A(BC_{\cdot 2}) & \ldots & A(BC_{\cdot s})\end{bmatrix} \\ &= A\begin{bmatrix}BC_{\cdot 1} & BC_{\cdot 2} & \ldots & BC_{\cdot s}\end{bmatrix} \\ &= A(BC)\ .\end{aligned}$$

Due to this associativity property, we may dispense with the parentheses and simply write ABC for this triple matrix product. Obviously longer products are possible.

EXERCISE A.7. *Consider the following matrices:*

$$A = \begin{bmatrix}2 & 3 & 1\\ 1 & 0 & -3\end{bmatrix} \quad B = \begin{bmatrix}4 & -1\\ 0 & -7\end{bmatrix} \quad C = \begin{bmatrix}-2 & 3 & 2\\ 1 & 1 & -3\\ 2 & 1 & 0\end{bmatrix}$$

$$D = \begin{bmatrix}2 & 3\\ 1 & 0\\ 8 & -5\end{bmatrix} \quad F = \begin{bmatrix}2 & 1 & 1 & 2\\ 1 & 0 & -4 & 0\\ 3 & 0 & -2 & 0\\ 5 & 1 & 1 & 1\end{bmatrix} \quad G = \begin{bmatrix}2 & 3 & 1 & -2\\ 1 & 0 & -3 & 0\end{bmatrix}\ .$$

Using these matrices, which pairs can be multiplied together and in what order? Which triples can be multiplied together and in what order (e.g. the triple product BAC is well defined)? Which quadruples can be multiplied together and in what order? Perform all of these multiplications.

3. Block Matrix Multiplication

To illustrate the general idea of block structures consider the following matrix.

$$A = \begin{bmatrix} 3 & -4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 3 \end{bmatrix}.$$

Visual inspection tells us that this matrix has structure. But what is it, and how can it be represented? We re-write the the matrix given above *blocking* out some key structures:

$$A = \left[\begin{array}{ccc|ccc} 3 & -4 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 2 & 1 & 4 \\ 0 & 0 & 0 & 1 & 0 & 3 \end{array}\right] = \left[\begin{array}{c|c} B & I_{3\times 3} \\ \hline 0_{2\times 3} & C \end{array}\right],$$

where

$$B = \begin{bmatrix} 3 & -4 & 1 \\ 0 & 2 & 2 \\ 1 & 0 & -1 \end{bmatrix}, \quad C = \begin{bmatrix} 2 & 1 & 4 \\ 1 & 0 & 3 \end{bmatrix},$$

$I_{3\times 3}$ is the 3×3 identity matrix, and $0_{2\times 3}$ is the 2×3 zero matrix. Having established this structure for the matrix A, it can now be exploited in various ways. As a simple example, we consider how it can be used in matrix multiplication.

Consider the matrix

$$M = \begin{bmatrix} 1 & 2 \\ 0 & 4 \\ -1 & -1 \\ 2 & -1 \\ 4 & 3 \\ -2 & 0 \end{bmatrix}.$$

The matrix product AM is well defined since A is 5×6 and M is 6×2. We show how to compute this matrix product using the structure of A. To do this we must first *block decompose M conformally with the block decomposition of A.* Another way to say this is that we must give M a block structure that allows us to do block matrix multiplication with the blocks of A. The correct block structure for M is

$$M = \begin{bmatrix} X \\ Y \end{bmatrix},$$

where

$$X = \begin{bmatrix} 1 & 2 \\ 0 & 4 \\ -1 & -1 \end{bmatrix}, \quad \text{and} \quad Y = \begin{bmatrix} 2 & -1 \\ 4 & 3 \\ -2 & 0 \end{bmatrix},$$

since then X can multiply $\begin{bmatrix} B \\ 0_{2\times 3} \end{bmatrix}$ and Y can multiply $\begin{bmatrix} I_{3\times 3} \\ C \end{bmatrix}$. This gives

$$\begin{aligned} AM &= \begin{bmatrix} B & I_{3\times 3} \\ 0_{2\times 3} & C \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} BX+Y \\ CY \end{bmatrix} \\ &= \begin{bmatrix} \begin{bmatrix} 2 & -11 \\ 2 & 12 \\ -1 & -2 \end{bmatrix} + \begin{bmatrix} -2 & 6 \\ 4 & 3 \\ -2 & 0 \end{bmatrix} \\ \\ \begin{bmatrix} 0 & 1 \\ -4 & -1 \end{bmatrix} \end{bmatrix} \\ &= \begin{bmatrix} 4 & -12 \\ 2 & 9 \\ 0 & 3 \\ 0 & 1 \\ -4 & -1 \end{bmatrix}. \end{aligned}$$

Block structured matrices and their matrix product is a very powerful tool in matrix analysis. Consider the matrices $M \in \mathbb{R}^{n\times m}$ and $T \in \mathbb{R}^{m\times k}$ given by

$$M = \begin{bmatrix} A_{n_1\times m_1} & B_{n_1\times m_2} \\ C_{n_2\times m_1} & D_{n_2\times m_2} \end{bmatrix}$$

and

$$T = \begin{bmatrix} E_{m_1\times k_1} & F_{m_1\times k_2} & G_{m_1\times k_3} \\ H_{m_2\times k_1} & J_{m_2\times k_2} & K_{m_2\times k_3} \end{bmatrix},$$

where $n = n_1 + n_2,\ m = m_1 + m_2$, and $k = k_1 + k_2 + k_3$. The block structures for the matrices M and T are said to be *conformal* with respect to matrix multiplication since

$$MT = \begin{bmatrix} AE+BH & AF+BJ & AG+BK \\ CE+DH & CF+DJ & CG+DK \end{bmatrix}.$$

Similarly, one can conformally block structure matrices with respect to matrix addition (how is this done?).

Exercise A.8. *Consider the the matrix*

$$H = \begin{bmatrix} -2 & 3 & 2 & 0 & 0 & 0 & 0 \\ 1 & 1 & -3 & 0 & 0 & 0 & 0 \\ 2 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 2 & -7 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 3 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 8 & -5 \end{bmatrix}.$$

Does H have a natural block structure that might be useful in performing a matrix-matrix multiply, and if so describe it by giving the blocks? Describe a conformal block decomposition of the matrix

$$M = \begin{bmatrix} 1 & 2 \\ 3 & -4 \\ -5 & 6 \\ 1 & -2 \\ -3 & 4 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

that would be useful in performing the matrix product HM. Compute the matrix product HM using this conformal decomposition.

EXERCISE A.9. *Let $T \in \mathbb{R}^{m\times n}$ with $T \neq 0$ and let I be the $m \times m$ identity matrix. Consider the block structured matrix $A = [\, I \quad T \,]$.*

(i) If $A \in \mathbb{R}^{k\times s}$, what are k and s?

(ii) Construct a non-zero $s \times n$ matrix B such that $AB = 0$.

The examples given above illustrate how block matrix multiplication works and why it might be useful. One of the most powerful uses of block structures is in understanding and implementing standard *matrix factorizations* or reductions.

4. Gauss-Jordan Elimination Matrices and Reduction to Reduced Echelon Form

In this section, we show that Gaussian-Jordan elimination can be represented as a consequence of left multiplication by a specially designed matrix called a *Gaussian-Jordan elimination matrix*.

Consider the vector $v \in \mathbb{R}^m$ block decomposed as

$$v = \begin{bmatrix} a \\ \alpha \\ b \end{bmatrix}$$

where $a \in \mathbb{R}^s$, $\alpha \in \mathbb{R}$, and $b \in \mathbb{R}^t$ with $m = s + 1 + t$. In this vector we refer to the α entry as the *pivot* and assume that $\alpha \neq 0$. We wish to determine a matrix G such that

$$Gv = e_{s+1}$$

where for $j = 1, \ldots, n$, e_j is the unit coordinate vector having a one in the jth position and zeros elsewhere. We claim that the matrix

$$G = \begin{bmatrix} I_{s\times s} & -\alpha^{-1}a & 0 \\ 0 & \alpha^{-1} & 0 \\ 0 & -\alpha^{-1}b & I_{t\times t} \end{bmatrix}$$

does the trick. Indeed,

$$Gv = \begin{bmatrix} I_{s\times s} & -\alpha^{-1}a & 0 \\ 0 & \alpha^{-1} & 0 \\ 0 & -\alpha^{-1}b & I_{t\times t} \end{bmatrix} \begin{pmatrix} a \\ \alpha \\ b \end{pmatrix} = \begin{bmatrix} a - a \\ \alpha^{-1}\alpha \\ -b + b \end{bmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = e_{s+1}. \tag{A.3}$$

The matrix G is called a *Gaussian-Jordan Elimination Matrix*, or GJEM for short. Note that G is invertible since

$$G^{-1} = \begin{bmatrix} I & a & 0 \\ 0 & \alpha & 0 \\ 0 & b & I \end{bmatrix},$$

Moreover, for any vector of the form $w = \begin{pmatrix} x \\ 0 \\ y \end{pmatrix}$ where $x \in \mathbb{R}^s$ $y \in \mathbb{R}^t$, we have

$$Gw = w.$$

The GJEM matrices perform precisely the operations required in order to execute Gauss-Jordan elimination. That is, each elimination step can be realized as left multiplication of the augmented matrix by the appropriate GJEM.

For example, consider the linear system

$$\begin{array}{rcrcrcr} 2x_1 & + & x_2 & + & 3x_3 & = & 5 \\ 2x_1 & + & 2x_2 & + & 4x_3 & = & 8 \\ 4x_1 & + & 2x_2 & + & 7x_3 & = & 11 \\ 5x_1 & + & 3x_2 & + & 4x_3 & = & 10 \end{array}$$

and its associated augmented matrix

$$A = \begin{bmatrix} 2 & 1 & 3 & 5 \\ 2 & 2 & 4 & 8 \\ 4 & 2 & 7 & 11 \\ 5 & 3 & 4 & 10 \end{bmatrix}.$$

The first step of Gauss-Jordan elimination is to transform the first column of this augmented matrix into the first unit coordinate vector. The procedure described in (A.3) can be employed for this purpose. In this case the pivot is the $(1,1)$ entry of the augmented matrix and so

$$s = 0,\ a \text{ is void},\ \alpha = 2,\ t = 3,\ \text{and } b = \begin{bmatrix} 2 \\ 4 \\ 5 \end{bmatrix},$$

which gives

$$G_1 = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ -5/2 & 0 & 0 & 1 \end{bmatrix}.$$

Multiplying these two matrices gives

$$G_1 A = \begin{bmatrix} 1/2 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ -5/2 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 3 & 5 \\ 2 & 2 & 4 & 8 \\ 4 & 2 & 7 & 11 \\ 5 & 3 & 4 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & 3/2 & 5/2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 1/2 & -7/2 & -5/2 \end{bmatrix}.$$

We now repeat this process to transform the second column of this matrix into the second unit coordinate vector. In this case the $(2,2)$ position becomes the pivot so that

$$s = 1,\ a = 1/2,\ \alpha = 1,\ t = 2,\ \text{and } b = \begin{bmatrix} 0 \\ 1/2 \end{bmatrix}$$

yielding

$$G_2 = \begin{bmatrix} 1 & -1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1/2 & 0 & 1 \end{bmatrix}.$$

Again, multiplying these two matrices gives

$$G_2G_1A = \begin{bmatrix} 1 & -1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1/2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 3/2 & 5/2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 1/2 & -7/2 & -5/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -4 & -4 \end{bmatrix}.$$

Repeating the process on the third column transforms it into the third unit coordinate vector. In this case the pivot is the $(3,3)$ entry so that

$$s = 2,\ a = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ \alpha = 1,\ t = 1,\ \text{and } b = -4$$

yielding

$$G_3 = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix}.$$

Multiplying these matrices gives

$$G_3G_2G_1A = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -4 & -4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

which is in reduced echelon form. Therefore the system is consistent and the unique solution is

$$x = \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}.$$

Observe that

$$G_3G_2G_1 = \begin{bmatrix} 3 & -1/2 & -1 & 0 \\ 1 & 1 & -1 & 0 \\ -2 & 0 & 1 & 0 \\ -10 & -1/2 & 4 & 1 \end{bmatrix}$$

and that

$$\begin{aligned} (G_3G_2G_1)^{-1} &= G_1^{-1}G_2^{-1}G_3^{-1} \\ &= \begin{bmatrix} 2 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 5 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -4 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 1 & 3 & 0 \\ 2 & 2 & 4 & 0 \\ 4 & 2 & 7 & 0 \\ 5 & 3 & 4 & 1 \end{bmatrix}. \end{aligned}$$

In particular, reduced Gauss-Jordan form can always be achieved by multiplying the augmented matrix on the left by an invertible matrix which can be written as a product of Gauss-Jordan elimination matrices.

EXERCISE A.10. *What are the Gauss-Jordan elimination matrices that transform the vector* $\begin{bmatrix} 2 \\ 3 \\ -2 \\ 5 \end{bmatrix}$ *in to* e_j *for* $j = 1, 2, 3, 4$*, and what are the inverses of these matrices?*

5. Some Special Square Matrices

We say that a matrix A is square if there is a positive integer n such that $A \in \mathbb{R}^{n\times n}$. For example, the Gauss-Jordan elimination matrices are a special kind of square matrix. Below we give a list of some square matrices with special properties that are very useful to our future work.

Diagonal Matrices: The diagonal of a matrix $A = [A_{ij}]$ is the vector $(A_{11}, A_{22}, \ldots, A_{nn})^T \in \mathbb{R}^n$. A matrix in $\mathbb{R}^{n\times n}$ is said to be diagonal if the only non-zero entries of the matrix are the diagonal entries. Given a vector $v \in \mathbb{R}^n$, we write $\mathrm{diag}(v)$ the denote the diagonal matrix whose diagonal is the vector v.

The Identity Matrix: The identity matrix is the diagonal matrix whose diagonal entries are all ones. We denote the identity matrix in $\mathbb{R}^k$ by I_k. If the dimension of the identity is clear, we simply write I. Note that for any matrix $A \in \mathbb{R}^{m\times n}$ we have $I_m A = A = AI_n$.

Inverse Matrices: The inverse of a matrix $X \in \mathbb{R}^{n\times n}$ is any matrix $Y \in \mathbb{R}^{n\times n}$ such that $XY = I$ in which case we write $X^{-1} := Y$. It is easily shown that if Y is an inverse of X, then Y is unique and $YX = I$.

Permutation Matrices: A matrix $P \in \mathbb{R}^{n\times n}$ is said to be a permutation matrix if P is obtained from the identity matrix by either permuting the columns of the identity matrix or permuting its rows. It is easily seen that $P^{-1} = P^T$.

Unitary Matrices: A matrix $U \in \mathbb{R}^{n\times n}$ is said to be a unitary matrix if $U^TU = I$, that is $U^T = U^{-1}$. Note that every permutation matrix is unitary. But the converse is note true since for any vector u with $\|u\|_2 = 1$ the matrix $I - 2uu^T$ is unitary.

Symmetric Matrices: A matrix $M \in \mathbb{R}^{n\times n}$ is said to be symmetric if $M^T = M$.

Skew Symmetric Matrices: A matrix $M \in \mathbb{R}^{n\times n}$ is said to be skew symmetric if $M^T = -M$.

6. The LU Factorization

In this section we revisit the reduction to echelon form, but we incorporate permutation matrices into the pivoting process. Recall that a matrix $P \in \mathbb{R}^{m\times m}$ is a *permutation matrix* if it can be obtained from the identity matrix by permuting either its rows or columns. It is straightforward to show that $P^TP = I$ so that the inverse of a permutation matrix is its transpose. Multiplication of a matrix on the left permutes the rows of the matrix while multiplication on the right permutes the columns. We now apply permutation matrices in the Gaussian elimination process in order to avoid zero pivots.

Let $A \in \mathbb{R}^{m\times n}$ and assume that $A \neq 0$. Set $\widetilde{A}_0 := A$. If the $(1,1)$ entry of $\widetilde{A}_0$ is zero, then apply permutation matrices P_{l0} and P_{r0} to the left and and right of $\widetilde{A}_0$, respectively, to bring *any* non-zero element of $\widetilde{A}_0$ into the $(1,1)$ position (e.g., the one with largest magnitude) and set $A_0 := P_{l0}\widetilde{A}_0P_{r0}$. Write A_0 in block form as

$$A_0 = \begin{bmatrix} \alpha_1 & v_1^T \\ u_1 & \widehat{A}_1 \end{bmatrix} \in \mathbb{R}^{m\times n},$$

with $0 \neq \alpha_1 \in \mathbb{R}$, $u_1 \in \mathbb{R}^{n-1}$, $v_1 \in \mathbb{R}^{m-1}$, and $\widetilde{A}_1 \in \mathbb{R}^{(m-1)\times(n-1)}$. Then using α_1 to zero out u_1 amounts to left multiplication of the matrix A_0 by the Gaussian elimination matrix

$$\begin{bmatrix} 1 & 0 \\ -\frac{u_1}{\alpha_1} & I \end{bmatrix}$$

to get

$$\begin{bmatrix} 1 & 0 \\ -\frac{u_1}{\alpha_1} & I \end{bmatrix} \begin{bmatrix} \alpha_1 & v_1^T \\ u_1 & \widehat{A}_1 \end{bmatrix} = \begin{bmatrix} \alpha_1 & v_1^T \\ 0 & \widetilde{A}_1 \end{bmatrix} \in \mathbb{R}^{m\times n}\ , \tag{A.4}$$

where

$$\widetilde{A}_1 = \widehat{A}_1 - u_1v_1^T/\alpha_1\ .$$

Define

$$\widetilde{L}_1 = \begin{bmatrix} 1 & 0 \\ \frac{u_1}{\alpha_1} & I \end{bmatrix} \in \mathbb{R}^{m\times m} \quad \text{and} \quad \widetilde{U}_1 = \begin{bmatrix} \alpha_1 & v_1^T \\ 0 & \widetilde{A}_1 \end{bmatrix} \in \mathbb{R}^{m\times n} .$$

and observe that

$$\widetilde{L}_1^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{u_1}{\alpha_1} & I \end{bmatrix}$$

Hence (A.4) becomes

$$\widetilde{L}_1^{-1} P_{l0} \widetilde{A}_0 P_{r0} = \widetilde{U}_1, \text{ or equivalently,} \quad A = P_{l0} \widetilde{L}_1 \widetilde{U}_1 P_{r0}^T . \tag{A.5}$$

Note that $\widetilde{L}_1$ is *unit* lower triangular (ones on the mail diagonal) and $\widetilde{U}_1$ is block upper-triangular with one nonsingular 1×1 block and one $(m-1) \times (n-1)$ block on the block diagonal.

Next consider the matrix $\widetilde{A}_1$ in $\widetilde{U}_1$. If the $(1,1)$ entry of $\widetilde{A}_1$ is zero, then apply permutation matrices $\widetilde{P}_{l1} \in \mathbb{R}^{(m-1)\times(m-1)}$ and $\widetilde{P}_{r1} \in \mathbb{R}^{(n-1)\times(n-1)}$ to the left and and right of $\widetilde{A}_1 \in \mathbb{R}^{(m-1)\times(n-1)}$, respectively, to bring *any* non-zero element of $\widetilde{A}_0$ into the $(1,1)$ position (e.g., the one with largest magnitude) and set $A_1 := \widetilde{P}_{l1} \widetilde{A}_1 P_{r1}$. If the element of $\widetilde{A}_1$ is zero, then stop. Define

$$P_{l1} := \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{l1} \end{bmatrix} \quad \text{and} \quad P_{r1} := \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{r1} \end{bmatrix}$$

so that P_{l1} and P_{r1} are also permutation matrices and

$$P_{l1} \widetilde{U}_1 P_{r1} = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{l1} \end{bmatrix} \begin{bmatrix} \alpha_1 & v_1^T \\ 0 & \widetilde{A}_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{r1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & v_1^T P_{r1} \\ 0 & \widetilde{P}_{l1} \widetilde{A}_1 P_{r1} \end{bmatrix} = \begin{bmatrix} \alpha_1 & \tilde{v}_1^T \\ 0 & A_1 \end{bmatrix}, \tag{A.6}$$

where $\tilde{v}_1 := P_{r1}^T v_1$. Define

$$U_1 := \begin{bmatrix} \alpha_1 & \tilde{v}_1^T \\ 0 & A_1 \end{bmatrix}, \quad \text{where} \quad A_1 = \begin{bmatrix} \alpha_2 & v_2^T \\ u_2 & \widetilde{A}_2 \end{bmatrix} \in \mathbb{R}^{(m-1)\times(n-1)},$$

with $0 \neq \alpha_2 \in \mathbb{R}$, $u_2 \in \mathbb{R}^{n-2}$, $v_1 \in \mathbb{R}^{m-2}$, and $\widetilde{A}_2 \in \mathbb{R}^{(m-2)\times(n-2)}$. In addition, define

$$L_1 := \begin{bmatrix} 1 & 0 \\ \widetilde{P}_{l1} \frac{u_1}{\alpha_1} & I \end{bmatrix},$$

so that

$$\begin{aligned} P_{l1}^T L_1 &= \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{l1}^T \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \widetilde{P}_{l1} \frac{u_1}{\alpha 1} & I \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \frac{u_1}{\alpha 1} & \widetilde{P}_{l1}^T \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ \frac{u_1}{\alpha 1} & I \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{P}_{l1}^T \end{bmatrix} \\ &= \widetilde{L}_1 P_{l1}^T, \end{aligned}$$

and consequently

$$L_1^{-1} P_{l1} = P_{l1} \widetilde{L}_1^{-1} .$$

Plugging this into (A.5) and using (A.6), we obtain

$$L_1^{-1} P_{l1} P_{l0} \widetilde{A}_0 P_{r0} P_{r1} = P_{l1} \widetilde{L}_1^{-1} P_{l0} \widetilde{A}_0 P_{r0} P_{r1} = P_{l1} \widetilde{U}_1 P_{r1} = U_1,$$

or equivalently,

$$P_{l1} P_{l0} A P_{r0} P_{r1} = L_1 U_1.$$

We can now repeat this process on the matrix A_1 since the $(1,1)$ entry of this matrix is non-zero. The process can run for no more than the number of rows of A which is m. However, it may terminate after $k < m$ steps if the matrix $\widehat{A}_k$ is the zero matrix. In either event, we obtain the following result.

THEOREM A.1. *[The LU Factorization] Let $A \in \mathbb{R}^{m\times n}$. If $k = \operatorname{rank}(A)$, then there exist permutation matrices $P_l \in \mathbb{R}^{m\times m}$ and $P_r \in \mathbb{R}^{n\times n}$ such that*

$$P_l A P_r = LU,$$

where $L \in \mathbb{R}^{m\times m}$ is a lower triangular matrix having ones on its diagonal and

$$U = \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix}$$

with $U_1 \in \mathbb{R}^{k\times k}$ a nonsingular upper triangular matrix.

Note that a column permutation is only required if the first column of $\hat{A}_k$ is zero for some k before termination. In particular, this implies that the $\operatorname{rank}(A) < m$. Therefore, if $\operatorname{rank}(A) = m$, column permutations are not required, and $P_r = I$. If one implements the LU factorization so that a column permutation is *only* employed in the case when the first column of $\hat{A}_k$ is zero for some k, then we say the LU factorization is obtained through partial pivoting.

EXAMPLE A.2. *We now use the procedure outlined above to compute the LU factorization of the matrix*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 4 & 2 \\ -1 & 1 & 3 \end{bmatrix}.$$

$$\begin{aligned} L_1^{-1}A &= \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 2 & 4 & 2 \\ -1 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & -3 \\ 0 & 2 & 5 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} L_2^{-1}L_1^{-1}A &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & -3 \\ 0 & 2 & 5 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 8 \end{bmatrix} \end{aligned}$$

We now have

$$U = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 8 \end{bmatrix},$$

and

$$L = L_1 L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{bmatrix}.$$

7. Solving Equations with the LU Factorization

Consider the equation $Ax = b$. In this section we show how to solve this equation using the LU factorization. Recall from Theorem A.1 that the algorithm of the previous section produces a factorization of A of the form $P_l \in \mathbb{R}^{m\times m}$ and $P_r \in \mathbb{R}^{n\times n}$ such that

$$A = P_l^T L U P_r^T,$$

where $P_l \in \mathbb{R}^{m\times m}$ and $P_r \in \mathbb{R}^{n\times n}$ are permutation matrices, $L \in \mathbb{R}^{m\times m}$ is a lower triangular matrix having ones on its diagonal, and

$$U = \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix}$$

with $U_1 \in \mathbb{R}^{k\times k}$ a nonsingular upper triangular matrix. Hence we may write the equation $Ax = b$ as

$$P_l^T L U P_r^T x = b.$$

Multiplying through by P_l and replacing $UP_r^T x$ by w gives the equation

$$Lw = \hat{b}, \quad \text{where } \hat{b} := P_l b\ .$$

This equation is easily solved by forward substitution since L is a nonsingular lower triangular matrix. Denote the solution by $\overline{w}$. To obtain a solution x we must still solve $UP_r^T x = \overline{w}$. Set $y = P_r x$. The this equation becomes

$$\overline{w} = Uy = \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

where we have decomposed y to conform to the decomposition of U. Doing the same for $\overline{w}$ gives

$$\begin{pmatrix} \overline{w}_1 \\ \overline{w}_2 \end{pmatrix} = \begin{bmatrix} U_1 & U_2 \\ 0 & 0 \end{bmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

or equivalently,

$$\begin{aligned} \overline{w}_1 &= U_1 y_1 + U_2 y_2 \\ \overline{w}_2 &= 0. \end{aligned}$$

Hence, if $\overline{w}_2 \neq 0$, the system is inconsistent, i.e., no solution exists. On the other hand, if $\overline{w}_2 = 0$, we can take $y_2 = 0$ and solve the equation

$$\overline{w}_1 = U_1 y_1 \tag{A.7}$$

for $\overline{y}_1$, then

$$\overline{x} = P_r^T \begin{pmatrix} \overline{y}_1 \\ 0 \end{pmatrix}$$

is a solution to $Ax = b$. The equation (A.7) is also easy to solve since U_1 is an upper triangular nonsingular matrix so that (A.7) can be solved by back substitution.

8. The Four Fundamental Subspaces and Echelon Form

Recall that a subset W of $\mathbb{R}^n$ is a subspace if and only if it satisfies the following three conditions:

(1) The origin is an element of W.
(2) The set W is closed with respect to addition, i.e. if $u \in W$ and $v \in W$, then $u + v \in W$.
(3) The set W is closed with respect to scalar multiplication, i.e. if $\alpha \in \mathbb{R}$ and $u \in W$, then $\alpha u \in W$.

EXERCISE A.11. *Given $v^1, v^2, \ldots, v^k \in \mathbb{R}^n$, show that the linear span of these vectors,*

$$\text{span}\left(v^1, v^2, \ldots, v^k\right) := \left\{ \alpha_1 v^1 + \alpha_2 v^2 + \cdots + \alpha_k v^k \,\middle|\, \alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R} \right\}$$

is a subspace.

EXERCISE A.12. *Show that for any set S in $\mathbb{R}^n$, the set*

$$S^\perp = \{v \,:\, w^T v = 0 \text{ for all } w \in S\}$$

is a subspace. If S is itself a subspace, then $S^\perp$ is called the subspace orthogonal (or perpendicular) to the subspace S.

EXERCISE A.13. *If S is any suset of $\mathbb{R}^n$ (not necessarily a subspace), show that $(S^\perp)^\perp = \text{span}\,(S)$.*

EXERCISE A.14. *If $S \subset \mathbb{R}^n$ is a subspace, show that $S = (S^\perp)^\perp$.*

A set of vectors $v^1, v^2, \ldots, v^k \in \mathbb{R}^n$ are said to be *linearly independent* if $0 = a_1v^1 + \cdots + a_kv^k$ if and only if $0 = a_1 = a_2 = \cdots = a_k$. A *basis* for a subspace in any maximal linearly independent subspace. An elementary fact from linear algebra is that the subspace equals the linear span of any basis for the subspace and that every basis of a subspace has the same number of vectors in it. We call this number the *dimension* for the subspace. If S is a subspace, we denote the dimension of S by $\dim S$.

EXERCISE A.15. *If $s \subset \mathbb{R}^n$ is a subspace, then any basis of S can contain only finitely many vectors.*

EXERCISE A.16. *Show that every subspace can be represented as the linear span of a basis for that subspace.*

EXERCISE A.17. *Show that every basis for a subspace contains the same number of vectors.*

EXERCISE A.18. *If $S \subset \mathbb{R}^n$ is a subspace, show that*

$$\mathbb{R}^n = S + S^\perp \tag{A.8}$$

and that

$$n = \dim S + \dim S^\perp. \tag{A.9}$$

Let $A \in \mathbb{R}^{m\times n}$. We associate with A its four fundamental subspaces:

$$\begin{aligned} \operatorname{Ran}(A) &:= \{Ax \,|\, x \in \mathbb{R}^n\} & \operatorname{Null}(A) &:= \{x \,|\, Ax = 0\} \\ \operatorname{Ran}(A^T) &:= \left\{A^Ty \,\middle|\, y \in \mathbb{R}^m\right\} & \operatorname{Null}(A^T) &:= \left\{y \,\middle|\, A^Ty = 0\right\}. \end{aligned}$$

where

$$\begin{aligned} \operatorname{rank}(A) &:= \dim \operatorname{Ran}(A) & \operatorname{nullity}(A) &:= \dim \operatorname{Null}(A) \\ \operatorname{rank}(A^T) &:= \dim \operatorname{Ran}(A^T) & \operatorname{nullity}(A^T) &:= \dim \operatorname{Null}(A^T) \end{aligned} \tag{A.10}$$

EXERCISE A.19. *Show that the four fundamental subspaces associated with a matrix are indeed subspaces.*

Observe that

$$\begin{aligned} \operatorname{Null}(A) &:= \{x \,|\, Ax = 0\} \\ &= \{x \,|\, A_{i\cdot} \bullet x = 0,\ i = 1, 2, \ldots, m\} \\ &= \{A_{1\cdot}, A_{2\cdot}, \ldots, A_{m\cdot}\}^\perp \\ &= \operatorname{span}(A_{1\cdot}, A_{2\cdot}, \ldots, A_{m\cdot})^\perp \\ &= \operatorname{Ran}(A^T)^\perp\ . \end{aligned}$$

Since for any subspace $S \subset \mathbb{R}^n$, we have $(S^\perp)^\perp = S$, we obtain

$$\operatorname{Null}(A)^\perp = \operatorname{Ran}(A^T) \text{ and } \operatorname{Null}(A^T) = \operatorname{Ran}(A)^\perp. \tag{A.11}$$

The equivalences in (A.11) are called the *Fundamental Theorem of the Alternative.*

One of the big consequences of echelon form is that

$$n = \operatorname{rank}(A) + \operatorname{nullity}(A). \tag{A.12}$$

By combining (A.12), (A.9) and (A.11), we obtain the equivalence

$$\operatorname{rank}(A^T) = \dim \operatorname{Ran}(A^T) = \dim \operatorname{Null}(A)^\perp = n - \operatorname{nullity}(A) = \operatorname{rank}(A).$$

That is, the row rank of a matrix equals the column rank of a matrix, i.e., the dimensions of the row and column spaces of a matrix are the same!

9. Eigenvalue Decomposition of Symmetric Matrices

Given a matrix $A \in \mathbb{R}^{n\times n}$, we say that the scalar λ is an eigenvalue of A if there is a non-zero vector x such that $Ax = \lambda x$, or equivalently, $\text{Null}(\lambda I - A) \neq \{0\}$. Observe that $\text{Null}(\lambda I - A) \neq \{0\}$ if and only if $(\lambda I - A)$ is singular, that is, $\det(\lambda I - A) = 0$. Consequently, λ is an eigenvalue of A if and only if $\det(\lambda I - A) = 0$. If we now think of λ as a variable, this says that we can find all eigenvalues of A by finding all roots of the equation $\det(\lambda I - A) = 0$. The function $p(\lambda) := \det(\lambda I - A)$ is easily seen to be a polynomial of degree n in λ which we call the *characteristic polynomial* of A. By the Fundamental Theorem of Algebra, we know that $p(\lambda)$ has n roots over the complex numbers if we count the multiplicities of these roots. Hence, when we discuss eigenvalues and eigenvectors we are forced in the setting of complex numbers. For this reason we may as well assume that $A \in \mathbb{C}^{n\times n}$.

Working on $\mathbb{C}^n$ requires us to re-examine our notion of the Euclidean norm and its associated dot product. Recall that for a complex number $\zeta := x + \mathsf{i}y$, with $x, y \in \mathbb{R}$ and $\mathsf{i} := \sqrt{-1}$, the magnitude of ζ is given by $|\zeta| = \sqrt{\bar{\zeta}\zeta}$, where $\bar{\zeta} := x - \mathsf{i}y$ is the complex conjugate of ζ. If we now define the Euclidean norm of a vector $z \in \mathbb{C}^n$ to be the square root of the sum of the squares of magnitude of its components, then

$$\|z\|_2 = \sqrt{\sum_{k-1}^{n} |z_k|^2} = \sqrt{\sum_{k-1}^{n} \overline{z}_k z_k} = \sqrt{\overline{z}^T z} = \sqrt{z^* z},$$

where we define

$$z^* z = (\overline{z})^T z,$$

that is, z^* takes z to its *conjugate transpose*. When $z \in \mathbb{R}^n$, we have $z^* = z^T$, and we recover the usual formulas. With the $*$ operation, we can extend our notion of dot product (or, inner product) by writing

$$\langle z,\, y\rangle := z^* y \ \in \mathbb{C}\,.$$

When z and y are real vectors we recover usual notion of dot product for such vectors. Finally, for matrices $A \in \mathbb{C}^{n\times n}$, we define

$$A^* := \overline{A}^T,$$

that is, we conjugate every element of A and then take the transpose. This notation is very helpful in a number of ways. For example, we have

$$\langle Ay,\, x\rangle = (Ay)^* x = y^* A^* x \quad \text{and} \quad \|Ax\|_2^2 = x^* A^* A x\ .$$

We call A^* the *adjoint* of A.

Recall that a matrix $H \in \mathbb{R}^{n\times n}$ is said to be symmetric of $H^T = H$. By extension, we say that an matrix $Q \in \mathbb{C}^{n\times n}$ is *self-adjoint* if $Q^* = Q$. Thus, in particular, every real symmetric matrix is self adjoint. We have the following remarkable fact about self-adjoint matrices.

LEMMA A.3. *If $Q \in \mathbb{C}^{n\times n}$ is self-adjoint, then Q has only real eigenvalues. In particular, if H is a real symmetric matrix, then H has only real eigenvalues and for each such eigenvalue there is a real eigenvector. Moreover, if (λ_1, v^1) and (λ_2, v^2) are two eigenvalue-eigenvectors pairs for H with $\lambda_1 \neq \lambda_2$, then $(v^1)^T v^2 = 0$.*

PROOF. Let $\lambda \in \mathbb{C}$ be an eigenvalue of Q. Then there is a non-zero eigenvector $x \in \mathbb{C}^n$ such that $Qx = \lambda x$. Therefore,

$$\lambda \|x\|_2^2 = \lambda x^* x = x^* Q x = x^* Q^* x = (x^* Q x)^* = (\lambda \|x\|_2^2)^* = \overline{\lambda}\, \|x\|_2^2\,,$$

so that $\lambda = \overline{\lambda}$ which can only occur if λ is a real number.

If H is real symmetrix, then it is self adjoint so all of its eigenvalues are real. If λ is one such eigenvalue with associated eigenvector $z = x + \mathsf{i}y$ with $x, y \in \mathbb{R}^n$, then

$$Hx + \mathsf{i}Hy = Hz = \lambda z = \lambda x + \mathsf{i}\lambda y.$$

Consequently, $Hx = \lambda x$ and $Hy = \lambda y$ since both Hx and Hy are real vectors. Since $z \neq 0$, either x or y or both are non-zero, in any case we have a real eigenvector for H corresponding to λ.

Next let (λ_1, v^1) and (λ_2, v^2) be eigenvalue-eigenvectors pairs for H with $\lambda_1 \neq \lambda_2$. Then

$$\lambda_1 (v^1)^T v^2 = (Hv^1)^T v^2 = (v^1)^T H v^2 = \lambda_2 (v^1)^T v^2,$$

since $\lambda_1 \neq \lambda_2$, we must have $(v^1)^T v^2 = 0$. □

Next, suppose λ_1 is an eigenvalue for the real symmetric matrix $H \in \mathbb{R}^{n\times n}$ and let the columns of the matrix $U_1 \in \mathbb{R}^{n\times k}$ form an orthonormal basis for the subspace $\text{Null}(\lambda_1 I - H)$, where $k = \dim(\text{Null}(\lambda_1 I - H)) \geq 1$. Let the columns of $U_2 \in \mathbb{R}^{n\times(n-k)}$ form an orthonormal basis for the subspace $\text{Null}(\lambda_1 I - H)^\perp$ and set $\widetilde{U} = [U_1\ U_2] \in \mathbb{R}^{n\times n}$. Then $\widetilde{U}^T\widetilde{U} = I$, that is, $\widetilde{U}$ is a *unitary matrix*. In particular, $\widetilde{U}^{-1} = \widetilde{U}^T$ and so $\widetilde{U}\widetilde{U}^T = I$ as well. We have the following relationships between U_1, U_2, and H:

$$HU_1 = \lambda_1 U_1, \quad U_1^T H U_1 = \lambda_1 U_1^T U_1 = \lambda_1 I_k \quad \text{and} \quad (U_1^T H U_2)^T = U_2^T H U_1 = \lambda_1 U_2^T U_1 = 0_{(n-k)\times k}.$$

Consequently,

$$\widetilde{U}^T H \widetilde{U} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} H \begin{bmatrix} U_1 & U_2 \end{bmatrix} = \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \begin{bmatrix} HU_1 & HU_2 \end{bmatrix} = \begin{bmatrix} U_1^T H U_1 & U_1^T H U_2 \\ U_2^T H U_1 & U_2^T H U_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 I_k & 0 \\ 0 & U_2^T H U_2 \end{bmatrix}, \tag{A.13}$$

and so

$$H = \widetilde{U}\widetilde{U}^T H \widetilde{U}\widetilde{U}^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \lambda_1 I_k & 0 \\ 0 & U_2^T H U_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}. \tag{A.14}$$

These observations provide the foundation for the following eigenvalue theorem for real symmetric matrices.

THEOREM A.4. *[Eigenvalue Decomposition for Symmetric Matrices] Let $H \in \mathbb{R}^{n\times n}$ be a real symmetric matrix. Then there is a unitary matrix U such that*

$$H = U\Lambda U^T,$$

where $\Lambda := \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ with $\lambda_1, \lambda_2, \ldots, \lambda_n$ being the eigenvalues of H repeated according to multiplicity.

PROOF. We proceed by induction on the dimension. The result is trivially true for $n = 1$. Assume that the result is true for all dimensions $k < n$ with $n > 1$ and show it is true for all $n \times n$ symmetric matrices. Let $H \in \mathbb{R}^{n\times n}$ be symmetric and let λ_1 be any eigenvalue of H with $k = \dim(\text{Null}(\lambda_1 I - H)) \geq 1$. Let $U_1 \in \mathbb{R}^{n\times k}$ and $U_2 \in \mathbb{R}^{n\times(n-k)}$ be as in (A.13) and (A.14) above. If $k = n$, the result follows from (A.13) so we can assume that $k < n$.

Since (A.13) is a similarity transformation of H, $\widetilde{U}^T H \widetilde{U}$ has the same characteristic polynomial as H:

$$\det(\lambda I_n - H) = (\lambda - \lambda_1)^k q(\lambda), \quad \text{where} \quad q(\lambda) = \det(\lambda I_{n-k} - U_2^T H U_2).$$

Therefore, the eigenvalues of $U_2^T H U_2$ are necessarily those of H that are not equal to λ_1 and each has the same multiplicity as they have for H.

Apply the induction hypothesis to the $(n-k)\times(n-k)$ matrix $U_2^T H U_2$ to obtain a real unitary matrix $V \in \mathbb{R}^{(n-k)\times(n-k)}$ such that

$$U_2^T H U_2 = V\widetilde{\Lambda}V^T,$$

where $\widetilde{\Lambda} = \text{diag}(\mu_1, \mu_2, \ldots, \mu_{(n-k)})$ with $\mu_1, \mu_2, \ldots, \mu_{(n-k)}$ being the eigenvalues of H that are not equal to λ_1 with each having the same multiplicity as they have for H. Then, by (A.14)

$$H = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \lambda_1 I_k & 0 \\ 0 & U_2^T H U_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & V \end{bmatrix} \begin{bmatrix} \lambda_1 I_k & 0 \\ 0 & \widetilde{\Lambda} \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & V^T \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix}.$$

The result is obtained by setting

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & V \end{bmatrix} = \begin{bmatrix} U_1 & U_2 V \end{bmatrix}$$

and observing that $U^TU = I$. □

One important consequence of this result is the following theorem

THEOREM A.5. *[The Rayleigh-Ritz Theorem] Let the symmetric matrix $H \in \mathbb{R}^{n\times n}$ have smallest eigenvalue $\lambda_{min}(H)$ and largest eigenvalue $\lambda_{max}(H)$. Then, for all $u \in \mathbb{R}^n$,*

$$\lambda_{min}(H)\,\|u\|_2^2 \le u^T H u \le \lambda_{max}(H)\,\|u\|_2^2\,,$$

with equality holding on the left for every eigenvector u for $\lambda_{min}(H)$ and equality holding on the right for every eigenvector u for $\lambda_{max}(H)$.

PROOF. Let $H = U\Lambda U^T$ be the eigenvalue decomposition of H in Theorem A.4 with $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Then the columns of U form an orthonormal basis for $\mathbb{R}^n$. Therefore, given any $u \in \mathbb{R}^n \setminus \{0\}$, there is a $z \in \mathbb{R}^n \setminus \{0\}$ such that $u = Uz$. Hence

$$u^T H u = (Uz)^T U\Lambda U^T (Uz) = z^T \Lambda z = \sum_{j=1}^{n} \lambda_j z_j^2.$$

Clearly,

$$\lambda_{\min}(H)\,\|z\|_2^2 = \sum_{j=1}^{n} \lambda_{\min}(H) z_j^2 \le \sum_{j=1}^{n} \lambda_j z_j^2 \le \sum_{j=1}^{n} \lambda_{\max}(H) z_j^2 = \lambda_{\max}(H)\,\|z\|_2^2\,.$$

The result now follows since $\|z\|_2^2 = z^Tz = z^TU^TUz = u^Tu = \|u\|_2^2$. □

Theorem A.5 provides necessary and sufficient conditions under which a symmetric matrix H is positive/negative definite/semi-definite. For example, since $\lambda_{\min}(H)\,\|u\|_2^2 \le u^THu$ with equality when u is an eigenvector associted with $\lambda_{\min}(H)$, we have that H is positive definite if and only if $\lambda_{\min}(H) > 0$. Similar results can be obtained for the other cases.

An additional property of positive semi-definite matrices is that they possess *square roots*. If $H \in \mathbb{R}^{n\times n}$ is symmetric and positive semi-definite, then Theorem A.4 tells us that $H = U\Lambda U^T$, where U is unitary and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with $\lambda_i \ge 0,\ i = 1, \ldots, n$. If we define $\Lambda^{1/2} := \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$ and $H^{1/2} = U\Lambda^{1/2}U^T$, then $H = U\Lambda U^T = U\Lambda^{1/2}U^T = H^{1/2}H^{1/2}$, so $H^{1/2}$ provides a natural notion of the square root of a matrix. However, $H^{1/2}$ is not uniquely defined since we can always re-order the diagonal elements and their corresponding columns to produce the same effect. In addition, $H^{1/2}$ is always symmetric while in some instances choosing a non-symmetric square root may be beneficial. For example, if we consider the linear least squares problem (3.2), then $H = A^TA$. Should A be considered a square root of H? In order to cover the full range of possible considerations, we make the following definition for the square root of a symmetric matrix.

DEFINITION A.6. *[Square Roots of Positive Semi-Definite Matrices] Let $H \in \mathbb{R}^{n\times n}$ be a symmetric positive semi-definite matrix. We say that the matrix $L \in \mathbb{R}^{n\times n}$ is a square root of H if $H = LL^T$.*

APPENDIX B

Elements of Multivariable Calculus

1. Norms and Continuity

As we have seen the 2-norm gives us a measure of the magnitude of a vector v in $\mathbb{R}^n$, $\|v\|_2$. As such it also gives us a measure of the distance between to vectors $u,\ v \in \mathbb{R}^n,\ \|u-v\|_2$. Such measures of magnitude and distance are very useful tools for measuring model misfit as is the case in linear least squares problem. They are also essential for analyzing the behavior of sequences and functions on $\mathbb{R}^n$ as well as on the space of matrices $\mathbb{R}^{m\times n}$. For this reason, we formalize the notion of a norm to incorporate other measures of magnitude and distance.

DEFINITION B.1. *[Vector Norm] A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is a vector norm on $\mathbb{R}^n$ if*

(1) $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if $x = 0$,
(2) $\|\alpha x\| = |\alpha| \, \|x\|$ for all $x \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$, and
(3) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$.

EXAMPLE B.2. *Perhaps the most common examples of norms are the p-norms for $1 \leq p \leq \infty$. Given $1 \leq p < \infty$, the ℓ_p-norm on $\mathbb{R}^n$ is defined as*

$$\|x\|_p := \left[\sum_{j=1}^{n} |x_j|^p \right]^{1/p} .$$

For $p = \infty$, we define

$$\|x\|_\infty := \max \{ |x_i| \mid i = 1, 2, \ldots, n \} .$$

This choice of notation for the ∞-norm comes from the relation

$$\lim_{p \uparrow \infty} \|x\|_p = \|x\|_\infty \quad \forall\ x \in \mathbb{R}^n.$$

In applications, the most important of these norms are the $p = 1, 2, \infty$ norms as well as variations on these norms.

In finite dimensions all norms are said the *equivalent* in the sense that one can show that for any two norms $\|\cdot\|_{(a)}$ and $\|\cdot\|_{(b)}$ on $\mathbb{R}^n$ there exist positive constants α and β such that

$$\alpha \|x\|_a \leq \|x\|_b \leq \beta \|x\|_a \quad \forall\, x \in \mathbb{R}^n \ .$$

But we caution that in practice the numerical behavior of these norms differ greatly when the dimension is large.

Since norms can be used to measure the distance between vectors, they can be used to form notions of continuity for functions mapping $\mathbb{R}^n$ to $\mathbb{R}^m$ that parallel those established for mappings from $\mathbb{R}$ to $\mathbb{R}$.

DEFINITION B.3. *[Continuous Functions] Let $F : \mathbb{R}^n \to \mathbb{R}^n$.*

(1) F is said to be continuous at a point $\overline{x} \in \mathbb{R}^n$ if for all $\epsilon > 0$ there is a $\delta > 0$ such that

$$\|F(x) - F(\overline{x})\| \leq \epsilon \quad \text{whenever} \quad \|x - \overline{x}\| \leq \delta \ .$$

(2) F is said to be continuous on a set $S \subset \mathbb{R}^n$ if it is continuous at every point of S.
(3) The function F is said to be continuous relative to a set $S \subset \mathbb{R}^n$ if

$$\|F(x) - F(\overline{x})\| \leq \epsilon \quad \text{whenever} \quad \|x - \overline{x}\| \leq \delta \quad \text{and} \quad x \in S \ .$$

(4) *The function F is said to be uniformly continuous on a set $S \subset \mathbb{R}^n$ if if for all $\epsilon > 0$ there is a $\delta > 0$ such that*

$$\|F(x) - F(y)\| \leq \epsilon \quad \textit{whenever} \quad \|x - y\| \leq \delta \quad \text{and} \quad x, y \in S.$$

Norms allow us to define certain topological notions that are very helpful in analizing the behavior of sequences and functions. Since we will make frequent use of these concepts, it is helpful to have certain notational conventions associated with norms. We list a few of these below:

the closed unit ball	$\mathbb{B} := \{x \mid \|x\| \leq 1\}$
the unit vectors	$\mathbb{S} := \{x \mid \|x\| = 1\}$
ϵ-ball about $\overline{x}$	$\overline{x} + \epsilon\mathbb{B} := \{x + \epsilon u \mid u \in \mathbb{B}\} = \{x \mid \|x - \overline{x}\| \leq \epsilon\}$

The unit ball associated with the 1, 2, and ∞ norms will be denoted by $\mathbb{B}_1$, $\mathbb{B}_2$, and $\mathbb{B}_\infty$, respectively.

A few basic topological notions are listed in the following definition. The most important of these for our purposes is *compactness*.

Definition B.4. *Let S be a subset of $\mathbb{R}^n$, and let $\|\cdot\|$ be a norm on $\mathbb{R}^n$.*

(1) *The set S is said to be an open set if for every $\overline{x} \in S$ there is an $\epsilon > 0$ such that $\overline{x} + \epsilon\mathbb{B} \subset S$.*
(2) *The set S is said to be a closed set if S contains every point $\overline{x} \in \mathbb{R}^n$ for which there is a sequence $\{x^k\} \subset S$ with $\lim_{k\to\infty} \left\|x^k - \overline{x}\right\| = 0$.*
(3) *The set S is said to be a bounded set set if there is a $\beta > 0$ such that $S \subset \beta\mathbb{B}$.*
(4) *The set S is said to be a compact set if it is both closed and bounded.*
(5) *A point $\overline{x} \in \mathbb{R}^n$ is a cluster point of the set S if there is a sequence $\{x^k\} \subset S$ with $\lim_{k\to\infty} \left\|x^k - \overline{x}\right\| = 0$.*
(6) *A point $\overline{x} \in \mathbb{R}^n$ is said to be a boundary point of the set S if for all $\epsilon > 0$, $(\overline{x} + \epsilon\mathbb{B}) \cap S \neq \emptyset$ while $(\overline{x} + \epsilon\mathbb{B}) \not\subset S$, i.e., every ϵ ball about $\overline{x}$ contains points that are in S and points that are not in S.*

The importance of the notion of compactness in optimization is illustrated in following basic theorems from analysis that we make extensive use of, but do not prove.

Theorem B.5. *[Compactness implies Uniform Continuity] Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function on an open set $S \subset \mathbb{R}^n$. Then F is uniformly continuous on every compact subset of S.*

Theorem B.6. *[Weierstrass Compactness Theorem] A set $D \subset \mathbb{R}^n$ is compact if and only if every infinite sequence in D has a cluster point in D.*

Theorem B.7. *[Weierstrass Extreme Value Theorem] Every continuous function on a compact set attains its extreme values on that set. That is, there are points in the set at which both the infimum and the supremum of the function relative to the set are attained.*

We will also have need of a norm on the space of matrices. First note that the space of matrices $\mathbb{R}^{m\times n}$ is itself a vector space since it is closed with respect to addition and real scalar multiplication with both operations being distributive and commutative and $\mathbb{R}^{m\times n}$ contains the zero matrix. In addition, we can embed $\mathbb{R}^{m\times n}$ in $\mathbb{R}^{mn}$ by stacking one column on top of another to get a long vector of length mn. This process of stacking the columns is denoted by the vec operator (column vec): given $A \in \mathbb{R}^{m\times n}$,

$$\operatorname{vec}(A) = \begin{pmatrix} A_{\cdot 1} \\ A_{\cdot 2} \\ \vdots \\ A_{\cdot n} \end{pmatrix} \in \mathbb{R}^{mn} \ .$$

EXAMPLE B.8.

$$\operatorname{vec}\begin{bmatrix} 1 & 2 & -3 \\ 0 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -1 \\ -3 \\ 4 \end{bmatrix}$$

Using the vec operation, we define an inner product on $\mathbb{R}^{m\times n}$ by taking the inner product of these vectors of length mn. Given $A, B \in \mathbb{R}^{m\times n}$ we write this inner product as $\langle A,\, B\rangle$. It is easy to show that this inner product obeys the formula

$$\langle A,\, B\rangle = \operatorname{vec}(A)^T \operatorname{vec}(B) = \operatorname{tr}\left(A^T B\right).$$

This is known as the *Frobenius inner product*. It generates a corresponding norm, called the *Frobenius norm*, by setting

$$\|A\|_F := \|\operatorname{vec}(A)\|_2 = \sqrt{\langle A,\, A\rangle}.$$

Note that for a given $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$ we have

$$\|Ax\|_2^2 = \sum_{i=1}^{m} (A_{i\cdot} \bullet x)^2 \le \sum_{i=1}^{m} (\|A_{i\cdot}\|_2 \|x\|_2)^2 = \|x\|_2^2 \sum_{i=1}^{m} \|A_{i\cdot}\|_2^2 = \|A\|_F^2 \|x\|_2^2,$$

and so

$$\|Ax\|_2 \le \|A\|_F \|x\|_2. \tag{B.1}$$

This relationship between the Frobenius norm and the 2-norm is very important and is used extensively in our development. In particular, this implies that for any two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times k}$ we have

$$\|AB\|_F \le \|A\|_F \|B\|_F \ .$$

2. Differentiation

In this section we use our understanding of differentiability for mappings from $\mathbb{R}$ to $\mathbb{R}$ to build a theory of differentiation for mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. Let F be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ which we denote by $F : \mathbb{R}^n \to \mathbb{R}^m$. Let the component functions of F be denoted by $F_i : \mathbb{R}^n \to \mathbb{R}$:

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_m(x) \end{pmatrix} .$$

EXAMPLE B.9.

$$F(x) = F\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3x_1^2 + x_1x_2x_3 \\ 2\cos(x_1)\sin(x_2x_3) \\ \ln[\exp(x_1^2 + x_2^2 + x_3^2)] \\ 1/\sqrt{1 + (x_2x_3)^2} \end{pmatrix} .$$

In this case, $n = 3$, $m = 4$, *and*

$$F_1(x) = 3x_1^2 + x_1x_2x_3, \quad F_2(x) = 2\cos(x_1)\sin(x_2x_3), \quad F_3(x) = \ln[\exp(x_1^2 + x_2^2 + x_3^2)], \quad F_4(x) = 1/\sqrt{1 + (x_2x_3)^2}\,.$$

The first step in understanding the differentiability of mappings on $\mathbb{R}^n$ is to study their one dimensional properties. For this, consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and let x and d be elements of $\mathbb{R}^n$. We define the *directional derivative* of f in the direction d, when it exits, to be the one sided limit

$$f'(x; d) := \lim_{t\downarrow 0} \frac{f(x + td) - f(x)}{t}.$$

EXAMPLE B.10. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x_1, x_2) := x_1\,|x_2|$, and let $x = (1,0)^T$ and $d = (2,2)$. Then,*

$$f'(x;d) = \lim_{t\downarrow 0} \frac{f(x+td)-f(x)}{t} = \lim_{t\downarrow 0} \frac{(1+2t)\,|0+2t| - 1\,|0|}{t} = \lim_{t\downarrow 0} \frac{2(1+2t)t}{t} = 2\ ,$$

while, for $d = -(2,2)^T$,

$$f'(x;d) = \lim_{t\downarrow 0} \frac{f(x+td)-f(x)}{t} = \lim_{t\downarrow 0} \frac{(1-2t)\,|0-2t| - 1\,|0|}{t} = \lim_{t\downarrow 0} \frac{2(1-2t)t}{t} = 2\ .$$

In general, we have

$$f'((1,0);(d_1,d_2)) = \lim_{t\downarrow 0} \frac{(1+d_1t)\,|d_2t|}{t} = |d_2|.$$

For technical reasons, we allow this limit to take the values $\pm\infty$. For example, if $f(x) = x^{1/3}$, then

$$f'(0;1) = \lim_{t\downarrow 0} t^{-2/3} = +\infty \quad \text{and} \quad f'(0;-1) = \lim_{t\downarrow 0} -t^{-2/3} = -\infty.$$

This example as well as the one given in Example B.10 show that the directional derivative $f'(x;d)$ is not necessarily either continuous or smooth in the d argument even if it exists for all choices of d. However, the directional derivative is always *positively homogeneous* in the sense that, given $\lambda \geq 0$, we have

$$f'(x;\lambda d) = \lim_{t\downarrow 0} \frac{f(x+\lambda td)-f(x)}{t} = \lambda \lim_{t\downarrow 0} \frac{f(x+\lambda td)-f(x)}{\lambda t} = \lambda f'(x;d)\ .$$

The directional derivative idea can be extended to functions F mapping $\mathbb{R}^n$ into $\mathbb{R}^m$ by defining it componentwise: if the limit

$$F'(x;d) := \lim_{t\downarrow 0} \frac{F(x+td)-F(x)}{t} = \begin{pmatrix} \lim_{t\downarrow 0} \frac{F_1(x+td)-F_1(x)}{t} \\ \lim_{t\downarrow 0} \frac{F_2(x+td)-F_2(x)}{t} \\ \vdots \\ \lim_{t\downarrow 0} \frac{F_m(x+td)-F_m(x)}{t} \end{pmatrix}$$

exists, it is called the directional derivative of F at x in the direction d.

These elementary ideas lead to the following notions of differentiability.

DEFINITION B.11. *[Differentiable Functions] Let $f : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^n \to \mathbb{R}^m$.*

(1) If $f'(x;d) = \lim_{t\to 0} \frac{f(x+\lambda td)-f(x)}{t}$, then we say that f is differentiable in the direction d, in which case $f'(x;-d) = -f'(x;d)$.

(2) Let e_j $j = 1,\ldots,n$ denote the unit coordinate vectors. If f is differentiable in the direction e_j, we say that the partial derivative of f with respect to the component x_j exists and write

$$\frac{\partial f(x)}{\partial x_j} := f'(x;e_j).$$

In particular, we have

$$f(x+te_j) = f(x) + t\frac{\partial f(x)}{\partial x_j} + o(t), \quad \text{where } \lim_{t\to 0} \tfrac{o(t)}{t} = 0.$$

Note that $\frac{\partial f(\cdot)}{\partial x_j} : \mathbb{R}^n \to \mathbb{R}$.

(3) We say that f is (Fréchet) differentiable at $x \in \mathbb{R}^n$ if there is a vector $g \in \mathbb{R}^n$ such that

$$\lim_{y\to x} \frac{|f(y) - f(x) - g^T(y-x)|}{\|y-x\|} = 0.$$

If such a vector g exists, we write $g = \nabla f(x)$ and call $\nabla f(x)$ the gradient of f at x. In particular, the differentiability of f at x is equivalent to the following statement:

$$f(y) = f(x) + \nabla f(x)^T(y - x) + o(\|y - x\|)$$

for all y near x, where $\lim_{y\to x} \frac{o(\|y-x\|)}{\|y-x\|} = 0$.

(4) We say that F is (Fréchet) differentiable at $x \in \mathbb{R}^n$ if there is a matrix $J \in \mathbb{R}^{m\times n}$ such that

$$\lim_{y\to x} \frac{\|F(y) - F(x) - J(y - x)\|}{\|y - x\|} = 0.$$

If such a matrix J exists, we write $J = \nabla F(x)$ and call $\nabla F(x)$ the Jacobian of F at x. In particular, the differentiability of f at x is equivalent to the following statement:

$$F(y) = F(x) + \nabla F(x)^T(y - x) + o(\|y - x\|)$$

for all y near x, where $\lim_{y\to x} \frac{o(\|y-x\|)}{\|y-x\|} = 0$.

REMARK B.12. *Note that there is an inconsistency here in the use of the ∇ notation when $F : \mathbb{R}^n \to \mathbb{R}^m$ with $m = 1$. The inconsistency arises due to the presense of g^T in Part (3) of Definition B.11 and the absence of a transpose in Part (4) of this definition. For this reason, we must take extra care in interpreting this notation in this case.*

REMARK B.13. *[Little-o Notation] In these notes we use the notation $o(t)$ to represent any element of a function class for which $\lim_{t\to 0} \frac{o(t)}{t} = 0$. In particular, this implies that for all $\alpha \in \mathbb{R}$*

$$\alpha o(t) = o(t),\ o(t) + o(t) = o(t), \quad \text{and} \quad t^s o(t^r) = o(t^{r+s}).$$

Several observations about these notions of differentiability are in order. First, the existence of the directional derivative $f'(x; d)$ nor the differentiability of f at x in the direction d requires the continuity of the function at that point. Second, the existence of $f'(x; d)$ in all directions d does imply the continuity of the mapping $d \mapsto f'(x; d)$. Therefore, the directional derivative, although useful, is a very weak object to describe the local variational properties of a function. On the other hand, differentiability is a very powerful statement. A few consequences of differentiability are listed in the following theorem.

THEOREM B.14. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and $F : \mathbb{R}^n \to \mathbb{R}^m$.*

(1) If f is differentiable at x, then

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix},$$

and $f'(x; d) = \nabla f(x)^T d$ for all $d \in \mathbb{R}^n$.

(2) If F is differentiable at x, then

$$(\nabla F(x))_{ij} = \frac{\partial F_i(x)}{\partial x_j} \quad i = 1, \ldots, m \quad \text{and} \quad j = 1, 2, \ldots, n.$$

(3) If F is differentiable at a point x, then it is necessarily continuous at x.

Higher order derivatives are obtained by applying these notions of differentiability to the derivatives themselves. For example, to compute the second derivative, the derivative needs to exist at all points near the point at which the second derivative needs to be computed so that the necessary limit is well defined. From the above, we know that the partial derivative $\frac{\partial F_i(x)}{\partial x_j}$, when it exists, is a mapping from

$\mathbb{R}^n$ to $\mathbb{R}$. Therefore, it is possible to consider the partial derivatives of these partial derivatives. For such partial derivatives we use the notation

$$\frac{\partial^2 F_i(x)}{\partial x_j \partial x_k} := \frac{\partial \left(\frac{\partial F_i(x)}{\partial x_k} \right)}{\partial x_j}. \tag{B.2}$$

The second derivative of $f : \mathbb{R}^n \to \mathbb{R}$ is the derivative of the mapping $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, and we write $\nabla(\nabla f(x)) =: \nabla^2 f(x)$. We call $\nabla^2 f(x)$ the *Hessian* of f at x. By (B.2), we have

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial^2 x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_1} \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial^2 x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f(x)}{\partial^2 x_n} \end{bmatrix}.$$

We have the following key property of the Hessian.

THEOREM B.15. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be such that all of the second partials $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$, i, $j = 1, 2, \ldots, n$ exist and are continuous near $x \in \mathbb{R}^n$. Then $\nabla^2 f(x)$ is a real symmetric matrix, i.e., $\nabla^2 f(x) = \nabla^2 f(x)^T$.*

The partial derivative representations of the gradient, Hessian, and Jacobian matrices is a convenient tool for computing these objects. For example, if we have

$$f(x) := 3x_1^2 + x_1 x_2 x_3,$$

then

$$\nabla f(x) = \begin{pmatrix} 6x_1 + x_2 x_3 \\ x_1 x_3 \\ x_1 x_2 \end{pmatrix} \quad \text{and} \quad \nabla^2 f(x) = \begin{bmatrix} 6 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \end{bmatrix}.$$

However, the partial derivatives are not the only tool for computing derivatives. In many cases, it is easier to compute the gradient, Hessian, and/or Jacobian directly from the definition using the little-o notation.

3. The Delta Method for Computing Derivatives

Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be differentiable at a point x if there is a vector $g \in \mathbb{R}^n$ such that

$$f(x + \Delta x) = f(x) + g^T \Delta x + o(\|\Delta x\|). \tag{B.3}$$

Hence, if we can write $f(x + \Delta x)$ in this form, then $g = \nabla f(x)$. To see how to use this idea, consider the least squares objective function

$$f(x) = \tfrac{1}{2} \|Ax - b\|_2^2, \qquad \text{where } A \in \mathbb{R}^{m \times n},\ b \in \mathbb{R}^m.$$

Then

$$\begin{aligned} f(x + \Delta x) &= \tfrac{1}{2} \|A(x + \Delta x) - b\|_2^2 \\ &= \tfrac{1}{2} \|(Ax - b) + A\Delta x\|_2^2 \\ &= \tfrac{1}{2} \|Ax - b\|_2^2 + (Ax - b)^T A \Delta x + \tfrac{1}{2} \|A\Delta x\|_2^2 \\ &= f(x) + (A^T(Ax - b))^T \Delta x + \tfrac{1}{2} \|A\Delta x\|_2^2. \end{aligned} \tag{B.4}$$

In this expression, $\frac{1}{2} \|A\Delta x\|_2^2 = o(\|\Delta x\|_2)$ since

$$\frac{\frac{1}{2} \|A\Delta x\|_2^2}{\|\Delta x\|_2} = \tfrac{1}{2} \|A\Delta x\|_2 \left\| A \frac{\Delta x}{\|\Delta x\|_2} \right\|_2 \to 0 \quad \text{as} \quad \|\Delta x\|_2 \to 0\ .$$

Therefore, by (B.3), the expression (B.4) tells us that

$$\nabla f(x) = A^T(Ax - b).$$

This approach to computing the derivative of a function is called the *delta method*. In a similar manner it can be used to compute the Hessian of f by applying the approach to ∇f:

$$\nabla f(x + \Delta x) = A^T(A(x + \Delta x) - b) = A^T(Ax - b) + A^T A\Delta x = \nabla f(x) + A^T A\Delta x,$$

and, hence, $\nabla^2 f(x) = A^T A$.

Let us now apply the delta method to compute the gradient and Hessian of the quadratic function

$$f(x) := \tfrac{1}{2}x^T Hx + g^T x, \quad \text{where } \; H \in \mathcal{S}^n, \; g \in \mathbb{R}^n.$$

Then

$$\begin{aligned} f(x + \Delta x) &= \tfrac{1}{2}(x + \Delta x)^T H(x + \Delta x) + g^T(x + \Delta x) \\ &= \tfrac{1}{2}x^T Hx + g^T x + (Hx + g)^T \Delta x + \tfrac{1}{2}\Delta x^T H\Delta x \\ &= f(x) + (Hx + g)^T \Delta x + \tfrac{1}{2}\Delta x^T H\Delta x, \end{aligned}$$

where $\frac{1}{2}\Delta x^T H\Delta x = o(\|\Delta x\|_2)$ since

$$\frac{\frac{1}{2}\Delta x^T H\Delta x}{\|\Delta x\|_2} = \tfrac{1}{2}\Delta x^T H\frac{\Delta x}{\|\Delta x\|_2} \to 0 \; .$$

Therefore, by (B.3), we must have

$$\nabla f(x) = Hx + g \; .$$

Again, we compute the Hessian by applying the delta method to the gradient:

$$\nabla f(x + \Delta x) = H(x + \Delta x) + g = (Hx + g) + H\Delta x = \nabla f(x) + H\Delta x,$$

and so

$$\nabla^2 f(x) = H \; .$$

4. Differential Calculus

There are many further tools for computing derivatives that do not require a direct appeal to either the partial derivatives or the delta method. These tools allow us to compute new derivatives from derivatives that are already known based on a *calculus* of differentiation. We are familiar with this differential calculus for functions mapping $\mathbb{R}$ to $\mathbb{R}$. Here we show how a few of these calculus rules extend to mappings from $\mathbb{R}^n$ to $\mathbb{R}^m$. The most elementary of these are the facts that the derivative of the scalar multiple of a function equals the scalar multiple of the derivative and the derivative of a sum is the sum of derivatives: given $F : \mathbb{R}^n \to \mathbb{R}^m$, $G : \mathbb{R}^n \to \mathbb{R}^m$ and $\alpha \in \mathbb{R}$,

$$\nabla(\alpha F) = \alpha \nabla F \quad \text{and} \quad \nabla(F + G) = \nabla F + \nabla G \; .$$

These rules are themselves derivable from the much more powerful *chain rule*.

THEOREM B.16. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ and $H : \mathbb{R}^m \to \mathbb{R}^k$ be such that F is differentiable at x and H is differentiable at $F(x)$. Then $G := H \circ F$ is differentiable at x with*

$$\nabla G(x) = \nabla H(F(x)) \circ \nabla F(x) \; .$$

REMARK B.17. *As noted in Remark B.12, one must take special care in the interpretation of this chain rule when $k = 1$ due to the presence of an additional transpose. In this case,*

$$\nabla G(x) = \nabla F(x)^T \nabla H(F(x)) \; .$$

For example, let $F : \mathbb{R}^n \to \mathbb{R}^m$ and consider the function

$$f(x) := \tfrac{1}{2} \|F(x)\|_2^2 = (\tfrac{1}{2} \|\cdot\|_2^2) \circ F(x),$$

that is, we are composing half the 2-norm squared with F. Since $\nabla(\frac{1}{2} \|\cdot\|_2^2)(y) = y$, we have

$$\nabla f(x) = \nabla F(x)^T F(x) \ .$$

This chain rule computation can be verified using the delta method:

$$\begin{aligned}
f(x + \Delta x) &= \tfrac{1}{2} \|F(x + \Delta x)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x + o(\|\Delta x\|_2)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x\|_2^2 + (F(x) + \nabla F(x)\Delta x)^T (o(\|\Delta x\|_2)) + \tfrac{1}{2} \|o(\|\Delta x\|_2)\|_2^2 \\
&= \tfrac{1}{2} \|F(x) + \nabla F(x)\Delta x\|_2 + o(\|\Delta x\|_2) \\
&= \tfrac{1}{2} \|F(x)\|_2^2 + (\nabla F(x)^T F(x))^T \Delta x + \tfrac{1}{2} \|\nabla F(x)\Delta x\|_2^2 + o(\|\Delta x\|_2) \\
&= f(x) + (\nabla F(x)^T F(x))^T \Delta x + o(\|\Delta x\|_2),
\end{aligned}$$

where $\lim_{t \to 0} \frac{o(t)}{t} = 0$ and we have used this notation as described in Remark B.17. Hence, again $\nabla f(x) = \nabla F(x)^T F(x)$.

5. The Mean Value Theorem

Given $f : \mathbb{R}^n \to \mathbb{R}$, the defining formula for the derivative,

$$f(y) = f(x) + \nabla f(x)(y - x) + o(\|y - x\|),$$

is a powerful tool for understanding the local behavior of the function f near x. If we drop the little-o term from the right hand side, we obtain the *first-order Taylor expansion* of f at x. This is called a *first-order approximation* to f at x due to the fact that the power of $\|y - x\|$ in the *error term* $o(\|y - x\|)$ is 1. Higher order approximations to f can be obtained using higher order derivatives. But before turning to these approximations, we make a closer study of the first-order expansion. In particular, we wish to extend the *Mean Value Theorem* to functions of many variables.

THEOREM B.18. *[1-Dimensional Mean Value Theorem]*
Let $\phi : \mathbb{R} \to \mathbb{R}$ be $k + 1$ times differentiable on an open interval $(a, b) \subset \mathbb{R}$. Then, for every $x, y \in (a, b)$ with $x \neq y$, there exists a $z \in (a, b)$ strictly between x and y such that

$$\phi(y) = \phi(x) + \phi'(x)(y - x) + \cdots + \frac{1}{k!}\phi^{(k)}(x)(y - x)^k + \frac{1}{(k + 1)!}\phi^{(k+1)}(z)(y - x)^{(k+1)} \ .$$

REMARK B.19. *Theorem B.18 is also called Taylor's Theorem with remainder, where we have chosen the Lagrange form for the remainder. Other forms for the remainder include the integral form and Cauchy's form. In most texts the name Mean Value Theorem is reserved for the first-order case alone.*

We use this results to easily obtain the following mean value theorem for function mapping $\mathbb{R}^n$ to $\mathbb{R}$.

THEOREM B.20. *[n-Dimensional Mean Value Theorem]*
Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Define the closed and open line segments connecting x and y by

$$[x, y] := \{(1 - \lambda)x + \lambda y \,|\, 0 \leq \lambda \leq 1\} \quad \text{and} \quad (x, y) := \{(1 - \lambda)x + \lambda y \,|\, 0 < \lambda < 1\},$$

respectively. Then there exists a $z, w \in (x, y)$ such that

$$f(y) = f(x) + \nabla f(z)^T (y - x) \quad \text{and} \quad f(y) = f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(z)(y - x).$$

PROOF. Define the function $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(t) := f(x + t(y - x))$. Since f is differentiable, so is ϕ and the chain rule tells us that

$$\phi'(t) = \nabla f(x + t(y - x))^T (y - x) \quad \text{and} \quad \phi'(t) = (y - x)^T \nabla^2 f(x + t(y - x))(y - x).$$

By applying the Mean Value Theorem B.18 to ϕ we obtain the existence of $t, s \in (0, 1)$ such that

$$f(y) = \phi(1) = \phi(0) + \phi'(t)(1 - 0) = f(x) + \nabla f(x + t(y - x))^T (y - x)$$

and

$$f(y) = \phi(1) = \phi(0) + \phi'(0)(1-0) + \tfrac{1}{2}\phi''(s)(1-0)^2 = f(x) + \nabla f(x)^T (y-x) + \tfrac{1}{2}(y-x)^T \nabla^2 f(x+s(y-x))(y-x).$$

By setting $z := x + t(y - x)$ and $w := x + s(y - x)$ we obtain the result. □

In a similar manner we can apply the Fundamental Theorem of Calculus to such functions.

THEOREM B.21. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Then*

$$f(y) = f(x) + \int_0^1 \nabla f(x + t(y - x))^T (y - x)\, dt \ .$$

PROOF. Apply the Fundamental Theorem of Calculus to the function ϕ defined in the proof of Theorem B.20. □

Unfortunately, the Mean Value Theorem does not extend to general differentiable function mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$ for $m > 1$. Nonetheless, we have the following approximate result.

THEOREM B.22. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable on an open set containing the two points $x, y \in \mathbb{R}^n$ with $x \neq y$. Then*

$$\|F(y) - F(x)\|_2 \leq \left[\max_{z \in [x,y]} \left\|F'(z)\right\|_F\right] \|y - x\|_2 \ . \tag{B.5}$$

PROOF. By the Fundamental Theorem of Calculus, we have

$$F(y) - F(x) = \begin{pmatrix} \int_0^1 \nabla F_1(x + t(y - x))^T (y - x)\, dt \\ \vdots \\ \int_0^1 \nabla F_m(x + t(y - x))^T (y - x)\, dt \end{pmatrix} = \int_0^1 \nabla F(x + t(y - x))(y - x)\, dt \ .$$

Therefore,

$$\begin{aligned} \|F(y) - F(x)\|_2 &= \left\| \int_0^1 \nabla F(x + t(y - x))(y - x)\, dt \right\|_2 \\ &\leq \int_0^1 \|\nabla F(x + t(y - x))(y - x)\|_2 \, dt \\ &\leq \int_0^1 \|\nabla F(x + t(y - x))\|_F \, \|y - x\|_2 \, dt \\ &\leq \left[\max_{z \in [x,y]} \left\|F'(z)\right\|_F\right] \|y - x\|_2 \ . \end{aligned}$$

□

The bound (B.5) is very useful in many applications. But it can be simplified in cases where ∇F is known to be continuous since in this case the Weierstrass extreme value theorem says that, for every $\beta > 0$,

$$\max_{z \in \beta\mathbb{B}} \left\|F'(z)\right\|_F =: K < \infty \ .$$

Hence, by Theorem B.22,

$$\|F(x) - F(y)\|_2 \leq K \|x - y\|_2 \quad \forall\ x, y \in \beta\mathbb{B} \ .$$

This kind of inequality is extremely useful and leads to the following notion of continuity.

DEFINITION B.23. *[Lipschitz Continuity]*
We say that $F : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on a set $S \subset \mathbb{R}^n$ if there exists a constant $K > 0$ such that

$$\|F(x) - F(y)\| \leq K \|x - y\| \quad \forall\ x, y \in S\ .$$

The constant K is called the modulus of Lipschitz continuity for F over S, and depends on the choice of norms for $\mathbb{R}^n$ and $\mathbb{R}^m$.

As one application of Lipschitz continuity, we give the following lemma concerning the accuracy of the first-order Taylor approximation of a function.

LEMMA B.24. *[Quadratic Bound Lemma]*
Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be such that ∇F is Lipschitz continuous on the set $S \subset \mathbb{R}^n$. If $x, y \in S$ are such that $[x, y] \subset S$, then

$$\|F(y) - (F(x) + \nabla F(x)(y - x))\|_2 \leq \frac{K}{2} \|y - x\|_2^2 ,$$

where K is the modulus of Lipschitz continuity for ∇F on S.

PROOF. Observe that

$$\begin{aligned} F(y) - F(x) - \nabla F(x)(y - x) &= \int_0^1 \nabla F(x + t(y - x))(y - x)dt - \nabla F(x)(y - x) \\ &= \int_0^1 [\nabla F(x + t(y - x)) - \nabla F(x)](y - x)dt. \end{aligned}$$

Hence

$$\begin{aligned} \|F(y) - (F(x) + \nabla F(x)(y - x))\|_2 &= \left\| \int_0^1 [\nabla F(x + t(y - x)) - \nabla F(x)](y - x)dt \right\|_2 \\ &\leq \int_0^1 \|(\nabla F(x + t(y - x) - \nabla F(x))(y - x)\|_2 \, dt \\ &\leq \int_0^1 \|\nabla F(x + t(y - x)) - \nabla F(x)\|_F \ \|y - x\|_2 \, dt \\ &\leq \int_0^1 Kt \|y - x\|_2^2 \, dt \\ &= \tfrac{K}{2} \|y - x\|_2^2 . \end{aligned}$$

□

The Mean Value Theorem also allows to obtain the following *second order* approximation.

THEOREM B.25. *Let $f : \mathbb{R}^n \to \mathbb{R}$ and suppose that $\nabla^2 f(x)$ exists and is continuous at x. Then*

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2). \tag{B.6}$$

PROOF. The mean value theorem tells us the for every $y \in x + \epsilon\mathbb{B}$ there is a $z \in (x, y)$ such that

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ &= f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + \tfrac{1}{2}(y - x)^T \left[\nabla^2 f(z) - \nabla^2 f(x)\right](y - x) \\ &= f(x) + \nabla f(x)^T (y - x) + \tfrac{1}{2}(y - x)^T \nabla^2 f(x)(y - x) + o(\|y - x\|^2). \end{aligned}$$

□

If we drop the $o(\|y - x\|^2)$ in the equation (B.6), we obtain the *second-order Taylor approximation* to f at x. This is a second-order approximation since the power of $\|y - x\|$ in the little-o term is 2, i.e., $o(\|y - x\|^2)$.

6. The Implicit Function Theorem

THEOREM B.26. *Let $F : \mathbb{R}^{n+m} \to \mathbb{R}^n$ be continuously differentiable on an open set $E \subset \mathbb{R}^{n+m}$. Further suppose that there is a point $(\bar{x}, \bar{y}) \in \mathbb{R}^{n+m}$ at which $F(\bar{x}, \bar{y}) = 0$. If $\nabla_x F(\bar{x}, \bar{y})$ is invertable, then there exist open sets $U \subset \mathbb{R}^{n+m}$ and $W \subset \mathbb{R}^m$, with $(\bar{x}, \bar{y}) \in U$ and $\bar{y} \in W$, having the following property:*

To every $y \in W$ corresponds a unique $x \in \mathbb{R}^n$ such that

$$(x, y) \in U \qquad \text{and} \qquad F(x, y) = 0\,.$$

Moreover, if x is defined to be $G(y)$, then G is a continuously differentiable mapping of W into $\mathbb{R}^n$ satisfying

$$G(\bar{y}) = \bar{x}, \quad F(G(y), x) = 0 \ \forall\ y \in W, \quad \text{and} \quad G'(\bar{y}) = -(\nabla_x F(\bar{x}, \bar{y}))^{-1} \nabla_y F(\bar{x}, \bar{y})\ .$$

APPENDIX C

Probability and Random Variables

1. Introduction

A probability space is a triple $(\Omega, \mathcal{A}, p)$, where Ω is called the event space, $\mathcal{A}$ is a collection of subsets of Ω, and p is a mapping from $\mathcal{A}$ to $[0,1]$. The collection $\mathcal{A}$ is called a σ-field since it satisfies the following three conditions:

$\mathcal{A}$ is a σ-field

(1) $\Omega \subset \mathcal{A}$
(2) If $S \in \mathcal{A}$, then $\Omega \backslash S = \{\omega \in \Omega : \omega \notin S\} \in \mathcal{A}$.
(3) If $S_n \in \mathcal{A},\ n = 1, 2, \ldots$, then $\bigcup_{n=1}^{\infty} S_n \in \mathcal{A}$.

The function $p : \mathcal{A} \to [0,1]$ is called a probability measure since it satisfied the following two conditions:

p is a Probability Measure

(1) $p(\Omega) = 1$
(2) If $S_n \in \mathcal{A},\ n = 1, 2, \ldots$ satisfy $S_i \cap S_j = \emptyset$ for $i \neq j$, then

$$p(\bigcup_{n=1}^{\infty} S_n) = \sum_{n-1}^{\infty} p(S_n).$$

A mapping $x : \Omega \to \mathbb{R}$ is called a *random variable* for the probability space $(\Omega, \mathcal{A}, p)$, if for every open set $U \subset \mathbb{R}$, we have

$$x^{-1}(U) = \{\omega \in \Omega : x(\omega) \in U\} \in \mathcal{A}\ .$$

Studying properties of random variables is the focus of this appendix. Our first important observation about random variables is that *any continuous function of a random variable is itself a random variable.* For example, if x is a random variable, then so are x^2 and e^x.

Example C.1. *(Fair Coin Toss)*
Consider the setting of a fair coin toss. Here the event space Ω is given by

$$\Omega = \{heads, tails\}.$$

The collection $\mathcal{A}$ is given by

$$\mathcal{A} = \{\{heads\}, \{tails\}, \{heads, tails\}, \emptyset\}\ .$$

The probability measure p is completely determined by the specification

$$p(heads) = \tfrac{1}{2} \quad and \quad p(tails) = \tfrac{1}{2}\ .$$

A random variable can be defined by describing a game wherein the payoff is \$1 if heads comes up and -\$1 if tails comes up. The random variable x is the payoff, and we have

$$x(heads) = 1 \quad and \quad x(tails) = -1.$$

Example C.2. *(Fair Die)*
Consider the setting of a fair die toss. Here we have

$$\begin{aligned} \Omega &= \{1, 2, 3, 4, 5, 6\} \\ \mathcal{A} &= \{\{1\}, \{2\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{5, 6\}, \ldots, \Omega, \emptyset\} \end{aligned}$$

and

$$p(1) = p(2) = \cdots = p(6) = \frac{1}{6} \ .$$

Again, one can define a random variable that describes a payoff depending on the roll of the die.

The two examples given above are examples of finite discrete probability spaces. Continuous probability spaces are also possible. For example, the temperature at a given place and time can be modeled as a continuous random variable. However, we reserve a more detailed discussion of continuous random variables until a more detailed understanding of them is required.

In these notes we introduce a few key tools for understanding the behavior of random variables. These are the *expectation*, *variance*, and *covariance*. These concepts are developed in the context of discrete random variables.

Consider the discrete probability space

$$\Omega = \{\omega_1, \omega_2, \omega_3, \dots\},$$

with probability measure given by

$$p(\omega_i) = p_i, \text{ for } i = 1, 2, \dots .$$

Let x be a random variable on this probability space given by

$$x(\omega_i) = x_i.$$

In this case the probability measure p defines a function $f : \mathbb{R} \to \mathbb{R}$ by the relation

$$f(x) = \begin{cases} p_i & , \text{ if } x{=}x_i \text{ for some } i = 1, 2, \dots ; \\ 0 & , \text{ otherwise.} \end{cases}$$

The function f is called the *probability density function* (or pdf) for the random variable x. Then the expectation of this random variable is defined to be

$$\begin{aligned} E(x) &= \sum_{\omega \in \Omega} f(x(\omega))x(\omega) \\ &= f(x_1)x_1 + f(x_2)x_2 + \cdots + f(x_n)x_n \\ &= p_1x_1 + p_2x_2 + p_3x_3 + \cdots + p_nx_n. \end{aligned}$$

Note that the mapping $E(\cdot)$ can be applied to any random variable over the discrete probability space associated with Ω and p. For this reason we call E the *expectation operator*. It *operates*, or acts, on random variables by mapping them into the real numbers $\mathbb{R}$. The expectation of a random variable is called the *mean* of the random variable. We use the notation $E(x) = \mu_x$.

Now that we have defined the expectation operator, we can use it to study other properties of a random variable by taking the expectation of functions of the random variable. But before doing so we catalogue for future reference three important properties of the expectation operator.

Three Properties of the Expectation Operator

(1) A random variable z that is constant over the event space Ω (i.e., $z(\omega) = z_0 \in \mathbb{R}$ for all $\omega \in \Omega$) is called a deterministic variable, or certain value. That is, its value is always determined and certain. There is nothing *random* about such variables. For example, the mean of a random variable is a certain value. For a certain value z_0, we have $E(z_0) = z_0$.

(2) The expectation operator is a *linear operator*. That is, for any to random variables x and y, and any two scalar values $\alpha, \beta \in \mathbb{R}$, we have

$$E(\, \alpha x + \beta y \,) \ = \ \alpha E(x) + \beta E(y) \ .$$

(3) The expectation operator is *sign preserving*. That is, if x is a random variable on Ω for which $x(\omega) \geq 0$ for all $\omega \in \Omega$, then $E(x) \geq 0$.

The mean of a random variable tells you its *average* value in the sense of expectation. We now consider another measure for a random variable that tells us something about how its values are dispersed about its mean. This measure is called the *variance* of a random variable. It is the expectation of the difference from the mean squared. That is, if x is a random variable, then the variance of x is the expectation of the random variable $(x - \mu_x)^2$. The notation for this is

$$\text{var}(x) \;=\; E[(x - \mu_x)^2].$$

By using the linearity of the expectation, we get the following alternative formula for the variance:

$$\begin{aligned} \text{var}(x) &= E[(x - \mu_x)^2] \\ &= E[x^2 - 2\mu_x x + \mu_x^2] \\ &= E[x^2] - 2\mu_x E[x] + \mu_x^2 \\ &= E[x^2] - \mu_x^2 \; . \end{aligned}$$

Note that the units of the variance are the square of the units of the random variable itself. To obtain a measure of *spread* of a random variable that is in the same units we simply take the square root of the variance. This measure is called the *standard deviation*, and is denoted with the letter σ:

$$\sigma_x \;=\; \sqrt{\text{var}(x)} \; .$$

Example C.3. *(Fair Die)*
For the fair die example, consider the random variable $x(\omega) = \omega$. In this case we get

$$\begin{aligned} E[x] &= \frac{1}{6}(1+2+3+4+5+6) = 3.5, \\ var(x) &= \frac{1}{6}(1+4+9+25+36) - (7/2)^2 = 2.92, \quad \textit{and} \\ \sigma_x &= 1.71 \; . \end{aligned}$$

Note that a deterministic, or certain, value always has zero variance since it equals its expectation everywhere.

Let x and y be two random variables having means μ_x and μ_y respectively. We define the *covariance* of these random variables as

$$\text{cov}(x, y) \;=\; E[(x - \mu_x)(y - \mu_y)] \; .$$

We will use the following notation for the covariance of two random variables:

$$\sigma_{xy} = \text{cov}(x, y).$$

Observe that $\sigma_{xy} = \sigma_{yx}$.

Again, using the linearity of the expectation operator we obtain the alternative expression

$$\begin{aligned} \text{cov}(x, y) &= E[(x - \mu_x)(y - \mu_y)] \\ &= E[xy - \mu_x y - \mu_y x + \mu_x \mu_y] \\ &= E[xy] - \mu_x E[y] - \mu_y E[x] + \mu_x \mu_y \\ &= E[xy] - \mu_x \mu_y \; . \end{aligned}$$

We say that two random variables are *stochastically independent* (or just independent) if the values they take are independent. For example, we can consider an event space where we simultaneously flip a fair coin and roll a fair die. The outcome of the coin flip and the roll of the die have nothing to do with each other and so if x is a random variable that only depends on the outcome of the coin toss and y is a random variable that only depends on the outcome of the die toss, then x and y are independent random variables. We have the following fundamental fact concerning independent random variables.

Fact: If x and y are independent random variables, then

$$E[xy] \;=\; E[x]\, E[y] \; .$$

EXAMPLE C.4. *(Independence)*
Let x be a random variable that depends on the outcome of a fair coin toss with $x(H) = 0$ and $x(T) = 1$ (H =heads, T =tails), and let y be a random variable that depends on the outcome of a fair die toss with $y(i) = i,\ i = 1, 2, 3, 4, 5, 6$. Then

$$E(xy) = \frac{1}{12}(0 \cdot 1 + 0 \cdot 2 + \cdots + 0 \cdot 6 + 1 \cdot 1 + 1 \cdot 2 + \cdots + 1 \cdot 6) = \frac{7}{4} = E[x]E[y].$$

Note that in the example given above, we can describe the *joint* probability space associated with the pair of random variables (x, y) as

$$\Omega = \{(H, 1), (H, 2), \ldots, (H, 6), (T, 1), (T, 2), \ldots, (T, 6)\},$$

with corresponding probability measure given by

$$p(\omega_i, \omega_j) = \frac{1}{12}, \text{ for } i = 1, 2,\ j = 1, 2, \ldots, 6\ .$$

The corresponding pdf for the paired random variables (x, y) is

$$f(x, y) = f_1(x) f_2(x),$$

where f_1 is the pdf for x and f_2 is the pdf for y. In this case, we call the pdf f the *joint probability density function* (joint pdf) for the random variables x and y.

An immediate consequence of independence is that the covariance is zero. That is, if x and y are independent random variables, then $\text{cov}(x, y) = E[xy] - E[x]E[y] = 0$.

We have the following useful bound on the covariance of two random variables.

Fact: For any two random variables x and y we have

$$|\sigma_{xy}| \leq \sigma_x \sigma_y\ .$$

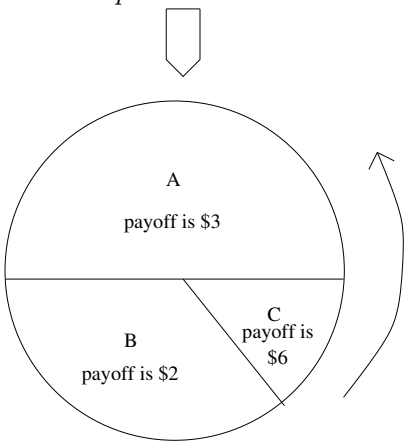
In connection with the covariance of two random variables x and y we also define their *correlation coefficient* ρ_{xy} by

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}\ .$$

By the fundamental bound stated above, we have

$$-1 \leq \rho_{xy} \leq 1\ .$$

EXAMPLE C.5. *In this example we study the properties of random variables associated with a betting wheel. A sample betting wheel is given in the picture.*



probability of A is 1/2, probability of B is 1/3, probability of C is1/6

The idea is that the wheel is spun and a payoff is given depending on the section of the wheel that the arrow at the top points to. In this setting we assume that the probability that the arrow points to any given region to proportional to the arc-length of the circular edge of that region divided by $2\pi r$ where r is the radius of the wheel. This gives a probability of $\frac{1}{2}$ that the arrow will point to the A region on a given spin of the wheel. The corresponding probabilities for the B and C regions are $\frac{1}{3}$ and $\frac{1}{6}$, respectively.

We now define three random variables associated with three different \$1 bets on the three regions of the wheel. The random variable x_1 is the payoff associated with a \$1 bet on region A:

$$x_1(A) = 3,\ x_1(B) = 0,\ \textit{and}\ x_1(C) = 0.$$

Correspondingly, random variables x_2 and x_3 are associated with the payoff for \$1 bets on regions B and C, respectively:

$$x_2(A) = 0,\ x_2(B) = 2,\ \textit{and}\ x_2(C) = 0$$
$$x_3(A) = 0,\ x_3(B) = 0,\ \textit{and}\ x_3(C) = 6\ .$$

We now compute the means, variances, and covariances for these random variables.

$$\begin{aligned} E(x_1) &= \tfrac{1}{2}x_1(A) + \frac{1}{3}x_1(B) + \frac{1}{6}x_1(C) = \frac{3}{2} \\ E(x_2) &= \tfrac{1}{2}x_2(A) + \frac{1}{3}x_2(B) + \frac{1}{6}x_2(C) = \frac{2}{3} \\ E(x_3) &= \tfrac{1}{2}x_3(A) + \frac{1}{3}x_3(B) + \frac{1}{6}x_3(C) = 1 \end{aligned}$$

$$\begin{aligned} var(x_1) &= \tfrac{1}{2}x_1^2(A) + \frac{1}{3}x_1^2(B) + \frac{1}{6}x_1^2(C) - (\frac{3}{2})^2 = \frac{9}{4} \\ var(x_2) &= \tfrac{1}{2}x_2^2(A) + \frac{1}{3}x_2^2(B) + \frac{1}{6}x_2^2(C) - (\frac{2}{3})^2 = \frac{4}{9} \\ var(x_3) &= \tfrac{1}{2}x_3^2(A) + \frac{1}{3}x_3^2(B) + \frac{1}{6}x_3^2(C) - (1)^2 = 5 \end{aligned}$$

$$\begin{aligned} cov(x_1, x_2) &= \tfrac{1}{2}x_1(A)x_2(A) + \frac{1}{3}x_1(B)x_2(B) + \frac{1}{6}x_1(C)x_2(C) - \mu_1\mu_2 \\ &= -\mu_1\mu_2 = -1 \\ cov(x_1, x_3) &= -\frac{3}{2} \\ cov(x_2, x_3) &= -\frac{2}{3}\ . \end{aligned}$$

The joint pdf for the random variables $x_1,\ x_2$, and x_3 is the function

$$f(x_1, x_2, x_3) = \begin{cases} \frac{1}{2} & ,\ \textit{if}\ (x_1, x_2, x_3) = (3, 0, 0); \\ \frac{1}{3} & ,\ \textit{if}\ (x_1, x_2, x_3) = (0, 2, 0); \\ \frac{1}{6} & ,\ \textit{if}\ (x_1, x_2, x_3) = (0, 0, 6); \\ 0 & ,\ \textit{otherwise.} \end{cases}$$

2. Linear Combinations of Random Variables and Random Vectors

We will also need to consider linear combinations of random variables. A linear combination of random variables is itself a random variable. We now show how to compute the mean and variance of such linear combinations. Let $x_i,\ i = 1, 2, \ldots, n$ be n random variables and set

$$\mu_i = E[x_i], \quad \sigma_i^2 = \mathrm{var}(x_i), \quad \text{and} \quad \mathrm{cov}(x_i, x_j) = \sigma_{ij},$$

for $i, j = 1, 2, \ldots, n$. Let $\alpha_i, i = 1, 2, \ldots, n$ be real scalars and consider the random variable

$$w = \sum_{i=1}^{n} \alpha_i x_i \ .$$

Due to the linearity of the expectation operator, we have

$$E(w) = E(\sum_{i=1}^{n} \alpha_i x_i) = \sum_{i=1}^{n} \alpha_i E(x_i) = a^T \mu,$$

and

$$\begin{aligned}
\text{var}(w) &= E[(w - E(w))^2] \\
&= E\left[\left(\sum_{i=1}^{n} \alpha_i x_i - \sum_{i=1}^{n} \alpha_i E(x_i)\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} \alpha_i (x_i - \mu_i)\right)^2\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j (x_i - \mu_i)(x_j - \mu_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j E[(x_i - \mu_i)(x_j - \mu_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j \text{cov}(x_i, x_j) \\
&= a^T \Sigma a,
\end{aligned}$$

where

$$a = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}, \text{ and } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{bmatrix} .$$

The matrix Σ in the last line of this derivation is called the covariance matrix for the vector of random variables

$$x = (x_1, x_2, \ldots, x_n)^T \ .$$

The covariance matrix Σ is clearly symmetric since $\sigma_{ij} = \text{cov}(x_i, x_j) = \text{cov}(x_j, x_i) = \sigma_{ji}$. But it is also easily seen to be positive semi-defiinite since

$$\Sigma = E[(x - \mu)(x - \mu)^T],$$

so that

$$a^T \Sigma a = E[a^T (x - \mu)(x - \mu)^T a] = E[((x - \mu)^T a)^2] \geq 0 \ .$$

3. The Mean - Standard Deviation Curve

Let x_1 and x_2 be two random variables, and consider the sum

$$w_\alpha = (1 - \alpha)x_1 + \alpha x_2 \ .$$

We assume that $\sigma_1 = \sigma_{x_1} \neq \sigma_{x_2} = \sigma_2$ with both being positive, and $|\rho_{12}| < 1$ where $\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2}$ is the correlation coefficient. We wish to graph the *mean-standard deviation curve* for w_α as α varies over the real numbers. That is, we wish to graph all of the points

$$(\sqrt{\text{var}(w_\alpha)}\,,\, E(w_\alpha)) \quad \text{for all} \quad \alpha \in \mathbb{R}. \tag{C.1}$$

For this recall that

$$\begin{aligned} \mu_\alpha = E(w_\alpha) &= (1-\alpha)\mu_1 + \alpha\mu_2 = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}^T \begin{pmatrix} 1-\alpha \\ \alpha \end{pmatrix}, \\ \text{and} & \\ \sigma_\alpha^2 = \text{var}(w_\alpha) &= (1-\alpha)^2\sigma_1^2 + 2\alpha(1-\alpha)\sigma_{12} + \alpha^2\sigma_2^2 \\ &= \begin{pmatrix} 1-\alpha \\ \alpha \end{pmatrix}^T \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{pmatrix} 1-\alpha \\ \alpha \end{pmatrix}. \end{aligned}$$

We use this parametrization of $\mu_\alpha = E(w_\alpha)$ and $\sigma_\alpha = \text{var}(w_\alpha)$ in α to write σ_α as a function of μ_α that can be graphed. First observe that

$$\alpha = \frac{\mu_\alpha - \mu_1}{\mu_2 - \mu_1}, \tag{C.2}$$

and

$$\sigma_\alpha^2 = (1-\alpha)^2\sigma_1^2 + 2\alpha(1-\alpha)\sigma_{12} + \alpha^2\sigma_2^2\ .$$

The second equation establishes a hyperbolic relationship between σ_α and α. Rearranging the second equation gives

$$\frac{\sigma_\alpha^2}{(b/a)^2} - \frac{\left(\alpha - \frac{\sigma_1^2 - \sigma_{12}}{a^2}\right)^2}{(b/a^2)^2} = 1,$$

where

$$\begin{aligned} b^2 &= (\sigma_1\sigma_2)^2 - \sigma_{12}^2 > 0 \quad \text{and} \\ a^2 &= \sigma_1^2 - 2\sigma_{12} + \sigma_2^2 = (\sigma_1 - \sigma_2)^2 + 2(1-\rho_{12})\sigma_1\sigma_2 > 0\ . \end{aligned}$$

This is the Cartesian equation for a hyperbola. Substituting in for α using (C.2) gives

$$\frac{\sigma_\alpha^2}{(b/a)^2} - \frac{(\mu_\alpha - ((1-\bar\alpha)\mu_1 + \bar\alpha\mu_2))^2}{(\mu_2 - \mu_1)^2(b/a^2)^2} = 1,$$

where

$$\bar\alpha = \frac{\sigma_1^2 - \sigma_{12}}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}.$$

Therefore, the mean-standard deviation curve (C.1) is the right half of a hyperbola (since $\sigma_\alpha \geq 0$) with axis of symmetry parallel to the standard deviation axis. The vertex of this half of the hyperbola occurs at

$$\begin{aligned} \mu_\alpha &= ((1-\bar\alpha)\mu_1 + \bar\alpha\mu_2)\,, \\ \sigma_\alpha &= \sqrt{\frac{(\sigma_1\sigma_2)^2 - \sigma_{12}}{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}}, \end{aligned}$$

and so it is this value of σ_α that gives the smallest standard deviation (and, hence the smallest variance) from among all possible linear combinations of x_1 and x_2. The slope of the asymptotes to the hyperbola are

$$\pm\frac{|\mu_2 - \mu_1|}{\sqrt{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}}.$$

Exercises

(1) A card is selected from and ordinary deck of playing cards. The outcome c is one of the 52 cards. This defines an event space

$$\Omega = \{A\clubsuit, 2\clubsuit, 3\clubsuit, \ldots, Q\clubsuit, K\clubsuit, A\diamondsuit, \ldots, A\spadesuit, \ldots, K\spadesuit\}$$

(A = ace, J = jack, Q = queen, K = king). We assume that the probability of each event is equally likely ($p(c) = 1/52$ for all $c \in \Omega$). Define the random variable x on Ω by

$$x(c) = \begin{cases} 4 & , \text{ if } c \text{ is an ace;} \\ 3 & , \text{ if } c \text{ is a king;} \\ 2 & , \text{ if } c \text{ is a queen;} \\ 1 & , \text{ if } c \text{ is a jack;} \\ 0 & , \text{ otherwise.} \end{cases}$$

What is the probability density function associated with the random variable x?

(2) Suppose x and y are two random variables with joint pdf $f(x, y) = 1/3$ if $(x, y) = (0, 0), (0, 1), (1, 1)$.
   (a) What is the pdf for x?
   (b) What is the pdf for y?
   (c) Compute $\mu_x,\ \mu_y,\ \sigma_x,\ \sigma_y$, and σ_{xy}.

(3) Suppose x and y are two random variables with joint pdf

(x, y)	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$
$f(x, y)$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{3}{15}$	$\frac{1}{15}$	$\frac{1}{15}$	$\frac{4}{15}$

with $f(x, y) = 0$ otherwise.
   (a) What is the pdf for x?
   (b) What is the pdf for y?
   (c) Compute $\mu_x,\ \mu_y,\ \sigma_x,\ \sigma_y$, and σ_{xy}.

(4) Sketch the mean-standard deviation curve for the two random variables x_1 and x_3 in Example C.5.

(5) Sketch the mean-standard deviation curve for the two random variables x_1 and x_2 in Example C.5 (be careful in this one!).

Index