Introduction to Optimization

Vincent Leclere

Autumn 2026

Foreword

This document is a textbook for the course *Optimisation 1A* at *École des Ponts et Chaussées, Institut Polytechnique Paris*. It is designed to be self-contained, but it assumes familiarity with basic concepts from linear algebra, calculus, and real analysis.

These notes will be available during the exam, and you will be invited to cite them in your answers.

Reading conventions. Throughout the book, we use the following symbols to help organize your reading and highlight important results:

- No symbol indicates core material.
- ♡ for *very important* results (key ideas and tools you should remember).
- ♢ for results that can be skipped on a first reading, but are part of the curriculum.
- ✠ for sections or results that are *not* part of the curriculum.
- `(*)` marks a more advanced question or exercise part.

Thanks. These notes are largely based on the lecture notes of previous years, written by Frédéric Meunier. I am grateful to him for sharing his notes and for his help in preparing the course.

Notation / Quick reference

$[n]$	The index set $\{1, \ldots, n\}$.
$\mathbb{R}_+$	The nonnegative real numbers.
$\operatorname{dom} f$	The effective domain $\{x \mid f(x) < +\infty\}$.
$\operatorname{lev}_M(f)$	The sublevel set $\{x \mid f(x) \leq M\}$.
$T_X(x)$	The tangent cone to X at x.
$T_X^\ell(x)$	The linearized tangent cone at x.
$J_0(x)$	The indices of the inequality constraints active at x.
$\mathcal{L}$	The Lagrangian.
$x^\sharp$	An optimal point or optimal solution.
$p^\star, d^\star, f^\star, \ldots$	Optimal values (primal, dual, objective, and so on).
$\operatorname{conv}(S)$	The convex hull of a set S.
$\operatorname{cone}(S)$	The conic hull of a set S.

Contents

Chapter 1

Introduction

1.1 What is optimization?

Optimization is the mathematical discipline concerned with making the best decision under constraints. In mathematical terms, it means minimizing (or maximizing) an *objective function* over a *feasible set*.

To work with real-world questions, we first model them as mathematical problems, then study the properties of optimal solutions and how to compute them.

An optimization problem is defined by a *decision variable* $\mathbf{x}$ (what we can choose), a set X of *feasible* decisions (the constraints), and an *objective function* f (also called a criterion) that measures the quality of a decision. In this course we mostly consider minimization problems of the form[1]

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) && \text{(1.1a)} \\ \text{s.t.} \quad & \mathbf{x} \in X && \text{(1.1b)} \end{aligned}$$

where *s.t.* stands for *subject to*. Any $\mathbf{x} \in X$ is a *feasible solution*; a global minimizer $\mathbf{x}^\sharp \in X$ is an *optimal solution*; and

$$\inf\{f(\mathbf{x}) \mid \mathbf{x} \in X\}$$

is the *optimal value*. In particular, the optimal value may be $-\infty$ if the problem is unbounded below, or $+\infty$ if it is infeasible[2]. Maximization is equivalent to minimizing $-f$.

1.1.1 Motivating examples

Optimization problems arise in many different contexts. In engineering design, one may want to minimize the material used to build a device subject to geometry or performance constraints (e.g. the shape of a chimney with fixed inlet/outlet diameters). In physics and mechanics, one often characterizes equilibria or trajectories via energy principles (e.g. minimizing a potential energy, or, more generally, optimizing an action functional). In statistics and machine learning, one typically fits a model by minimizing a loss function (e.g. maximum likelihood or regularized empirical risk minimization). In economics and operations management, agents may maximize a utility or profit subject to budget and resource constraints. Finally, in logistics and networks, one may plan a transportation scheme, a delivery tour, or a routing strategy that minimizes a total cost or congestion.

[1] Throughout this text, scalars are denoted by plain letters (e.g. $t \in \mathbb{R}$) and vectors by bold letters (e.g. $\mathbf{x} \in \mathbb{R}^n$); the i-th component of $\mathbf{x}$ is written x_i. For a positive integer k, we use the shorthand $[k] := \{1, \dots, k\}$.

[2] With the classical convention that $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$.

Here are a few representative examples, spanning several modeling paradigms that will reappear throughout the course.

♡ *Example* 1.1 (Production planning (LP)). A company produces n goods. Each unit of good j consumes a_{ij} units of resource i and yields a profit p_j. Let $x_j \geq 0$ denote the quantity of good j produced. The goal is to maximize total profit subject to resource availabilities b_i and demand limits d_j :

$$\underset{\mathbf{x}}{\text{Max}} \quad \sum_{j\in[n]} p_j x_j \tag{1.2a}$$

$$\text{s.t.} \quad \sum_{j\in[n]} a_{ij} x_j \leq b_i \qquad \forall i \in [m] \tag{1.2b}$$

$$0 \leq x_j \leq d_j \qquad \forall j \in [n] \tag{1.2c}$$

This is naturally modeled as a *linear program* (LP).

♡ *Example* 1.2 (Knapsack (0–1 ILP)). We have n items with values v_i and weights w_i, and a bag with capacity W. Let $x_i \in \{0, 1\}$ indicate whether item i is selected. The goal is to maximize total value without exceeding the capacity:

$$\underset{\mathbf{x}}{\text{Max}} \quad \sum_{i\in[n]} v_i x_i \tag{1.3a}$$

$$\text{s.t.} \quad \sum_{i\in[n]} w_i x_i \leq W \tag{1.3b}$$

$$x_i \in \{0, 1\} \qquad \forall i \in [n] \tag{1.3c}$$

This is a simple and intuitive *integer linear program* (0–1 ILP).

Example 1.3 (Traveling salesman problem (TSP)). Given n cities and pairwise travel costs c_{ij}, find a shortest tour visiting every city exactly once and returning to the start. Let $x_{ij} \in \{0, 1\}$ indicate whether arc (i, j) is used in the tour:

$$\underset{\mathbf{x}}{\text{Min}} \quad \sum_{i\neq j} c_{ij} x_{ij} \tag{1.4a}$$

$$\text{s.t.} \quad \sum_{j\in[n]\setminus\{i\}} x_{ij} = 1 \qquad \forall i \in [n] \tag{1.4b}$$

$$\sum_{j\in[n]\setminus\{i\}} x_{ji} = 1 \qquad \forall i \in [n] \tag{1.4c}$$

$$\sum_{i\in S} \sum_{j\in[n]\setminus S} x_{ij} \geq 1 \qquad \forall \emptyset \subsetneq S \subsetneq [n] \tag{1.4d}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in [n],\ i \neq j \tag{1.4e}$$

The third family of constraints (subtour elimination) ensures the solution forms a single connected tour. The TSP is computationally hard in general and motivates mixed-integer modeling and exact algorithms such as branch-and-bound. It is also a special case of the vehicle routing problem (VRP), which has many practical applications in logistics and transportation. Alternative TSP formulations are presented in Example 5.5 and Exercise 5.3.

Example 1.4 (Spring equilibrium (QP)). Consider a system of masses connected by linear springs. The equilibrium displacement vector $\mathbf{q} \in \mathbb{R}^n$ minimizes the total elastic potential energy minus

the work of external forces $\mathbf{f}$:

$$\min_{\mathbf{q}} \quad \tfrac{1}{2}\mathbf{q}^\top K\mathbf{q} - \mathbf{f}^\top \mathbf{q} \tag{1.5a}$$

$$\text{s.t.} \quad A\mathbf{q} = \mathbf{b} \tag{1.5b}$$

where $K \in \mathbb{R}^{n\times n}$ is the symmetric positive semidefinite stiffness matrix and $A\mathbf{q} = \mathbf{b}$ encodes boundary conditions (e.g. fixed supports). The objective is quadratic in $\mathbf{q}$, so this is a *quadratic program* (QP).

Example 1.5 (Maximum entropy at equilibrium (statistical physics)). At thermal equilibrium, a physical system can occupy microscopic states i with energies E_i, but we only observe coarse macroscopic information (e.g. the average energy U). A natural modeling principle is to choose the least biased probability distribution (p_i) consistent with these constraints, i.e. the one maximizing the Shannon entropy:

$$\max_{(p_i)} \quad -\sum_i p_i \log p_i \tag{1.6a}$$

$$\text{s.t.} \quad \sum_i p_i = 1 \tag{1.6b}$$

$$\sum_i p_i E_i = U \tag{1.6c}$$

$$p_i \geq 0 \qquad \forall i \tag{1.6d}$$

This yields the Gibbs–Boltzmann distribution $p_i \propto e^{-\beta E_i}$.

Despite the variety of settings, the modeling step is similar: choose decision variables, write constraints describing feasible decisions, and define an objective to optimize.

1.1.2 ♢ Some real-world applications of optimization

Let us briefly mention three representative application areas that fit particularly well with the finite-dimensional models studied in this course.

Energy dispatch and marginal prices. In electric power systems, one repeatedly solves *dispatch* problems to decide how much each generator should produce in order to meet demand at minimum cost, subject to operational constraints (capacity limits, ramping limits, reserve requirements, and possibly network constraints). In its simplest form ("economic dispatch"), a key constraint is the balance equation "total generation equals total demand". The Lagrange multiplier associated with this balance constraint has a direct economic meaning: it is the *marginal price* of energy, i.e. the cost increase induced by tightening the balance by one unit. In market settings, this "shadow price" interpretation underlies the way clearing prices are computed; with network constraints, one obtains the more refined notion of *locational* marginal prices.

Transportation and logistics. Many logistics decisions can be formulated as optimization problems on networks: how to ship goods from plants to warehouses and stores, how to route vehicles, how to select depots, or how to schedule deliveries. When the decisions are continuous flows, these models often lead to linear programs (transportation and minimum-cost flow). When discrete choices are essential (vehicle routing, facility location, "open/close" decisions), one naturally arrives at mixed-integer linear programs. This modeling paradigm is widely used in supply-chain design, last-mile delivery, and production-distribution planning.

Structural design. In mechanics and engineering design, one may want to minimize weight or material usage subject to stress, displacement, or stability constraints. After discretization (e.g. by finite elements), many such questions become finite-dimensional optimization problems

where the decision variables are geometric parameters, material properties, or design coefficients. Quadratic energy models yield quadratic programs, while more detailed models lead to nonlinear constrained optimization. The resulting multipliers can again be interpreted as sensitivity indicators (e.g. how the optimal objective would change if a load limit or a displacement tolerance were tightened).

1.1.3 ✠ Beyond finite-dimensional optimization

The focus of this course is on finite-dimensional problems (decision variables in $\mathbb{R}^n$ with finitely many constraints). Many important areas of modern optimization go beyond this setting, but they often lead to finite-dimensional problems after suitable approximations (discretization, sampling, or relaxation). We briefly mention a few of these extensions.

Optimal control. When the decision variable is a function of time, we enter the realm of *optimal control*. A typical problem consists in finding a state trajectory $\mathbf{x}(t)$ and a control $\mathbf{u}(t)$ that minimize an integral cost (and possibly a terminal cost), for instance

$$\int_0^T \ell(\mathbf{x}(t), \mathbf{u}(t))\, dt + \phi(\mathbf{x}(T)),$$

subject to a dynamical constraint $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t))$, pointwise constraints on $\mathbf{u}(t)$ and/or $\mathbf{x}(t)$, and boundary conditions on $\mathbf{x}(0)$ and $\mathbf{x}(T)$. After time discretization (direct transcription), one typically obtains a large but finite-dimensional constrained optimization problem. Applications include trajectory optimization in aerospace and robotics, as well as energy management in dynamical systems.

Shape optimization. In *shape optimization*, the decision variable is a geometric object (a domain, a boundary, or a shape parameterization), and the objective depends on the solution of a physical model on this domain (often a PDE). Typical goals include minimizing drag, maximizing lift, or minimizing compliance (maximizing stiffness) under a volume constraint. After discretization of both the geometry and the governing equations, these problems lead to high-dimensional constrained optimization with strong links to sensitivity analysis. Applications include aerodynamic design, lightweight structures, and topology optimization in mechanical engineering.

Optimal transport. *Optimal transport* studies how to transform one distribution into another at minimum cost. In the discrete setting, given masses (a_i) and (b_j) and a transportation cost matrix (c_{ij}), one seeks a coupling (π_{ij}) minimizing $\sum_{i,j} c_{ij}\pi_{ij}$ under marginal constraints $\sum_j \pi_{ij} = a_i$ and $\sum_i \pi_{ij} = b_j$; this is a linear program. Beyond its classical logistics interpretation, optimal transport defines distances between probability measures (Wasserstein distances) that are used in statistics and machine learning (e.g. for comparing datasets, domain adaptation, or generative modeling).

Stochastic programming. In *stochastic programming*, the data depend on a random variable (demand, prices, renewable production, ...), and one aims at decisions that perform well on average. A standard template is a two-stage model: a first-stage decision is taken before uncertainty is revealed, then a second-stage *recourse* decision adapts after observing the outcome, and the objective involves an expectation. In practice, expectations are often approximated by sample averages (scenario models), leading to large but finite-dimensional optimization problems. Applications include supply-chain planning and energy procurement under uncertain demand or renewable generation.

Robust optimization. *Robust optimization* takes a worst-case viewpoint: uncertain data are assumed to lie in a prescribed uncertainty set, and one seeks decisions that remain feasible for all realizations while minimizing the worst-case objective. For many uncertainty sets, robust

counterparts of linear or convex models remain tractable and provide explicit guarantees. Typical use cases include production planning with uncertain demands, portfolio selection with uncertain returns, and safe engineering design under parameter uncertainty.

Game theory. When several decision-makers interact, optimization naturally leads to *equilibrium* concepts. In a zero-sum game, one player minimizes a payoff while the other maximizes it, leading to a min–max problem closely connected to duality and saddle points. In general (nonzero-sum) settings, one looks for Nash equilibria, where each player is optimal given the others. Applications include market models, congestion/traffic equilibria, and adversarial formulations in learning.

1.1.4 ✠ A short history of optimization

Optimization problems were studied long before the field was identified as a discipline. In the late 17th century (1690s), the calculus of variations emerged from questions such as the brachistochrone. In the 18th century (1780s), Monge formulated the transportation problem.

Optimization (often called *mathematical programming*) developed rapidly in the 20th century, driven by applications and the rise of computing. Linear programming and the simplex method appeared in the 1940s, and first-order optimality conditions for constrained nonlinear problems were formalized in the early 1950s (KKT conditions). In the 1960s–1970s, convex analysis and duality became a unifying language: convexity makes local optimality global, and Lagrangian/Fenchel duality yields bounds and optimality certificates.

By the 1960s, space guidance and navigation pushed optimal control and optimal estimation (LQR/Kalman filtering, later LQG) into high-stakes large-scale engineering. From the late 1970s to the 1980s, practical continuous optimization was shaped by robust numerical methods: trust-region and line-search globalization for Newton-type methods, and quasi-Newton updates (notably BFGS, widely adopted in the 1980s) made smooth unconstrained optimization reliable at scale.

In the 1980s (especially after 1984), interior-point methods provided an alternative to simplex for LP, with strong polynomial-time theory and excellent performance on many large structured problems. On the discrete side, branch-and-bound and cutting planes matured into *branch-and-cut*; combined with better presolve, cutting strategies, and computing power, MILP solvers made a qualitative jump in the late 1990s, turning many integer models into routinely solvable tools for practice.

Algorithm design remains active well beyond the "classical" era. In the 2000s, first-order methods for nonsmooth convex problems (proximal and splitting methods, accelerated gradients) became mainstream for large-scale learning and inverse problems. In the 2010s, stochastic-gradient methods (mini-batch SGD and variants) underpinned deep learning at massive scale, while modern conic solvers and renewed interest in nonconvex optimization expanded the frontier. Across these developments, the same core ideas keep recurring in new forms: geometry/convexity, duality and certificates, and scalable algorithms that exploit structure.

1.2 Practice and tools in optimization

Solving an optimization problem in practice is an iterative loop between modeling, computation, and evaluation. It is rarely a linear pipeline: an unsatisfactory solution may mean going back to the model (variables/constraints/objective), revisiting assumptions, or collecting better data. In particular, it is very rare to have a closed-form solution to an optimization problem; optimization is mostly about designing and using numerical algorithms that compute (often approximate) solutions reliably.

1.2.1 Modeling

Modeling an optimization problem is a delicate task. It consists of (i) identifying decision variables (those we can act upon, as opposed to parameters that are given data), (ii) making constraints explicit, which may be *hard* (must be satisfied, typically physical constraints) or *soft* (may be violated but at a cost, typically economic constraints), and (iii) defining the objective criterion, i.e. a function measuring the quality of a solution.

This step often requires iterations between engineers and domain experts to refine the model. Common pitfalls are well-known. A first one is to confuse *prior beliefs* about what a "reasonable" solution should look like with actual constraints that must be enforced. A second one is to forget constraints altogether, which can yield unrealistic solutions and then trigger a long sequence of ad-hoc constraint additions. A third one is to use an objective that does not faithfully capture the intended trade-offs, especially when several criteria must be balanced.

1.2.2 Formalization

Formalization turns a real-world question into a precise mathematical statement by specifying decision variables, constraints, and an objective function.

It is useful to first write the formal problem on paper or in LaTeX, then translate it into code. This is typically done using a *modeling language* (or *modeler*) to express the problem in a concise, human-readable form that can be passed to an optimization solver. Examples include Pyomo or PuLP in Python, JuMP in Julia, or modeling languages such as AMPL or GAMS.

1.2.3 Tools (modeler, solver, post-processing)

A typical optimization toolchain combines four complementary elements.

Modeler. A modeling language (e.g. Pyomo, JuMP, AMPL) expresses variables, constraints, and objectives in a way that stays close to the mathematical formulation, while handling indices, data, and common modeling patterns.

Solver. The solver (e.g. Gurobi, MOSEK) is the numerical engine that takes a model in a supported class (LP, QP, SOCP, MILP, ...) and computes a solution together with diagnostic information (termination status, feasibility/optimality certificates when available). Commercial solvers are often expensive, although free academic licenses (including for students) are widely available.

Post-processing. Once a solution is produced, post-processing tools (often in Python or Julia) are used to interpret the output: check residuals and constraints, compute derived quantities, perform sensitivity analyses, and visualize results.

Programming language. Finally, a general-purpose programming language orchestrates the whole workflow: reading data, building and solving a model repeatedly (e.g. in a simulation or parameter sweep), and integrating optimization as a component of a larger pipeline.

1.2.4 Method selection

Depending on the optimization problem type and the goal, one can use very diverse approaches. To keep things simple, one can distinguish methods that aim at producing a good solution efficiently (often local in nonconvex problems), and methods that provide guarantees or certificates (e.g. global optimality in convex optimization, or proven optimality for LP/MILP via duality and branch-and-bound).

Even though this course (and others) will present various solution algorithms, it is highly recommended not to implement them from scratch, but to rely on existing libraries. Indeed,

theoretical algorithms are often not efficient in practice and require many safeguards and improvements to become usable.

In a nutshell, it is usually best to start by writing the optimization problem clearly (on paper or in LaTeX), then code it in a modeler, and rely on a solver rather than re-implementing algorithms.

1.2.5 Numerical evaluation

Once the problem is formalized and a method is selected, it must be implemented to compute a solution. Real-world problems are often too complex to be solved exactly, or even approximated efficiently. In some cases, approximations are needed (e.g. discretizing a continuous problem, neglecting some constraints, or ignoring some variables).

After a solution is obtained, it is important to evaluate it numerically and verify that it is acceptable (including with respect to what may have been neglected in the modeling). This typically includes checking feasibility residuals, solver termination status and tolerances, and, when available, optimality measures or certificates (e.g. KKT residuals, duality gaps). It is also useful to assess the sensitivity of the solution to changes in the data or to small modifications of the objective function. In data-driven settings, one should also evaluate the solution out-of-sample to avoid optimizing for noise. If one wants solutions that remain valid under data perturbations, one may turn to *robust optimization*, an active research area.

1.3 ♢ Some elements of complexity theory

1.3.1 Discrete complexity: P, NP, and reductions

The word "efficient" has a precise meaning in theoretical computer science. Roughly speaking, an algorithm is considered efficient if its running time is bounded by a polynomial function of the input size (the number of bits needed to encode the data). Polynomial-time bounds are robust under reasonable changes of computational model, whereas exponential-time bounds typically become prohibitive as the instance size grows.

A *decision problem* is a yes/no question with an input instance (data). Many optimization problems are studied through such associated decision versions. For example, given a minimization problem $\min\{f(x) \mid x \in X\}$ and a threshold B, one can ask: "does there exist $x \in X$ such that $f(x) \leq B$?". This translation is useful because complexity theory is traditionally formulated for decision problems; intuitively, the optimization problem is "hard" whenever its decision version is hard.

In this course, the relevant class is NP, which consists of decision problems whose "yes" answers can be *verified* in polynomial time given a suitable *certificate* (also called a witness). The class P consists of decision problems that we know how to *solve* in polynomial time. One always has $\mathsf{P} \subseteq \mathsf{NP}$, and it is widely believed that $\mathsf{P} \neq \mathsf{NP}$ (this is one of the central open problems in theoretical computer science). The P versus NP question is one of the Clay Mathematics Institute's Millennium Prize Problems. For optimization, this perspective helps set expectations: linear programming has polynomial-time algorithms, while many discrete optimization problems (including standard forms of 0–1 linear programming) are believed not to admit polynomial-time algorithms in the worst case.

The formal way to express "at least as hard as" is through *reductions*. A polynomial-time (many-one) reduction from a decision problem A to a decision problem B is an algorithm that transforms any instance of A into an instance of B in polynomial time, in such a way that the answer is preserved (yes maps to yes, no maps to no). If A reduces to B and B can be solved

in polynomial time, then A can also be solved in polynomial time. A problem is called NP-hard if every problem in NP reduces to it; if it is NP-hard and also belongs to NP, it is called NP-complete.

To make these notions concrete, here are three classical NP-complete problems, stated in elementary terms. *(3-Partition)* Given $3m$ integers and a target sum B, can one partition the integers into m triples, each summing exactly to B? *(Knapsack, decision version)* Given items with weights and values, a capacity W, and a target value V, is there a subset of items of total weight at most W and total value at least V? *(Traveling salesman, decision version)* Given pairwise distances between cities and a bound B, is there a tour visiting every city exactly once (and returning to the start) of total length at most B? Their associated optimization versions (maximize value, minimize tour length, ...) are therefore NP-hard.

1.3.2 Continuous optimization: accuracy and iteration complexity

In continuous optimization, another notion of "complexity" is routinely used: instead of asking for an exact solution, one asks for an ε-*optimal* solution, i.e. a point $\mathbf{x}_\varepsilon$ such that $f(\mathbf{x}_\varepsilon) - f^\star \leq \varepsilon$ (when an optimal value $f^\star$ exists and is finite). For iterative methods, one is then interested in the number of iterations (and the total computational effort) required to reach such an accuracy.

This viewpoint naturally leads to *iteration complexity* results. For instance, for smooth convex minimization, first-order methods such as gradient descent admit rates of the form $f(\mathbf{x}^k) - f^\star = O(1/k)$, and in the strongly convex case one typically obtains linear rates (roughly, an error that decreases geometrically with k). In nonconvex smooth optimization, where global minimizers may be hard to locate, it is common to measure accuracy through *stationarity* (e.g. $\|\nabla f(\mathbf{x}^k)\| \leq \varepsilon$). These results quantify the behavior of algorithms in terms of an accuracy parameter ε, and complement (rather than replace) the discrete complexity viewpoint above.

1.4 Course objectives

The main objective of the course is to provide tools to study and, when possible, solve *differentiable* constrained optimization problems, i.e. problems of the form (1.1) where $X \subseteq \mathbb{R}^n$ is described by finitely many equalities and inequalities involving differentiable functions. More precisely, we focus on problems of the form

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \quad & f(\mathbf{x}) && && \text{(1.7a)} \\ \text{s.t.} \quad & g_i(\mathbf{x}) = 0 && \forall i \in [n_E] && \text{(1.7b)} \\ & h_j(\mathbf{x}) \leq 0 && \forall j \in [n_I] && \text{(1.7c)} \end{aligned}$$

where f, the g_i, and the h_j are differentiable functions from $\mathbb{R}^n$ to $\mathbb{R}$.

The course is organized around four main pillars:

- **Constrained optimization:** Lagrange multipliers and KKT conditions.
- **Convex optimization:** why convexity leads to stronger guarantees and tractable duality.
- **Linear programming:** polyhedra, simplex, and duality.
- **Integer optimization (intro):** mixed-integer linear programming and branch-and-bound (overview).

Chapter 2

Optimization under constraints

In this chapter we study first-order optimality conditions for differentiable constrained problems: Lagrange multipliers for equality constraints and the Karush–Kuhn–Tucker (KKT) conditions for problems with inequalities. We also explain why some form of *constraint qualification* is needed for KKT to hold, and we introduce the Lagrangian viewpoint and weak duality.

Roadmap. We start by fixing the basic vocabulary (extended-value functions, optimal value versus optimal solution) and the topological tools behind existence results. We then introduce tangent cones and constraint qualifications, derive the Lagrange multiplier rule and the KKT conditions, and conclude with a first look at Lagrangian duality.

2.1 Vocabulary and notation

We collect here a few basic notions (notation, optimality, and extended-value functions) that will be used throughout the course. Some of the vocabulary below was introduced informally in Chapter 1; here we give precise definitions.

Notation conventions. For a positive integer k, we write $[k] := \{1, \dots, k\}$. Vectors with all entries equal to 0 (resp. 1) are denoted by $\mathbf{0}$ (resp. $\mathbf{1}$). We define the extended real line by $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$, and we set $\overline{\mathbb{R}}^{\circ} := \mathbb{R} \cup \{+\infty\}$. In minimization, $\overline{\mathbb{R}}^{\circ}$ is the natural codomain for *extended-value* objective functions (the value $+\infty$ encodes infeasibility), while $\overline{\mathbb{R}}$ will occasionally appear for optimal values (infima/suprema) and dual functions.

2.1.1 Optimization problems

Recall from Chapter 1 that an optimization problem consists in minimizing an objective function f over a feasible set $X \subseteq \mathbb{R}^n$:

$$\operatorname*{Min}_{\mathbf{x}} \quad f(\mathbf{x}) \tag{2.1a}$$

$$\text{s.t.} \quad \mathbf{x} \in X \tag{2.1b}$$

In this chapter we allow f to take values in $\overline{\mathbb{R}}^{\circ}$ (see extended-value functions below). We denote the *optimal value* by

$$p^\star := \inf\{f(\mathbf{x}) \mid \mathbf{x} \in X\} \in \overline{\mathbb{R}},$$

and any $\mathbf{x}^\sharp \in X$ such that $f(\mathbf{x}^\sharp) = p^\star$ is an *optimal solution*. The problem is *feasible* if $X \neq \emptyset$.

We will also need local optimality: a point $\mathbf{x}^\sharp \in X$ is a *local minimum of f on X* if there exists a neighborhood V of $\mathbf{x}^\sharp$ in $\mathbb{R}^n$ such that $f(\mathbf{x}^\sharp) \leq f(\mathbf{x})$ for all $\mathbf{x} \in V \cap X$.

2.1.2 Extended-value functions

Allowing f to take the value $+\infty$ is a convenient way to encode constraints. For a function $f : E \to \overline{\mathbb{R}}$ on a vector space E, its *domain* is

$$\operatorname{dom} f := \{x \in E \mid f(x) < +\infty\},$$

and its *sublevel set* at level $M \in \mathbb{R}$ is

$$\operatorname{lev}_M(f) := \{x \in E \mid f(x) \leq M\}.$$

Differentiability convention. When an extended-value function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be differentiable at x, we mean that f is finite on an open neighbourhood of x and differentiable there in the usual sense. The same convention applies to twice differentiable functions. When we say that f is differentiable (or twice differentiable) on its domain, we therefore implicitly assume that its domain is open. For a set $A \subseteq E$, we denote by $\mathbb{I}_A : E \to \overline{\mathbb{R}}$ the indicator function defined by $\mathbb{I}_A(x) = 0$ if $x \in A$ and $\mathbb{I}_A(x) = +\infty$ otherwise.

The indicator function provides a simple way to rewrite constrained problems as unconstrained ones. Indeed, for any set $X \subseteq E$ and any function $f : E \to \overline{\mathbb{R}}$, one has

$$\inf_{x \in X} f(x) = \inf_{x \in E} \big(f(x) + \mathbb{I}_X(x)\big).$$

The right-hand side is an ordinary unconstrained infimum on E, but it enforces the constraint $x \in X$ implicitly since $\mathbb{I}_X(x) = +\infty$ outside X. This viewpoint will be used later to derive the Lagrangian and weak duality by replacing constraints with suitable supremum representations of indicator functions.

2.2 Elementary results on $\mathbb{R}^n$

We start with some results on unconstrained optimization in $\mathbb{R}^n$.

Remark 2.1 (Optimal value versus optimal solution). An optimization problem may have a finite optimal value but no optimal solution. For example,

$$\operatorname*{Min}_{x} \quad \frac{1}{x} \tag{2.2a}$$

$$\text{s.t.} \quad x > 0 \tag{2.2b}$$

has optimal value 0 (as $x \to +\infty$), but the value is not attained by any feasible x.

2.2.1 Existence results

The previous example shows that an optimal solution need not exist, even when the optimal value is finite. We now give a few sufficient conditions for existence of optimal solutions.

We refer to Appendix A, Section A.1 for the basic notions of compactness and continuity, and for the extreme value theorem.

♡ **Theorem 2.1.** *Let C be a nonempty compact subset of $\mathbb{R}^n$. If $f : C \to \mathbb{R}$ is continuous, then*

$$\operatorname*{Min}_{\mathbf{x}} \quad f(\mathbf{x}) \tag{2.3a}$$

$$s.t. \quad \mathbf{x} \in C \tag{2.3b}$$

admits at least one optimal solution.

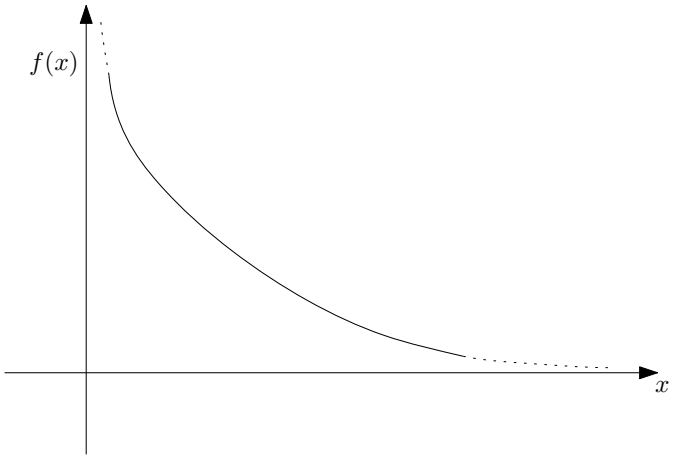

Figure 2.1: A problem may have a finite optimal value but no optimal solution.

Proof. This is an immediate consequence of the extreme value theorem; see Corollary A.4 in Appendix A. □

We say that a function $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is *coercive* if $f(\mathbf{x}) \to +\infty$ as $\|\mathbf{x}\| \to +\infty$.

♡ **Theorem 2.2.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive, then the problem*

$$\operatorname*{Min}_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{2.4a}$$

admits at least one optimal solution.

Proof. Since f is coercive, there exists $r \in \mathbb{R}_+$ such that $\|\mathbf{x}\| \geq r$ implies $f(\mathbf{x}) \geq f(\mathbf{0})$. The global minimum of f restricted to the closed ball centered at $\mathbf{0}$ with radius r exists by Theorem 2.1, and is therefore an optimal solution to the unconstrained problem. □

2.2.2 Unconstrained first- and second-order conditions

If f is differentiable, a *critical point* is a point where the gradient vanishes. An unconstrained local minimum of a differentiable function must be a critical point (a first-order necessary condition). It is not sufficient in general, as illustrated by Figure 2.2.

♡ **Theorem 2.3** (FONC)**.** *Let $\mathbf{x}^\sharp$ be a local minimum of $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. If f is differentiable at $\mathbf{x}^\sharp$, then $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$.*

Proof. Let $\mathbf{d} \in \mathbb{R}^n$. Since $\mathbf{x}^\sharp$ is a local minimum, for all sufficiently small $t > 0$ we have

$$\frac{1}{t}\Big(f(\mathbf{x}^\sharp + t\mathbf{d}) - f(\mathbf{x}^\sharp)\Big) \geq 0,$$

which yields $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0$. Similarly, for all sufficiently small $t > 0$,

$$\frac{1}{t}\Big(f(\mathbf{x}^\sharp) - f(\mathbf{x}^\sharp - t\mathbf{d})\Big) \leq 0,$$

which yields $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \leq 0$. Therefore $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} = 0$ for all $\mathbf{d} \in \mathbb{R}^n$, hence $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$. □

♢ **Theorem 2.4** (SOSC)**.** *Assume $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is twice differentiable on its domain. If $\mathbf{x}^\sharp \in \operatorname{dom} f$ satisfies $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^\sharp)$ is positive definite, then $\mathbf{x}^\sharp$ is a local minimum.*

Proof. Write the second-order Taylor expansion at $\mathbf{x}^\sharp$:

$$f(\mathbf{x}^\sharp + \mathbf{h}) = f(\mathbf{x}^\sharp) + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}^\sharp)\mathbf{h} + o(\|\mathbf{h}\|^2).$$

Since $\nabla^2 f(\mathbf{x}^\sharp)$ is positive definite, there exists $\alpha > 0$ such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^\sharp)\mathbf{h} \geq \alpha\|\mathbf{h}\|^2$ for all $\mathbf{h} \in \mathbb{R}^n$. Choose $\eta > 0$ such that the remainder satisfies $o(\|\mathbf{h}\|^2) \geq -\frac{\alpha}{2}\|\mathbf{h}\|^2$ whenever $\|\mathbf{h}\| \leq \eta$. Then $f(\mathbf{x}^\sharp + \mathbf{h}) \geq f(\mathbf{x}^\sharp)$ for $\|\mathbf{h}\| \leq \eta$, so $\mathbf{x}^\sharp$ is a local minimum. □

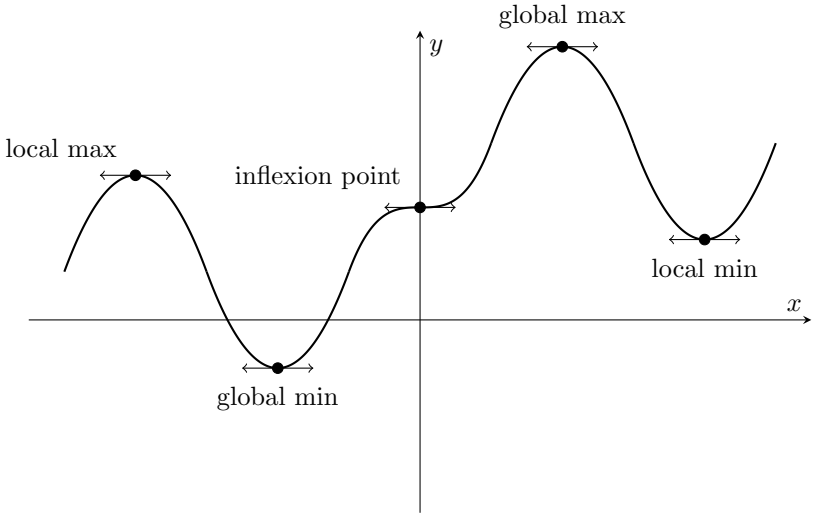

Figure 2.2: Illustration of critical points.

2.3 ♢ Feasible directions and tangent cones

We now return to constrained optimization in a general form:

$$\begin{aligned} \text{Min} \quad & f(\mathbf{x}) && \text{(2.5a)} \\ \text{s.t.} \quad & \mathbf{x} \in X && \text{(2.5b)} \end{aligned}$$

where $X \subseteq \mathbb{R}^n$ is the feasible set and $f : \mathbb{R}^n \to \overline{\mathbb{R}}$.

2.3.1 Existence under constraints

♢ **Theorem 2.5.** *Assume $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is continuous on its domain and coercive, and assume $X \cap \operatorname{dom} f$ is a nonempty closed subset of $\mathbb{R}^n$. Then the constrained problem* (2.5) *admits at least one optimal solution.*

Proof. Pick any $\mathbf{x}_0 \in X \cap \operatorname{dom} f$. Since f is coercive, there exists $r > 0$ such that $\|\mathbf{x}\| \geq r$ implies $f(\mathbf{x}) \geq f(\mathbf{x}_0)$. By restricting f to the nonempty compact set $(X \cap \operatorname{dom} f) \cap \bar{B}(\mathbf{0}, r)$ and applying Theorem 2.1, we obtain an optimal solution. □

2.3.2 Tangent cone

♢ **Definition 2.6** (Feasible direction and tangent cone)**.** *Let $X \subseteq \mathbb{R}^n$ and let $\mathbf{x} \in X$. A vector $\mathbf{d} \in \mathbb{R}^n$ is a* feasible direction *at $\mathbf{x}$ if there exist $t_k \searrow 0$ and $\mathbf{d}^k \to \mathbf{d}$ such that $\mathbf{x} + t_k \mathbf{d}^k \in X$ for all k. The set of all feasible directions at $\mathbf{x}$ is denoted by $T_X(\mathbf{x})$ and is called the* tangent cone *of X at $\mathbf{x}$:*

$$T_X(\mathbf{x}) := \{\mathbf{d} \in \mathbb{R}^n \mid \exists t_k \searrow 0,\ \exists \mathbf{d}^k \to \mathbf{d} \text{ such that } \mathbf{x} + t_k \mathbf{d}^k \in X\ \forall k\}.$$

✠ **Lemma 2.7.** *For any $X \subseteq \mathbb{R}^n$ and any $\mathbf{x} \in X$, the tangent cone $T_X(\mathbf{x})$ is closed.*

Proof. Let $(\mathbf{d}^k)_k$ be a sequence of feasible directions at $\mathbf{x}$ converging to $\mathbf{d}$. By definition, for each k there is a sequence witnessing that $\mathbf{d}^k$ is feasible. Choose $\widehat{\mathbf{d}}^k$ and t_k from that sequence recursively so that

$$\|\widehat{\mathbf{d}}^k - \mathbf{d}^k\| < 1/k, \qquad t_k < \begin{cases} 1, & k = 1, \\ \min\{1/k, t_{k-1}/2\}, & k \geq 2, \end{cases}$$

and $\mathbf{x} + t_k \widehat{\mathbf{d}}^k \in X$. Then $\widehat{\mathbf{d}}^k \to \mathbf{d}$ and $t_k \searrow 0$, so $\mathbf{d}$ is feasible. □

2.3.3 A first-order necessary condition

♢ **Theorem 2.8.** *Let $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and let $\mathbf{x}^\sharp$ be a local minimum of f on X. If f is differentiable at $\mathbf{x}^\sharp$, then*

$$\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0 \qquad \text{for all } \mathbf{d} \in T_X(\mathbf{x}^\sharp).$$

Proof. Let $\mathbf{d} \in T_X(\mathbf{x}^\sharp)$ and choose sequences (t_k) and $(\mathbf{d}^k)$ from the definition, so that $\mathbf{x}^\sharp + t_k \mathbf{d}^k \in X$, $t_k > 0$, $t_k \to 0$ and $\mathbf{d}^k \to \mathbf{d}$. Since $\mathbf{x}^\sharp$ is a local minimum, for k large enough we have $f(\mathbf{x}^\sharp + t_k \mathbf{d}^k) \geq f(\mathbf{x}^\sharp)$. Using differentiability at $\mathbf{x}^\sharp$,

$$f(\mathbf{x}^\sharp + t_k \mathbf{d}^k) = f(\mathbf{x}^\sharp) + t_k \, \nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d}^k + o(t_k \|\mathbf{d}^k\|).$$

Dividing by t_k and letting $k \to \infty$ yields $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0$. □

2.4 Differentiable constraints and constraint qualifications

We now focus on problems described by finitely many differentiable equalities and inequalities:

$$\begin{aligned}
\underset{\mathbf{x} \in \mathbb{R}^n}{\text{Min}} \quad & f(\mathbf{x}) && && \text{(2.6a)}\\
\text{s.t.} \quad & g_i(\mathbf{x}) = 0 && \forall i \in [n_E] && \text{(2.6b)}\\
& h_j(\mathbf{x}) \leq 0 && \forall j \in [n_I] && \text{(2.6c)}
\end{aligned}$$

The feasible set is

$$X := \{\mathbf{x} \in \mathbb{R}^n \mid g_i(\mathbf{x}) = 0,\ \forall i \in [n_E],\ h_j(\mathbf{x}) \leq 0,\ \forall j \in [n_I]\}.$$

2.4.1 Linearized tangent cone

Definition 2.9 (Active set)**.** *For $\mathbf{x} \in X$, the set of active inequality constraints is*

$$J_0(\mathbf{x}) := \{j \in [n_I] \mid h_j(\mathbf{x}) = 0\}.$$

♢ **Definition 2.10** (Linearized tangent cone)**.** *Assume g_i and h_j are differentiable at $\mathbf{x} \in X$. The* linearized tangent cone *at $\mathbf{x}$ is*

$$T_X^\ell(\mathbf{x}) := \Big\{\mathbf{d} \in \mathbb{R}^n \mid \nabla g_i(\mathbf{x}) \cdot \mathbf{d} = 0,\ \forall i \in [n_E],\ \nabla h_j(\mathbf{x}) \cdot \mathbf{d} \leq 0,\ \forall j \in J_0(\mathbf{x})\Big\}.$$

One always has $T_X(\mathbf{x}) \subseteq T_X^\ell(\mathbf{x})$:

♢ **Lemma 2.11.** *Assume g_i and h_j are differentiable at a feasible point $\mathbf{x} \in X$. Then $T_X(\mathbf{x}) \subseteq T_X^\ell(\mathbf{x})$.*

Proof. Let $\mathbf{d} \in T_X(\mathbf{x})$ and pick sequences $t_k \searrow 0$ and $\mathbf{d}^k \to \mathbf{d}$ such that $\mathbf{x} + t_k \mathbf{d}^k \in X$ for all k. For each equality constraint $g_i(\mathbf{x} + t_k \mathbf{d}^k) = 0$ and $g_i(\mathbf{x}) = 0$, so by differentiability at $\mathbf{x}$,

$$0 = \frac{g_i(\mathbf{x} + t_k \mathbf{d}^k) - g_i(\mathbf{x})}{t_k} = \nabla g_i(\mathbf{x}) \cdot \mathbf{d}^k + \frac{o(\|t_k \mathbf{d}^k\|)}{t_k}.$$

Since $\mathbf{d}^k \to \mathbf{d}$, the sequence $(\mathbf{d}^k)$ is bounded, so $\|t_k \mathbf{d}^k\|/t_k = \|\mathbf{d}^k\|$ stays bounded and $\frac{o(\|t_k \mathbf{d}^k\|)}{t_k} \to 0$. Letting $k \to \infty$ yields $\nabla g_i(\mathbf{x}) \cdot \mathbf{d} = 0$. Similarly, for each active inequality constraint $j \in J_0(\mathbf{x})$ we have $h_j(\mathbf{x} + t_k \mathbf{d}^k) \leq 0$ and $h_j(\mathbf{x}) = 0$, hence

$$0 \geq \frac{h_j(\mathbf{x} + t_k \mathbf{d}^k) - h_j(\mathbf{x})}{t_k} = \nabla h_j(\mathbf{x}) \cdot \mathbf{d}^k + \frac{o(\|t_k \mathbf{d}^k\|)}{t_k},$$

which yields $\nabla h_j(\mathbf{x}) \cdot \mathbf{d} \leq 0$. Therefore $\mathbf{d} \in T_X^\ell(\mathbf{x})$. □

The true tangent cone may be difficult to compute, whereas the linearized cone is described explicitly by the gradients of the constraints. The possible difficulty is that linearization may introduce spurious first-order directions, so the inclusion may be strict. Constraint qualification is precisely the condition that no such directions remain.

♢ **Definition 2.12** (Constraint qualification)**.** *We say that the constraints are* qualified *at a feasible point* $\mathbf{x} \in X$ *if*

$$T_X(\mathbf{x}) = T_X^\ell(\mathbf{x}).$$

Example 2.1 (A constraint that is not qualified). Consider $\min\{x \mid x^3 \geq 0\}$ on $\mathbb{R}$. The unique minimizer is $x^\sharp = 0$, but the constraint is not qualified at 0: one can check that $T_X(0) = \mathbb{R}_+$ while $T_X^\ell(0) = \mathbb{R}$.

2.4.2 Sufficient qualification conditions

Checking $T_X(\mathbf{x}) = T_X^\ell(\mathbf{x})$ directly is often difficult. Here are standard sufficient conditions.

♡ **Proposition 2.13.** *If all the functions g_i and h_j are affine, then the constraints of* (2.6) *are qualified at every feasible point.*

Proof. Let $\mathbf{x} \in X$ and $\mathbf{d} \in T_X^\ell(\mathbf{x})$. Since the equalities are affine, $g_i(\mathbf{x} + t\mathbf{d}) = 0$ for all $t \in \mathbb{R}$. Write each inequality as $h_j(\mathbf{x}) = a_j^\top \mathbf{x} + b_j$. If $a_j^\top \mathbf{d} > 0$, then necessarily $h_j(\mathbf{x}) < 0$ (otherwise $\mathbf{d} \notin T_X^\ell(\mathbf{x})$). For each $j \in [n_I]$, define a radius $\varepsilon_j > 0$ as follows:

- if $a_j^\top \mathbf{d} \leq 0$, set $\varepsilon_j := 1$ (then $h_j(\mathbf{x} + t\mathbf{d}) = h_j(\mathbf{x}) + t\, a_j^\top \mathbf{d} \leq h_j(\mathbf{x}) \leq 0$ for all $t \in [0, 1]$);
- if $a_j^\top \mathbf{d} > 0$ (hence $h_j(\mathbf{x}) < 0$), set $\varepsilon_j := -\dfrac{h_j(\mathbf{x})}{a_j^\top \mathbf{d}} > 0$ (then $h_j(\mathbf{x} + t\mathbf{d}) \leq 0$ for all $t \in [0, \varepsilon_j]$).

Since there are finitely many constraints, letting $\varepsilon := \min_{j \in [n_I]} \varepsilon_j$ gives $\varepsilon > 0$ and $\mathbf{x} + t\mathbf{d} \in X$ for all $t \in [0, \varepsilon]$. Thus $\mathbf{d} \in T_X(\mathbf{x})$ and $T_X^\ell(\mathbf{x}) \subseteq T_X(\mathbf{x})$. □

♡ **Proposition 2.14** (Slater's condition)**.** *Assume g_i are affine and h_j are convex and differentiable. If there exists $\mathbf{x}_S \in \mathbb{R}^n$ such that $g_i(\mathbf{x}_S) = 0$ for all i and $h_j(\mathbf{x}_S) < 0$ for all j (strict feasibility), then the constraints of* (2.6) *are qualified at every feasible point.*

Proof. Fix a feasible point $\mathbf{x} \in X$. We already know $T_X(\mathbf{x}) \subseteq T_X^\ell(\mathbf{x})$ (Lemma 2.11); we prove the reverse inclusion.

Let $\mathbf{d} \in T_X^\ell(\mathbf{x})$ and define $s := \mathbf{x}_S - \mathbf{x}$. Since the g_i are affine and $g(\mathbf{x}_S) = g(\mathbf{x}) = 0$, one has $\nabla g_i(\mathbf{x}) \cdot s = 0$ for all i. Moreover, for any active inequality constraint $j \in J_0(\mathbf{x})$, convexity of h_j gives

$$h_j(\mathbf{x}_S) \geq h_j(\mathbf{x}) + \nabla h_j(\mathbf{x})^\top (\mathbf{x}_S - \mathbf{x}) = \nabla h_j(\mathbf{x})^\top s,$$

so $\nabla h_j(\mathbf{x})^\top s < 0$ because $h_j(\mathbf{x}) = 0$ and $h_j(\mathbf{x}_S) < 0$.

For $\varepsilon > 0$, set $\mathbf{d}_\varepsilon := \mathbf{d} + \varepsilon s$. Then $\nabla g_i(\mathbf{x}) \cdot \mathbf{d}_\varepsilon = 0$ for all i, and for $j \in J_0(\mathbf{x})$ we have

$$\nabla h_j(\mathbf{x})^\top \mathbf{d}_\varepsilon = \nabla h_j(\mathbf{x})^\top \mathbf{d} + \varepsilon \nabla h_j(\mathbf{x})^\top s < 0,$$

since $\nabla h_j(\mathbf{x})^\top \mathbf{d} \leq 0$ by $\mathbf{d} \in T_X^\ell(\mathbf{x})$. By differentiability of h_j at $\mathbf{x}$, this strict inequality implies that $h_j(\mathbf{x} + t\mathbf{d}_\varepsilon) < 0$ for all sufficiently small $t > 0$. For inactive constraints $h_j(\mathbf{x}) < 0$, feasibility of $\mathbf{x} + t\mathbf{d}_\varepsilon$ holds for small $t > 0$ by continuity. Finally, because the g_i are affine and $\nabla g_i(\mathbf{x}) \cdot \mathbf{d}_\varepsilon = 0$, one has $g_i(\mathbf{x} + t\mathbf{d}_\varepsilon) = 0$ for all t. Therefore $\mathbf{x} + t\mathbf{d}_\varepsilon \in X$ for all $t \in (0, t_\varepsilon]$, so $\mathbf{d}_\varepsilon \in T_X(\mathbf{x})$.

Since $T_X(\mathbf{x})$ is closed (Lemma 2.7) and $\mathbf{d}_\varepsilon \to \mathbf{d}$ as $\varepsilon \downarrow 0$, we conclude that $\mathbf{d} \in T_X(\mathbf{x})$. Thus $T_X^\ell(\mathbf{x}) \subseteq T_X(\mathbf{x})$ and qualification holds. □

✠ **Proposition 2.15** (Mangasarian–Fromovitz constraint qualification (MFCQ))**.** *Assume g_i and h_j are continuously differentiable and let $\mathbf{x}^\sharp \in X$. Assume that the family $(\nabla g_i(\mathbf{x}^\sharp))_{i \in [n_E]}$ is linearly independent. If there exists $\widehat{\mathbf{d}} \in \mathbb{R}^n$ such that*

$$\nabla g_i(\mathbf{x}^\sharp) \cdot \widehat{\mathbf{d}} = 0, \ \forall i \in [n_E], \qquad \nabla h_j(\mathbf{x}^\sharp) \cdot \widehat{\mathbf{d}} < 0, \ \forall j \in J_0(\mathbf{x}^\sharp),$$

then the constraints of (2.6) *are qualified at $\mathbf{x}^\sharp$.*

Proof. (Proof sketch.) We already know $T_X(\mathbf{x}^\sharp) \subseteq T_X^\ell(\mathbf{x}^\sharp)$ (Lemma 2.11). We prove the reverse inclusion.

Let $\mathbf{d} \in T_X^\ell(\mathbf{x}^\sharp)$ and, for $\varepsilon > 0$, set $\mathbf{d}_\varepsilon := \mathbf{d} + \varepsilon\widehat{\mathbf{d}}$. Then $\nabla g_i(\mathbf{x}^\sharp) \cdot \mathbf{d}_\varepsilon = 0$ for all i, and

$$\nabla h_j(\mathbf{x}^\sharp) \cdot \mathbf{d}_\varepsilon = \nabla h_j(\mathbf{x}^\sharp) \cdot \mathbf{d} + \varepsilon\, \nabla h_j(\mathbf{x}^\sharp) \cdot \widehat{\mathbf{d}} < 0 \qquad \forall j \in J_0(\mathbf{x}^\sharp).$$

By differentiability, this implies $h_j(\mathbf{x}^\sharp + t\mathbf{d}_\varepsilon) < 0$ for all active j and all sufficiently small $t > 0$; inactive inequalities remain satisfied by continuity.

It remains to satisfy the equalities exactly. Since the gradients $(\nabla g_i(\mathbf{x}^\sharp))_{i\in[n_E]}$ are linearly independent, the Jacobian $Dg(\mathbf{x}^\sharp)$ has full row rank, and the set $\{\mathbf{x} \mid g(\mathbf{x}) = 0\}$ is a C^1 manifold near $\mathbf{x}^\sharp$ (implicit function theorem). Moreover, every direction $\mathbf{d}_\varepsilon$ with $Dg(\mathbf{x}^\sharp)\mathbf{d}_\varepsilon = 0$ is a tangent direction to this manifold: there exists a C^1 curve $\gamma_\varepsilon(t)$ such that $\gamma_\varepsilon(0) = \mathbf{x}^\sharp$, $g(\gamma_\varepsilon(t)) = 0$, and $\gamma_\varepsilon'(0) = \mathbf{d}_\varepsilon$. Replacing $\mathbf{x}^\sharp + t\mathbf{d}_\varepsilon$ with $\gamma_\varepsilon(t) = \mathbf{x}^\sharp + t\mathbf{d}_\varepsilon + o(t)$ preserves the strict inequalities for small t, hence $\gamma_\varepsilon(t) \in X$ for $t \in (0, t_\varepsilon]$. Therefore $\mathbf{d}_\varepsilon \in T_X(\mathbf{x}^\sharp)$.

Finally, since $T_X(\mathbf{x}^\sharp)$ is closed and $\mathbf{d}_\varepsilon \to \mathbf{d}$ as $\varepsilon \downarrow 0$, we obtain $\mathbf{d} \in T_X(\mathbf{x}^\sharp)$. Thus $T_X^\ell(\mathbf{x}^\sharp) \subseteq T_X(\mathbf{x}^\sharp)$ and qualification holds. □

Corollary 2.16 (LICQ implies qualification)**.** *Assume g_i and h_j are continuously differentiable and let $\mathbf{x}^\sharp \in X$. If the family*

$$\{\nabla g_i(\mathbf{x}^\sharp)\}_{i\in[n_E]} \cup \{\nabla h_j(\mathbf{x}^\sharp)\}_{j\in J_0(\mathbf{x}^\sharp)}$$

is linearly independent (LICQ), then the constraints of (2.6) *are qualified at $\mathbf{x}^\sharp$.*

Proof. Let $J_0(\mathbf{x}^\sharp) = \{j_1, \dots, j_m\}$. Consider the matrix $A \in \mathbb{R}^{(n_E+m)\times n}$ whose rows are $\nabla g_1(\mathbf{x}^\sharp)^\top$, ...,$\nabla g_{n_E}(\mathbf{x}^\sharp)^\top, \nabla h_{j_1}(\mathbf{x}^\sharp)^\top, \dots,$ LICQ means that A has full row rank, hence the linear map $\mathbf{d} \mapsto A\mathbf{d}$ is surjective onto $\mathbb{R}^{n_E+m}$. Therefore there exists $\widehat{\mathbf{d}} \in \mathbb{R}^n$ such that

$$A\widehat{\mathbf{d}} = (\underbrace{0, \dots, 0}_{n_E \text{ entries}}, \underbrace{-1, \dots, -1}_{m \text{ entries}})^\top.$$

In particular, $\nabla g_i(\mathbf{x}^\sharp) \cdot \widehat{\mathbf{d}} = 0$ for all i and $\nabla h_j(\mathbf{x}^\sharp) \cdot \widehat{\mathbf{d}} < 0$ for all $j \in J_0(\mathbf{x}^\sharp)$, so MFCQ holds at $\mathbf{x}^\sharp$ and Proposition 2.15 yields qualification. □

2.5 Equality constraints: Lagrange multipliers

For pedagogy, we first consider equality constraints only:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) \tag{2.7a}$$

$$\text{s.t.} \quad g_i(\mathbf{x}) = 0 \qquad \forall i \in [n_E] \tag{2.7b}$$

Definition 2.17 (Lagrangian (equality constraints))**.** *Define $g(\mathbf{x}) := (g_1(\mathbf{x}), \dots, g_{n_E}(\mathbf{x})) \in \mathbb{R}^{n_E}$. The* Lagrangian *is*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := f(\mathbf{x}) + \boldsymbol{\lambda}^\top g(\mathbf{x}), \qquad \boldsymbol{\lambda} \in \mathbb{R}^{n_E}.$$

♢ **Lemma 2.18.** *For any matrix M, one has* $\mathrm{Im}(M) = (\mathrm{Ker}(M^\top))^\perp$.

♡ **Theorem 2.19** (Lagrange multipliers, equality case)**.** *Assume f and g_i are differentiable and let $\mathbf{x}^\sharp$ be a local minimum of f on $X = \{\mathbf{x} \mid g_i(\mathbf{x}) = 0,\ i \in [n_E]\}$. If the constraints are qualified at $\mathbf{x}^\sharp$ (in particular, this holds if $(\nabla g_i(\mathbf{x}^\sharp))_{i\in[n_E]}$ is linearly independent), then there exists $\boldsymbol{\lambda}^\sharp \in \mathbb{R}^{n_E}$ such that*

$$\nabla f(\mathbf{x}^\sharp) + \sum_{i=1}^{n_E} \lambda_i^\sharp \nabla g_i(\mathbf{x}^\sharp) = \mathbf{0}, \qquad g(\mathbf{x}^\sharp) = 0.$$

Proof. By Theorem 2.8, $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0$ for all $\mathbf{d} \in T_X(\mathbf{x}^\sharp)$. For equality constraints, $T_X^\ell(\mathbf{x}^\sharp) = \{\mathbf{d} \mid \nabla g_i(\mathbf{x}^\sharp) \cdot \mathbf{d} = 0,\ \forall i \in [n_E]\}$, and qualification gives $T_X(\mathbf{x}^\sharp) = T_X^\ell(\mathbf{x}^\sharp)$. Thus $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0$ for all $\mathbf{d}$ in a linear subspace, which forces $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} = 0$ for all $\mathbf{d}$ in that subspace. Let G be the Jacobian matrix whose rows are $\nabla g_i(\mathbf{x}^\sharp)^\top$. The subspace above is $\mathrm{Ker}(G)$, so $\nabla f(\mathbf{x}^\sharp) \in (\mathrm{Ker}(G))^\perp = \mathrm{Im}(G^\top)$ by Lemma 2.18. Therefore $\nabla f(\mathbf{x}^\sharp) = -G^\top \boldsymbol{\lambda}^\sharp$ for some $\boldsymbol{\lambda}^\sharp$, which is the desired relation. □

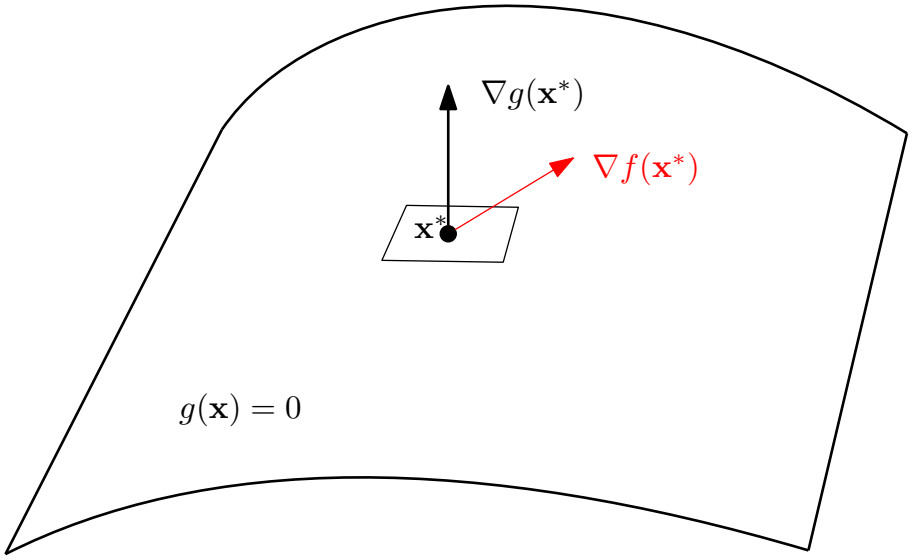

Figure 2.3: If $\mathbf{x}^\sharp$ is not a constrained local minimum, the gradient $\nabla f(\mathbf{x}^\sharp)$ may have a component tangent to the constraint manifold, yielding a feasible direction that decreases the objective.

Example 2.2 (A simple equality constraint). Minimize $\frac{1}{2}\|\mathbf{x}\|^2$ subject to $a^\top \mathbf{x} = 1$ with $a \neq 0$. The Lagrangian is $\mathcal{L}(\mathbf{x}, \lambda) = \frac{1}{2}\|\mathbf{x}\|^2 + \lambda(a^\top \mathbf{x} - 1)$. Stationarity gives $\mathbf{x} + \lambda a = 0$, hence $\mathbf{x} = -\lambda a$. Using $a^\top \mathbf{x} = 1$, we obtain

$$-\lambda \|a\|^2 = 1, \qquad \lambda = -\frac{1}{\|a\|^2},$$

and therefore $\mathbf{x}^\sharp = \dfrac{a}{\|a\|^2}$.

Example 2.3 (A constrained polynomial on the sphere). Consider $\min\{x^3 + y^3 \mid x^2 + y^2 = 1\}$. At an optimal solution $(x^\sharp, y^\sharp)$ there exists $\lambda \in \mathbb{R}$ such that

$$3x^{\sharp 2} + 2\lambda x^\sharp = 0, \qquad 3y^{\sharp 2} + 2\lambda y^\sharp = 0, \qquad x^{\sharp 2} + y^{\sharp 2} = 1.$$

This yields finitely many candidates (e.g. $(\pm 1, 0)$, $(0, \pm 1)$, $\pm\frac{1}{\sqrt{2}}(1, 1)$), among which the minimum can be selected by evaluation.

Example 2.4 (Connected vessels and equal water levels). Consider n vertical vessels connected at their bases. Let $a_k(y) > 0$ be the cross-sectional area of vessel k at height y. If its water level is x_k, then its volume is

$$V_k(x_k) = \int_0^{x_k} a_k(y)\, dy,$$

and, up to a common positive physical constant, its gravitational potential energy is

$$E_k(x_k) = \int_0^{x_k} y\, a_k(y)\, dy.$$

For a fixed total volume $V > 0$, equilibrium is modeled by

$$\begin{aligned} \operatorname*{Min}_{x_1,\dots,x_n \geq 0} \quad & \sum_{k=1}^{n} \int_0^{x_k} y\, a_k(y)\, dy \\ \text{s.t.} \quad & \sum_{k=1}^{n} \int_0^{x_k} a_k(y)\, dy = V. \end{aligned}$$

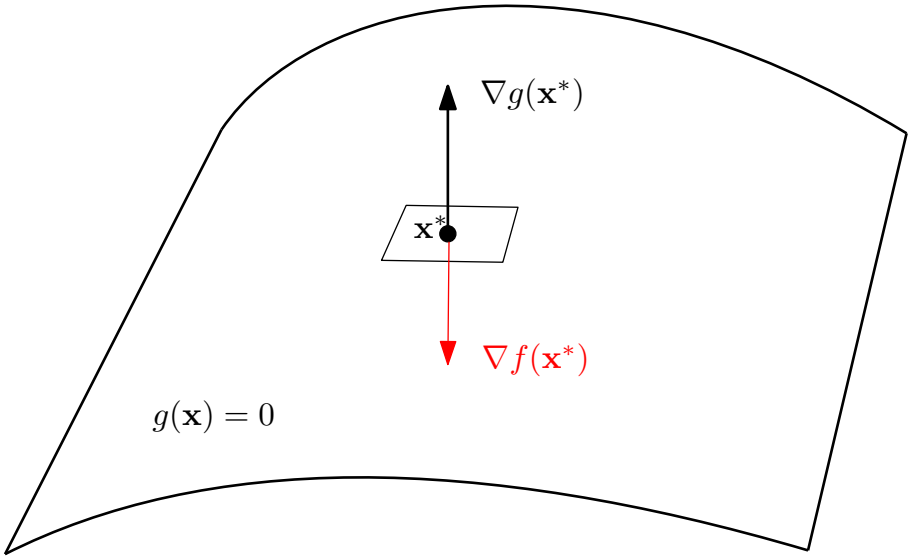

Figure 2.4: At a constrained local minimum (under qualification), $\nabla f(\mathbf{x}^\sharp)$ belongs to the span of the constraint gradients $\{\nabla g_i(\mathbf{x}^\sharp)\}$, i.e. it is orthogonal to the tangent directions.

Assume that the vessels are sufficiently tall, the functions are regular enough, and the equilibrium has $x_k > 0$ for every k. With $g(\mathbf{x}) = \sum_k V_k(x_k) - V$, the constraint is qualified because $\nabla g(\mathbf{x}) = (a_1(x_1), \dots, a_n(x_n)) \neq \mathbf{0}$. The Lagrangian is $\mathcal{L}(\mathbf{x}, \lambda) = \sum_k E_k(x_k) + \lambda g(\mathbf{x})$. By the fundamental theorem of calculus, stationarity gives

$$x_k a_k(x_k) + \lambda a_k(x_k) = 0 \qquad \forall k.$$

Since $a_k(x_k) > 0$, we obtain

$$x_k = -\lambda \qquad \forall k.$$

Thus all connected vessels have the same equilibrium water level, independently of their shapes. The multiplier is minus this common level.

2.6 ♡ Karush–Kuhn–Tucker conditions

We return to the general constrained problem (2.6). We have the following first order necessary condition, known as the Karush–Kuhn–Tucker (KKT) theorem.

♡ **Theorem 2.20** (KKT theorem (first-order necessary condition))**.** *Assume that f, g_i and h_j are differentiable and let $\mathbf{x}^\sharp \in X \cap \operatorname{dom} f$ be a local minimum of f on X. If the constraints are qualified at $\mathbf{x}^\sharp$, then there exist multipliers $\boldsymbol{\lambda}^\sharp \in \mathbb{R}^{n_E}$ and $\boldsymbol{\mu}^\sharp \in \mathbb{R}^{n_I}_+$ such that*

(stationarity)	$\nabla f(\mathbf{x}^\sharp) + \sum_{i=1}^{n_E} \lambda_i^\sharp \nabla g_i(\mathbf{x}^\sharp) + \sum_{j=1}^{n_I} \mu_j^\sharp \nabla h_j(\mathbf{x}^\sharp) = \mathbf{0},$	
(primal feasibility)	$g_i(\mathbf{x}^\sharp) = 0, \ \forall i \in [n_E],$	$h_j(\mathbf{x}^\sharp) \leq 0, \ \forall j \in [n_I],$
(dual feasibility)	$\boldsymbol{\mu}^\sharp \geq 0,$	
(complementary slackness)	$\mu_j^\sharp \, h_j(\mathbf{x}^\sharp) = 0, \ \forall j \in [n_I].$	

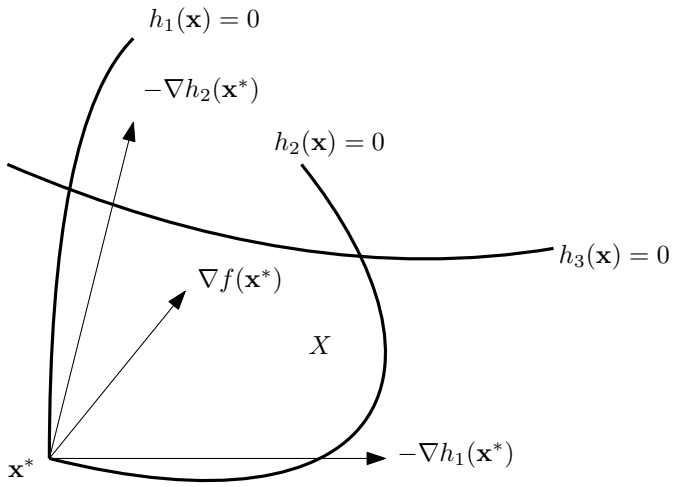

Figure 2.5: Geometric interpretation: at a local minimum $\mathbf{x}^\sharp$, the negative gradient $-\nabla f(\mathbf{x}^\sharp)$ belongs to the cone generated by the gradients of the active constraints (under qualification).

2.6.1 How to use the KKT theorem in practice

The KKT theorem provides a system of equations and inequalities that any qualified local minimizer must satisfy. The key difficulty is *complementary slackness*: for each inequality constraint $h_j(\mathbf{x}) \leq 0$, one must have

$$\mu_j \geq 0, \qquad h_j(\mathbf{x}) \leq 0, \qquad \mu_j h_j(\mathbf{x}) = 0,$$

so at a KKT point each constraint is either *inactive* (then $h_j(\mathbf{x}) < 0$ and necessarily $\mu_j = 0$) or *active* (then $h_j(\mathbf{x}) = 0$ and μ_j is an additional unknown).

In the small-dimensional problems of this course, this naturally leads to a *case distinction* (equivalently, an enumeration of active sets): choose, for each $j \in [n_I]$, whether $\mu_j = 0$ or $h_j(\mathbf{x}) = 0$, solve the resulting stationarity/feasibility system, and keep the solutions that satisfy $\boldsymbol{\mu} \geq 0$ and all constraints.

Since KKT conditions are only necessary in general, the surviving KKT points are *candidates* (local maxima and saddle points may also satisfy them), and one then compares objective values (or uses second-order information[1]) to identify minima.

In the convex setting of Chapter 3, KKT conditions become sufficient under suitable regularity, so any feasible KKT point is globally optimal; in that case, the same active-set reasoning becomes an effective way to solve exercises, but not a general-purpose numerical method for large instances.

Example 2.5 (A box-and-halfspace constrained quadratic). Consider the problem in $\mathbb{R}^2$

$$\operatorname*{Min}_{\mathbf{x}\in\mathbb{R}^2} \ \frac{1}{2}\|\mathbf{x}\|^2 \qquad \text{s.t.} \qquad x_1 + x_2 \geq 1, \quad x_1 \geq 0, \quad x_2 \geq 0.$$

Rewrite the inequalities in the form $h(\mathbf{x}) \leq 0$:

$$h_0(\mathbf{x}) = 1 - x_1 - x_2, \qquad h_1(\mathbf{x}) = -x_1, \qquad h_2(\mathbf{x}) = -x_2.$$

All constraints are affine, hence qualified at every feasible point. The KKT system reads: there

[1] Constrained second-order optimality conditions are beyond the scope of this course.

exist $\boldsymbol{\mu} = (\mu_0, \mu_1, \mu_2) \in \mathbb{R}^3_+$ such that

$$\begin{aligned}
&\textbf{(stationarity)} && \mathbf{x} - \mu_0 \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \mu_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \mu_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \mathbf{0}, \\
&\textbf{(feasibility)} && x_1 + x_2 \geq 1, \quad x_1 \geq 0, \quad x_2 \geq 0, \\
&\textbf{(complementary slackness)} && \mu_0(1 - x_1 - x_2) = 0, \quad \mu_1 x_1 = 0, \quad \mu_2 x_2 = 0.
\end{aligned}$$

We now distinguish the four cases corresponding to whether the nonnegativity constraints are active.

- **Case 1:** $x_1 > 0$ **and** $x_2 > 0$**.** Then complementary slackness gives $\mu_1 = \mu_2 = 0$. Stationarity yields $x_1 = x_2 = \mu_0$.
  - If $\mu_0 = 0$, then $\mathbf{x} = (0, 0)$, which is infeasible since $x_1 + x_2 \geq 1$.
  - If $\mu_0 > 0$, then complementary slackness for h_0 enforces $1 - x_1 - x_2 = 0$, i.e. $2\mu_0 = 1$. Hence $\mu_0 = \frac{1}{2}$ and $\mathbf{x}^\sharp = (\frac{1}{2}, \frac{1}{2})$.
- **Case 2:** $x_1 = 0$ **and** $x_2 > 0$**.** Then $\mu_2 = 0$ and stationarity gives $x_1 - \mu_0 - \mu_1 = 0$, hence $\mu_1 = -\mu_0$. Since $\mu_0, \mu_1 \geq 0$, we get $\mu_0 = \mu_1 = 0$, and then stationarity forces $x_2 = 0$, contradicting $x_2 > 0$. Thus this case yields no KKT point.
- **Case 3:** $x_1 > 0$ **and** $x_2 = 0$**.** By symmetry with Case 2, this case yields no KKT point.
- **Case 4:** $x_1 = 0$ **and** $x_2 = 0$**.** This case is infeasible because $x_1 + x_2 \geq 1$.

Therefore the only KKT point is $\mathbf{x}^\sharp = (\frac{1}{2}, \frac{1}{2})$ (with $\mu_0 = \frac{1}{2}$ and $\mu_1 = \mu_2 = 0$), which is indeed the unique global minimum of this problem.

2.6.2 Some remarks about KKT conditions

The KKT conditions extend the unconstrained first-order condition $\nabla f(\mathbf{x}^\sharp) = 0$ to constrained problems. In the unconstrained case, a nonzero gradient provides an immediate descent direction (typically $-\nabla f(\mathbf{x}^\sharp)$), so local optimality forces $\nabla f(\mathbf{x}^\sharp) = 0$. With constraints, descent directions must be compatible with feasibility; the multiplier form of stationarity expresses that there is no *feasible* first-order descent direction and that $-\nabla f(\mathbf{x}^\sharp)$ can be balanced by the gradients of the active constraints.

Outside convexity, these conditions are generally necessary but not sufficient: local maxima and saddle points may also satisfy them. Moreover, without constraint qualification, a local minimum may fail to satisfy KKT conditions (see the example in Section 2.4). They nevertheless remain a central tool for analyzing solutions, deriving certificates, and motivating algorithms.

2.6.3 Proof sketch of the KKT theorem

The following geometric lemma is the key ingredient for passing from the tangent-cone condition of Theorem 2.8 to the KKT multiplier form. Geometrically, it says that c has nonnegative inner product with every vector satisfying the cone inequalities if and only if $-c$ belongs to the conic hull of the a_k's.

♢ **Lemma 2.21** (Farkas lemma (cone form))**.** *Let $a_1, \ldots, a_m \in \mathbb{R}^n$ and let $A = (a_1, \ldots, a_m) \in \mathbb{R}^{n \times m}$. For $c \in \mathbb{R}^n$, the following are equivalent:*

1. *$c \cdot d \geq 0$ for all $d \in \mathbb{R}^n$ such that $a_k \cdot d \leq 0$ for all $k = 1, \ldots, m$;*

2. there exists $y \in \mathbb{R}^m_+$ such that $c = -Ay$.

In the KKT proof below, c plays the role of $\nabla f(\mathbf{x}^\sharp)$ and the a_k's are the gradients of the active constraints; the Farkas lemma then produces the Lagrange multipliers.

Proof sketch of Theorem 2.20. Theorem 2.8 gives $\nabla f(\mathbf{x}^\sharp)\cdot\mathbf{d} \geq 0$ for all $\mathbf{d} \in T_X(\mathbf{x}^\sharp)$. Qualification gives $T_X(\mathbf{x}^\sharp) = T_X^\ell(\mathbf{x}^\sharp)$, which is defined by linear equalities and inequalities on $\mathbf{d}$ involving the gradients of the constraints. To apply Lemma 2.21, write each equality $\nabla g_i(\mathbf{x}^\sharp)\cdot\mathbf{d} = 0$ as the pair of inequalities $\nabla g_i(\mathbf{x}^\sharp)\cdot\mathbf{d} \leq 0$ and $(-\nabla g_i(\mathbf{x}^\sharp))\cdot\mathbf{d} \leq 0$, which leads to a free-signed multiplier $\lambda_i^\sharp$ when the two nonnegative multipliers are combined. Applying Lemma 2.21 to this linear system yields multipliers $\boldsymbol{\lambda}^\sharp$ and $\boldsymbol{\mu}^\sharp \geq 0$ such that the stationarity condition holds, with $\mu_j^\sharp = 0$ for inactive constraints. Complementary slackness follows from $\mu_j^\sharp = 0$ for inactive constraints and $h_j(\mathbf{x}^\sharp) = 0$ for active ones. □

Example 2.6 (One-sided constraint in $\mathbb{R}$). Minimize $f(x) = (x-1)^2$ subject to $x \geq 0$, i.e. $h(x) = -x \leq 0$. The unconstrained minimizer $x = 1$ is feasible, so $x^\sharp = 1$ is optimal. KKT holds with $\mu^\sharp = 0$.

2.7 Lagrangian viewpoint and weak duality

This section introduces a complementary viewpoint on constrained problems: instead of working directly with feasibility and tangent cones, we embed the constraints into the objective through *Lagrange multipliers*. This produces (i) a family of *lower bounds* on the optimal value (weak duality) and (ii) a language that, in favorable settings, reinterprets KKT conditions as a *saddle-point* condition.

2.7.1 From constrained problems to a min–sup formulation

Let $g(\mathbf{x}) := (g_i(\mathbf{x}))_{i\in[n_E]} \in \mathbb{R}^{n_E}$ and $h(\mathbf{x}) := (h_j(\mathbf{x}))_{j\in[n_I]} \in \mathbb{R}^{n_I}$. Define the *Lagrangian* of Problem (2.6) by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := f(\mathbf{x}) + \boldsymbol{\lambda}^\top g(\mathbf{x}) + \boldsymbol{\mu}^\top h(\mathbf{x}), \qquad (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+.$$

For the feasible set X of (2.6), the original *primal* problem (2.6) can be written as[2]

$$\operatorname*{Min}_{\mathbf{x}\in\mathbb{R}^n} \; f(\mathbf{x}) + \mathbb{I}_{g(\mathbf{x})=0} + \mathbb{I}_{h(\mathbf{x})\leq 0}.$$

This reformulation is *exact* and does not require any assumption: the minimizers and the optimal value are unchanged, since infeasible points are simply assigned value $+\infty$.

Using the dual representations

$$\mathbb{I}_{g(\mathbf{x})=0} = \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{n_E}} \boldsymbol{\lambda}^\top g(\mathbf{x}), \qquad \mathbb{I}_{h(\mathbf{x})\leq 0} = \sup_{\boldsymbol{\mu}\in\mathbb{R}^{n_I}_+} \boldsymbol{\mu}^\top h(\mathbf{x}),$$

we obtain, again without any assumption, the following *min–sup* representation of (2.6):

$$\operatorname*{Min}_{\mathbf{x}\in\mathbb{R}^n} \; \sup_{\boldsymbol{\lambda}\in\mathbb{R}^{n_E},\, \boldsymbol{\mu}\in\mathbb{R}^{n_I}_+} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

Indeed, for a fixed $\mathbf{x}$, the inner supremum is finite if and only if $\mathbf{x} \in X \cap \operatorname{dom} f$ (in which case it equals $f(\mathbf{x})$), and it is $+\infty$ otherwise.

The *Lagrangian dual problem* is obtained by exchanging the order of the Min and sup, i.e. by replacing "Min sup" with "Max inf". As we will see, this exchange always yields a valid bound (weak duality), and in convex settings it may even be exact (strong duality).

[2]Recall that $\mathbb{I}_{\mathcal{S}}$ is the indicator function of the set/condition $\mathcal{S}$ (equal to 0 when it holds and $+\infty$ otherwise).

2.7.2 Dual function and weak duality

Define the *dual function*

$$d(\boldsymbol{\lambda}, \boldsymbol{\mu}) := \inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \in \overline{\mathbb{R}}.$$

The associated *Lagrangian dual problem* is

$$\begin{aligned} \operatorname*{Max}_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \quad & d(\boldsymbol{\lambda}, \boldsymbol{\mu}) && \text{(2.8a)} \\ \text{s.t.} \quad & \boldsymbol{\lambda} \in \mathbb{R}^{n_E} && \text{(2.8b)} \\ & \boldsymbol{\mu} \in \mathbb{R}^{n_I}_+ && \text{(2.8c)} \end{aligned}$$

♢ **Lemma 2.22** (Min–max inequality)**.** *Let* $\Phi : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$*. Then*

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \Phi(x, y) \leq \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \Phi(x, y).$$

Proof. Fix any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. By definition of infimum and supremum,

$$\inf_{x' \in \mathcal{X}} \Phi(x', y) \leq \Phi(x, y) \leq \sup_{y' \in \mathcal{Y}} \Phi(x, y').$$

Taking the supremum over $y \in \mathcal{Y}$ in the left inequality and the infimum over $x \in \mathcal{X}$ in the right inequality yields the claim. □

♡ **Proposition 2.23** (Weak duality)**.** *Let* $p^\star$ *be the optimal value of the primal problem* (2.6) *and let* $d^\star$ *be the optimal value of the dual problem* (2.8)*. Then* $p^\star \geq d^\star$*.*

Proof. Apply Lemma 2.22 to $\Phi(\mathbf{x}, (\boldsymbol{\lambda}, \boldsymbol{\mu})) = \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ with $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+$. The right-hand side is exactly the primal value in its Min sup form, hence equals $p^\star$. The left-hand side is $\sup_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = d^\star$ by definition of the dual problem. □

2.7.3 Lower and upper bounds

Weak duality (Proposition 2.23) gives, for any *dual-feasible multipliers* $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+$,

$$d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq p^\star.$$

Thus dual feasibility provides a *lower bound* on $p^\star$. Conversely, any primal feasible point $\mathbf{x} \in X$ provides an *upper bound*, since

$$p^\star \leq f(\mathbf{x}).$$

The difference between an upper bound and a lower bound is called a *duality gap*.

These bounds provide certificates of solution quality: if an algorithm produces a primal feasible point $\mathbf{x}$ and dual-feasible multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ such that $f(\mathbf{x}) - d(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \varepsilon$, then $\mathbf{x}$ is certified ε-optimal.

2.7.4 Strong duality and saddle points

When the duality gap is 0 one speaks of *strong duality*; in that case, optimal primal and dual solutions form a saddle point of the Lagrangian. Strong duality is not guaranteed in general: in Chapter 3 we will see that convexity and Slater's condition provide a powerful framework where strong duality holds and KKT conditions become sufficient for global optimality.

Definition 2.24 (Saddle point)**.** *Let* $\Phi : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$*. A pair* $(x^\sharp, y^\sharp) \in \mathcal{X} \times \mathcal{Y}$ *is a* saddle point *of* Φ *on* $\mathcal{X} \times \mathcal{Y}$ *if*

$$\Phi(x^\sharp, y) \leq \Phi(x^\sharp, y^\sharp) \leq \Phi(x, y^\sharp) \qquad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

♢ **Proposition 2.25** (Saddle point implies strong duality)**.** *Assume there exists* $(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \in \mathbb{R}^n \times \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+$ *such that* $(\mathbf{x}^\sharp, (\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp))$ *is a saddle point of the Lagrangian* $\mathcal{L}$ *on* $\mathbb{R}^n \times (\mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+)$*. Then strong duality holds* ($p^\star = d^\star$)*,* $\mathbf{x}^\sharp$ *is primal optimal, and* $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ *is dual optimal. Moreover,*

$$p^\star = d^\star = \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp).$$

Proof. By the saddle property,

$$\inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = \sup_{\boldsymbol{\lambda} \in \mathbb{R}^{n_E},\ \boldsymbol{\mu} \in \mathbb{R}^{n_I}_+} \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}, \boldsymbol{\mu}).$$

The left-hand side is $d(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \leq d^\star$ by definition of $d^\star$. If $\mathbf{x}^\sharp$ is not feasible, then either $g(\mathbf{x}^\sharp) \neq 0$ or some component of $h(\mathbf{x}^\sharp)$ is positive; in both cases one can send a multiplier to $\pm\infty$ (for equalities) or $+\infty$ (for inequalities) and obtain $\sup_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0} \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}, \boldsymbol{\mu}) = +\infty$, which contradicts the finite saddle value. Hence $\mathbf{x}^\sharp \in X$, and then $\sup_{\boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0} \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}^\sharp) \geq p^\star$. Together with weak duality $d^\star \leq p^\star$, this forces $p^\star = d^\star$ and shows that both $\mathbf{x}^\sharp$ and $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ are optimal. □

2.7.5 Sensitivity interpretation of multipliers

In many problems, Lagrange multipliers quantify how the optimal value changes when perturbing the constraints: roughly, $\lambda_i^\sharp$ (or $\mu_j^\sharp$) measures the marginal cost of tightening the corresponding constraint. This is sometimes called the *shadow price* interpretation.

Example 2.7 (Shadow price of a bound constraint). Consider the parametric family of problems indexed by $b \in \mathbb{R}$:

$$p^\star(b) := \min_{x \in \mathbb{R}} x^2 \qquad \text{subject to} \qquad x \geq b,$$

i.e. $h(x) = b - x \leq 0$.

- If $b \leq 0$: the unconstrained minimizer $x^\sharp = 0$ is feasible, so $p^\star(b) = 0$; KKT gives $\mu^\sharp = 0$.
- If $b > 0$: the constraint is active at $x^\sharp = b$, so $p^\star(b) = b^2$; KKT stationarity $2b - \mu^\sharp = 0$ gives $\mu^\sharp = 2b$.

In both regimes, $\mu^\sharp = \dfrac{dp^\star}{db}(b)$: the multiplier equals the rate at which the optimal value increases when the constraint is tightened (i.e. when b increases).

The connection between multipliers and sensitivity becomes particularly clean under strong duality and regularity assumptions (see Chapter 3).

2.8 Exercises

Exercise 2.1
Let $\mathbf{p}^1, \dots, \mathbf{p}^s \in \mathbb{R}^n$. Consider the problem

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \sum_{i=1}^{s} \|\mathbf{x} - \mathbf{p}^i\|_2^2.$$

1. Show that an optimal solution always exists, and that it can be obtained from a simple explicit formula.
2. This optimal point is well-known: what is it called?

Solution

Exercise 2.2
With the same data as in the previous exercise, consider

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \max_{i=1,\dots,s} \|\mathbf{x} - \mathbf{p}^i\|_2. \tag{R}$$

The goal is not to compute the optimizer, but to prove existence and give an alternative modeling.

1. Explain why this is also called the *smallest enclosing ball* problem.
2. Explain why (R) has at least one optimal solution.[3]
3. By introducing a variable t, rewrite (R) with a linear objective, no equality constraints, and only polynomial inequality constraints.

Solution

Exercise 2.3
Let $A \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$.

1. Explain why the function $\mathbf{x} \in \mathbb{R}^n \mapsto \|A\mathbf{x} - \mathbf{b}\|_2$ attains a global minimum on $\mathbb{R}^n$ (this is the classical least-squares problem).
2. Deduce that the linear system $A^\top A\mathbf{x} = A^\top \mathbf{b}$ always has at least one solution.

Solution

Exercise 2.4 *Fermat's principle and Snell's law*
Two points lie on opposite sides of a horizontal interface:

$$A = (0, a), \qquad B = (L, -b),$$

where $a > 0$, $b > 0$, and $L > 0$. Light travels at speed $v_1 > 0$ above the interface and at speed $v_2 > 0$ below it. Assume that the ray consists of two straight segments meeting the interface at $C = (x, 0)$. According to Fermat's principle, the physical trajectory minimizes travel time.

1. Show that the travel time is

$$T(x) = \frac{\sqrt{x^2 + a^2}}{v_1} + \frac{\sqrt{(L - x)^2 + b^2}}{v_2}.$$

[3] In fact it is unique, but this is not the point here.

2. Show that T admits a minimizer.

3. Show that $T''(x) > 0$ and deduce that the minimizer is unique.

4. Let θ_1 and θ_2 be the angles made by the two segments with the normal to the interface. Derive Snell's law from the first-order optimality condition.

5. If $v_1 = v_2$, what does the optimality condition mean geometrically?

Solution

Exercise 2.5

Consider

$$\begin{array}{ll} \text{Min} & x^3 - y^3 + z^3 \\ \text{s.t.} & x^2 + y^2 + z^2 = 1. \end{array}$$

1. Show that an optimal solution exists.

2. Find an optimal solution using KKT conditions.

Solution

Exercise 2.6

Let $n \geq 1$ and consider

$$\begin{array}{rl} \text{Max} & \sum_{i=1}^n s_i \ln(x_i) \\ \text{s.t.} & \mathbf{a} \cdot \mathbf{x} = 1 \\ & \mathbf{x} \in (\mathbb{R}_+^*)^n, \end{array}$$

where $s_i > 0$ for all $i \in [n]$.

1. Determine the optimal value when there exists j such that $a_j \leq 0$.

From now on assume that $a_i > 0$ for all i.

2. Using a compactness argument, show that an optimal solution exists.

3. Find it using KKT conditions.

Solution

Exercise 2.7

Consider the problem

$$\begin{array}{rl} \text{Min} & \mathbf{x}^\top M \mathbf{x} \\ \text{s.t.} & \|\mathbf{x}\|_2 = 1, \end{array}$$

where M is a square matrix.

1. Show that an optimal solution exists.

2. Characterize it using KKT conditions.

3. Deduce that every real symmetric matrix has an eigenvector.

Solution

Applying the KKT conditions

Exercise 2.8
Solve

$$\begin{array}{rl} \text{Min} & x_1^2 + x_2^2 \\ \text{s.t.} & -2x_1 - x_2 + 10 \leq 0 \\ & -x_1 \leq 0. \end{array}$$

Solution

Exercise 2.9
Solve

$$\begin{array}{rl} \text{Min} & 5x_1^2 + 6x_2^2 \\ \text{s.t.} & x_1 - 4 \leq 0 \\ & 25 - x_1^2 - x_2^2 \leq 0. \end{array}$$

Solution

Exercise 2.10
For the problem

$$\begin{array}{rl} \text{Min} & (x_1 + 1)^2 + x_2^2 \\ \text{s.t.} & -x_1^3 + x_2^2 \leq 0, \end{array}$$

compute the tangent cone $T_X(x_1, x_2)$ (a justification based on a drawing is enough) and the linearized cone $T_X^\ell(x_1, x_2)$ at $(x_1, x_2) = (0, 0)$.

Deduce that constraint qualification fails, and check that KKT conditions do not apply at the optimum.

Solution

Exercise 2.11
Solve the problem ($n \geq 2$)

$$\begin{array}{rl} \text{Max} & \displaystyle\prod_{i=1}^{n} \ln(x_i) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{n} x_i = n^2 \\ & x_i \geq 1 \qquad \forall i \in [n]. \end{array}$$

1. Show that an optimal solution exists.
2. Show that any optimal solution satisfies $x_i^\sharp > 1$ for all i.
3. Show that there exists $\lambda \in \mathbb{R}$ such that for all i,

$$\lambda x_i^\sharp \ln(x_i^\sharp) = \prod_{j=1}^{n} \ln(x_j^\sharp).$$

4. Explain why $\lambda \neq 0$ and deduce an optimal solution.
5. Is it unique?

Solution

Exercise 2.12
By studying the problem

$$\begin{array}{ll} \text{Min} & \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^n - \prod_{i=1}^n x_i \\ \text{s.t.} & \|\mathbf{x}\|^2 = 1 \\ & x_i \geq 0 \qquad i \in [n], \end{array}$$

show that the arithmetic mean is always at least the geometric mean (for positive numbers).

Solution

Duality

Exercise 2.13
Find the dual of the following problem:

$$\begin{array}{ll} \text{Min} & \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} \leq \mathbf{b}. \end{array}$$

Solution

Exercise 2.14
Consider

$$\begin{array}{ll} \text{Min} & 5x + 11y + z \\ \text{s.t.} & x + y - z \geq 19 \\ & -y + z \geq 13 \\ & x + y \geq 24 \\ & x, y, z \geq 0. \end{array}$$

1. Write the dual problem.
2. A heuristic algorithm returns the primal point $(32, 0, 13)$ and the dual point $(5, 6, 0)$. Did it actually find an optimal solution of the primal?

Solution

Exercise 2.15
Let M be positive definite. Find the dual of

$$\begin{array}{ll} \text{Min} & \mathbf{x}^\top M\mathbf{x} + \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} \leq \mathbf{b}. \end{array}$$

Solution

Exercise 2.16
Consider

$$\begin{array}{ll} \text{Min} & e^{x_2} \\ \text{s.t.} & x_1 \leq x_2 \\ & x_2 \geq 0. \end{array}$$

1. Compute its optimal value.
2. Find the dual problem and its optimal value.
3. Does strong duality hold?

Solution

Exercise 2.17
Consider again the problem from Exercise 2.10:

$$\begin{array}{rl} \text{Min} & (x_1+1)^2 + x_2^2 \\ \text{s.t.} & -x_1^3 + x_2^2 \leq 0. \end{array}$$

1. Compute its optimal value.
2. Find the dual problem and its optimal value.
3. Does strong duality hold?

Solution

Chapter 3

Convex optimization

Convex optimization is the study of problems where both the feasible set and the objective enjoy convexity properties. This structure is fundamental: for convex problems, any local minimum is global; first-order conditions become both necessary and sufficient; and duality often provides certificates of optimality and sensitivity information.

We continue to consider problems of the form

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \quad f(\mathbf{x}) \tag{3.1a}$$

$$\text{s.t.} \quad \mathbf{x} \in X \tag{3.1b}$$

and the differentiable constrained problems of the form (1.7) from Chapter 1. We focus on the *convex* case: f is convex and X is convex in (3.1); and for (1.7), the equality constraints are affine while the inequality constraints are convex.

Beyond these theoretical simplifications, convex formulations arise in many applications (e.g. portfolio optimization, statistical learning, signal processing, optimal control). For a broader reference, see e.g. Boyd and Vandenberghe [1].

3.1 Convexity

We begin with the basic geometric notions (convex sets and convex functions) that underpin the theory and the algorithms of convex optimization.

3.1.1 Convex sets

♡ **Definition 3.1** (Convex set)**.** *A set $C \subseteq \mathbb{R}^n$ is* convex *if for all $\mathbf{x}, \mathbf{y} \in C$ and all $t \in [0,1]$,*

$$t\mathbf{x} + (1-t)\mathbf{y} \in C.$$

The whole space $\mathbb{R}^n$ is convex, and the intersection of any family of convex sets is convex. Figure 3.1 illustrates convexity.

Definition 3.2 (Convex hull)**.** *For a set $C \subseteq \mathbb{R}^n$, the* convex hull *of C is*

$$\text{conv}(C) := \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mid k \in \mathbb{N},\ \mathbf{x}_i \in C,\ \lambda_i \geq 0,\ \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

It is the smallest convex set containing C.

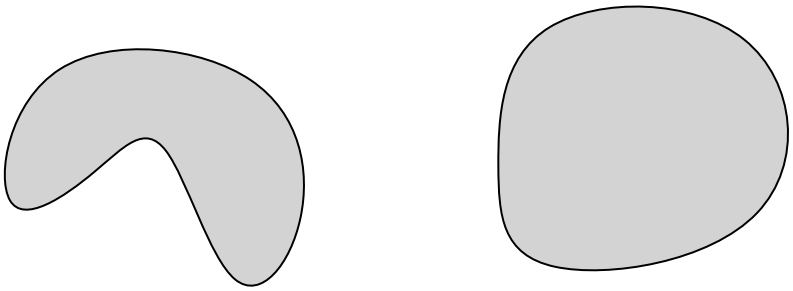

Figure 3.1: Two subsets of the plane. The left one is not convex, the right one is convex.

Proposition 3.3 (Operations preserving convexity)**.** *If $C_1, C_2 \subseteq \mathbb{R}^n$ are convex, then the following sets are convex:*

- *the intersection $C_1 \cap C_2$ and, more generally, $\bigcap_{i \in I} C_i$ for any index set I;*
- *the Cartesian product $C_1 \times C_2$ and the Minkowski sum $C_1 + C_2 = \{x_1 + x_2 \mid x_i \in C_i\}$;*
- *the image $A(C_1) = \{Ax \mid x \in C_1\}$ and the inverse image $A^{-1}(C_2) = \{x \mid Ax \in C_2\}$ under an affine map A;*
- *translations $C_1 + x_0$ and scalings tC_1 for $t \geq 0$; in particular, the projection of a convex set onto any affine subspace is convex.*

Proof. Each claim follows from the defining property of convexity: stability under convex combinations. For instance, if $\mathbf{x}, \mathbf{y} \in C_1 \cap C_2$, then $t\mathbf{x} + (1-t)\mathbf{y}$ belongs to both C_1 and C_2, hence to $C_1 \cap C_2$. If $\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2$ and $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ belong to $C_1 + C_2$ with $\mathbf{x}_i, \mathbf{y}_i \in C_i$, then

$$t\mathbf{x} + (1-t)\mathbf{y} = \big(t\mathbf{x}_1 + (1-t)\mathbf{y}_1\big) + \big(t\mathbf{x}_2 + (1-t)\mathbf{y}_2\big) \in C_1 + C_2,$$

since each C_i is convex. For an affine map $A(\mathbf{x}) = M\mathbf{x} + b$, one has $A(t\mathbf{x} + (1-t)\mathbf{y}) = tA(\mathbf{x}) + (1-t)A(\mathbf{y})$, which proves convexity of images and inverse images. Translations, scalings, and projections are special cases of affine images. □

Example 3.1. Affine subspaces, half-spaces, Euclidean balls and polyhedra are convex. The cone of symmetric positive semidefinite matrices is also convex.

Example 3.2 (Some nonconvex sets). The set of integers $\mathbb{Z}$ is not convex. In $\mathbb{R}^n$, the set $\{x \mid \|x\|_2 \geq 1\}$ is not convex, and neither is the set of orthogonal matrices.

3.1.2 Convex functions

Let $C \subseteq \mathbb{R}^n$ be a convex set and let $f : C \to \overline{\mathbb{R}}$. The *epigraph* of f is the subset of $\mathbb{R}^{n+1}$ defined by

$$\operatorname{epi}(f) := \{(\mathbf{x}, t) \in C \times \mathbb{R} \mid f(\mathbf{x}) \leq t\}.$$

We define convexity through this geometric object.

♡ **Definition 3.4** (Convex function)**.** *The function f is* convex *if and only if* $\operatorname{epi}(f)$ *is convex. Equivalently, for all $\mathbf{x}, \mathbf{y} \in C$ and all $t \in [0,1]$,*

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}).$$

In particular, $\operatorname{dom} f$ is convex. The function f is *concave* if $-f$ is convex.

Example 3.3. Figure 3.2 shows two functions of one real variable. The left one is not convex; the right one is convex.

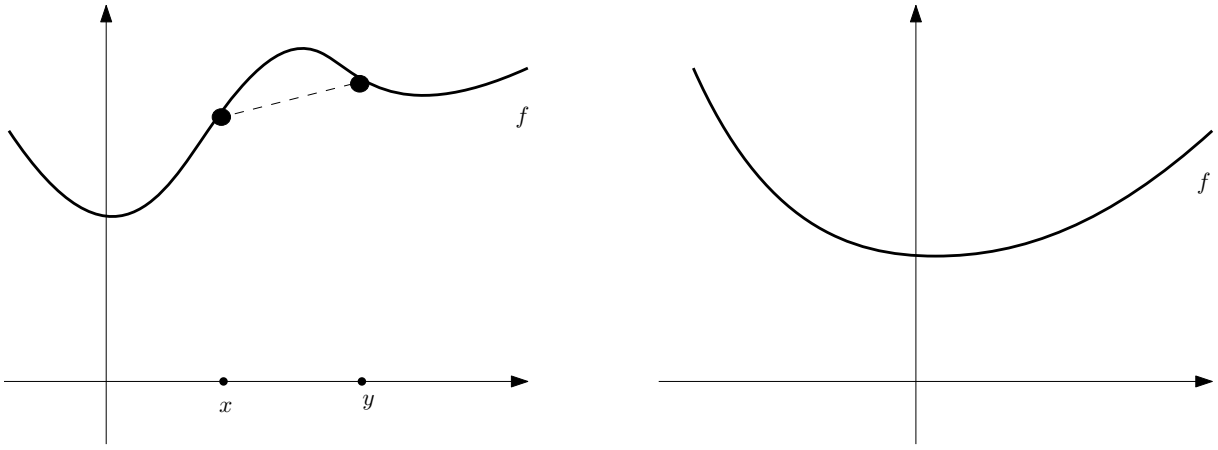

Figure 3.2: A nonconvex function (left) and a convex one (right).

Proposition 3.5 (Convex combinations)**.** *A function f is convex if and only if for every integer $k \geq 2$,*

$$f\left(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i)$$

for all $\mathbf{x}_i \in \operatorname{dom} f$ and all $\lambda_i \geq 0$ with $\sum_i \lambda_i = 1$.

Proof. If f is convex, the inequality follows by induction on k. The case $k = 2$ is Definition 3.4. Assume it holds up to $k-1$. If $\lambda_k = 1$, then $\sum_{i=1}^{k} \lambda_i \mathbf{x}_i = \mathbf{x}_k$ and the inequality holds with equality. Otherwise, write for $k \geq 3$,

$$\sum_{i=1}^{k} \lambda_i \mathbf{x}_i = (1-\lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1-\lambda_k} \mathbf{x}_i + \lambda_k \mathbf{x}_k.$$

Apply convexity with weights $(1-\lambda_k, \lambda_k)$ and then apply the induction hypothesis to the first term. □

♡ **Proposition 3.6** (Above its tangents (differentiable case))**.** *Assume that f is differentiable. Then f is convex if and only if* $\operatorname{dom} f$ *is convex and*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y} \in \operatorname{dom} f.$$

Proof. Assume f is convex. For $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $t \in (0, 1]$, convexity gives $f(\mathbf{y} + t(\mathbf{x}-\mathbf{y})) \leq (1-t)f(\mathbf{y}) + tf(\mathbf{x})$, i.e. $\frac{f(\mathbf{y}+t(\mathbf{x}-\mathbf{y})) - f(\mathbf{y})}{t} \leq f(\mathbf{x}) - f(\mathbf{y})$. Letting $t \downarrow 0$ yields $\nabla f(\mathbf{y})^\top (\mathbf{x}-\mathbf{y}) \leq f(\mathbf{x}) - f(\mathbf{y})$.

Conversely, assume the inequality holds and let $\mathbf{z} = t\mathbf{x} + (1-t)\mathbf{y}$ with $t \in [0, 1]$. Applying the inequality twice at $\mathbf{z}$ gives $f(\mathbf{x}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{x} - \mathbf{z})$ and $f(\mathbf{y}) \geq f(\mathbf{z}) + \nabla f(\mathbf{z})^\top (\mathbf{y} - \mathbf{z})$. Taking a convex combination yields $tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \geq f(\mathbf{z})$, i.e. convexity. □

Proposition 3.6 can be read as: a differentiable convex function lies above all its tangent hyperplanes; see Figure 3.3.

Definition 3.7 (Strict and strong convexity)**.** *A convex function f is* strictly convex *if the inequality in Definition 3.4 is strict for $\mathbf{x} \neq \mathbf{y}$ and $t \in (0, 1)$. It is* α-strongly convex *(or* α-convex*) if there exists $\alpha > 0$ such that for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $t \in [0, 1]$,*

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) - \frac{\alpha}{2} t(1-t)\|\mathbf{x} - \mathbf{y}\|^2.$$

♢ **Lemma 3.8** (Strong convexity via a quadratic shift)**.** *Let $C \subseteq \mathbb{R}^n$ be convex and let $f : C \to \overline{\mathbb{R}}$. Fix $\alpha \geq 0$ and define*

$$g(\mathbf{x}) := f(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}\|^2, \qquad \mathbf{x} \in C.$$

Then f is α-strongly convex on C if and only if g is convex on C.

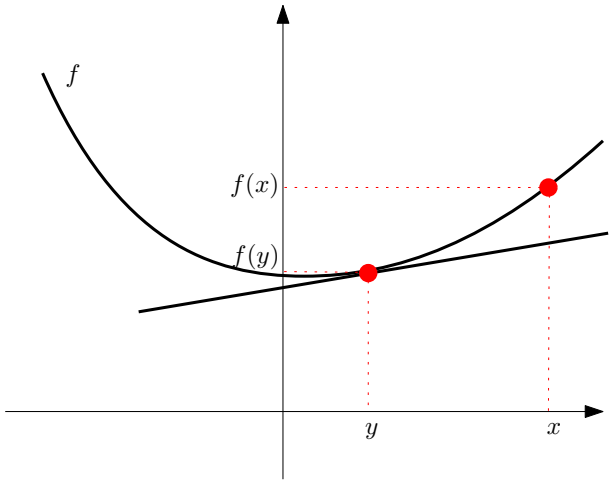

Figure 3.3: A convex differentiable function lies above all its tangents.

Proof. For $\mathbf{x}, \mathbf{y} \in C$ and $t \in [0,1]$, the identity

$$\|t\mathbf{x} + (1-t)\mathbf{y}\|^2 = t\|\mathbf{x}\|^2 + (1-t)\|\mathbf{y}\|^2 - t(1-t)\|\mathbf{x}-\mathbf{y}\|^2$$

implies that

$$g(t\mathbf{x} + (1-t)\mathbf{y}) \leq tg(\mathbf{x}) + (1-t)g(\mathbf{y}) \iff f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) - \frac{\alpha}{2}t(1-t)\|\mathbf{x}-\mathbf{y}\|^2.$$

The left-hand inequality is convexity of g on C, and the right-hand inequality is α-strong convexity of f on C. □

Proposition 3.9 (Closure properties)**.** *Let f, g be convex functions and let $t > 0$.*

i) $tf + g$ is convex.

ii) If $(f_i)_{i\in I}$ is a family of convex functions, then $\sup_{i\in I} f_i$ is convex.

iii) If a is affine and f is convex, then $f \circ a$ is convex.

iv) If f is convex and nondecreasing and g is convex, then $f \circ g$ is convex (when defined).

v) $\operatorname{dom} f$ and all sublevel sets $\operatorname{lev}_M(f)$ are convex.

Proof. Let $\mathbf{x}, \mathbf{y}$ belong to the relevant domains and let $\lambda \in [0,1]$.

i) By convexity,
$$(tf+g)(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = tf(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) + g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda(tf+g)(\mathbf{x}) + (1-\lambda)(tf+g)(\mathbf{y}).$$

ii) The epigraph of the supremum is the intersection of the epigraphs, which are convex by assumption.

iii) If a is affine, $a(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \lambda a(\mathbf{x}) + (1-\lambda)a(\mathbf{y})$, so convexity of f yields convexity of $f \circ a$.

iv) Convexity of g gives $g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})$. Since f is nondecreasing,
$$f(g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})) \leq f(\lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})),$$
and convexity of f gives $f(\lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})) \leq \lambda f(g(\mathbf{x})) + (1-\lambda)f(g(\mathbf{y}))$.

v) If $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, then $f(\mathbf{x}), f(\mathbf{y}) < +\infty$, so convexity implies $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y}) < +\infty$, hence $\lambda\mathbf{x}+(1-\lambda)\mathbf{y} \in \operatorname{dom} f$. If $\mathbf{x}, \mathbf{y} \in \operatorname{lev}_M(f)$, then $f(\mathbf{x}), f(\mathbf{y}) \leq M$ and convexity gives $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq M$, so $\operatorname{lev}_M(f)$ is convex.

□

Example 3.4 (A warning). The pointwise product of convex functions is *not* convex in general.

Example 3.5 (Indicator function). For a set $C \subseteq \mathbb{R}^n$, the indicator function $\mathbb{I}_C$ is convex if and only if C is convex.

Example 3.6 (Other classical examples). For any $a \in \mathbb{R}$, the function $x \mapsto e^{ax}$ is convex on $\mathbb{R}$. For any norm $\|\cdot\|$ and any $q \geq 1$, the function $x \mapsto \|x\|^q$ is convex. The function $x \mapsto \ln(x)$ is concave on $(0, +\infty)$, whereas $x \mapsto x\ln(x)$ is convex on $(0, +\infty)$.

3.1.3 Regularity properties

♢ **Lemma 3.10.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}^{\circ}$ be convex. Then f is continuous on the interior of* $\operatorname{dom} f$.

Proof. Let $\mathbf{y} \in \operatorname{int}(\operatorname{dom} f)$. Choose $r > 0$ such that the closed cube

$$C := \mathbf{y} + [-r, r]^n$$

is contained in $\operatorname{dom} f$. Let V be the finite set of vertices of C and set

$$M := \max\left(f(\mathbf{y}), \max_{\mathbf{z}\in V} f(\mathbf{z})\right).$$

Every point of C is a convex combination of its vertices, so convexity gives $f(\mathbf{z}) \le M$ for all $\mathbf{z} \in C$.

Take $\mathbf{x} \neq \mathbf{y}$ sufficiently close to $\mathbf{y}$, and define

$$\rho := \|\mathbf{x} - \mathbf{y}\|_\infty, \qquad t := \frac{\rho}{r}, \qquad \mathbf{u} := \frac{\mathbf{x} - \mathbf{y}}{\rho}.$$

Then $0 < t < 1$, $\|\mathbf{u}\|_\infty = 1$, and $\mathbf{z}_+ := \mathbf{y} + r\mathbf{u}$ and $\mathbf{z}_- := \mathbf{y} - r\mathbf{u}$ belong to C. Since $\mathbf{x} = (1-t)\mathbf{y} + t\mathbf{z}_+$, convexity gives

$$f(\mathbf{x}) - f(\mathbf{y}) \le t\big(M - f(\mathbf{y})\big).$$

Also, $\mathbf{y} = \frac{1}{1+t}\mathbf{x} + \frac{t}{1+t}\mathbf{z}_-$, so convexity gives

$$f(\mathbf{x}) - f(\mathbf{y}) \ge -t\big(M - f(\mathbf{y})\big).$$

Hence

$$|f(\mathbf{x}) - f(\mathbf{y})| \le \frac{\|\mathbf{x} - \mathbf{y}\|_\infty}{r}\big(M - f(\mathbf{y})\big) \longrightarrow 0$$

as $\mathbf{x} \to \mathbf{y}$. Therefore f is continuous at $\mathbf{y}$. □

Example 3.7. Convex functions may be discontinuous at the boundary of their domain. For example,

$$f(x) = \begin{cases} 0 & \text{if } x \in [0, 1), \\ 1 & \text{if } x = 1, \\ +\infty & \text{otherwise,} \end{cases}$$

is convex and discontinuous at $x = 1$.

✠ **Lemma 3.11.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}^{\circ}$ be convex. Then f is continuous on $\mathbb{R}^n$ if and only if* $\operatorname{dom} f = \mathbb{R}^n$ *(i.e. f is finite everywhere).*

✠ **Lemma 3.12.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}^{\circ}$ be convex. Then f is Lipschitz continuous on every compact subset of* $\operatorname{int}(\operatorname{dom} f)$.

♢ **Lemma 3.13.** *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex (in particular, f is finite everywhere). Then f is coercive.*

Proof. Up to translating f, we may assume $f(0) < +\infty$. Then there exists $\delta > 0$ such that $f(\mathbf{x}) < +\infty$ for all $\mathbf{x}$ with $\|\mathbf{x}\| \le \delta$. Let $\alpha > 0$ be a strong convexity parameter for f.

Fix $\mathbf{x}$ with $\|\mathbf{x}\| \ge 2\delta$ and apply strong convexity with $\mathbf{y} = \mathbf{0}$ and $t = \delta/\|\mathbf{x}\|$:

$$f\left(\frac{\delta}{\|\mathbf{x}\|}\mathbf{x}\right) \le \left(1 - \frac{\delta}{\|\mathbf{x}\|}\right) f(\mathbf{0}) + \frac{\delta}{\|\mathbf{x}\|} f(\mathbf{x}) - \frac{\alpha}{2}\frac{\delta}{\|\mathbf{x}\|}\left(1 - \frac{\delta}{\|\mathbf{x}\|}\right)\|\mathbf{x}\|^2.$$

Rearranging yields

$$f(\mathbf{x}) \ge \left(f\left(\frac{\delta}{\|\mathbf{x}\|}\mathbf{x}\right) - \left(1 - \frac{\delta}{\|\mathbf{x}\|}\right) f(\mathbf{0})\right)\frac{\|\mathbf{x}\|}{\delta} + \frac{\alpha}{2}\left(1 - \frac{\delta}{\|\mathbf{x}\|}\right)\|\mathbf{x}\|^2.$$

Let $r := \max_{\|\mathbf{z}\| \le \delta} |f(\mathbf{z})|$, which is finite since f is continuous on $\{\mathbf{z} \mid \|\mathbf{z}\| \le \delta\}$ by Lemma 3.10. Using $\|\mathbf{x}\| \ge 2\delta$ (hence $1 - \delta/\|\mathbf{x}\| \ge 1/2$) gives

$$f(\mathbf{x}) \ge -2\frac{r}{\delta}\|\mathbf{x}\| + \frac{\alpha}{4}\|\mathbf{x}\|^2,$$

so $f(\mathbf{x}) \to +\infty$ when $\|\mathbf{x}\| \to +\infty$. □

3.1.4 Second-order characterizations

When f is twice differentiable, convexity can be tested through its Hessian.

♢ **Definition 3.14** (Positive (semi-)definite matrices)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. We say that A is* positive semidefinite *(PSD), and write $A \succeq 0$, if $z^\top Az \geq 0$ for all $z \in \mathbb{R}^n$. We say that A is* positive definite *(PD), and write $A \succ 0$, if $z^\top Az > 0$ for all $z \neq 0$.*

See Appendix A.3 for more on PSD matrices and diagonalization.

♢ **Proposition 3.15.** *If A is symmetric, then $A \succeq 0$ (resp. $A \succ 0$) if and only if all its eigenvalues are nonnegative (resp. positive). Moreover, $A - \alpha I \succeq 0$ (equivalently, $A \succeq \alpha I$) if and only if all eigenvalues of A are at least α.*

Proof. By the spectral theorem, there exists an orthonormal basis of eigenvectors $(u_i)_{i=1}^n$ and eigenvalues $(\lambda_i)_{i=1}^n$ such that $A = \sum_{i=1}^n \lambda_i u_i u_i^\top$. For any $z \in \mathbb{R}^n$,

$$z^\top Az = \sum_{i=1}^n \lambda_i (u_i^\top z)^2.$$

If all $\lambda_i \geq 0$ (resp. > 0), then $z^\top Az \geq 0$ for all z (resp. > 0 for $z \neq 0$), hence $A \succeq 0$ (resp. $A \succ 0$). Conversely, if some $\lambda_i < 0$, taking $z = u_i$ gives $z^\top Az = \lambda_i < 0$, contradicting $A \succeq 0$. The last statement follows by applying the same argument to $A - \alpha I$, whose eigenvalues are $\lambda_i - \alpha$. □

♢ **Proposition 3.16** (Second-order characterizations)**.** *Assume f is twice differentiable and that* $\operatorname{dom} f$ *is convex. Then:*

- *f is convex on its domain if and only if $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \operatorname{dom} f$;*
- *if $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \operatorname{dom} f$, then f is strictly convex;*
- *f is α-strongly convex if and only if $\nabla^2 f(\mathbf{x}) - \alpha I \succeq 0$ for all $\mathbf{x} \in \operatorname{dom} f$.*

Proof. Fix $\mathbf{x} \in \operatorname{dom} f$ and $v \in \mathbb{R}^n$, and consider the one-dimensional restriction $\varphi(t) := f(\mathbf{x} + tv)$ on the interval where it is well-defined. Then $\varphi''(0) = v^\top \nabla^2 f(\mathbf{x})\, v$.

Convexity ⇒ PSD Hessian. If f is convex, then φ is convex, hence $\varphi''(0) \geq 0$ for every v, which means $\nabla^2 f(\mathbf{x}) \succeq 0$.

PSD Hessian ⇒ convexity. Assume $\nabla^2 f(\cdot) \succeq 0$ on $\operatorname{dom} f$. For $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, set $\varphi(t) := f((1-t)\mathbf{x} + t\mathbf{y})$ for $t \in [0,1]$. Since $\operatorname{dom} f$ is convex, $(1-t)\mathbf{x} + t\mathbf{y} \in \operatorname{dom} f$ for all $t \in [0,1]$. Then

$$\varphi''(t) = (\mathbf{y} - \mathbf{x})^\top \nabla^2 f((1-t)\mathbf{x} + t\mathbf{y})\,(\mathbf{y} - \mathbf{x}) \geq 0,$$

so φ is convex on $[0,1]$, which yields $f((1-t)\mathbf{x} + t\mathbf{y}) = \varphi(t) \leq (1-t)\varphi(0) + t\varphi(1) = (1-t)f(\mathbf{x}) + tf(\mathbf{y})$.

Strict convexity. If $\nabla^2 f \succ 0$ everywhere, then for $\mathbf{x} \neq \mathbf{y}$ one has $\varphi''(t) > 0$ for all t, hence φ is strictly convex and the inequality above is strict for $t \in (0,1)$.

Strong convexity. Define $g(\mathbf{x}) := f(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}\|^2$. Then $\nabla^2 g(\mathbf{x}) = \nabla^2 f(\mathbf{x}) - \alpha I$. By the first part, g is convex if and only if $\nabla^2 f(\mathbf{x}) - \alpha I \succeq 0$ on $\operatorname{dom} f$. Moreover, by Lemma 3.8, g is convex if and only if f is α-strongly convex. □

Remark 3.1. The converse of the strict-convexity implication is false: a strictly convex C^2 function may have a Hessian that is not positive definite everywhere. For instance, $f(x) = x^4$ is strictly convex on $\mathbb{R}$ but satisfies $f''(0) = 0$.

♢ **Proposition 3.17.** *Let $C \subseteq \mathbb{R}^n$ be convex and let f be differentiable on an open neighbourhood of C. Then f is α-strongly convex on C if and only if*

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \alpha \|\mathbf{x} - \mathbf{y}\|^2 \qquad \forall \mathbf{x}, \mathbf{y} \in C.$$

Proof. Define $g(\mathbf{x}) := f(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x}\|^2$. By Lemma 3.8, f is α-strongly convex on C if and only if g is convex on C.

Assume first that g is convex and differentiable. Applying Proposition 3.6 to g and swapping $(\mathbf{x}, \mathbf{y})$ gives

$$(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq 0 \qquad \forall \mathbf{x}, \mathbf{y} \in C.$$

Since $\nabla g(\mathbf{x}) = \nabla f(\mathbf{x}) - \alpha\mathbf{x}$, this is exactly $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq \alpha\|\mathbf{x} - \mathbf{y}\|^2$.

Conversely, assume the inequality in the proposition. Then $(\nabla g(\mathbf{x}) - \nabla g(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in C$. Fix $\mathbf{x}, \mathbf{y} \in C$ and set $\varphi(t) := g((1-t)\mathbf{x} + t\mathbf{y})$. For $0 \leq s \leq t \leq 1$, monotonicity of ∇g applied to the pair $(\mathbf{x}_s, \mathbf{x}_t)$ with $\mathbf{x}_t := (1-t)\mathbf{x} + t\mathbf{y}$ yields $(\nabla g(\mathbf{x}_t) - \nabla g(\mathbf{x}_s))^\top (\mathbf{y} - \mathbf{x}) \geq 0$, hence $\varphi'(t) \geq \varphi'(s)$. Thus φ' is nondecreasing and φ is convex on $[0, 1]$, which implies convexity of g on C. Therefore f is α-strongly convex on C. □

3.2 Optimality in convex problems

We now exploit convexity to turn first-order necessary conditions into exact characterizations of global optimality.

Consider the convex problem (3.1) with f convex differentiable and X convex. In this case, the first-order necessary condition of Chapter 2 becomes sufficient.

♡ **Theorem 3.18** (First-order optimality (convex case))**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex and differentiable and let $X \subseteq \mathbb{R}^n$ be convex. Then $\mathbf{x}^\sharp \in X \cap \operatorname{dom} f$ is a global minimizer of (3.1) if and only if*

$$\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \geq 0 \qquad \text{for all } \mathbf{d} \in T_X(\mathbf{x}^\sharp).$$

Proof. Necessity is Theorem 2.8. For sufficiency, take any $\mathbf{x} \in X$ and use Proposition 3.6: $f(\mathbf{x}) \geq f(\mathbf{x}^\sharp) + \nabla f(\mathbf{x}^\sharp)^\top (\mathbf{x} - \mathbf{x}^\sharp)$. Since X is convex, $\mathbf{x} - \mathbf{x}^\sharp \in T_X(\mathbf{x}^\sharp)$, hence the right-hand side is $\geq f(\mathbf{x}^\sharp)$. □

Proposition 3.19 (Unconstrained case)**.** *If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex and differentiable, then $\mathbf{x}^\sharp \in \operatorname{dom} f$ is a global minimizer if and only if $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$.*

Proof. If f is convex and differentiable, then $f(\mathbf{y}) \geq f(\mathbf{x}^\sharp) + \nabla f(\mathbf{x}^\sharp)^\top (\mathbf{y} - \mathbf{x}^\sharp)$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

Thus, if $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$, then $f(\mathbf{y}) \geq f(\mathbf{x}^\sharp)$ for all $\mathbf{y} \in \operatorname{dom} f$, so $\mathbf{x}^\sharp$ is a global minimizer. Conversely, if $\mathbf{x}^\sharp$ is a global minimizer, it is a local minimizer, so $\nabla f(\mathbf{x}^\sharp) = \mathbf{0}$ by Theorem 2.8. □

Proposition 3.20 (Uniqueness)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be strictly convex and let X be convex. If (3.1) has a finite optimal value and admits an optimal solution, then this solution is unique.*

Proof. Assume by contradiction that $\mathbf{x}_1 \neq \mathbf{x}_2$ are two optimal solutions. For any $t \in (0, 1)$, convexity of X gives $\mathbf{x}_t := t\mathbf{x}_1 + (1-t)\mathbf{x}_2 \in X$, and strict convexity gives

$$f(\mathbf{x}_t) < tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2).$$

Since $f(\mathbf{x}_1) = f(\mathbf{x}_2)$ equals the optimal value, the right-hand side equals the optimal value, hence $f(\mathbf{x}_t)$ would be strictly smaller than the optimal value, a contradiction. □

♢ **Proposition 3.21** (Convexity of the minimizer set)**.** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex and let X be convex. If $\mathbf{x}_1$ and $\mathbf{x}_2$ are global minimizers of (3.1), then $t\mathbf{x}_1 + (1-t)\mathbf{x}_2$ is also a global minimizer for every $t \in [0, 1]$.*

Proof. For any $t \in [0, 1]$, convexity gives $f(t\mathbf{x}_1 + (1-t)\mathbf{x}_2) \leq tf(\mathbf{x}_1) + (1-t)f(\mathbf{x}_2)$. Since $f(\mathbf{x}_1) = f(\mathbf{x}_2) = \inf_{x \in X} f(x)$, the right-hand side equals the optimal value, hence $t\mathbf{x}_1 + (1-t)\mathbf{x}_2$ is also optimal. □

Proposition 3.22. *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be strongly convex and continuous, and assume $X \cap \operatorname{dom} f$ is nonempty, closed, and convex. Then (3.1) admits a unique optimal solution.*

Proof. Strong convexity implies strict convexity, hence uniqueness by Proposition 3.20. Moreover, strong convexity implies coercivity by Lemma 3.13, so existence follows from Theorem 2.5. □

$f : \mathbb{R} \to \overline{\mathbb{R}}^{\circ}$

Regularity	Convexity	Strict convexity	α-strong convexity
C^0	$f(M^{\lambda}_{x,y}) \le M^{\lambda}_{f(x),f(y)}$	$f(M^{\lambda}_{x,y}) < M^{\lambda}_{f(x),f(y)}$	$f(M^{\lambda}_{x,y}) \le M^{\lambda}_{f(x),f(y)} - \frac{\alpha}{2}\lambda(1-\lambda)(x-y)^2$
C^1	f' is nondecreasing $f(z) \ge f(x) + f'(x)(z-x)$	f' is strictly increasing (sufficient) $f(z) > f(x) + f'(x)(z-x)\ \forall z \ne x$ (sufficient)	$(f'(x) - f'(y))(x-y) \ge \alpha(x-y)^2$ $f(z) \ge f(x) + f'(x)(z-x) + \frac{\alpha}{2}(z-x)^2$
C^2	$f'' \ge 0$	$f'' > 0$ (sufficient)	$f'' \ge \alpha$

$f : \mathbb{R}^n \to \overline{\mathbb{R}}^{\circ}$

Regularity	Convexity	Strict convexity	α-strong convexity
C^0	$f(M^{\lambda}_{\mathbf{x},\mathbf{y}}) \le M^{\lambda}_{f(\mathbf{x}),f(\mathbf{y})}$	$f(M^{\lambda}_{\mathbf{x},\mathbf{y}}) < M^{\lambda}_{f(\mathbf{x}),f(\mathbf{y})}$	$f(M^{\lambda}_{\mathbf{x},\mathbf{y}}) \le M^{\lambda}_{f(\mathbf{x}),f(\mathbf{y})} - \frac{\alpha}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$
C^1	$f(\mathbf{z}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{z}-\mathbf{x})$ $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \ge 0$	$f(\mathbf{z}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{z}-\mathbf{x})\ \forall \mathbf{z} \ne \mathbf{x}$ (sufficient)	$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x}-\mathbf{y}) \ge \alpha\|\mathbf{x}-\mathbf{y}\|^2$ (Prop. 3.17) $f(\mathbf{z}) \ge f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{z}-\mathbf{x}) + \frac{\alpha}{2}\|\mathbf{z}-\mathbf{x}\|^2$
C^2	$\nabla^2 f \succeq 0$	$\nabla^2 f \succ 0$ (sufficient)	$\nabla^2 f - \alpha I \succeq 0$

Table 3.1: Summary of convexity conditions in dimension 1 and n. Here $M^{\lambda}_{x,y} := \lambda x + (1-\lambda)y$ denotes a convex combination; the same notation applies to scalar or vector arguments and to function values. See Definition 3.14 for the meaning of $A \succeq 0$ and $A \succ 0$.

3.3 KKT and strong duality in convex optimization

We next specialize the KKT framework to convex problems, where (under suitable regularity) KKT conditions become both necessary and sufficient and duality provides optimality certificates.

Consider the convex constrained problem

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \quad & f(\mathbf{x}) && && \text{(3.2a)}\\ \text{s.t.} \quad & g_i(\mathbf{x}) = 0 && \forall i \in [n_E] && \text{(3.2b)}\\ & h_j(\mathbf{x}) \le 0 && \forall j \in [n_I] && \text{(3.2c)} \end{aligned}$$

where f and the h_j are convex differentiable and the g_i are affine. Let X denote its feasible set.

Definition 3.23 (Slater point)**.** *A point $\mathbf{x}_S \in \mathbb{R}^n$ is a* Slater point *for* (3.2) *if it satisfies the equality constraints and strictly satisfies the inequality constraints:*

$$g_i(\mathbf{x}_S) = 0,\ \forall i \in [n_E], \qquad h_j(\mathbf{x}_S) < 0,\ \forall j \in [n_I].$$

Recall Proposition 2.14, which states that the existence of a Slater point implies constraint qualification at any feasible point, hence the KKT conditions of Theorem 2.20 hold at any optimal solution.

♡ **Theorem 3.24** (KKT conditions (convex case))**.** *Let $\mathbf{x}^\sharp \in X \cap \operatorname{dom} f$.*

- *(Necessity) If $\mathbf{x}^\sharp$ is optimal and the constraints are qualified at $\mathbf{x}^\sharp$, then there exist multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{n_E}$ and $\boldsymbol{\mu} \in \mathbb{R}^{n_I}_+$ such that the KKT conditions hold:*

$$\begin{cases} \nabla f(\mathbf{x}^\sharp) + \displaystyle\sum_{i\in[n_E]} \lambda_i \nabla g_i(\mathbf{x}^\sharp) + \sum_{j\in[n_I]} \mu_j \nabla h_j(\mathbf{x}^\sharp) = \mathbf{0}, \\ g_i(\mathbf{x}^\sharp) = 0,\ \forall i \in [n_E], \qquad h_j(\mathbf{x}^\sharp) \le 0,\ \forall j \in [n_I], \\ \boldsymbol{\mu} \ge 0, \qquad \mu_j h_j(\mathbf{x}^\sharp) = 0,\ \forall j \in [n_I]. \end{cases}$$

- *(Sufficiency) Conversely, if there exist $\boldsymbol{\lambda} \in \mathbb{R}^{n_E}$ and $\boldsymbol{\mu} \in \mathbb{R}^{n_I}_+$ satisfying the KKT conditions at $\mathbf{x}^\sharp$, then $\mathbf{x}^\sharp$ is optimal.*

Proof. Necessity is Theorem 2.20. For sufficiency, let $\boldsymbol{\lambda}, \boldsymbol{\mu}$ satisfy KKT at $\mathbf{x}^\sharp$ and let $\mathbf{d} \in T_X(\mathbf{x}^\sharp)$ be any feasible direction. Then $\nabla g_i(\mathbf{x}^\sharp) \cdot \mathbf{d} = 0$ for all i, and $\nabla h_j(\mathbf{x}^\sharp) \cdot \mathbf{d} \le 0$ for $j \in J_0(\mathbf{x}^\sharp)$. Moreover, complementary slackness gives $\mu_j = 0$ for inactive constraints, so stationarity implies $\nabla f(\mathbf{x}^\sharp) \cdot \mathbf{d} \ge 0$. Theorem 3.18 then yields optimality. □

The KKT system can also be read directly from the Lagrangian dual viewpoint: stationarity means that $\mathbf{x}^\sharp$ minimizes $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$; primal feasibility is the original constraint system; dual feasibility is $\boldsymbol{\mu} \ge 0$; and complementary slackness expresses that each inequality constraint is either active ($h_j(\mathbf{x}^\sharp) = 0$) or has zero multiplier ($\mu_j = 0$).

♢ **Theorem 3.25** (Strong duality, saddle points, and KKT (Slater))**.** *Assume that* (3.2) *admits a Slater point and that the primal admits at least one optimal solution. Let $\mathbf{x}^\sharp \in X$. Then the following assertions are equivalent:*

1. *$\mathbf{x}^\sharp$ is an optimal solution of the primal problem;*
2. *there exist multipliers $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \in \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+$ such that $(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ is a saddle point of the Lagrangian $\mathcal{L}$ on $\mathbb{R}^n \times (\mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+)$;*

3. *there exist multipliers* $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \in \mathbb{R}^{n_E} \times \mathbb{R}^{n_I}_+$ *such that* $(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ *satisfies the KKT conditions.*

Moreover, in this case the duality gap is zero ($p^\star = d^\star$) *and* $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ *is an optimal solution of the dual problem.*

Slater's condition gives necessity of KKT, convexity makes KKT sufficient, and the KKT conditions are precisely the first-order expression of the Lagrangian saddle-point property.

Proof. (1)⇒(3). If $\mathbf{x}^\sharp$ is optimal, Slater's condition implies constraint qualification at $\mathbf{x}^\sharp$, hence KKT multipliers exist by Theorem 3.24.

(3)⇒(1). This is the sufficiency part of Theorem 3.24.

(3)⇒(2). Let $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ satisfy KKT at $\mathbf{x}^\sharp$. Since f and the h_j are convex, the g_i are affine, and $\boldsymbol{\mu}^\sharp \geq 0$, the map $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ is convex. Stationarity gives $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = 0$, hence $\mathbf{x}^\sharp$ minimizes $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ by Proposition 3.19, i.e. $\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \leq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ for all $\mathbf{x}$. On the other hand, for any multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ with $\boldsymbol{\mu} \geq 0$, primal feasibility gives

$$\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}^\sharp) + \boldsymbol{\mu}^\top h(\mathbf{x}^\sharp) \leq f(\mathbf{x}^\sharp) = \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp),$$

where the last equality follows from complementary slackness. Therefore $(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ is a saddle point.

(2)⇒(3). Let $(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ be a saddle point with $\boldsymbol{\mu}^\sharp \geq 0$. Moreover, if some equality constraint were violated at $\mathbf{x}^\sharp$ (say $g_i(\mathbf{x}^\sharp) \neq 0$), then letting $\lambda_i \to \pm\infty$ would make $\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}, \boldsymbol{\mu}^\sharp)$ arbitrarily large, contradicting the saddle inequality; hence $g(\mathbf{x}^\sharp) = 0$. Similarly, if some inequality constraint were violated (say $h_j(\mathbf{x}^\sharp) > 0$), then letting $\mu_j \to +\infty$ would make $\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu})$ arbitrarily large, again contradicting the saddle inequality; hence $h(\mathbf{x}^\sharp) \leq 0$. Since $\mathbf{x}^\sharp$ minimizes $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ and this function is differentiable, one has $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = 0$. Finally, combining $h(\mathbf{x}^\sharp) \leq 0$ with the saddle inequality at fixed $\mathbf{x}^\sharp$ yields $\boldsymbol{\mu}^{\sharp\top} h(\mathbf{x}^\sharp) = 0$, i.e. complementary slackness.

Strong duality and dual optimality. Let $\mathbf{x}^\sharp$ be primal optimal and let $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ satisfy KKT (which exists by (1)⇒(3)). The argument above shows that $\mathbf{x}^\sharp$ minimizes $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$, hence

$$\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = \inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = d(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) \leq d^\star.$$

But $\mathcal{L}(\mathbf{x}^\sharp, \boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = f(\mathbf{x}^\sharp) = p^\star$, so $p^\star \leq d^\star$, which together with weak duality $d^\star \leq p^\star$ gives $p^\star = d^\star$. Therefore $d(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp) = d^\star$ and $(\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ is dual optimal. □

3.3.1 Marginal interpretation of multipliers

Consider the perturbed inequality constraints $h(\mathbf{x}) \leq q$ for $q \in \mathbb{R}^{n_I}$, and define the associated value function

$$v(q) := \operatorname*{Min}_{\mathbf{x} \in \mathbb{R}^n} \{ f(\mathbf{x}) \mid g(\mathbf{x}) = 0,\ h(\mathbf{x}) \leq q \}.$$

♢ **Proposition 3.26** (Lower affine support)**.** *Assume that the original problem* $v(0)$ *is convex and satisfies Slater's condition, and let* $\boldsymbol{\mu}^\sharp$ *be an optimal multiplier (associated with the constraints* $h(\mathbf{x}) \leq 0$*). Then, for all* $q \in \mathbb{R}^{n_I}$*, one has*

$$v(q) \geq v(0) - (\boldsymbol{\mu}^\sharp)^\top q.$$

In particular, if v *is differentiable at* 0*, then* $\nabla v(0) = -\boldsymbol{\mu}^\sharp$*.*

Proof. Fix any q and any dual-feasible multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ with $\boldsymbol{\mu} \geq 0$. For every feasible $\mathbf{x}$ of the perturbed problem ($g(\mathbf{x}) = 0$, $h(\mathbf{x}) \leq q$), one has

$$f(\mathbf{x}) \geq f(\mathbf{x}) + \boldsymbol{\lambda}^\top g(\mathbf{x}) + \boldsymbol{\mu}^\top (h(\mathbf{x}) - q).$$

Taking the infimum over all such feasible $\mathbf{x}$ yields

$$v(q) \geq \inf_{\substack{\mathbf{x} \in \mathbb{R}^n \\ g(\mathbf{x}) = 0,\ h(\mathbf{x}) \leq q}} \big(f(\mathbf{x}) + \boldsymbol{\lambda}^\top g(\mathbf{x}) + \boldsymbol{\mu}^\top h(\mathbf{x}) \big) - \boldsymbol{\mu}^\top q.$$

Relaxing the constraints can only decrease the infimum, so

$$v(q) \geq \inf_{\mathbf{x} \in \mathbb{R}^n} \big(f(\mathbf{x}) + \boldsymbol{\lambda}^\top g(\mathbf{x}) + \boldsymbol{\mu}^\top h(\mathbf{x}) \big) - \boldsymbol{\mu}^\top q.$$

Now take $(\boldsymbol{\lambda}, \boldsymbol{\mu}) = (\boldsymbol{\lambda}^\sharp, \boldsymbol{\mu}^\sharp)$ optimal for the original problem. Strong duality (Theorem 3.25) gives

$$v(0) = \inf_{\mathbf{x}\in\mathbb{R}^n} \big(f(\mathbf{x}) + (\boldsymbol{\lambda}^\sharp)^\top g(\mathbf{x}) + (\boldsymbol{\mu}^\sharp)^\top h(\mathbf{x})\big),$$

which implies $v(q) \geq v(0) - (\boldsymbol{\mu}^\sharp)^\top q$. If v is differentiable at 0, then for any direction $u \in \mathbb{R}^{n_I}$ and any $t \in \mathbb{R}$, $v(tu) \geq v(0) - t(\boldsymbol{\mu}^\sharp)^\top u$. Dividing by t and letting $t \to 0^+$ and $t \to 0^-$ yields $\nabla v(0) \cdot u = -(\boldsymbol{\mu}^\sharp)^\top u$ for all u, hence $\nabla v(0) = -\boldsymbol{\mu}^\sharp$. □

The minus sign in $\nabla v(0) = -\boldsymbol{\mu}^\sharp$ reflects the convention $h(\mathbf{x}) \leq q$: increasing q relaxes the constraints, so the optimal value of a minimization problem can decrease.

For equality constraints, the sensitivity sign likewise depends on the chosen convention. If the perturbation is written $g(\mathbf{x}) = q$ and

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; q) = f(\mathbf{x}) + \boldsymbol{\lambda}^\top (g(\mathbf{x}) - q),$$

then, under the usual regularity and differentiability assumptions, $\nabla v(q) = -\boldsymbol{\lambda}^\sharp$. Writing the same equality as $q - g(\mathbf{x}) = 0$ gives a multiplier with the opposite sign. In Exercise 3.10, the balance equation is $D - p_1 - p_2 - p_3 = 0$, so this latter convention yields $v'(D) = \lambda$ whenever the value function is differentiable and the active set is locally unchanged.

3.4 Algorithms

This final section gives a brief algorithmic complement to the theory developed so far. In many convex problems, even when existence and uniqueness are well understood, an optimal solution has no closed-form expression and must be approximated numerically. Iterative methods therefore return a sequence of points and stop once a prescribed accuracy level ε is reached.

Besides a candidate primal point, it is often useful to monitor quantities that play the role of certificates: objective decrease, a measure of stationarity or feasibility, and sometimes a dual lower bound (hence a duality gap) when such a bound is available.

3.4.1 Descent-direction methods

Many algorithms for smooth optimization fit into the following template. At an iterate $\mathbf{x}^k$, one chooses a direction $\mathbf{d}^k$ and a step size $\tau^k > 0$, and updates

$$\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k + \tau^k \mathbf{d}^k.$$

A vector $\mathbf{d}^k$ is a *descent direction* at $\mathbf{x}^k$ if $\nabla f(\mathbf{x}^k) \cdot \mathbf{d}^k < 0$.

The step size controls both robustness and speed. Three classical choices are: a *fixed step* $\tau^k = \tau$ (often based on smoothness information such as a Lipschitz constant of ∇f); an *exact line search* $\tau^k \in \arg\min_{\tau>0} f(\mathbf{x}^k + \tau \mathbf{d}^k)$ when this one-dimensional problem is tractable; and *backtracking* (Armijo-type) rules that start from a candidate step and decrease it until a sufficient decrease condition holds. Stopping tests are similarly flexible; typical ones involve stationarity (e.g. $\|\nabla f(\mathbf{x}^k)\| \leq \varepsilon$ in the unconstrained case) or a small change between consecutive iterates.

3.4.2 Fixed-step gradient method

The simplest choice is $\mathbf{d}^k = -\nabla f(\mathbf{x}^k)$ with a constant step size $\tau > 0$. When f is smooth (for instance when ∇f is L-Lipschitz), taking τ small enough guarantees that each iteration decreases the objective, and one can quantify the convergence rate (stated later in this section).

Algorithm 1: Gradient method (fixed step size)

Data: Initial point $\mathbf{x}^0$, accuracy $\varepsilon > 0$, step size $\tau > 0$
Result: Approximate solution
$k \leftarrow 0$;
while $\|\nabla f(\mathbf{x}^k)\| > \varepsilon$ **do**
  $\mathbf{x}^{k+1} \leftarrow \mathbf{x}^k - \tau \nabla f(\mathbf{x}^k)$;
  $k \leftarrow k + 1$;
return $\mathbf{x}^k$

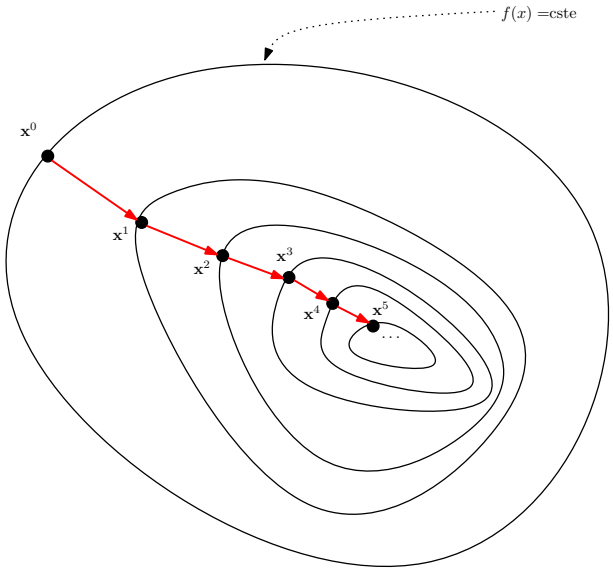

Figure 3.4: A few iterations of the gradient method on a two-dimensional objective.

The gradient method is robust but can be slow; it nevertheless plays a central role as a building block (and a fallback step) for more advanced methods. In particular, it already provides the basic $O(1/k)$ rate for smooth convex minimization.

3.4.3 Projection on a closed convex set

Definition 3.27 (Projection)**.** *Let $K \subseteq \mathbb{R}^n$ be nonempty, closed, and convex. For fixed $\mathbf{x} \in \mathbb{R}^n$, the function $\mathbf{z} \mapsto \|\mathbf{z} - \mathbf{x}\|^2$ is strongly convex and continuous. Proposition 3.22 therefore guarantees that it has a unique minimizer on K. The* projection *of $\mathbf{x}$ onto K is this point:*

$$P_K(\mathbf{x}) := \operatorname*{arg\,min}_{\mathbf{z} \in K} \|\mathbf{z} - \mathbf{x}\|^2.$$

Hence $P_K : \mathbb{R}^n \to K$ is well-defined.

♢ **Lemma 3.28.** *Let K be closed and convex. For $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in K$, one has $\mathbf{y} = P_K(\mathbf{x})$ if and only if*

$$(\mathbf{y} - \mathbf{x})^\top (\mathbf{z} - \mathbf{y}) \geq 0 \qquad \forall \mathbf{z} \in K.$$

♢ **Lemma 3.29.** *If K is closed and convex, then P_K is 1-Lipschitz: $\|P_K(\mathbf{x}) - P_K(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y}$.*

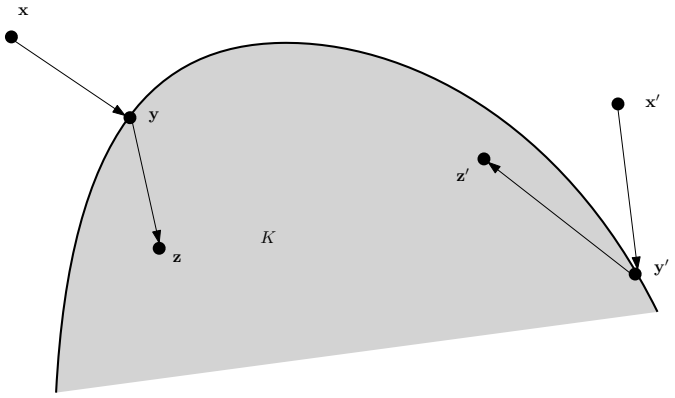

Figure 3.5: Projection on a closed convex set: $\mathbf{y} = P_K(\mathbf{x})$, whereas $\mathbf{y}'$ is not the projection of $\mathbf{x}'$.

3.4.4 Projected gradient method

The projected gradient method applies to minimizing a differentiable convex function over a closed convex set X:

$$\operatorname*{Min}_{\mathbf{x}\in X} f(\mathbf{x}).$$

It updates by a gradient step followed by a projection.

Algorithm 2: Projected gradient method

Data: Accuracy $\varepsilon > 0$, initial point $\mathbf{x}^0 \in X$, step size $\tau > 0$
Result: Approximate solution

$k \leftarrow 0$;
while $\|P_X(\mathbf{x}^k - \tau\nabla f(\mathbf{x}^k)) - \mathbf{x}^k\| > \varepsilon$ **do**
  $\mathbf{x}^{k+1/2} \leftarrow \mathbf{x}^k - \tau\nabla f(\mathbf{x}^k)$;
  $\mathbf{x}^{k+1} \leftarrow P_X(\mathbf{x}^{k+1/2})$;
  $k \leftarrow k + 1$;
return $\mathbf{x}^k$

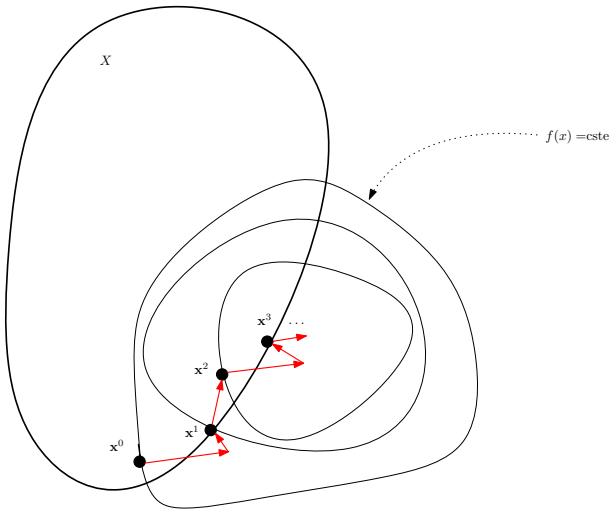

Figure 3.6: One step of projected gradient descent: a gradient step followed by a projection on the feasible set.

This method can be very effective when the projection P_X is easy to compute (e.g. boxes, balls, half-spaces), but it may be as hard as the original problem for more complex sets (e.g. general polyhedra), since computing P_X amounts to solving a constrained quadratic program.

3.4.5 Convergence rates for (projected) gradient descent

We now state the classical convergence rates for smooth convex problems. Throughout, we assume that f is differentiable and that its gradient is L-Lipschitz on $\mathbb{R}^n$, i.e.

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\| \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

♢ **Theorem 3.30** (Projected gradient: sublinear and linear rates)**.** *Let $X \subseteq \mathbb{R}^n$ be nonempty, closed, and convex, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, with L-Lipschitz gradient. Let $(\mathbf{x}^k)$ be generated by*

$$\mathbf{x}^{k+1} = P_X\big(\mathbf{x}^k - \tau \nabla f(\mathbf{x}^k)\big) \qquad (k \ge 0),$$

with a constant step size $\tau \in (0, 1/L)$. Assume that the problem $\min_{\mathbf{x}\in X} f(\mathbf{x})$ admits at least one minimizer $\mathbf{x}^\sharp$. Then, for every $k \ge 1$,

$$f(\mathbf{x}^k) - f(\mathbf{x}^\sharp) \le \frac{\|\mathbf{x}^0 - \mathbf{x}^\sharp\|^2}{2\tau k}.$$

If, in addition, f is α-strongly convex on X for some $\alpha > 0$ and $\tau \in \big(0, \frac{2\alpha}{L^2}\big)$, then the minimizer $\mathbf{x}^\sharp$ is unique and

$$\|\mathbf{x}^k - \mathbf{x}^\sharp\| \le \big(1 - 2\tau\alpha + \tau^2 L^2\big)^{k/2} \, \|\mathbf{x}^0 - \mathbf{x}^\sharp\| \qquad (k \ge 0).$$

Proof sketch. Fix a minimizer $\mathbf{x}^\sharp \in X$. The Lipschitz property of ∇f implies the standard descent inequality

$$f(\mathbf{y}) \le f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

Apply it with $\mathbf{x} = \mathbf{x}^k$ and $\mathbf{y} = \mathbf{x}^{k+1}$, then use convexity to bound $f(\mathbf{x}^k)$ from above by $f(\mathbf{x}^\sharp) + \nabla f(\mathbf{x}^k)^\top(\mathbf{x}^k - \mathbf{x}^\sharp)$. Finally, use Lemma 3.28 with $\mathbf{x} = \mathbf{x}^k - \tau\nabla f(\mathbf{x}^k)$ and $\mathbf{z} = \mathbf{x}^\sharp$ to obtain

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^\sharp) \le \frac{1}{2\tau}\Big(\|\mathbf{x}^k - \mathbf{x}^\sharp\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\sharp\|^2\Big),$$

which telescopes to yield the $O(1/k)$ bound.

For the strongly convex case, define $\theta(\mathbf{x}) := \mathbf{x} - \tau\nabla f(\mathbf{x})$. Using Proposition 3.17 and the Lipschitz property of ∇f gives $\|\theta(\mathbf{x}) - \theta(\mathbf{y})\| \le \sqrt{1 - 2\tau\alpha + \tau^2 L^2}\,\|\mathbf{x} - \mathbf{y}\|$. Since P_X is 1-Lipschitz (Lemma 3.29) and $\mathbf{x}^\sharp$ is a fixed point of $P_X \circ \theta$, one gets the contraction estimate on $\|\mathbf{x}^k - \mathbf{x}^\sharp\|$. □

♢ **Corollary 3.31** (Gradient method (rates))**.** *Under the assumptions of Theorem 3.30, taking $X = \mathbb{R}^n$ yields the fixed-step gradient method. In particular, for $\tau \in (0, 1/L)$ and any minimizer $\mathbf{x}^\sharp$ of f,*

$$f(\mathbf{x}^k) - f(\mathbf{x}^\sharp) \le \frac{\|\mathbf{x}^0 - \mathbf{x}^\sharp\|^2}{2\tau k} \qquad (k \ge 1).$$

If f is α-strongly convex and

$$\tau \in \left(0, \min\left\{\frac{1}{L}, \frac{2\alpha}{L^2}\right\}\right),$$

then $\mathbf{x}^\sharp$ is unique and $\|\mathbf{x}^k - \mathbf{x}^\sharp\|$ converges to 0 at a linear rate.

This algorithmic section is only a brief introduction to the vast landscape of convex optimization algorithms. Even the term "gradient method" refers to many variants (adaptive steps, line-search, accelerated methods, etc.). For constrained problems, other approaches can be particularly effective, such as Frank–Wolfe [2, Chapter 3] or interior-point methods [1, Chapter 11].

3.5 Exercises

Convexity

Exercise 3.1

Under what condition is the function $\mathbf{x} \mapsto \mathbf{x}^\top M \mathbf{x}$ convex?

Solution

Exercise 3.2 *Least-squares regression and an outlier*

Consider the observations

$$(0,1), \quad (1,3), \quad (2,5), \quad (3,7), \quad (4,15).$$

For an affine model $y = ax + b$, define

$$F(a,b) = \sum_{i=1}^{5} (ax_i + b - y_i)^2.$$

1. Write this problem as $\min_\theta \|A\theta - \mathbf{y}\|_2^2$, with $\theta = (a,b)$.
2. Explain why the objective is convex. Is its minimizer unique?
3. Derive the normal equations.
4. Compute the fitted line.
5. Compute the residuals and the optimal squared error. Comment briefly on the influence of the final observation.

Solution

Exercise 3.3

Let $C \subset \mathbb{R}^n$ be convex. Show that $f : C \to \mathbb{R}$ is convex if and only if, for every $\mathbf{x} \in C$ and every $\mathbf{d} \in \mathbb{R}^n$ such that $\mathbf{x} + \mathbf{d} \in C$, the function $g : t \in [0,1] \mapsto f(\mathbf{x} + t\mathbf{d})$ is convex.

Solution

Exercise 3.4

Using the concavity of the logarithm, prove that the arithmetic mean is always greater than or equal to the geometric mean.

Solution

Exercise 3.5

Show that the function

$$f(\mathbf{x}) = \ln\left(\sum_{i=1}^{n} e^{x_i}\right)$$

is convex.

Solution

Convex optimization

Exercise 3.6
Solve

$$\begin{array}{ll} \text{Min} & x_1^2 + x_2^2 - 6x_1 - 2x_2 + 10 \\ \text{s.t.} & 2x_1 + x_2 - 2 \leq 0 \\ & x_2 - 1 \leq 0 \\ & x_1, x_2 \in \mathbb{R}\,. \end{array}$$

Solution

Exercise 3.7
Solve

$$\begin{array}{ll} \text{Min} & -2x_1 + x_2 \\ \text{s.t.} & x_1^2 - x_2 \leq 0 \\ & x_2 - 4 \leq 0 \\ & x_1, x_2 \in \mathbb{R}\,. \end{array}$$

Solution

Exercise 3.8
Let M be symmetric positive definite and let $\mathbf{c}$ and $\mathbf{a}$ be two vectors. Solve

$$\begin{array}{ll} \text{Min} & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} & (\mathbf{x} - \mathbf{a})^\top M (\mathbf{x} - \mathbf{a}) \leq 1\,. \end{array}$$

Solution

Exercise 3.9 *A scarce-resource procurement problem*
An industry plans its annual procurement of scarce resources. For each resource $i \in [m]$, the yearly need is $a_i > 0$. Each resource can be shipped from several regions. The more a region is used, the longer it takes to deliver: if x_{ij} denotes the quantity of resource i shipped from region j, then the delivery time is $c_j x_{ij}^2$, with $c_j > 0$.

We want to minimize the maximum delivery time among all resources/regions.

1. Explain why this can be modeled as

$$\begin{array}{lll} \text{Min} & \max_{i\in[m],\, j\in[n]} c_j x_{ij}^2 & \\ \text{s.t.} & \sum_{j=1}^n x_{ij} = a_i & \forall i \in [m] \\ & x_{ij} \geq 0 & \forall i \in [m],\ \forall j \in [n]\,. \end{array}$$

2. Reformulate the problem as a convex optimization problem without the max (introduce an auxiliary variable and additional constraints).

3. Write the dual of your reformulation, using multipliers λ_i for the equalities $\sum_j x_{ij} = a_i$, ω_{ij} for the nonnegativity constraints $x_{ij} \geq 0$, and μ_{ij} for your new constraints.

4. Set $\mu_{ij} = 1/(mn)$ and $\omega_{ij} = 0$. Show that the optimal value is bounded below by

$$\frac{\sum_{i=1}^m a_i^2}{mn \sum_{j=1}^n \frac{1}{c_j}}\,.$$

Solution

Exercise 3.10 *Economic dispatch and marginal electricity prices*
Three generators produce powers p_1, p_2, p_3, with costs

$$C_1(p_1) = \frac{1}{2}p_1^2 + 2p_1, \qquad C_2(p_2) = \frac{1}{4}p_2^2 + 4p_2, \qquad C_3(p_3) = \frac{1}{6}p_3^2 + 6p_3.$$

Their capacities are

$$0 \le p_1 \le 4, \qquad 0 \le p_2 \le 6, \qquad 0 \le p_3 \le 10,$$

and demand D must be met exactly.

1. Formulate the cost-minimization problem, writing the balance equation as $D - p_1 - p_2 - p_3 = 0$.
2. Explain why this is a convex optimization problem and why its solution is unique whenever the feasible set is nonempty.
3. Write the KKT conditions, including multipliers for the lower and upper generation bounds.
4. Solve the dispatch for $D = 6$.
5. Solve it again for $D = 10$.
6. For each demand level, identify which generators operate strictly inside their bounds, which are off or capacity-constrained, and compute the multiplier λ of the balance constraint.
7. Let $v(D)$ be the minimum generation cost. Using the marginal interpretation of multipliers (see Proposition 3.26), explain why $v'(D) = \lambda$ whenever v is differentiable and the active set is locally unchanged.
8. Interpret λ economically.

Solution

Exercise 3.11 *Electrical networks, energy minimization, and Kirchhoff's law*
Consider electrical nodes $0, 1, 2, 3$, with fixed boundary voltages $v_0 = 10$ and $v_3 = 0$. The unknown voltages are v_1, v_2. There are resistors on the edges

$$r_{01} = 1, \qquad r_{02} = 2, \qquad r_{12} = 1, \qquad r_{13} = 2, \qquad r_{23} = 1.$$

For each edge $\{i, j\}$, let $\sigma_{ij} = 1/r_{ij}$ be its conductance and define the Dirichlet energy (one half of the total dissipated power)

$$E(v_1, v_2) = \frac{1}{2} \sum_{\{i,j\}} \sigma_{ij}(v_i - v_j)^2,$$

where the sum is over the five listed edges.

1. Write $E(v_1, v_2)$ explicitly.
2. Show that E is strictly convex. You may compute its Hessian.
3. Deduce that E has a unique minimizer.
4. Compute the first-order optimality conditions.

5. Show that the condition at each internal node $i = 1, 2$ has the form

$$\sum_{j:\{i,j\} \text{ is an edge}} \sigma_{ij}(v_i - v_j) = 0.$$

6. Define the current from i to j by Ohm's law, $I_{ij} = (v_i - v_j)/r_{ij}$. Interpret the stationarity equations.

7. Compute the internal voltages and all edge currents, using the direction from the first listed node to the second. Verify current conservation at both internal nodes.

Solution

Algorithms in convex optimization

Exercise 3.12 *A production planning problem*
Consider the optimization problem

$$\begin{array}{lll} \text{Min} & a\displaystyle\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} b_k \frac{1}{x_k} & \\ \text{s.t.} & \displaystyle\sum_{k=1}^{n} x_k \leq N & \\ & x_k > 0 \qquad & k \in [n]\,, \end{array}$$

where $a > 0$, $N > 0$, and $b_k > 0$.

Find an optimal production plan. Is it unique?

Solution

Exercise 3.13 *Marginal interpretation of multipliers*
Consider the parametric problem

$$\begin{array}{lll} v(\mathbf{u}) = \text{Min} & f(\mathbf{x}) & \\ \text{s.t.} & h_j(\mathbf{x}) \leq u_j \qquad & j \in [q]\,, \end{array}$$

where $\mathbf{u} \in \mathbb{R}^q$ is a parameter vector. Assume f and the h_j are convex and differentiable, constraints are qualified, and the optimal value is always finite.

1. Show that v is convex.

2. Let $\boldsymbol{\mu}^0$ be the vector of Lagrange multipliers associated with an optimal solution at $\mathbf{u} = \mathbf{0}$.
   (a) Show that $v(\mathbf{u}) \geq v(\mathbf{0}) - \boldsymbol{\mu}^0 \cdot \mathbf{u}$.
   (b) Conclude that if v is differentiable at $\mathbf{0}$, then $\nabla v(\mathbf{0}) = -\boldsymbol{\mu}^0$.

Solution

Exercise 3.14
Consider the unconstrained problem

$$\text{Min } f(\mathbf{x}) = \frac{1}{2}\left(x_1^2 + \gamma x_2^2\right),$$

where $\gamma > 0$ is given. Describe the iterates of the (constant-step) gradient method starting from $\mathbf{x}^0 = (\gamma, 1)^\top$, and verify convergence in this case.

Solution

Chapter 4

Linear optimization

Linear programs arise naturally in a wide range of contexts and share, like convex optimization, remarkable structural properties. One may wonder why we devote a separate chapter to this particular subclass of convex problems. One reason is that the geometry of feasible sets becomes especially transparent: feasible regions are polyhedra, and optimal solutions can be sought among their vertices. Another reason is algorithmic: many results from Chapter 3 (in particular, the convergence guarantees for first-order methods) rely on strong convexity, while linear objectives are not even strictly convex. Nevertheless, linear programs can be solved extremely efficiently in practice, most notably with the simplex method. Among modern references, the short book by Matoušek and Gärtner [5] provides a very accessible account of the theory and many applications.

4.1 Geometric interpretation

We start from the geometry of feasible sets (polyhedra) and the key consequence for linear objectives: when an optimum exists, it can be found at a vertex.

4.1.1 Polyhedra

♡ **Definition 4.1** (Half-spaces, polyhedra, polytopes)**.** *A* (closed) half-space *of* $\mathbb{R}^n$ *is a set of the form*

$$H^- = \{\mathbf{x} \in \mathbb{R}^n \mid a^\top \mathbf{x} \leq b\},$$

where $a \in \mathbb{R}^n \setminus \{0\}$ *and* $b \in \mathbb{R}$*. A* polyhedron *is an intersection of finitely many closed half-spaces. A* polytope *is a bounded polyhedron.*

Throughout this chapter, it is helpful to keep in mind a simple two-dimensional example.

Example 4.1 (Running example: a triangle). Consider the set $P \subseteq \mathbb{R}^2$ defined by

$$P = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq 1\}.$$

This is a polytope (a bounded polyhedron). Its vertices are $(0, 0)$, $(1, 0)$, and $(0, 1)$, and its edges are the three line segments joining these points. For instance, the edge $\{(x_1, x_2) \mid x_1 \geq 0,\ x_2 = 0,\ x_1 \leq 1\}$ is the face obtained by intersecting P with the supporting line $\{(x_1, x_2) \mid x_2 = 0\}$.

The running example is illustrated in Figure 4.1.

Definition 4.2 (Supporting hyperplanes and faces)**.** *Let* $P \subseteq \mathbb{R}^n$ *be a polyhedron. A hyperplane* $H = \{\mathbf{x} \mid a^\top \mathbf{x} = \beta\}$ *is* supporting *for* P *if* $P \subseteq \{\mathbf{x} \mid a^\top \mathbf{x} \leq \beta\}$ *and* $P \cap H \neq \emptyset$*. A (nonempty)* face *of* P *is a set of the form* $F = P \cap H$ *for some supporting hyperplane* H*.*

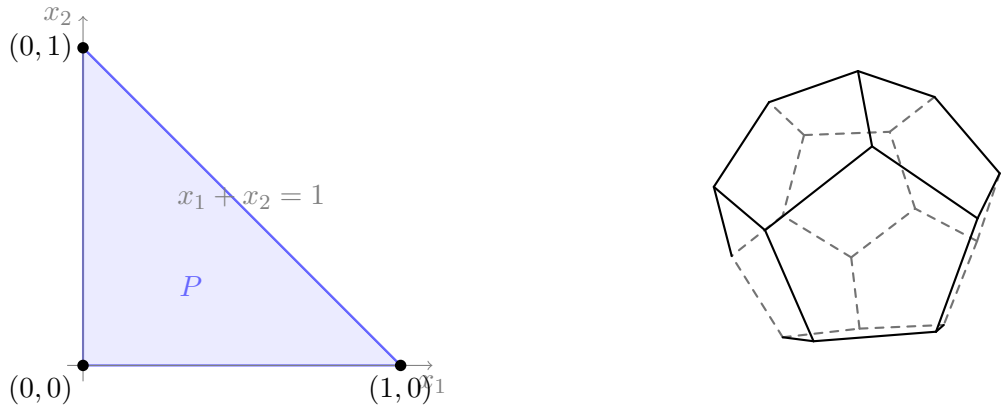

Figure 4.1: Left: the triangle polytope of Example 4.1. Right: a three-dimensional polyhedron (dodecahedron, illustration).

We use $\dim(F)$ for the affine dimension of F.

A *vertex* is a face of dimension 0 (a singleton), and an *edge* is a face of dimension 1 (a line segment, a ray, or a line). Two vertices are *adjacent* if they are connected by an edge.

Remark 4.1. For polyhedra, a point is a vertex (a 0-dimensional face) if and only if it is an extreme point (*i.e.,* it is not a strict convex combination of two distinct points of the polyhedron).

♡ **Proposition 4.3** (Vertex characterizations)**.** *Let $P \subseteq \mathbb{R}^n$ be a polyhedron and $\mathbf{x} \in P$. The following are equivalent:*

1. $\mathbf{x}$ *is a vertex of* P.
2. *There exists* $\mathbf{c} \in \mathbb{R}^n$ *such that* $\mathbf{x}$ *is the unique solution of* $\min\{\mathbf{c}^\top \mathbf{y} : \mathbf{y} \in P\}$.
3. *Writing* $P = \{\mathbf{y} \in \mathbb{R}^n : A\mathbf{y} \leq \mathbf{b}\}$ *and* $I(\mathbf{x}) := \{i : (A\mathbf{x})_i = b_i\}$, *one has* $\operatorname{rank}(A_{I(\mathbf{x})}) = n$.

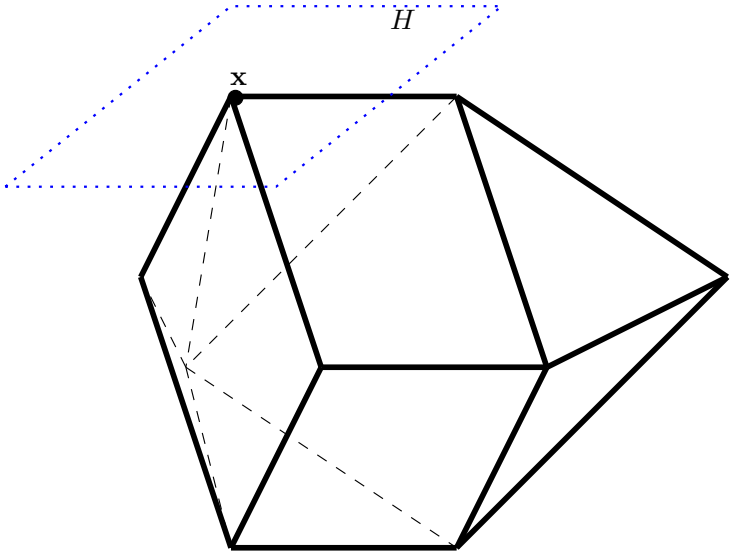

Figure 4.2: A point is a vertex if it is the unique intersection of the polyhedron with some supporting hyperplane.

Polyhedral geometry is a major branch of convex geometry with many applications in optimization, combinatorics, and game theory.

4.1.2 ✠ H- and V-representations of polyhedra

Definition 4.1 corresponds to an *H-representation* (half-space description). Polyhedra also admit *V-representations*, which separate bounded and unbounded directions. Here, for a finite set $V = \{\mathbf{v}_1, \dots, \mathbf{v}_k\}$, we write

$$\operatorname{conv}(V) := \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{v}_i \mid \lambda_i \geq 0,\ \sum_{i=1}^{k} \lambda_i = 1 \right\},$$

and for a finite set of directions $D = \{\mathbf{d}_1, \dots, \mathbf{d}_\ell\}$,

$$\operatorname{cone}(D) := \left\{ \sum_{j=1}^{\ell} \mu_j \mathbf{d}_j \mid \mu_j \geq 0 \right\}.$$

♢ **Theorem 4.4** (Minkowski–Weyl theorem)**.** *A nonempty set $P \subseteq \mathbb{R}^n$ is a polyhedron if and only if there exist points $v_1, \dots, v_k \in \mathbb{R}^n$ and directions $d_1, \dots, d_\ell \in \mathbb{R}^n$ such that*

$$P = \operatorname{conv}\{v_1, \dots, v_k\} + \operatorname{cone}\{d_1, \dots, d_\ell\}.$$

Proof. (Proof sketch.) *(⇐)* If $P = \operatorname{conv}(V) + \operatorname{cone}(D)$ with finite $V = \{v_i\}$ and $D = \{d_j\}$, then P is the image of the polyhedron

$$\{(\lambda, \mu) : \lambda \geq 0,\ \mathbf{1}^\top \lambda = 1,\ \mu \geq 0\}$$

under the affine map $(\lambda, \mu) \mapsto \sum_i \lambda_i v_i + \sum_j \mu_j d_j$. Since affine images of polyhedra are polyhedra, P is a polyhedron.

(⇒) If $P = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}\}$, its recession cone is the polyhedral cone $R := \{\mathbf{d} \in \mathbb{R}^n \mid A\mathbf{d} \leq 0\}$. One can show that $P = Q + R$, where Q is a (possibly lower-dimensional) polytope obtained by intersecting P with an appropriate supporting halfspace in directions transverse to R (equivalently, after factoring out the lineality space). Finally, R is a polyhedral cone, hence finitely generated: $R = \operatorname{cone}\{d_1, \dots, d_\ell\}$ for finitely many directions. Since Q is a polytope, $Q = \operatorname{conv}\{v_1, \dots, v_k\}$ for finitely many points. This yields the desired V-representation. □

The two descriptions are mathematically equivalent, but converting between them can be computationally expensive in the worst case (for instance, it may require exponentially many facets or vertices).

4.1.3 Linear programs

We first consider the inequality form

$$\operatorname*{Min}_{\mathbf{x} \in \mathbb{R}^n} \quad \mathbf{c}^\top \mathbf{x} \tag{4.1a}$$

$$\text{s.t.} \quad A\mathbf{x} \leq \mathbf{b} \tag{4.1b}$$

where $A \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^n$. The feasible set of (4.1) is a polyhedron, and linear optimization has a particularly pleasant geometric interpretation.

A fundamental result, formalized below, is that it suffices to look at vertices of the feasible polyhedron to find an optimal solution.

♡ **Theorem 4.5** (Optimal solutions at vertices)**.** *Assume that* (4.1) *has at least one feasible point and that the objective function is bounded below on the feasible set. Then* (4.1) *admits at least one optimal solution.*

Moreover, if the feasible polyhedron has at least one vertex, then there exists an optimal solution which is a vertex of the feasible polyhedron.

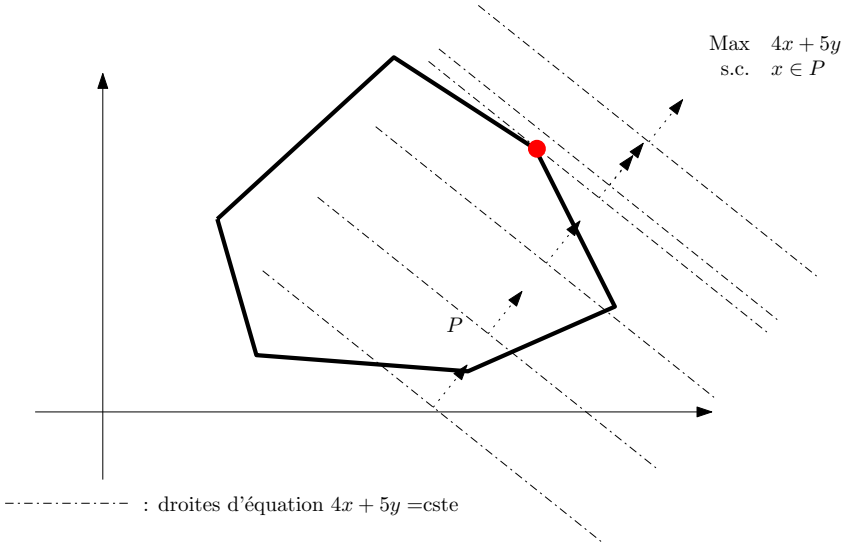

Figure 4.3: A geometric view of a linear objective over a polyhedron: an optimum (when it exists) can be reached at a vertex.

Proof sketch. Convert (4.1) to standard form (Section 4.2) and apply Theorem 4.6. The vertex statement follows since optimal basic feasible solutions correspond to vertices (Theorem 4.9). □

In principle, one could solve (4.1) by enumerating all vertices of the feasible polyhedron and selecting the best one. This is not efficient in practice because the number of vertices can be exponential in the problem size. The simplex method exploits Theorem 4.5 by moving along edges from vertex to vertex until reaching optimality.

4.2 Equivalent formulations and standing assumptions

We now introduce the standard forms used in theory and in algorithms, and state a standing full-row-rank assumption for the constraint matrix.

Problem (4.1) is in *inequality form*. Other standard forms are often used, such as the *canonical form*

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \mathbf{c}^\top \mathbf{x} && \text{(4.2a)} \\ \text{s.t.} \quad & A\mathbf{x} \le \mathbf{b} && \text{(4.2b)} \\ & \mathbf{x} \ge \mathbf{0} && \text{(4.2c)} \end{aligned}$$

and the *standard form*

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \mathbf{c}^\top \mathbf{x} && \text{(4.3a)} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} && \text{(4.3b)} \\ & \mathbf{x} \ge \mathbf{0} && \text{(4.3c)} \end{aligned}$$

Up to changing the dimension and redefining A, $\mathbf{b}$ and $\mathbf{c}$, any linear program written in one of these forms (inequality, canonical, standard) can be rewritten in any of the others; see exercises. For instance, a canonical-form problem can be put into standard form by introducing

slack variables:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \mathbf{c}^\top \mathbf{x} && \text{(4.4a)} \\ \text{s.t.} \quad & A\mathbf{x} + \mathbf{y} = \mathbf{b} && \text{(4.4b)} \\ & \mathbf{x} \geq \mathbf{0} && \text{(4.4c)} \\ & \mathbf{y} \geq \mathbf{0} && \text{(4.4d)} \end{aligned}$$

where the matrix (A, I) plays the role of the constraint matrix.

In the rest of this chapter, we focus on the standard form (4.3), and we denote by m the number of rows of A and by n its number of columns. If A does not have full row rank, then either the system $A\mathbf{x} = \mathbf{b}$ has no solution (so (4.3) is infeasible), or it has redundant equations that can be removed without changing the feasible set. We therefore assume for the remainder of the chapter:

$$\boxed{\operatorname{rank}(A) = m}.$$

In particular, $m \leq n$.

It is useful to keep in mind that (4.3) is a smooth constrained optimization problem with an affine feasible set. In particular, the constraints are qualified at every feasible point (Proposition 2.13), so any optimal solution satisfies the KKT conditions of Chapter 2 (Theorem 2.20). In linear programming, these optimality conditions can be reformulated as primal–dual feasibility plus complementary slackness, and they lead to strong duality results; see Section 4.5.

4.3 Existence and characterization of an optimal solution

We next connect the geometric notion of a vertex to an algebraic description through bases, which leads to a finite characterization of optimal solutions.

4.3.1 Bases and basic feasible solutions

For a subset $K \subseteq [n] := \{1, \ldots, n\}$, we denote by A_K the submatrix of A made of the columns $(A_i)_{i \in K}$. Similarly, for a vector $\mathbf{x} \in \mathbb{R}^n$, we denote by $\mathbf{x}_K$ the subvector $(\mathbf{x}_i)_{i \in K}$.

A subset $B \subseteq [n]$ is a *basis* of (4.3) if A_B is invertible. Note that this is not a basis in the linear-algebra sense for $\mathbb{R}^n$; rather, $(A_i)_{i \in B}$ is a basis of $\mathbb{R}^m$. Such a basis B is a *feasible basis* if, in addition, $A_B^{-1}\mathbf{b} \geq \mathbf{0}$.

Let B be a basis and let $N := [n] \setminus B$. Define $\mathbf{y} \in \mathbb{R}^n$ by

$$\mathbf{y}_B = A_B^{-1}\mathbf{b}, \qquad \mathbf{y}_N = \mathbf{0}.$$

Then $A\mathbf{y} = \mathbf{b}$, hence $\mathbf{y}$ automatically satisfies the equality constraints of (4.3). The vector $\mathbf{y}$ is called a *basic solution* (associated with B). If B is feasible, then $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y}$ is a feasible point of (4.3); it is then called a *basic feasible solution* (BFS).

4.3.2 Existence theorem

The following result is specific to linear programming. The first statement already fails for general convex optimization: a convex problem can be feasible and have a finite lower bound, yet fail to have any minimizer. For instance, minimizing y over $\{(x, y) \in (\mathbb{R}_+ \setminus \{0\})^2 \mid y \geq 1/x\}$ yields an infimum equal to 0 but no minimizer. It is also not a consequence of the coercivity-based existence theorem (Theorem 2.5), since $\mathbf{x} \mapsto \mathbf{c}^\top \mathbf{x}$ is not coercive.

♡ **Theorem 4.6** (Fundamental theorem of linear programming)**.** *Assume that* (4.3) *is feasible and that the objective is bounded below on the feasible set. Then* (4.3) *admits at least one optimal solution.*

Moreover, if an optimal solution exists, then there exists an optimal solution which is a basic feasible solution.

The basis associated with an optimal basic feasible solution is called an *optimal basis*.

The proof repeatedly moves along feasible null-space directions without worsening the objective, eliminating positive components until the remaining support is linearly independent and therefore defines a basic feasible solution.

Proof of Theorem 4.6. Let X be the feasible set of (4.3). We show that for any $\mathbf{x} \in X$, there exists a BFS $\tilde{\mathbf{x}}$ such that $\mathbf{c}^\top \tilde{\mathbf{x}} \le \mathbf{c}^\top \mathbf{x}$. Since the number of feasible bases is finite, this implies existence of an optimal BFS, hence the theorem.

Fix any $\mathbf{x} \in X$ and consider a feasible point $\tilde{\mathbf{x}} \in X$ satisfying $\mathbf{c}^\top \tilde{\mathbf{x}} \le \mathbf{c}^\top \mathbf{x}$ and having as many zero components as possible. Let $K := \{i \in [n] \mid \tilde{\mathbf{x}}_i > 0\}$. If the columns of A_K are linearly independent, then $\tilde{\mathbf{x}}$ is a BFS: one can extend K to a basis $B \supseteq K$ such that A_B is invertible, and then $\tilde{\mathbf{x}}$ is exactly the basic solution associated with B (with $\mathbf{x}_N = \mathbf{0}$).

Assume now that the columns of A_K are linearly dependent. Then there exists $\mathbf{y} \in \mathbb{R}^K$, $\mathbf{y} \neq 0$, such that $A_K \mathbf{y} = 0$. Define $\mathbf{z} \in \mathbb{R}^n$ by $\mathbf{z}_i = \mathbf{y}_i$ if $i \in K$ and $\mathbf{z}_i = 0$ otherwise. Then $A\mathbf{z} = 0$ and $\tilde{\mathbf{x}} + t\mathbf{z}$ remains feasible (i.e. $A(\tilde{\mathbf{x}} + t\mathbf{z}) = \mathbf{b}$ and $\tilde{\mathbf{x}} + t\mathbf{z} \ge 0$) exactly for t in an interval

$$J = \left[\max_{i \in I_+} \left(-\frac{\tilde{\mathbf{x}}_i}{\mathbf{z}_i} \right),\ \min_{i \in I_-} \left(-\frac{\tilde{\mathbf{x}}_i}{\mathbf{z}_i} \right) \right], \qquad I_+ = \{i \mid \mathbf{z}_i > 0\},\ I_- = \{i \mid \mathbf{z}_i < 0\}.$$

If I_+ (resp. I_-) is empty, the lower (resp. upper) endpoint of J is $-\infty$ (resp. $+\infty$).

If $\mathbf{c}^\top \mathbf{z} \neq 0$, we consider the endpoint of J whose sign is opposite to that of $\mathbf{c}^\top \mathbf{z}$. This endpoint must be finite; otherwise the objective value would go to $-\infty$ along the ray $\tilde{\mathbf{x}} + t\mathbf{z}$, contradicting boundedness from below. If $\mathbf{c}^\top \mathbf{z} = 0$, we choose any finite endpoint of J (at least one exists since I_+ and I_- cannot both be empty). At this endpoint, the point $\tilde{\mathbf{x}} + t\mathbf{z}$ has at least one more zero component than $\tilde{\mathbf{x}}$ while remaining feasible and not worsening the objective, contradicting the maximality of the number of zeros in $\tilde{\mathbf{x}}$. Therefore the columns of A_K must be linearly independent. □

Theorem 4.6 shows that, at least in principle, one can compute an exact optimal solution in finite time (when it exists) by enumerating all bases. Since the number of bases is bounded by $\binom{n}{m}$, one could loop over all subsets $B \subseteq [n]$ of size m, test whether A_B is invertible, then whether $A_B^{-1}\mathbf{b} \ge 0$, and keep the best objective value among the corresponding basic solutions. This is not practical for large instances: $\binom{n}{m}$ is typically huge (e.g. on the order of 4^m when $m \approx n/2$).

4.4 The simplex method

We can now describe the simplex method, which navigates along adjacent vertices by pivoting between bases. In standard form, feasible points are described by $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \ge \mathbf{0}$, and we saw that vertices correspond to basic feasible solutions. The simplex method exploits this structure by maintaining a feasible basis (hence a BFS) and iteratively moving to a neighboring BFS that improves the objective.

Two practical remarks are worth keeping in mind. First, the method needs an initial feasible basis; when none is obvious, one typically computes one by solving an auxiliary "Phase I" problem (see Section 4.2 and Section 4.3.1 for the relevant standard forms). Second, when the current BFS is *degenerate* (some basic components are zero), a pivot may change the basis without changing the point, and special care is needed to avoid cycling; this motivates pivot rules with finite-termination guarantees.

4.4.1 Preliminaries

The simplex method, proposed by Dantzig in 1947 (published later [3]), is an extremely effective algorithm for solving linear programs. Because it is widely implemented and heavily optimized in modern solvers, engineers and researchers rarely need to re-implement it. We therefore focus on the core principles, while also making explicit the basic computations behind one pivot step.

Let B be a basis, $N = [n] \setminus B$, and denote by $\tilde{\mathbf{x}}$ the associated basic solution. Define the *reduced cost* vector

$$\mathbf{r} := \mathbf{c}_N - A_N^\top (A_B^{-1})^\top \mathbf{c}_B \in \mathbb{R}^{n-m}.$$

Reduced costs quantify how the objective changes when increasing a nonbasic variable x_k away from 0 while preserving the equality constraints. In particular, in a minimization problem, a negative reduced cost $r_k < 0$ indicates that increasing x_k can improve the objective, at least locally along an edge of the feasible polyhedron.

♢ **Proposition 4.7** (A simplex pivot lemma)**.** *Assume that B is a feasible basis.*

1. *If $\mathbf{r} \geq \mathbf{0}$, then B is an optimal basis.*
2. *Let $k \in N$ be such that $r_k < 0$. Then exactly one of the following two cases holds:*
   (a) *there exists a feasible basis $B' \subseteq B \cup \{k\}$ with $B' \neq B$, and the BFS associated with B' is at least as good as the one associated with B;*
   (b) *there is no feasible basis $B' \subseteq B \cup \{k\}$ with $B' \neq B$, and (4.3) is unbounded below (its optimal value is $-\infty$).*

Proof of Proposition 4.7. Let $\tilde{\mathbf{x}}$ be the basic solution associated with B and take any feasible $\mathbf{x}$. Write $\mathbf{c}^\top \mathbf{x} = \mathbf{c}_B^\top \mathbf{x}_B + \mathbf{c}_N^\top \mathbf{x}_N$. Since $\mathbf{x}_B = A_B^{-1}\mathbf{b} - A_B^{-1}A_N\mathbf{x}_N$ and $\mathbf{c}^\top \tilde{\mathbf{x}} = \mathbf{c}_B^\top A_B^{-1}\mathbf{b}$, one obtains

$$\mathbf{c}^\top \mathbf{x} = \mathbf{c}^\top \tilde{\mathbf{x}} + \mathbf{r}^\top \mathbf{x}_N.$$

If $\mathbf{r} \geq 0$, this implies $\mathbf{c}^\top \mathbf{x} \geq \mathbf{c}^\top \tilde{\mathbf{x}}$ for all feasible $\mathbf{x}$, hence $\tilde{\mathbf{x}}$ is optimal.

Assume now that there exists $k \in N$ with $r_k < 0$. Consider the matrix $A_{B\cup\{k\}}$: its columns are linearly dependent, so there exists $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z} \neq 0$, supported on $B \cup \{k\}$, such that $A\mathbf{z} = 0$. Up to replacing $\mathbf{z}$ by $-\mathbf{z}$, we may assume $\mathbf{z}_k > 0$. Then $\tilde{\mathbf{x}} + t\mathbf{z}$ is feasible precisely for

$$t \in \left[0,\ \min_{j \in J}\left(-\frac{\tilde{\mathbf{x}}_j}{\mathbf{z}_j}\right)\right], \qquad J = \{j \in B \mid \mathbf{z}_j < 0\}.$$

Moreover, $\mathbf{c}^\top \mathbf{z} = r_k \mathbf{z}_k < 0$ (using $\mathbf{z}_B = -A_B^{-1}A_N\mathbf{z}_N$).

If J is nonempty, let $j^\star \in J$ attain the minimum ratio. Then $B' := B \cup \{k\} \setminus \{j^\star\}$ is a feasible basis, and the associated BFS is $\tilde{\mathbf{x}} - \frac{\tilde{\mathbf{x}}_{j^\star}}{\mathbf{z}_{j^\star}}\mathbf{z}$, which is at least as good as $\tilde{\mathbf{x}}$ because $\mathbf{c}^\top \mathbf{z} < 0$. This yields case 2a.

If J is empty, then $\tilde{\mathbf{x}} + t\mathbf{z}$ is feasible for all $t \geq 0$ and the objective decreases strictly along this ray, so (4.3) is unbounded below. This yields case 2b. □

4.4.2 Algorithm outline

Conceptually, assuming we are given an initial feasible basis B_0, the simplex method repeatedly performs the following steps. From the current feasible basis B (hence the current BFS $\tilde{\mathbf{x}}$), compute reduced costs $\mathbf{r}$. If $\mathbf{r} \geq \mathbf{0}$, Proposition 4.7 shows that $\tilde{\mathbf{x}}$ is optimal. Otherwise, select an entering index $k \in N$ with $r_k < 0$ (this choice is the *pivot rule*), and move along the corresponding feasible edge until one basic variable reaches 0; this ratio test identifies the leaving index and produces a new feasible basis. If the ratio test is unbounded (no basic variable blocks the move), then the objective can decrease indefinitely and the problem is unbounded below.

The pseudocode below summarizes this logic.

Algorithm 3: Simplex method (conceptual form)

Data: A feasible basis B_0
Result: The optimal value, and (if finite) an optimal basis

$B \leftarrow B_0$;
$N \leftarrow [n] \setminus B$;
compute $\mathbf{r} = \mathbf{c}_N - A_N^\top (A_B^{-1})^\top \mathbf{c}_B$;
while *there exists* $k \in N$ *with* $r_k < 0$ **do**
  Choose an entering index $k \in N$ with $r_k < 0$;
  if *no basic variable blocks the move in the entering direction* **then**
    return $-\infty$;
  else
    Perform a pivot step (ratio test) to obtain a new feasible basis $B' \subseteq B \cup \{k\}$ with $B' \neq B$;
    $B \leftarrow B'$;
    $N \leftarrow [n] \setminus B$;
    compute $\mathbf{r} = \mathbf{c}_N - A_N^\top (A_B^{-1})^\top \mathbf{c}_B$;
return $(\mathbf{c}_B^\top A_B^{-1}\mathbf{b},\ B)$;

In general, there are several choices of an entering index k with $r_k < 0$. The corresponding selection rule is called the *pivot rule*, and performance can depend strongly on this choice. In practice, reduced costs and basis updates are computed efficiently via Gaussian elimination (without explicitly inverting A_B at each iteration). Equivalently, one can view a pivot as a rank-one update of the factorization of A_B (e.g. via Sherman–Morrison-type formulas), which is the basis of fast implementations.

To ensure that the method always terminates (i.e. to avoid cycling in degenerate cases), one can use pivot rules with finite-termination guarantees; see, e.g., the classical work of Dantzig, Orden, and Wolfe [4].

♢ **Theorem 4.8** (Finite-termination pivot rules)**.** *There exist pivot rules for the simplex method that guarantee finite termination (no cycling), even in degenerate cases.*

The simplex method is extremely efficient on most instances but it is not a polynomial-time algorithm in the worst case (there are instances on which it takes an exponential number of pivots). This worst-case behavior is not inconsistent with practical performance: *smoothed analysis* shows that, under small random perturbations of the input data, the expected running time (or expected number of pivots) of certain simplex variants becomes polynomial in the dimension and in the inverse perturbation size [6]. At a high level, this suggests that worst-case hard instances are "fragile" and typically disappear under tiny noise, which better matches real-world data.
In contrast, linear programming can be solved in polynomial time by the ellipsoid method or by interior-point methods. In practice, modern solvers combine highly optimized simplex implementations with interior-point methods, depending on the structure and size of the instance.

4.4.3 One pivot step (ratio test and basis update)

Let B be a feasible basis and $\tilde{\mathbf{x}}$ the associated BFS. Choose an entering index $k \in N$ with $r_k < 0$. To keep the equality constraints satisfied, we vary x_k and compensate on the basic variables: the *pivot direction* $d \in \mathbb{R}^n$ is defined by

$$d_k = 1, \qquad d_i = 0,\ \forall i \in N \setminus \{k\}, \qquad d_B = -A_B^{-1} A_k.$$

Then $A(\tilde{\mathbf{x}}+td) = \mathbf{b}$ for all t, and feasibility of the nonnegativity constraints requires $\tilde{\mathbf{x}}_B + td_B \geq 0$ and $t \geq 0$. Therefore the admissible step sizes are exactly

$$t \in \left[0,\ \min_{i \in B \mid d_i < 0} \frac{\tilde{\mathbf{x}}_i}{-d_i}\right].$$

If the set $\{i \in B \mid d_i < 0\}$ is empty, then $\tilde{\mathbf{x}} + td$ remains feasible for all $t \geq 0$ and the objective decreases strictly since

$$\mathbf{c}^\top d = r_k < 0,$$

so (4.3) is unbounded below. Otherwise, any index $i^\star$ attaining the minimum ratio is a *leaving index*, and the new basis is $B' = (B \cup \{k\}) \setminus \{i^\star\}$. The associated new BFS is $\tilde{\mathbf{x}}' = \tilde{\mathbf{x}} + t^\star d$ with $t^\star = \tilde{\mathbf{x}}_{i^\star}/(-d_{i^\star})$.

4.4.4 Finding an initial feasible basis (Phase I)

To find a feasible basis for (4.3), a standard approach is to solve the auxiliary *Phase I* problem

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{z}} \quad & \sum_{i \in [m]} z_i && \text{(4.5a)}\\
\text{s.t.} \quad & A\mathbf{x} + \mathbf{z} = \mathbf{b} && \text{(4.5b)}\\
& \mathbf{x} \geq \mathbf{0} && \text{(4.5c)}\\
& \mathbf{z} \geq \mathbf{0} && \text{(4.5d)}
\end{aligned}$$

assuming without loss of generality that $\mathbf{b} \geq \mathbf{0}$. One checks that (4.5) has optimal value 0 if and only if (4.3) is feasible. At an optimum of value 0, all artificial variables are zero, although some of them may remain basic at value zero in a degenerate solution. Under the standing assumption $\operatorname{rank}(A) = m$, such artificial variables can be pivoted out by degenerate pivots, yielding a basis consisting only of original variables and hence a feasible basis for (4.3). Since (4.5) has an obvious feasible basis (the artificial variables z_i), one can thus apply the simplex method to (4.5) to test feasibility and obtain a first feasible basis for the original problem.

4.4.5 ♢ Vertices and basic feasible solutions

The next two results connect the geometric notion of a vertex to the algebraic notion of a *basic feasible solution* introduced earlier in Section 4.3.1. This bridge is useful to keep in mind: it explains why simplex (an algebraic pivoting method) can be interpreted geometrically as moving from vertex to vertex on the feasible polyhedron.

♢ **Theorem 4.9.** *Consider a linear program written in standard form. Then a point* $\mathbf{x}$ *is a basic feasible solution if and only if* $\mathbf{x}$ *is a vertex of the feasible polyhedron.*

Proof. Consider the feasible polyhedron of the standard-form problem (4.3), i.e. $P = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}\}$.

(BFS ⇒ vertex). Let $\mathbf{x}$ be a basic feasible solution associated with a basis B and $N = [n] \setminus B$, so $\mathbf{x}_N = \mathbf{0}$ and $\mathbf{x}_B = A_B^{-1}\mathbf{b}$. Assume $\mathbf{x} = t\mathbf{y} + (1-t)\mathbf{z}$ with $t \in (0,1)$ and $\mathbf{y}, \mathbf{z} \in P$. Since $\mathbf{x}_N = \mathbf{0}$ and $\mathbf{y}_N, \mathbf{z}_N \geq \mathbf{0}$, one must have $\mathbf{y}_N = \mathbf{z}_N = \mathbf{0}$. Then $A_B\mathbf{y}_B = \mathbf{b} = A_B\mathbf{z}_B$, hence $\mathbf{y}_B = \mathbf{z}_B = \mathbf{x}_B$ because A_B is invertible. Thus $\mathbf{y} = \mathbf{z} = \mathbf{x}$, so $\mathbf{x}$ is a vertex.

(Vertex ⇒ BFS). Let $\mathbf{x}$ be a vertex of P and define its support $K := \{i \in [n] \mid \mathbf{x}_i > 0\}$. If the columns of A_K were linearly dependent, there would exist $\mathbf{d} \neq 0$ supported on K such that $A\mathbf{d} = \mathbf{0}$. For $\varepsilon > 0$ small enough, one would have $\mathbf{x} \pm \varepsilon\mathbf{d} \geq \mathbf{0}$ and $A(\mathbf{x} \pm \varepsilon\mathbf{d}) = \mathbf{b}$, so $\mathbf{x} = \frac{1}{2}(\mathbf{x} + \varepsilon\mathbf{d}) + \frac{1}{2}(\mathbf{x} - \varepsilon\mathbf{d})$ would be a strict convex combination of two distinct feasible points, contradicting that $\mathbf{x}$ is a vertex. Therefore the columns of A_K are linearly independent, so K can be extended to a basis B (with A_B invertible). Since $\mathbf{x}_N = \mathbf{0}$ and $A\mathbf{x} = \mathbf{b}$, one has $A_B\mathbf{x}_B = \mathbf{b}$ and thus $\mathbf{x}_B = A_B^{-1}\mathbf{b}$. Hence $\mathbf{x}$ is the basic solution associated with B, and since $\mathbf{x} \geq \mathbf{0}$, it is a basic feasible solution. □

♢ **Proposition 4.10.** *If two consecutive bases visited by the simplex method yield two distinct basic feasible solutions, then these two solutions are adjacent vertices of the feasible polyhedron.*

Proof. Consecutive bases in the simplex method differ by a single pivot: $B' = (B \cup \{k\}) \setminus \{i^\star\}$ for some entering index k and leaving index $i^\star$. The corresponding BFS $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ satisfy $A\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq 0$, and they share all nonbasic variables except possibly one: both have $\mathbf{x}_j = 0$ for every $j \notin B \cup \{k\}$. Thus the whole segment $\{\tilde{\mathbf{x}} + t(\tilde{\mathbf{x}}' - \tilde{\mathbf{x}}) \mid t \in [0,1]\}$ lies in the intersection of the equality constraints with the common nonnegativity constraints, which defines a one-dimensional face of the feasible polyhedron. Since $\tilde{\mathbf{x}} \neq \tilde{\mathbf{x}}'$, this face is an edge, so the two vertices are adjacent. □

4.5 Duality in linear programming

We now revisit duality for linear programs and highlight a key feature: unlike general convex optimization, strong duality holds in linear programming without any interior-point (Slater-type) assumption.

Consider again the primal problem (4.3):

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \quad & \mathbf{c}^\top \mathbf{x} && \text{(4.6a)} \\ \text{s.t.} \quad & A\mathbf{x} = \mathbf{b} && \text{(4.6b)} \\ & \mathbf{x} \geq \mathbf{0} && \text{(4.6c)} \end{aligned}$$

Its Lagrangian dual (see Chapter 2, Section 2.7) can be written in the classical form

$$\begin{aligned} \underset{\mathbf{y}\in\mathbb{R}^m}{\text{Max}} \quad & \mathbf{b}^\top \mathbf{y} && \text{(4.7a)} \\ \text{s.t.} \quad & A^\top \mathbf{y} \leq \mathbf{c} && \text{(4.7b)} \end{aligned}$$

As for any linear program, taking the dual of (4.7) yields a problem equivalent to the primal: the dual of the dual is the primal.

♡ **Theorem 4.11** (Strong duality in linear programming)**.** *In linear programming, if at least one of the two problems* (4.6) *(primal) or* (4.7) *(dual) is feasible, then there is strong duality: the optimal values coincide (possibly equal to* $\pm\infty$*).*

In particular, if the primal and the dual are both feasible, then both optimal values are finite and equal. If one problem is feasible and the other is infeasible, then the feasible one must be unbounded (its optimal value is $-\infty$ for the primal and $+\infty$ for the dual). The only situation where there is a "duality gap" is when both problems are infeasible, in which case the primal value is $+\infty$ and the dual value is $-\infty$.

The proof first treats a feasible primal: existence and KKT handle the bounded case, while weak duality handles the unbounded case; it then applies the same argument symmetrically starting from the dual.

Proof. We use weak duality: for any primal feasible $\mathbf{x}$ and any dual feasible $\mathbf{y}$, one has

$$\mathbf{b}^\top \mathbf{y} \leq \mathbf{c}^\top \mathbf{x}.$$

In particular, the dual value is always $\leq$ the primal value.

Assume first that the primal is feasible. If the primal is unbounded below, then the dual cannot be feasible (otherwise $\mathbf{b}^\top \mathbf{y}$ would provide a finite lower bound on $\mathbf{c}^\top \mathbf{x}$ for all feasible $\mathbf{x}$). Hence the dual value is $-\infty$ and the two values coincide.

Assume now that the primal is feasible and bounded below. By Theorem 4.6, the primal admits an optimal solution $\mathbf{x}^\sharp$. Consider (4.6) as a differentiable constrained problem with equality constraints $g(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = 0$

and inequality constraints $h(\mathbf{x}) = -\mathbf{x} \leq 0$. All constraints are affine, hence qualified at $\mathbf{x}^\sharp$ (Proposition 2.13), so the KKT theorem (Theorem 2.20) yields multipliers $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^n_+$ such that

$$\mathbf{c} + A^\top \lambda - \mu = \mathbf{0}, \qquad \mu_i \mathbf{x}^\sharp_i = 0, \ \forall i \in [n].$$

Let $\mathbf{y} := -\lambda$. Then $A^\top \mathbf{y} \leq \mathbf{c}$, so $\mathbf{y}$ is dual feasible. Moreover, taking the dot product of the stationarity condition with $\mathbf{x}^\sharp$ and using $A\mathbf{x}^\sharp = \mathbf{b}$ and $\mu^\top \mathbf{x}^\sharp = 0$ gives

$$\mathbf{c}^\top \mathbf{x}^\sharp = \mathbf{b}^\top \mathbf{y}.$$

Therefore the primal and the dual have feasible points with the same objective value, so by weak duality both are optimal and the optimal values are equal.

Assume now that the dual is feasible. If the dual is unbounded above, then the primal cannot be feasible (otherwise $\mathbf{c}^\top \mathbf{x}$ would provide a finite upper bound on $\mathbf{b}^\top \mathbf{y}$ for all feasible $\mathbf{y}$ by weak duality). If moreover the dual is bounded above, then its optimal value $d^\star$ is finite. Consider the equivalent minimization problem

$$\operatorname*{Min}_{\mathbf{y} \in \mathbb{R}^m} \{-\mathbf{b}^\top \mathbf{y} \mid A^\top \mathbf{y} \leq \mathbf{c}\},$$

whose optimal value is $-d^\star$. Introduce a slack variable $\mathbf{s} \in \mathbb{R}^n_+$ so that $A^\top \mathbf{y} + \mathbf{s} = \mathbf{c}$, and write the free variable $\mathbf{y}$ as $\mathbf{y} = \mathbf{y}^+ - \mathbf{y}^-$ with $\mathbf{y}^+, \mathbf{y}^- \geq \mathbf{0}$. This yields the standard-form linear program

$$\operatorname*{Min}_{\mathbf{y}^+, \mathbf{y}^-, \mathbf{s}} \{-\mathbf{b}^\top \mathbf{y}^+ + \mathbf{b}^\top \mathbf{y}^- \mid A^\top \mathbf{y}^+ - A^\top \mathbf{y}^- + \mathbf{s} = \mathbf{c}, \ \mathbf{y}^+, \mathbf{y}^-, \mathbf{s} \geq \mathbf{0}\},$$

which is feasible and bounded below. Applying the already proven case ("primal feasible and bounded below") to this standard-form problem shows that its dual is feasible and that there is no duality gap. The dual of this standard-form problem is

$$\operatorname*{Max}_{\mathbf{z} \in \mathbb{R}^n} \{\mathbf{c}^\top \mathbf{z} \mid A\mathbf{z} \leq -\mathbf{b}, \ -A\mathbf{z} \leq \mathbf{b}, \ \mathbf{z} \leq \mathbf{0}\}.$$

The two inequalities $A\mathbf{z} \leq -\mathbf{b}$ and $-A\mathbf{z} \leq \mathbf{b}$ are equivalent to $A\mathbf{z} = -\mathbf{b}$. With the change of variables $\mathbf{x} := -\mathbf{z} \geq \mathbf{0}$, this dual becomes the original primal problem (4.6), and its objective $\mathbf{c}^\top \mathbf{z}$ becomes $-\mathbf{c}^\top \mathbf{x}$. Therefore its optimal value is $-p^\star$, where $p^\star$ is the primal optimal value. Therefore $-d^\star = -p^\star$, i.e. $d^\star = p^\star$, and in particular the primal is feasible. □

4.6 Exercises

Exercise 4.1

Show that any linear optimization problem in inequality form can be converted to standard form, and conversely.

Solution

Exercise 4.2

Consider

$$\begin{array}{rl} \text{Min} & \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{array}$$

Assume A and $\mathbf{b}$ have rational entries. Show that if the problem has at least one optimal solution, then it has an optimal solution with rational entries.

Solution

Exercise 4.3 *Beverage production planning*

A company produces three beverages (XXX, YYY, ZZZ) with profits (per liter) of 1.55, 3.05 and 4.80 euros, respectively. Each beverage has a minimum weekly production level: 1020, 1450 and 750 liters.

Producing 20 liters requires the following times (in hours) in (i) processing, (ii) mixing, (iii) bottling:

	XXX	YYY	ZZZ
processing	2	3	6
mixing	4	5.5	7.5
bottling	2	1	3.5

Next week, the company has 708 hours of processing time, 932 hours of mixing, and 342 hours of bottling.

Formulate the profit-maximizing production plan as a linear program.

Solution

Exercise 4.4 *A refinery/blending problem*

A refinery must produce each day two grades of gasoline (A and B) from three components (1, 2, 3). The available quantities and unit costs are:

component	$Q_{\max}$	unit cost
1	3000	3
2	2000	6
3	4000	4

The selling prices and quality specifications are:

gasoline	specification	unit selling price
A	$\leq 30\%$ of component 1	5.5
	$\geq 40\%$ of component 2	
	$\leq 50\%$ of component 3	
B	$\leq 50\%$ of component 1	4.5
	$\geq 10\%$ of component 2	

Formulate a linear program that maximizes profit, assuming all production can be sold.

Solution

Exercise 4.5 *Best fitting line*
Given points $(x_1, y_1), \ldots, (x_n, y_n)$ in the plane, we seek the best fitting line $y = ax + b$ in the ℓ_∞ norm, i.e. we want to minimize

$$\sup_{i=1,\ldots,n} |ax_i + b - y_i|.$$

1. Show that this can be formulated as a linear program. Do the same if we minimize $\sum_{i=1}^n |ax_i + b - y_i|$ (an ℓ_1 fit).
2. Apply both formulations to the data

   $$(0, 1), \quad (1, 3), \quad (2, 5), \quad (3, 7), \quad (4, 15).$$

   Determine the ℓ_1-best and ℓ_∞-best affine fits and compare them with the least-squares line obtained in Exercise 3.2.

Solution

Exercise 4.6
For a vector x, let $x^{[i]}$ denote the i-th largest element of $\{x_1, \ldots, x_n\}$. For instance, $x^{[1]}$ is the largest component of $\mathbf{x}$ and $x^{[n]}$ the smallest. For $\mathbf{x} = (2, 1, 0, 1)$, we have $x^{[2]} = x^{[3]} = 1$.

1. For fixed k, show that the problem

   $$\begin{array}{ll} \text{Min} & \sum_{i=1}^k x^{[i]} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \in \mathbb{R}^n_+ \end{array}$$

   can be formulated as a linear program by introducing one constraint per subset of size k of $[n]$.
2. (*) Show that one can in fact write an LP with at most $m + n$ constraints (not counting sign constraints), where m is the number of rows of A.

Solution

Exercise 4.7 *Largest inscribed ball*
Show that the problem of finding the largest Euclidean ball inscribed in a polyhedron can be formulated as a linear program.

Solution

Exercise 4.8 *Fractional knapsack relaxation*
Consider the linear program

$$\begin{array}{lll} \text{Min} & \sum_{i=1}^n c_i x_i & \\ \text{s.t.} & \sum_{i=1}^n a_i x_i = b & \\ & 0 \le x_i \le 1 \qquad & i = 1, \ldots, n, \end{array}$$

where $a_i > 0$ for all $i \in [n]$, and assume the problem is feasible.

1. Primal approach

(a) Write it in standard form and justify the existence of an optimal basic solution.

(b) Deduce that there exists an optimal solution $\mathbf{x}^\sharp$ with at most one *fractional* component (strictly between 0 and 1).

(c) Show that if $c_k/a_k < c_j/a_j$, then $x_k^\sharp \geq x_j^\sharp$.

(d) Deduce an easy way to compute an optimal solution.

2. Dual approach

(a) Write the dual linear program.

(b) Show it reduces to maximizing a concave, nonsmooth function of a single real variable.

(c) (*) Solve this one-dimensional problem.

Solution

Exercise 4.9 *Bounding the optimality gap*
Consider the linear program

$$\begin{array}{ll} \text{Min} & \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A'\mathbf{x} = \mathbf{b}' \\ & \sum_{i=1}^n x_i = 1 \\ & \mathbf{x} \in \mathbb{R}^n_+. \end{array}$$

Equivalently, it is in standard form

$$\begin{array}{ll} \text{Min} & \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \in \mathbb{R}^n_+, \end{array}$$

with

$$A = \begin{pmatrix} A' \\ 1 \ \cdots \ 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} \mathbf{b}' \\ 1 \end{pmatrix}.$$

Let B be a feasible basis and $\mathbf{r}$ the associated reduced-cost vector. Show that if $\mathbf{r} \ngeq \mathbf{0}$, then the optimal value v satisfies

$$\mathbf{c}_B^\top A_B^{-1}\mathbf{b} + (\min_j r_j) \ \leq \ v \ \leq \ \mathbf{c}_B^\top A_B^{-1}\mathbf{b}.$$

Solution

Exercise 4.10 *A proof of strong duality via the simplex method*
It was shown that there exists a pivot rule ensuring that the simplex algorithm always terminates. Deduce from this an alternative proof of strong duality in linear optimization.

Solution

Exercise 4.11 *Column generation*
Consider the linear program (P)

$$\begin{array}{ll} \text{Min} & \mathbf{c} \cdot \mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \in \mathbb{R}^n_+, \end{array}$$

and let $I \subseteq [n]$. Consider the restricted problem (P^I)

$$\begin{array}{ll} \text{Min} & \mathbf{c}_I \cdot \mathbf{x}' \\ \text{s.t.} & A_I\mathbf{x}' = \mathbf{b} \\ & \mathbf{x}' \in \mathbb{R}^I_+. \end{array}$$

Let (D) and (D^I) be the duals of (P) and (P^I), and let $\tilde{\mathbf{y}}$ be an optimal solution of (D^I). Show that if $\tilde{\mathbf{y}}$ is feasible for (D), then the optimal value of (P) equals that of (P^I).

This remark is the starting point of *column generation*, useful for solving LPs with too many variables to be given to a solver directly.

Solution

Chapter 5

Mixed-integer linear programming

Mixed-integer linear programs (MILPs) extend linear programs by requiring some variables to take integer values. This discrete component dramatically increases the computational difficulty (in general, MILPs are NP-hard), but linear programming relaxations provide strong bounds and play a central role in modern algorithms.

We first introduce the MILP model and its LP relaxation, then present the branch-and-bound principle, and finally review common modeling patterns using binary variables.

5.1 Definition and motivation

♡ **Definition 5.1** (Mixed-integer linear program)**.** *Let* $[n] := \{1, \dots, n\}$ *and let* $I \subseteq [n]$*. A* mixed-integer linear program *(MILP) is an optimization problem of the form*

$$\begin{aligned} \underset{\mathbf{x}\in\mathbb{R}^n}{\text{Min}} \quad & \mathbf{c}^\top \mathbf{x} && && \text{(5.1a)} \\ s.t. \quad & A\mathbf{x} \leq \mathbf{b} && && \text{(5.1b)} \\ & \mathbf{x}_i \in \mathbb{Z} && \forall i \in I && \text{(5.1c)} \end{aligned}$$

When $I = [n]$ and $\mathbf{x}_i \in \{0, 1\}$ for all i, we speak of a *binary integer linear program* (BILP), also called a *binary program.* MILP is a very general modeling framework: many combinatorial optimization problems can be formulated in this way, and the availability of powerful solvers makes MILP a practical tool for large-scale instances. However, solvers are not magic: the same problem can be formulated in multiple ways with very different computational performances, and understanding the structure of a model is crucial for effective optimization.

♢ **Proposition 5.2** (Computational hardness)**.** *Even binary programs of the form* (5.1) *(with* $I = [n]$ *and* $\mathbf{x} \in \{0, 1\}^n$*) are NP-hard in general.*

Proof. The 0–1 knapsack problem (5.2) is a special case of (5.1) with binary variables, and its decision version is NP-complete (e.g. by reduction from Subset Sum). Therefore, solving binary linear programs in general is NP-hard. □

5.2 Examples

Typical examples of MILP models include knapsack and set covering problems, facility location and network design, scheduling and routing problems (TSP/VRP), unit commitment in power systems, production planning with setups, and many variants thereof. We will use the knapsack problem as a concrete running example for algorithms.

♡ *Example* 5.1 (Knapsack problem). We are given items with weights $w_i > 0$ and values $v_i \geq 0$, and a knapsack capacity $W > 0$. Choosing item i is modeled by a binary variable $\mathbf{x}_i \in \{0,1\}$. The knapsack problem can be written as a BILP:

$$\operatorname*{Max}_{\mathbf{x}\in\{0,1\}^n} \quad \sum_{i\in[n]} v_i \mathbf{x}_i \tag{5.2a}$$

$$\text{s.t.} \quad \sum_{i\in[n]} w_i \mathbf{x}_i \leq W \tag{5.2b}$$

Example 5.2 (Capacitated facility location). We are given a set of potential facilities J and a set of clients I. Opening facility $j \in J$ incurs a fixed cost $f_j \geq 0$ and provides a capacity $u_j > 0$. Client $i \in I$ has demand $d_i > 0$, and assigning i to facility j induces a service cost $c_{ij} \geq 0$. Introduce binary variables $y_j \in \{0,1\}$ indicating whether facility j is open, and $x_{ij} \in \{0,1\}$ indicating whether client i is served by facility j. A standard MILP is

$$\operatorname*{Min}_{x,y} \quad \sum_{j\in J} f_j y_j + \sum_{i\in I}\sum_{j\in J} c_{ij} x_{ij} \tag{5.3a}$$

$$\text{s.t.} \quad \sum_{j\in J} x_{ij} = 1 \qquad \forall i \in I \tag{5.3b}$$

$$x_{ij} \leq y_j \qquad \forall i \in I,\ \forall j \in J \tag{5.3c}$$

$$\sum_{i\in I} d_i x_{ij} \leq u_j y_j \qquad \forall j \in J \tag{5.3d}$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in I,\ \forall j \in J \tag{5.3e}$$

$$y_j \in \{0,1\} \qquad \forall j \in J \tag{5.3f}$$

The linking constraints $x_{ij} \leq y_j$ and the capacity constraints enforce that clients can only be assigned to open facilities.

♢ *Example* 5.3 (Lot-sizing with setup costs). Consider a single-item production planning problem over periods $t = 1, \ldots, T$ with deterministic demands $d_t \geq 0$. Let $x_t \geq 0$ denote the quantity produced in period t, $s_t \geq 0$ the inventory at the end of period t, and let $y_t \in \{0,1\}$ indicate whether the production line is set up in period t. If producing in period t incurs a variable cost $c_t x_t$, a holding cost $h_t s_t$, and a setup cost $K_t y_t$, one can write

$$\operatorname*{Min}_{x,s,y} \quad \sum_{t\in[T]} (c_t x_t + h_t s_t + K_t y_t) \tag{5.4a}$$

$$\text{s.t.} \quad s_t = s_{t-1} + x_t - d_t \qquad \forall t \in [T] \tag{5.4b}$$

$$0 \leq x_t \leq M_t y_t \qquad \forall t \in [T] \tag{5.4c}$$

$$s_t \geq 0 \qquad \forall t \in [T] \tag{5.4d}$$

$$y_t \in \{0,1\} \qquad \forall t \in [T] \tag{5.4e}$$

where s_0 is given and M_t is a valid upper bound on x_t. The constraint $x_t \leq M_t y_t$ is a typical fixed-charge (on/off) pattern: production is allowed only if the setup decision is active.

♢ *Example* 5.4 (Unit commitment (simplified)). In unit commitment, one schedules power generators over time. Let G be a set of units and $t = 1, \ldots, T$ time periods. For each unit $g \in G$ and period t, let $u_{gt} \in \{0,1\}$ indicate whether the unit is on, and let $p_{gt} \geq 0$ be the produced power. With minimum and maximum outputs $P_g^{\min} \leq P_g^{\max}$, a basic on/off model is

$$P_g^{\min} u_{gt} \leq p_{gt} \leq P_g^{\max} u_{gt} \qquad \forall g \in G,\ \forall t \in [T].$$

To account for strictly positive startup costs $s_{gt} > 0$, introduce $v_{gt} \in \{0,1\}$ indicating a startup at time t, and enforce $u_{gt} - u_{g,t-1} \le v_{gt}$ (with u_{g0} given). With a demand profile D_t, a simple MILP is

$$\begin{aligned}
\min_{p,u,v} \quad & \sum_{t\in[T]} \sum_{g\in G} \big(c_{gt}p_{gt} + f_{gt}u_{gt} + s_{gt}v_{gt}\big) && && \text{(5.5a)}\\
\text{s.t.} \quad & \sum_{g\in G} p_{gt} = D_t && \forall t \in [T] && \text{(5.5b)}\\
& P_g^{\min} u_{gt} \le p_{gt} \le P_g^{\max} u_{gt} && \forall g \in G,\ \forall t \in [T] && \text{(5.5c)}\\
& u_{gt} - u_{g,t-1} \le v_{gt} && \forall g \in G,\ \forall t \in [T] && \text{(5.5d)}\\
& p_{gt} \ge 0 && \forall g \in G,\ \forall t \in [T] && \text{(5.5e)}\\
& u_{gt} \in \{0,1\} && \forall g \in G,\ \forall t \in [T] && \text{(5.5f)}\\
& v_{gt} \in \{0,1\} && \forall g \in G,\ \forall t \in [T] && \text{(5.5g)}
\end{aligned}$$

The one-sided startup constraint permits $v_{gt} = 1$ even when no startup occurs. However, because v_{gt} occurs only through the strictly positive startup cost in the minimization objective, every optimal solution sets $v_{gt} = 1$ exactly when a startup requires it. More realistic models add ramping limits and minimum up/down times, but the core MILP structure already appears in (5.5).

✠ *Example* 5.5 (Traveling salesman problem (TSP)). Given cities $[n]$ and travel costs $c_{ij} \ge 0$ for $i \ne j$, the traveling salesman problem seeks a minimum-cost Hamiltonian tour. This is the problem introduced in Example 1.3; Exercise 5.3 gives another classical formulation. Introduce variables $x_{ij} \in \{0,1\}$ indicating whether the tour uses the arc $i \to j$. A classical formulation uses degree constraints together with subtour elimination:

$$\begin{aligned}
\min_{x} \quad & \sum_{i\ne j} c_{ij}x_{ij} && && \text{(5.6a)}\\
\text{s.t.} \quad & \sum_{j\in[n]\setminus\{i\}} x_{ij} = 1 && \forall i \in [n] && \text{(5.6b)}\\
& \sum_{i\in[n]\setminus\{j\}} x_{ij} = 1 && \forall j \in [n] && \text{(5.6c)}\\
& \sum_{i\in S} \sum_{j\in S,\ j\ne i} x_{ij} \le |S| - 1 && \forall \emptyset \subsetneq S \subsetneq [n] && \text{(5.6d)}\\
& x_{ij} \in \{0,1\} && \forall i,j \in [n],\ i \ne j && \text{(5.6e)}
\end{aligned}$$

The subtour constraints are exponentially many; in practice they are generated on the fly (separation) inside branch-and-bound.

Remark 5.1 (Formulation strength). Different MILP formulations, such as the TSP formulations just referenced, can describe the same integer feasible solutions while having LP relaxations of different strength. A stronger LP relaxation generally provides better bounds and can therefore reduce the amount of branching required in branch-and-bound.

Example 5.6 (Capacitated vehicle routing (VRP)). In a capacitated vehicle routing problem, a depot (indexed by 0) must serve clients $[n]$ with demands $d_i > 0$ using vehicles of capacity Q. Using arc variables $x_{ij} \in \{0,1\}$ for $i \ne j$ in $\{0,1,\dots,n\}$, the degree constraints enforce that each client has exactly one predecessor and one successor:

$$\sum_{j\ne i} x_{ij} = 1, \qquad \sum_{j\ne i} x_{ji} = 1 \qquad \forall i \in [n].$$

To enforce capacity and prevent subtours, one may introduce load variables $\ell_i \in [0, Q]$ and add, for all $i \neq j$ in $[n]$,

$$\ell_j \geq \ell_i + d_j - Q(1 - x_{ij}), \qquad d_i \leq \ell_i \leq Q.$$

Together with suitable depot constraints (a fixed number of routes or a fleet-size variable), this yields a MILP. The precise formulation depends on the chosen VRP variant (number of vehicles, time windows, multiple depots, ...).

5.3 LP relaxation, bounds, and the integer hull

The key algorithmic idea behind most MILP methods is to exploit linear programming relaxations. In the binary case, the most basic relaxation consists in replacing $\mathbf{x} \in \{0,1\}^n$ with $\mathbf{x} \in [0,1]^n$.

Definition 5.3 (LP relaxation)**.** *Consider the MILP* (5.1) *and assume for simplicity that all integrality constraints are binary:* $\mathbf{x}_i \in \{0,1\}$ *for* $i \in I$*. The* LP relaxation *is obtained by replacing these constraints with* $\mathbf{x}_i \in [0,1]$ *for* $i \in I$*.*

♡ **Proposition 5.4** (Bounding property (minimization))**.** *Let* $z^\star_{\text{MILP}}$ *be the optimal value of a minimization MILP, and let* $z^\star_{\text{LP}}$ *be the optimal value of its LP relaxation. Then*

$$z^\star_{\text{LP}} \leq z^\star_{\text{MILP}}.$$

Proof. The feasible set of the MILP is included in the feasible set of its relaxation, so the infimum over the larger set is no greater. □

From a geometric viewpoint, if $P = \{\mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} \leq \mathbf{b}\}$ is the feasible polyhedron of the relaxation, then the set of mixed-integer feasible points is

$$P_I := \{\mathbf{x} \in P \mid x_i \in \mathbb{Z}\ \forall i \in I\}.$$

Its convex hull $\text{conv}(P_I)$ is called the *mixed-integer hull* (or the *integer hull* when $I = [n]$). When the LP relaxation is tight (its optimum is attained at an integer point), solving the LP already solves the MILP. In general, there is an *integrality gap*, and the goal of algorithms is to close it efficiently.

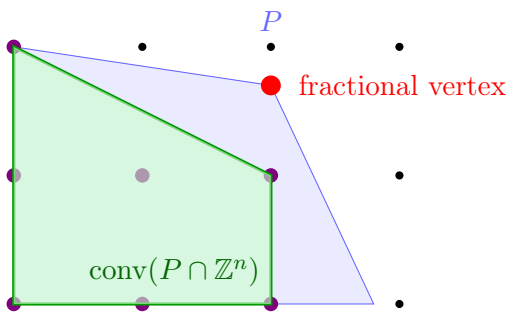

Figure 5.1: A polyhedron P (LP relaxation), integer-feasible points (violet), and the integer hull $\text{conv}(P \cap \mathbb{Z}^n)$.

5.4 Branch-and-bound

We now describe the branch-and-bound principle in the binary case, which already captures the main ideas used in general MILP solvers. The solution space $\{0,1\}^n$ can be organized as a binary tree by fixing variables one by one: each node corresponds to a subproblem where some variables are fixed to 0 or 1. At such a node, the LP relaxation provides a lower bound on the best integer value reachable below that node (for minimization problems).

5.4.1 Node subproblems and LP lower bounds

Consider the BILP

$$\min\{\mathbf{c}^\top \mathbf{x} \mid A\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \in \{0,1\}^n\}.$$

At a node ν, some indices are fixed to 0 or 1: we write $I_\nu = I_\nu^0 \cup I_\nu^1$ with

$$\mathbf{x}_i = 0,\ \forall i \in I_\nu^0, \qquad \mathbf{x}_i = 1,\ \forall i \in I_\nu^1.$$

The set I_ν is the set of variables whose values have been decided along the path from the root to ν. All remaining variables, with indices in $[n] \setminus I_\nu$, are still free to take values in $\{0,1\}$ (for the integer subproblem) or in $[0,1]$ (for the relaxation). In other words, a node corresponds to *restricting* the feasible set by adding simple bounds $x_i = 0$ or $x_i = 1$ for a subset of indices.

The associated integer subproblem is

$$(\mathcal{P}_\nu) \qquad \min\{\mathbf{c}^\top \mathbf{x} \mid A\mathbf{x} \leq \mathbf{b},\ \mathbf{x} \in \{0,1\}^n,\ \mathbf{x}_i = 0,\ \forall i \in I_\nu^0,\ \mathbf{x}_i = 1,\ \forall i \in I_\nu^1\},$$

and its LP relaxation is obtained by replacing $\mathbf{x} \in \{0,1\}^n$ with $\mathbf{x} \in [0,1]^n$. Denote by F_ν^{IP} and F_ν^{LP} the feasible sets of the node problem and its relaxation, respectively. Since

$$F_\nu^{\mathrm{IP}} \subseteq F_\nu^{\mathrm{LP}},$$

the minimum of the linear objective over the larger set is no greater: the optimal value z_ν^{LP} of the node LP relaxation is a *lower bound* on the optimal value $z_\nu^\star$ of $(\mathcal{P}_\nu)$ (for minimization problems). This is exactly the same set-inclusion argument as Proposition 5.4, applied after adding the fixing constraints that define the node.

This observation has two immediate algorithmic consequences. First, if the LP relaxation at ν is infeasible, then $F_\nu^{\mathrm{LP}} = \emptyset$, hence also $F_\nu^{\mathrm{IP}} = \emptyset$: the integer subproblem $(\mathcal{P}_\nu)$ is infeasible. Second, if the LP relaxation has an optimal solution $\mathbf{x}^{LP}$ that happens to be binary, then $\mathbf{x}^{LP} \in F_\nu^{\mathrm{IP}}$ and its objective value equals z_ν^{LP}; therefore $\mathbf{x}^{LP}$ is optimal for $(\mathcal{P}_\nu)$.

For a *maximization* BILP (such as knapsack in Example 5.1), the same reasoning gives the opposite type of bound: the node LP relaxation provides an *upper bound* on the best integer value reachable below the node.

5.4.2 Algorithm outline

Branch-and-bound maintains a set of unexplored nodes (often stored as a stack, a queue, or a priority queue), together with an *incumbent* integer feasible solution and its objective value $z^\star$. For minimization, the incumbent value $z^\star$ is an *upper bound* on the optimal value (it is the value of some feasible integer solution), while each node LP relaxation provides a *lower bound* on the best integer value reachable below that node.

At a given node ν, one solves the LP relaxation. If the relaxation is infeasible, then the integer subproblem is infeasible and the node can be discarded. If the relaxation is feasible but its lower bound is already worse than the incumbent (i.e. $z_\nu^{LP} \geq z^\star$), then the node cannot

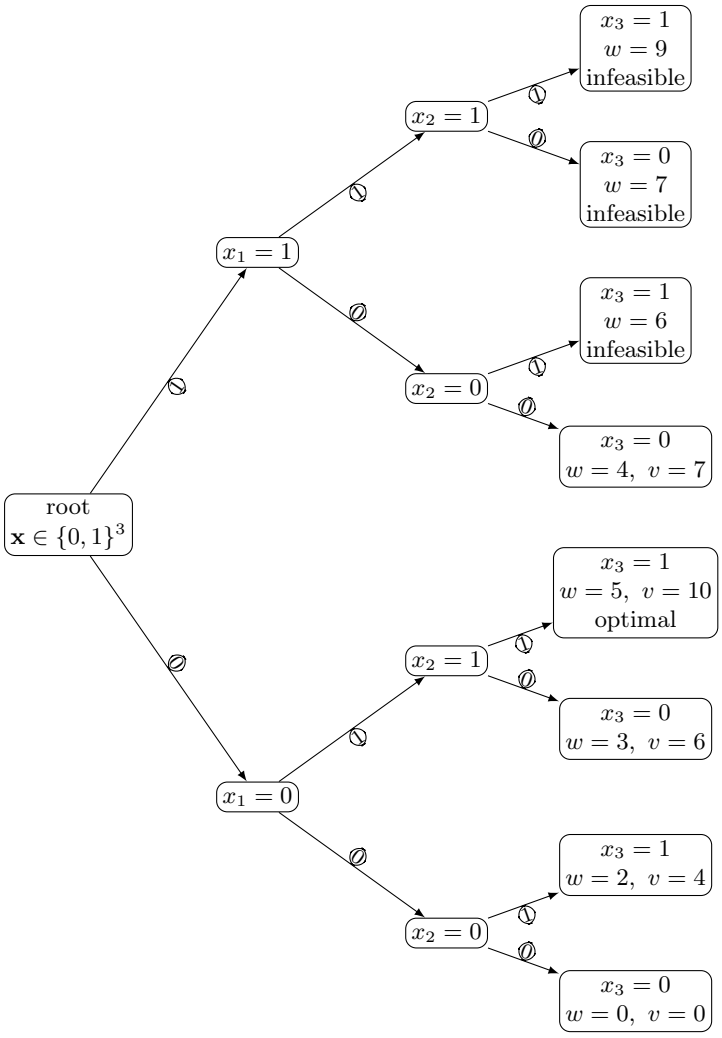

Figure 5.2: Binary search tree for a 3-item knapsack instance with capacity $W = 5$, weights $(w_1, w_2, w_3) = (4, 3, 2)$ and values $(v_1, v_2, v_3) = (7, 6, 4)$. Each root-to-leaf path fixes $(x_1, x_2, x_3) \in \{0,1\}^3$; leaves may be infeasible (capacity violated) or feasible with a corresponding objective value.

contain a better integer solution and it can be pruned by bound. If the LP optimal solution happens to be binary, then it is an optimal integer solution for that node and can be used to update the incumbent. Otherwise, one *branches* by selecting a fractional variable x_i and creating two children with additional constraints $x_i = 0$ and $x_i = 1$.

The following pseudocode summarizes this logic.

Algorithm 4: Branch-and-bound for binary integer programs

Data: BILP: $\min\{\mathbf{c}^\top \mathbf{x} \mid A\mathbf{x} \le \mathbf{b},\ \mathbf{x} \in \{0,1\}^n\}$
Result: An optimal solution and its value, or a declaration that the BILP is infeasible

$z^\star \leftarrow +\infty$, $\mathbf{x}^\sharp \leftarrow \varnothing$
Create root node ν_{root} with $I^0_{\nu_{\text{root}}} = I^1_{\nu_{\text{root}}} = \varnothing$ and push it into list L
while $L \neq \varnothing$ **do**
  Extract a node ν from L
  Solve the LP relaxation at ν and obtain $(\mathbf{x}^{LP}, z^{LP})$
  if *LP infeasible* **then**
    discard ν
  else if $z^{LP} \ge z^\star$ **then**
    discard ν `// prune by bound`
  else if $\mathbf{x}^{LP} \in \{0,1\}^n$ **then**
    Update $(\mathbf{x}^\sharp, z^\star) \leftarrow (\mathbf{x}^{LP}, z^{LP})$ `// new incumbent`
  else
    Choose an index i with $\mathbf{x}^{LP}_i \notin \{0,1\}$ `// branching rule`
    Create child ν_0 with $I^0_{\nu_0} = I^0_\nu \cup \{i\}$, $I^1_{\nu_0} = I^1_\nu$ and push ν_0 into L
    Create child ν_1 with $I^1_{\nu_1} = I^1_\nu \cup \{i\}$, $I^0_{\nu_1} = I^0_\nu$ and push ν_1 into L
if $\mathbf{x}^\sharp = \varnothing$ **then**
  return *"BILP infeasible"*
else
  return $(\mathbf{x}^\sharp, z^\star)$

♡ **Theorem 5.5** (Correctness of branch-and-bound)**.** *Algorithm 4 terminates after finitely many nodes and returns either an optimal solution of the BILP and its value, or a declaration that the BILP is infeasible.*

Proof sketch. At every node ν, the LP relaxation value z^{LP} is a lower bound on the optimal value of $(\mathcal{P}_\nu)$ (Proposition 5.4). Therefore pruning by the test $z^{LP} \ge z^\star$ never removes a node that could contain an improving integer solution. When a node LP solution is binary, it is feasible for the integer subproblem and its value equals the node lower bound, hence it is optimal for that node. Each branch fixes a previously unfixed binary variable, so the search tree has depth at most n and is finite. At termination, if an incumbent exists, all nodes have been explored or safely pruned, so no feasible integer solution has value smaller than $z^\star$ and the incumbent is optimal. If no incumbent exists, no node could have been pruned by bound; after all nodes have been processed, every leaf subproblem has therefore been shown infeasible, so the BILP is infeasible. □

5.4.3 Practical aspects

The efficiency of branch-and-bound depends heavily on the *quality* of the LP relaxations and on algorithmic choices such as which node to explore next (depth-first, best bound, ...) and which variable to branch on. A tighter relaxation gives bounds closer to the integer optimum, so more nodes can be pruned before further branching is necessary. In practice, modern solvers combine branch-and-bound with many additional techniques: valid inequalities (cutting planes) to tighten relaxations, primal heuristics to quickly find good incumbents, and preprocessing to reduce the problem size.

Example 5.7 (Knapsack relaxation). The LP relaxation of (5.2) replaces $\mathbf{x} \in \{0,1\}^n$ by $\mathbf{x} \in [0,1]^n$. This relaxation can be solved efficiently by sorting items by decreasing "efficiency" v_i/w_i

and filling the knapsack greedily, allowing one fractional item. In branch-and-bound, the fractional variable of this LP solution is a natural branching choice.

Example 5.8 (A tiny branch-and-bound run (knapsack)). Consider the 0–1 knapsack *maximization* problem

$$\max\{9x_1 + 6x_2 + 4x_3 \mid 4x_1 + 3x_2 + 2x_3 \le 5,\ \mathbf{x} \in \{0,1\}^3\}.$$

For maximization, the roles of bounds are reversed: a feasible integer solution gives a *lower bound* on the optimum, while the LP relaxation gives an *upper bound*.

Root node. The LP relaxation $\mathbf{x} \in [0,1]^3$ is solved by the fractional-knapsack rule. It takes $x_1 = 1$ (weight 4, value 9) and then fills the remaining capacity 1 by taking $x_2 = \frac{1}{3}$ (weight 1, value 2), yielding the fractional solution

$$\mathbf{x}^{LP} = \left(1, \tfrac{1}{3}, 0\right), \qquad z^{LP} = 11.$$

Thus the root upper bound is 11. A simple feasible integer solution is $(0,1,1)$ (weight 5, value 10), so we can start with the incumbent $z^\star = 10$. Since the LP solution is fractional, we branch on x_2.

Branch $x_2 = 0$. The relaxation becomes $\max\{9x_1 + 4x_3 \mid 4x_1 + 2x_3 \le 5,\ x_1, x_3 \in [0,1]\}$. Its optimum is $x_1 = 1$, $x_3 = \frac{1}{2}$ with value $9 + 2 = 11$ and weight 5. Since the LP solution is still fractional, we branch again (for instance on x_3): if $x_3 = 0$, the node LP optimum is the integer point $x_1 = 1$, with value 9. If $x_3 = 1$, the LP relaxation permits $x_1 = \frac{3}{4}$ and has upper bound

$$9 \cdot \tfrac{3}{4} + 4 = 10.75.$$

This bound exceeds the incumbent value 10, so the node cannot yet be pruned by bound. Branching on x_1, the child $x_1 = 1$ is infeasible, while the child $x_1 = 0$ gives the integer point $(0,0,1)$ of value 4. Thus the best integer value in the entire subtree under $x_2 = 0$ is 9.

Branch $x_2 = 1$. The remaining capacity is 2. Since $9/4 > 4/2$, the node LP relaxation takes $x_1 = \frac{1}{2}$ and $x_3 = 0$, giving the upper bound

$$6 + 9 \cdot \tfrac{1}{2} = 10.5.$$

Branching on x_1, the child $x_1 = 1$ is infeasible, while the child $x_1 = 0$ takes $x_3 = 1$ and gives the integer solution $\mathbf{x} = (0,1,1)$ of value 10. All nodes are now closed, so the incumbent value 10 is certified optimal.

This toy example illustrates the typical logic: LP relaxations provide bounds; branching enforces integrality; and pruning discards subtrees that cannot contain an improving solution.

5.5 Modeling with binary variables

A useful way to think about these formulations is that binary variables encode whether a decision is taken, continuous variables encode how much, and linking constraints make the second depend on the first. This viewpoint allows many nonlinear-looking requirements to be represented by linear inequalities. We review the resulting big-M, logical, cardinality, and piecewise-linear modeling patterns.

5.5.1 Big-M constraints and on/off decisions

Big-M constraints model conditional relationships by introducing a constant M large enough to make a constraint inactive when a binary variable takes a given value. For instance, if $b \in \{0,1\}$ and $x \in [0, M]$, the constraint $x \le Mb$ enforces $x = 0$ when $b = 0$ while allowing x to be positive when $b = 1$. The choice of M is crucial: it should be as small as possible while remaining valid, otherwise the relaxation becomes weak and numerical issues may arise.

Example 5.9 (Semi-continuous variable). Let x be a decision variable that must satisfy either $x = 0$ or $x \in [\underline{x}, \overline{x}]$. Introducing $b \in \{0, 1\}$, this can be modeled as

$$x \le \overline{x}\, b, \qquad x \ge \underline{x}\, b, \qquad b \in \{0, 1\}.$$

5.5.2 Logical constraints

Logical relationships between binary variables can be expressed by linear inequalities. For example, the implication "if $b_1 = 1$ then $b_2 = 1$" is modeled by $b_2 \ge b_1$. The constraint "$b_1 = 1$ or $b_2 = 1$" can be modeled by $b_1 + b_2 \ge 1$, whereas "exactly one of b_1, b_2 is 1" is modeled by $b_1 + b_2 = 1$.

5.5.3 Cardinality constraints (number of active variables)

In many applications, one wishes to limit how many decisions are active. If $x_i \ge 0$ are continuous variables and $b_i \in \{0, 1\}$ indicate whether x_i is allowed to be positive, then

$$x_i \le M b_i, \ \forall i \in [n]$$

links x_i to b_i. The constraint $\sum_{i=1}^n b_i \le k$ guarantees that at most k continuous variables can be positive. In contrast, $\sum_{i=1}^n b_i = k$ sets exactly k activation binaries to one, but does not by itself guarantee that exactly k variables x_i are positive. If known bounds satisfy $0 < \underline{x}_i \le \overline{x}_i$, the stronger linking constraints

$$\underline{x}_i b_i \le x_i \le \overline{x}_i b_i$$

make $b_i = 1$ equivalent to x_i being active and positive.

5.5.4 Piecewise-linear functions

Piecewise-linear functions are ubiquitous in modeling (cost curves, penalties, saturations). Whether one needs integer variables depends on the geometry of the function one wants to represent.

Convex case (LP formulation). Assume we are given breakpoints (x_i, f_i), $i \in [k]$, with $x_1 < \cdots < x_k$, and that these points define a *convex* piecewise-linear function (equivalently, the slopes between consecutive points are nondecreasing). Introducing weights $\lambda_i \ge 0$ with $\sum_{i \in [k]} \lambda_i = 1$, one can model the *epigraph* $t \ge f(x)$ by

$$x = \sum_{i \in [k]} x_i \lambda_i, \qquad t \ge \sum_{i \in [k]} f_i \lambda_i, \qquad \sum_{i \in [k]} \lambda_i = 1, \qquad \lambda_i \ge 0, \ \forall i \in [k].$$

This formulation is linear and does not require integer variables. In a minimization model where t appears in the objective (directly or indirectly), the inequality becomes tight at optimality, yielding $t = f(x)$. If the points do *not* define a convex function, this same construction models the *convex envelope* of the data, not the original nonconvex curve.

Example 5.10 (Convex piecewise-linear cost). Consider breakpoints $(x_1, f_1) = (0, 0)$, $(x_2, f_2) = (1, 1)$, $(x_3, f_3) = (2, 3)$ (slopes 1 then 2, hence convex). To minimize t subject to $t \ge f(x)$ and $x \in [0, 2]$, one can use variables $\lambda_1, \lambda_2, \lambda_3 \ge 0$ and write

$$x = \lambda_2 + 2\lambda_3, \qquad t \ge \lambda_2 + 3\lambda_3, \qquad \lambda_1 + \lambda_2 + \lambda_3 = 1.$$

At optimality, (x, t) lies on the lower hull of the three points, i.e. on the piecewise-linear graph of f.

Nonconvex case (MILP formulation). If the function is not convex, selecting the active segment typically requires binary variables. Consider breakpoints $(x_0, f_0), \dots, (x_k, f_k)$ and segments $i \in [k]$ between (x_{i-1}, f_{i-1}) and (x_i, f_i). Introducing binaries $b_i \in \{0,1\}$ and continuous variables $t_i \in [0,1]$, one can enforce "(x, f) lies on exactly one segment" by

$$\sum_{i \in [k]} b_i = 1, \qquad 0 \le t_i \le b_i, \ \forall i \in [k],$$

together with the linear identities

$$x = \sum_{i \in [k]} \big(x_{i-1} b_i + (x_i - x_{i-1}) t_i\big), \qquad f = \sum_{i \in [k]} \big(f_{i-1} b_i + (f_i - f_{i-1}) t_i\big).$$

Example 5.11 (Why binaries matter (relaxation effect)). Consider the nonconvex "$\wedge$"-shaped function defined by the breakpoints $(0,0)$, $(1,1)$, $(2,0)$. Using the formulation above with $k = 2$, the binary constraints ensure that (x, f) lies on exactly one of the two segments. If one relaxes $b_1, b_2 \in \{0,1\}$ to $b_1, b_2 \in [0,1]$, the formulation allows convex combinations of the two segments. For instance, taking $b_1 = b_2 = \frac{1}{2}$, $t_1 = 0$, and $t_2 = \frac{1}{2}$ gives $x = 1$ and $f = 0$, which is strictly below the true value $f(1) = 1$ on the graph. This illustrates why nonconvex piecewise-linear modeling typically needs integer variables if one wants the exact curve rather than its convex envelope.

5.5.5 Products and McCormick linearization

Products involving binaries can be linearized with additional variables and inequalities.

$\diamondsuit$ **Proposition 5.6** (Binary product). *Let $b_1, b_2 \in \{0,1\}$ and introduce a variable $y \in [0,1]$ intended to satisfy $y = b_1 b_2$. Then the equivalence*

$$y = b_1 b_2 \quad \Longleftrightarrow \quad y \le b_1, \ y \le b_2, \ y \ge b_1 + b_2 - 1, \ y \ge 0$$

holds when b_1, b_2 are binary.

Proof. If $y = b_1 b_2$ with $b_1, b_2 \in \{0,1\}$, then $y \le b_1$, $y \le b_2$, and $y \ge b_1 + b_2 - 1$ are immediate to check case by case; also $y \ge 0$.

Conversely, assume the four inequalities hold with $b_1, b_2 \in \{0,1\}$. If $b_1 = 0$ or $b_2 = 0$, then $y \le 0$ and $y \ge 0$, hence $y = 0 = b_1 b_2$. If $b_1 = b_2 = 1$, then $y \le 1$ and $y \ge 1$, hence $y = 1 = b_1 b_2$. □

$\diamondsuit$ **Proposition 5.7** (Binary–continuous product (McCormick inequalities)). *Let $b \in \{0,1\}$, let $x \in [\underline{x}, \overline{x}]$, and introduce y intended to satisfy $y = bx$. Then the following linear inequalities are valid and describe the convex hull of (b, x, y):*

$$y \le \overline{x}\, b, \qquad y \ge \underline{x}\, b, \qquad y \le x - \underline{x}(1 - b), \qquad y \ge x - \overline{x}(1 - b).$$

Proof. Each inequality is valid when $b \in \{0,1\}$ and $y = bx$: if $b = 0$ then $y = 0$ and the inequalities reduce to bounds consistent with $x \in [\underline{x}, \overline{x}]$; if $b = 1$ then $y = x$ and they reduce to $x \in [\underline{x}, \overline{x}]$.

To see that they describe the convex hull, let $S := \{(b, x, y) \mid b \in \{0,1\},\ x \in [\underline{x}, \overline{x}],\ y = bx\}$. Every convex combination of points in S satisfies the inequalities, so $\operatorname{conv}(S)$ is contained in the polyhedron they define. Conversely, take (b, x, y) satisfying the inequalities with $b \in [0,1]$. If $b = 0$ (resp. $b = 1$), the inequalities force $y = 0$ (resp. $y = x$), hence $(b, x, y) \in S \subseteq \operatorname{conv}(S)$. If $b \in (0,1)$, set $x_1 := y/b$ and $x_0 := (x - y)/(1 - b)$. The inequalities imply $\underline{x} \le x_0, x_1 \le \overline{x}$, and one checks that

$$(b, x, y) = (1 - b)\,(0, x_0, 0) + b\,(1, x_1, x_1),$$

which is a convex combination of two points of S. Thus the polyhedron defined by the inequalities is contained in $\operatorname{conv}(S)$, proving equality. □

5.6 Worked modeling examples

The modeling patterns above are most useful when combined in a single formulation. We conclude with two self-contained examples that illustrate how binary variables, linking constraints, and auxiliary continuous variables can encode fairly rich decisions.

Example 5.12 (Fixed-charge transportation with lane selection). Let S be a set of supply nodes and D a set of demand nodes. For $s \in S$ let $a_s \geq 0$ be the available supply, and for $d \in D$ let $b_d \geq 0$ be the required demand. Shipping one unit from s to d costs $c_{sd} \geq 0$. In addition, using the lane (s, d) at all incurs a fixed cost $f_{sd} \geq 0$ (e.g. contracting a carrier, opening a route, or activating a warehouse-to-store connection).

Introduce shipment variables $x_{sd} \geq 0$ and binary activation variables $y_{sd} \in \{0, 1\}$. If U_{sd} is a valid upper bound on x_{sd}, the fixed-charge transportation problem can be written as

$$\begin{aligned}
\min_{x,y} \quad & \sum_{s \in S} \sum_{d \in D} \big(c_{sd} x_{sd} + f_{sd} y_{sd}\big) && && \text{(5.7a)}\\
\text{s.t.} \quad & \sum_{d \in D} x_{sd} \leq a_s && \forall s \in S && \text{(5.7b)}\\
& \sum_{s \in S} x_{sd} \geq b_d && \forall d \in D && \text{(5.7c)}\\
& 0 \leq x_{sd} \leq U_{sd} y_{sd} && \forall s \in S,\ \forall d \in D && \text{(5.7d)}\\
& y_{sd} \in \{0, 1\} && \forall s \in S,\ \forall d \in D && \text{(5.7e)}
\end{aligned}$$

The linking constraints $x_{sd} \leq U_{sd} y_{sd}$ are of the big-M type: if $y_{sd} = 0$ then $x_{sd} = 0$, whereas if $y_{sd} = 1$ the lane can carry flow up to U_{sd}. If one wants to limit the number of used lanes, one can add the cardinality constraint $\sum_{s,d} y_{sd} \leq K$ for a chosen integer K.

Example 5.13 (Unit commitment with piecewise-linear heat rates). Consider a single generator over periods $t = 1, \ldots, T$. Let $u_t \in \{0, 1\}$ indicate whether the unit is on, $p_t \geq 0$ the power output, and $v_t \in \{0, 1\}$ a startup indicator. The generator has minimum and maximum outputs $P^{\min}$ and $P^{\max}$, and its fuel cost as a function of output is modeled by a *nonconvex* piecewise-linear heat rate curve with breakpoints

$$P^{\min} = \bar{p}_0 < \bar{p}_1 < \bar{p}_2 = P^{\max},$$
$$\text{slopes } \sigma_1 > \sigma_2 \geq 0 \quad \text{(concave, i.e. decreasing marginal cost).}$$

On/off and startup logic. The minimum/maximum output constraints are the on/off pattern of Example 5.9:

$$P^{\min} u_t \leq p_t \leq P^{\max} u_t \qquad \forall t.$$

A startup occurs at period t when the unit was off at $t-1$ and on at t, which requires $v_t = 1$ through the logical implication $u_t - u_{t-1} \leq v_t$. Assuming $c^s > 0$, because v_t appears only through the strictly positive startup cost in the minimization objective, an optimal solution sets $v_t = 1$ exactly when a startup requires it. Together with the convention $u_0 = 0$, this gives

$$u_t - u_{t-1} \leq v_t, \quad v_t \in \{0, 1\} \qquad \forall t.$$

Minimum up/down times can be added through logical temporal constraints. A minimum up-time requires $u_t - u_{t-1} \leq u_{t'}$ for $t' = t, \ldots, \min(t + L^{\text{up}} - 1, T)$, so that the unit remains on after a startup. Analogously, a minimum down-time requires $u_{t-1} - u_t \leq 1 - u_{t'}$ for $t' = t, \ldots, \min(t + L^{\text{down}} - 1, T)$, so that the unit remains off after a shutdown.

Piecewise-linear cost via binary segment selection. Since the heat rate curve is concave (not convex), the LP trick of Section 5.5 would only model its convex envelope, which underestimates costs. We therefore use binary segment selectors $b_{t,i} \in \{0,1\}$ for $i \in \{1,2\}$, with $\delta_{t,i} \in [0,1]$ the fractional position within segment i:

$$b_{t,1} + b_{t,2} = u_t, \qquad 0 \le \delta_{t,i} \le b_{t,i}, \quad i \in \{1,2\}.$$

Note that when $u_t = 0$ both selectors are forced to 0. The power output and the fuel cost h_t are then recovered by

$$p_t = \sum_{i=1}^{2} \big(\bar{p}_{i-1}\, b_{t,i} + (\bar{p}_i - \bar{p}_{i-1})\delta_{t,i}\big),$$
$$h_t = \sum_{i=1}^{2} \big(f_{i-1}\, b_{t,i} + (f_i - f_{i-1})\delta_{t,i}\big),$$

where the breakpoint costs, measured relative to the minimum-output operating point, are

$$f_0 = 0, \qquad f_1 = \sigma_1(\bar{p}_1 - \bar{p}_0), \qquad f_2 = f_1 + \sigma_2(\bar{p}_2 - \bar{p}_1).$$

Any baseline cost associated with operating at $P^{\min}$ can be included in the fixed running cost $c^f u_t$.

Full formulation. Denoting by $c^s > 0$ the startup cost and c^f the fixed running cost, the MILP reads

$$\begin{aligned}
\text{Min} \quad & \sum_{t=1}^{T} \big(h_t + c^f u_t + c^s v_t\big) && && \text{(5.8a)}\\
\text{s.t.} \quad & P^{\min} u_t \le p_t \le P^{\max} u_t && \forall t && \text{(5.8b)}\\
& u_t - u_{t-1} \le v_t && \forall t && \text{(5.8c)}\\
& b_{t,1} + b_{t,2} = u_t && \forall t && \text{(5.8d)}\\
& p_t = \sum_{i=1}^{2} \big(\bar{p}_{i-1} b_{t,i} + (\bar{p}_i - \bar{p}_{i-1})\delta_{t,i}\big) && \forall t && \text{(5.8e)}\\
& h_t = \sum_{i=1}^{2} \big(f_{i-1} b_{t,i} + (f_i - f_{i-1})\delta_{t,i}\big) && \forall t && \text{(5.8f)}\\
& 0 \le \delta_{t,i} \le b_{t,i} && \forall t,\ i && \text{(5.8g)}\\
& p_t, h_t \ge 0, \quad u_t, v_t, b_{t,i} \in \{0,1\} && \forall t,\ i && \text{(5.8h)}
\end{aligned}$$

This example combines on/off decisions (big-M), a logical startup implication, and binary segment selection for a nonconvex piecewise-linear cost — three of the core MILP modeling patterns of this chapter.

5.7 Exercises

Exercise 5.1 *Minimum-cost one-to-one assignment*
We have a set of agents U and a set of tasks V with $|U| = |V| = n$. Each agent must be assigned to exactly one task, and each task must be assigned to exactly one agent. The cost of assigning agent $u \in U$ to task $v \in V$ is $c_{uv} \geq 0$.

1. Model the problem as a MILP (variables, objective, constraints).
2. Adapt the model if some pairs (u, v) are forbidden.

Solution

Exercise 5.2 *Shortest path*
Let $G = (V, E)$ be a directed graph with arc costs $c_{uv} \geq 0$ for all $(u, v) \in E$. We seek a shortest path from $s \in V$ to $t \in V$.

Give a MILP formulation (variables, objective, constraints).

Solution

Exercise 5.3 *Traveling salesman problem (TSP)*
Let $G = (V, E)$ be a graph with arc costs c_{uv} for $(u, v) \in E$. We seek a tour visiting each node exactly once.

Give a MILP formulation of the problem.

Solution

Exercise 5.4 *Vehicle routing*
A set N of customers must be served by a set K of identical vehicles, each with capacity Q. Each customer $i \in N$ has demand $d_i > 0$ and must be served by exactly one vehicle. The travel cost between customers $i, j \in N$ is $c_{ij} \geq 0$. All vehicles start and end at a depot $0 \notin N$. Using a vehicle incurs a fixed cost $f > 0$.

Give a MILP that minimizes the total travel cost while serving all customers without exceeding vehicle capacities. Provide variables, objective, and essential constraints.

Solution

Exercise 5.5 *Capacitated facility location*
Sets I (customers) and J (potential facilities). Each customer $i \in I$ has demand $d_i > 0$. Opening facility $j \in J$ costs f_j and provides capacity Q_j. Shipping one unit from j to i costs $c_{ij} \geq 0$.

Give a compact MILP that minimizes fixed plus shipping costs while satisfying all demand. Provide variables, objective, and essential constraints. Propose a simple valid inequality that strengthens the LP relaxation.

Solution

Exercise 5.6 *0–1 knapsack*
We have a set of items I. Each item $i \in I$ has benefit $b_i \geq 0$ and weight $w_i > 0$. The knapsack capacity is $W > 0$. We want to select items to maximize the total benefit without exceeding capacity.

Give a compact MILP (variables, objective, constraints).

Then propose a valid inequality based on the heaviest items.

Solution

Exercise 5.7 *Uncapacitated lot sizing (ULS)*
A product must meet demand $d_t \geq 0$ over periods $t = 1, \ldots, T$. Producing $u_t \geq 0$ in period t costs $c_t u_t$ plus a fixed setup cost f_t if production is started (binary variable y_t). Inventory $s_t \geq 0$ costs $h_t s_t$. The initial stock s_0 is given, and s_T is free. Assume $c_t \geq 0$, $f_t \geq 0$, and $h_t \geq 0$ for all t.

1. Model the problem as a MILP (variables, objective, essential constraints).
2. Propose a simple strengthening inequality to improve the LP relaxation.

Solution

Exercise 5.8 *Single-machine scheduling*
We consider a set of jobs J processed on a single machine. For each job $j \in J$, we know a processing time $\tau_j > 0$, a release date $r_j \geq 0$ (when it can start), and a due date d_j (by which it should be completed). We want to minimize the total tardiness $\sum_{j \in J} \eta_j$, where η_j is the completion-time delay relative to d_j.

Formulate a MILP for the problem.

Solution

Exercise 5.9 *Pricing problem*
A producer chooses a unit price p for a product. Realized demand is $q = A - Bp$ (with $A > 0$, $B > 0$). There is a capacity $q \leq Q^{\max}$, and we impose bounds $p \in [\underline{p}, \overline{p}]$, $q \in [\underline{q}, \overline{q}]$ with $0 \leq \underline{q} < \overline{q} \leq Q^{\max}$. The unit cost is $c \geq 0$.

1. Write the optimization model.
2. Construct an LP relaxation using McCormick envelopes.
3. Is this relaxation exact in general?

Solution

Appendix A

Mathematical background

This appendix collects a few standard facts from topology, differential calculus, and linear algebra that are repeatedly used in the textbook. Throughout, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$, and $\mathbf{x}\cdot\mathbf{y} := \mathbf{x}^\top\mathbf{y}$ denotes the associated inner product.

A.1 Topology in $\mathbb{R}^n$

A.1.1 Compactness in $\mathbb{R}^n$

Definition A.1 (Compact set)**.** *A set $K \subseteq \mathbb{R}^n$ is* compact *if every sequence $(\mathbf{x}^k)_{k\in\mathbb{N}}$ in K admits a subsequence that converges to a point in K.*

In Euclidean spaces, compactness admits a particularly simple characterization.

♢ **Theorem A.2** (Heine–Borel theorem)**.** *A set $K \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

♢ **Theorem A.3** (Continuous image of a compact set)**.** *Let $K \subseteq \mathbb{R}^n$ be compact and let $f : K \to \mathbb{R}^p$ be continuous. Then $f(K) \subseteq \mathbb{R}^p$ is compact.*

Proof. Let (y^k) be any sequence in $f(K)$. For each k, pick $\mathbf{x}^k \in K$ such that $y^k = f(\mathbf{x}^k)$. Since K is compact, there exists a subsequence $\mathbf{x}^{k_\ell} \to \mathbf{x}^\star \in K$. By continuity, $y^{k_\ell} = f(\mathbf{x}^{k_\ell}) \to f(\mathbf{x}^\star) \in f(K)$, hence $f(K)$ is compact. □

♢ **Corollary A.4** (Extreme value theorem)**.** *Let $K \subseteq \mathbb{R}^n$ be compact and let $f : K \to \mathbb{R}$ be continuous. Then f attains its minimum and its maximum on K.*

Proof. By Theorem A.3, the set $f(K) \subseteq \mathbb{R}$ is compact, hence closed and bounded. Therefore $\min f(K)$ and $\max f(K)$ exist and belong to $f(K)$. □

A.2 Differential calculus in $\mathbb{R}^n$

A.2.1 Derivative, gradient, and Hessian

Definition A.5 (Differentiability)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$. We say that f is* (Fréchet) differentiable *at $\mathbf{x} \in U$ if there exists a linear form $\ell : \mathbb{R}^n \to \mathbb{R}$ such that*

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \ell(\mathbf{h}) + o(\|\mathbf{h}\|) \qquad \text{as } \mathbf{h} \to 0.$$

The linear form ℓ is unique and is denoted by $Df(\mathbf{x})$.

In finite dimension, linear forms identify with vectors via the Euclidean inner product.

Definition A.6 (Gradient)**.** *If f is differentiable at $\mathbf{x}$, the* gradient $\nabla f(\mathbf{x}) \in \mathbb{R}^n$ *is the unique vector such that*

$$Df(\mathbf{x})[\mathbf{h}] = \nabla f(\mathbf{x})^\top \mathbf{h} \qquad \forall \mathbf{h} \in \mathbb{R}^n.$$

♢ **Proposition A.7** (Coordinate formula for the gradient)**.** *Assume that $f : U \to \mathbb{R}$ is differentiable at $\mathbf{x} \in U$. Let $(e_i)_{i=1}^n$ be the canonical basis of $\mathbb{R}^n$. Then each partial derivative $\partial_i f(\mathbf{x}) := \frac{\partial f}{\partial x_i}(\mathbf{x})$ exists and*

$$\nabla f(\mathbf{x}) = \big(\partial_1 f(\mathbf{x}), \dots, \partial_n f(\mathbf{x})\big)^\top.$$

Proof. For each i, consider the one-dimensional function $\varphi_i(t) := f(\mathbf{x} + te_i)$. Differentiability of f at $\mathbf{x}$ gives

$$\varphi_i(t) = f(\mathbf{x}) + t\,\nabla f(\mathbf{x})^\top e_i + o(|t|),$$

so $\varphi_i'(0)$ exists and equals $\nabla f(\mathbf{x})^\top e_i$. By definition, $\varphi_i'(0) = \partial_i f(\mathbf{x})$ and $\nabla f(\mathbf{x})^\top e_i$ is the i-th coordinate of $\nabla f(\mathbf{x})$. □

Definition A.8 (Hessian)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable on U. We say that f is* twice differentiable *at $\mathbf{x} \in U$ if ∇f is differentiable at $\mathbf{x}$. The corresponding Jacobian matrix of ∇f is the* Hessian *and is denoted by $\nabla^2 f(\mathbf{x}) \in \mathbb{R}^{n\times n}$.*

♢ **Proposition A.9** (Coordinate formula for the Hessian)**.** *Assume that $f : U \to \mathbb{R}$ is twice differentiable at $\mathbf{x} \in U$. Then all second-order partial derivatives exist and*

$$\nabla^2 f(\mathbf{x}) = \Big(\partial_{ij}^2 f(\mathbf{x})\Big)_{i,j\in[n]}, \qquad \partial_{ij}^2 f(\mathbf{x}) := \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}).$$

Proof. Since ∇f is differentiable at $\mathbf{x}$, each component $\partial_i f$ is differentiable at $\mathbf{x}$. By Proposition A.7 applied to the mapping $\partial_i f$, its gradient at $\mathbf{x}$ is the vector of partial derivatives $(\partial_{1i}^2 f(\mathbf{x}), \dots, \partial_{ni}^2 f(\mathbf{x}))^\top$. These are exactly the entries of the Jacobian of ∇f, hence of $\nabla^2 f(\mathbf{x})$. □

♢ **Theorem A.10** (Symmetry of the Hessian)**.** *If f is twice continuously differentiable on an open set $U \subseteq \mathbb{R}^n$, then for all $\mathbf{x} \in U$,*

$$\nabla^2 f(\mathbf{x}) \text{ is symmetric, i.e. } \partial_{ij}^2 f(\mathbf{x}) = \partial_{ji}^2 f(\mathbf{x}) \quad \forall i, j.$$

Proof. This is the classical Schwarz (Clairaut) theorem on the equality of mixed partial derivatives for C^2 functions. □

A.2.2 Taylor expansion and identification of derivatives

♢ **Theorem A.11** (Second-order Taylor expansion)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be twice differentiable at $\mathbf{x} \in U$. Then, as $\mathbf{h} \to 0$,*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2}\,\mathbf{h}^\top \nabla^2 f(\mathbf{x})\,\mathbf{h} + o(\|\mathbf{h}\|^2).$$

Conversely, in many computations one starts from an expansion of $f(\mathbf{x} + \mathbf{h})$ and *reads off* the coefficients of $\mathbf{h}$ and $\mathbf{h}^\top(\cdot)\mathbf{h}$. This provides a convenient and common way of identifying the gradient and Hessian at $\mathbf{x}$.

♢ **Proposition A.12** (First-order expansion determines the gradient)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$. Fix $\mathbf{x} \in U$. If there exists $g \in \mathbb{R}^n$ such that*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + g^\top \mathbf{h} + o(\|\mathbf{h}\|) \qquad \text{as } \mathbf{h} \to 0,$$

then f is differentiable at $\mathbf{x}$ and $\nabla f(\mathbf{x}) = g$.

Proof. The mapping $\mathbf{h} \mapsto g^\top \mathbf{h}$ is linear, so the displayed relation is exactly Definition A.5 with $Df(\mathbf{x})[\mathbf{h}] = g^\top \mathbf{h}$. By Definition A.6, the unique representing vector is $\nabla f(\mathbf{x}) = g$. □

♢ **Proposition A.13** (Second-order expansion determines the Hessian)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$. Fix $\mathbf{x} \in U$. If there exist $g \in \mathbb{R}^n$ and a symmetric matrix $H \in \mathbb{R}^{n\times n}$ such that*

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + g^\top \mathbf{h} + \frac{1}{2}\,\mathbf{h}^\top H \mathbf{h} + o(\|\mathbf{h}\|^2) \qquad \text{as } \mathbf{h} \to 0,$$

then f is differentiable at $\mathbf{x}$ with $\nabla f(\mathbf{x}) = g$. Moreover, the matrix H is uniquely determined by the expansion; when such an H exists, it is natural to call it the *Hessian of f at $\mathbf{x}$. If f is twice differentiable at $\mathbf{x}$ in the sense of Definition A.8, then this H coincides with the Hessian matrix $\nabla^2 f(\mathbf{x})$ appearing in Theorem A.11.*

Proof. The differentiability statement follows from Proposition A.12 by ignoring the $O(\|\mathbf{h}\|^2)$ term.

For uniqueness, assume that two symmetric matrices H_1, H_2 satisfy

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + g^\top \mathbf{h} + \frac{1}{2}\,\mathbf{h}^\top H_1 \mathbf{h} + o(\|\mathbf{h}\|^2) = f(\mathbf{x}) + g^\top \mathbf{h} + \frac{1}{2}\,\mathbf{h}^\top H_2 \mathbf{h} + o(\|\mathbf{h}\|^2).$$

Subtracting the two expansions yields $\mathbf{h}^\top (H_1 - H_2)\mathbf{h} = o(\|\mathbf{h}\|^2)$ as $\mathbf{h} \to 0$. Fix any $v \in \mathbb{R}^n$ and set $\mathbf{h} = tv$. Dividing by t^2 and letting $t \to 0$ gives $v^\top (H_1 - H_2) v = 0$ for all v. With $v = e_i$ we get $(H_1 - H_2)_{ii} = 0$. With $v = e_i + e_j$ we get $2(H_1 - H_2)_{ij} = 0$ for $i \neq j$. Hence $H_1 = H_2$.

If f is twice differentiable at $\mathbf{x}$ in the sense of Definition A.8, then Theorem A.11 provides an expansion with $H = \nabla^2 f(\mathbf{x})$, and uniqueness implies that this matrix must coincide with the one in the present statement. □

A.2.3 Chain rule

♢ **Theorem A.14** (Chain rule)**.** *Let $U \subseteq \mathbb{R}^p$ and $V \subseteq \mathbb{R}^n$ be open. Let $g : U \to V$ be differentiable at $u \in U$, and let $f : V \to \mathbb{R}$ be differentiable at $g(u)$. Then $f \circ g$ is differentiable at u and*

$$\nabla (f \circ g)(u) = Dg(u)^\top \nabla f(g(u)),$$

where $Dg(u) \in \mathbb{R}^{n\times p}$ denotes the Jacobian matrix of g at u.

♢ **Corollary A.15** (Chain rule for an affine map)**.** *Let $A \in \mathbb{R}^{n\times p}$ and $b \in \mathbb{R}^n$, and define $g(u) := Au + b$. If f is differentiable at $Au + b$, then $f \circ g$ is differentiable at u and*

$$\nabla (f(Au+b)) = A^\top \nabla f(Au+b).$$

Proof. Here $Dg(u) = A$ for all u, so the result follows from Theorem A.14. □

Example A.1 (Least squares: gradient and Hessian). Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and consider

$$f(\mathbf{x}) := \frac{1}{2}\|A\mathbf{x} - b\|^2 = \frac{1}{2}(A\mathbf{x} - b)^\top (A\mathbf{x} - b).$$

Set $g(\mathbf{x}) := A\mathbf{x} - b$ and $\phi(z) := \frac{1}{2}\|z\|^2$. Then $f = \phi \circ g$. Since $\nabla\phi(z) = z$ and $Dg(\mathbf{x}) = A$, Corollary A.15 yields

$$\nabla f(\mathbf{x}) = A^\top (A\mathbf{x} - b).$$

Moreover, ∇f is affine, hence differentiable everywhere, and

$$\nabla^2 f(\mathbf{x}) = A^\top A \qquad \forall \mathbf{x} \in \mathbb{R}^n.$$

Example A.2 (Quadratic form). Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and define $f(\mathbf{x}) := \mathbf{x}^\top Q \mathbf{x}$. Expanding at $\mathbf{x}$ gives, for $\mathbf{h} \to 0$,

$$f(\mathbf{x}+\mathbf{h}) = \mathbf{x}^\top Q \mathbf{x} + 2\mathbf{x}^\top Q \mathbf{h} + \mathbf{h}^\top Q \mathbf{h}.$$

Comparing with Proposition A.13 shows that

$$\nabla f(\mathbf{x}) = 2Q\mathbf{x}, \qquad \nabla^2 f(\mathbf{x}) = 2Q.$$

A.2.4 Quadratic functions

♢ **Proposition A.16** (Gradient and Hessian of a quadratic)**.** *Let $Q \in \mathbb{R}^{n\times n}$ be symmetric, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Define*

$$f(\mathbf{x}) := \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} + b^\top \mathbf{x} + c.$$

Then f is twice differentiable on $\mathbb{R}^n$ and

$$\nabla f(\mathbf{x}) = Q\mathbf{x} + b, \qquad \nabla^2 f(\mathbf{x}) = Q \qquad \forall \mathbf{x} \in \mathbb{R}^n.$$

Proof. For $\mathbf{h} \in \mathbb{R}^n$,

$$f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) = \frac{1}{2}\big(2\mathbf{x}^\top Q\mathbf{h} + \mathbf{h}^\top Q\mathbf{h}\big) + b^\top \mathbf{h} = (Q\mathbf{x}+b)^\top \mathbf{h} + \frac{1}{2}\,\mathbf{h}^\top Q\mathbf{h}.$$

This is of the form $\nabla f(\mathbf{x})^\top \mathbf{h} + o(\|\mathbf{h}\|)$, and differentiating the affine map $\mathbf{x} \mapsto Q\mathbf{x} + b$ gives $\nabla^2 f(\mathbf{x}) = Q$. □

A.3 Linear algebra reminders

A.3.1 Orthogonal matrices

Definition A.17 (Orthogonal matrix)**.** *A matrix $U \in \mathbb{R}^{n\times n}$ is* orthogonal *if*

$$U^\top U = I$$

(equivalently, $UU^\top = I$). Equivalently, the columns of U form an orthonormal basis of $\mathbb{R}^n$.

A.3.2 Eigenvalues and spectrum

Definition A.18 (Eigenvalue, eigenvector, spectrum)**.** *Let $A \in \mathbb{R}^{n\times n}$. A scalar $\lambda \in \mathbb{R}$ is an* eigenvalue *of A if there exists a nonzero vector $u \in \mathbb{R}^n$ such that $Au = \lambda u$. Any such vector u is an* eigenvector *associated with λ. The set of eigenvalues of A is called the* spectrum *of A and is denoted by $\sigma(A)$.*

A.3.3 Spectral theorem

♢ **Theorem A.19** (Spectral theorem (real symmetric matrices))**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then there exist an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ and real numbers $\lambda_1, \dots, \lambda_n$ such that*

$$A = U\,\mathrm{diag}(\lambda_1, \dots, \lambda_n)\,U^\top.$$

The numbers λ_i are the eigenvalues of A (counted with multiplicity), and the columns of U are corresponding orthonormal eigenvectors.

A.3.4 Positive (semi-)definite matrices and the spectrum

Definition A.20 (Positive semidefinite / positive definite)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. We say that A is* positive semidefinite*, and write $A \succeq 0$, if $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. We say that A is* positive definite*, and write $A \succ 0$, if $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$.*

♢ **Proposition A.21** (Spectrum characterization)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\lambda_1, \dots, \lambda_n$. Then:*

$$A \succeq 0 \iff \lambda_i \geq 0\ \forall i, \qquad A \succ 0 \iff \lambda_i > 0\ \forall i.$$

Proof. Let $A = U\Lambda U^\top$ be a spectral decomposition with $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. For any $\mathbf{x} \in \mathbb{R}^n$, set $\mathbf{z} := U^\top \mathbf{x}$. Then $\mathbf{x}^\top A\mathbf{x} = \mathbf{z}^\top \Lambda \mathbf{z} = \sum_{i=1}^n \lambda_i \mathbf{z}_i^2$. If all $\lambda_i \geq 0$ (resp. > 0), then $\mathbf{x}^\top A\mathbf{x} \geq 0$ for all $\mathbf{x}$ (resp. > 0 for $\mathbf{x} \neq 0$), so $A \succeq 0$ (resp. $A \succ 0$). Conversely, if some $\lambda_i < 0$, taking $\mathbf{x}$ equal to the corresponding eigenvector gives $\mathbf{x}^\top A\mathbf{x} = \lambda_i < 0$, contradicting $A \succeq 0$; similarly if some $\lambda_i = 0$, A cannot be positive definite. □

Solutions

Solution 2.1

Back to Exercise

1. The objective is continuous and coercive on $\mathbb{R}^n$, hence it attains a global minimum. Differentiating and setting the gradient to zero yields, for each coordinate j,

$$\sum_{i=1}^{s} 2(x_j - p_j^i) = 0 \quad \Rightarrow \quad x_j = \frac{1}{s}\sum_{i=1}^{s} p_j^i.$$

2. This is the barycenter (centroid) of the points $\mathbf{p}^i$.

Solution 2.2

Back to Exercise

1. The variable $\mathbf{x}$ represents the center of a ball and $\max_i \|\mathbf{x} - \mathbf{p}^i\|_2$ its radius. Minimizing this radius is exactly finding the smallest-radius ball containing all points.
   More formally, introducing a radius variable $r \geq 0$, the smallest enclosing ball problem is

$$\begin{array}{ll} \text{Min} & r \\ \text{s.t.} & \|\mathbf{x} - \mathbf{p}^i\|_2^2 \leq r^2 \qquad \forall i, \\ & r \geq 0. \end{array}$$

   It is equivalent to (R) since for any feasible $\mathbf{x}$ in (R), $(\mathbf{x}, \max_i \|\mathbf{x} - \mathbf{p}^i\|_2)$ is feasible with the same value, and conversely any feasible $(\mathbf{x}, r)$ gives a point $\mathbf{x}$ with objective value $\leq r$ in (R).
2. The objective $\mathbf{x} \mapsto \max_i \|\mathbf{x} - \mathbf{p}^i\|_2$ is continuous and coercive, hence it attains a minimum on $\mathbb{R}^n$.
3. Introduce $t \geq 0$ and impose $\|\mathbf{x} - \mathbf{p}^i\|_2 \leq t$ for all i. Squaring gives polynomial inequalities:

$$\begin{array}{ll} \text{Min} & t \\ \text{s.t.} & \sum_j (x_j - p_j^i)^2 - t^2 \leq 0 \qquad \forall i, \\ & t \geq 0. \end{array}$$

Solution 2.3

Back to Exercise

1. If $A = 0$, the problem is trivial. Otherwise, consider $g(\mathbf{y}) = \|\mathbf{y} - \mathbf{b}\|_2$ on $\mathbb{R}^m$, which is continuous and coercive. Let $X := \mathrm{Im}(A)$, which is a nonempty closed subset of $\mathbb{R}^m$ (it is a finite-dimensional subspace). Therefore g attains a minimum on X; call a minimizer $\mathbf{y}^\sharp \in X$. Any $\mathbf{x}^\sharp$ satisfying $A\mathbf{x}^\sharp = \mathbf{y}^\sharp$ is a minimizer of $\|A\mathbf{x} - \mathbf{b}\|_2$.
2. Squaring the norm does not change the set of minimizers and makes the objective differentiable:

$$\operatorname*{Min}_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2^2.$$

   First-order optimality conditions give $A^\top(A\mathbf{x} - \mathbf{b}) = \mathbf{0}$, i.e. $A^\top A\mathbf{x} = A^\top \mathbf{b}$.

Solution 2.4

Back to Exercise

1. The lengths of the two segments are $\sqrt{x^2+a^2}$ and $\sqrt{(L-x)^2+b^2}$, respectively. Dividing each length by the corresponding speed gives the stated expression.
2. The function T is continuous and coercive on $\mathbb{R}$: as $|x| \to \infty$, its first term already tends to $+\infty$. Hence T attains a minimum.
3. Direct differentiation gives
$$T'(x) = \frac{x}{v_1\sqrt{x^2+a^2}} - \frac{L-x}{v_2\sqrt{(L-x)^2+b^2}}$$
and
$$T''(x) = \frac{a^2}{v_1(x^2+a^2)^{3/2}} + \frac{b^2}{v_2((L-x)^2+b^2)^{3/2}} > 0.$$
Thus T is strictly convex, so its minimizer is unique.
4. Since $T'(0) < 0$ and $T'(L) > 0$, the minimizer $x^\sharp$ lies in $(0, L)$. Therefore
$$\sin\theta_1 = \frac{x^\sharp}{\sqrt{(x^\sharp)^2+a^2}}, \qquad \sin\theta_2 = \frac{L-x^\sharp}{\sqrt{(L-x^\sharp)^2+b^2}}.$$
The condition $T'(x^\sharp) = 0$ becomes
$$\frac{\sin\theta_1}{v_1} = \frac{\sin\theta_2}{v_2},$$
which is Snell's law.
5. If $v_1 = v_2$, then $\theta_1 = \theta_2$. Equivalently, $x^\sharp/a = (L-x^\sharp)/b$, so A, C, and B are collinear: the optimal trajectory is the straight segment from A to B.

Solution 2.5

Back to Exercise

1. The objective is continuous and the feasible set $\{(x,y,z) : x^2+y^2+z^2 = 1\}$ is nonempty and compact, so a minimizer exists.
2. The constraint qualification holds since the gradient of $x^2+y^2+z^2-1$ is never zero on the sphere. KKT conditions yield the existence of $\lambda \in \mathbb{R}$ such that, at an optimum (x,y,z),
$$3x^2 + 2\lambda x = 0, \qquad -3y^2 + 2\lambda y = 0, \qquad 3z^2 + 2\lambda z = 0.$$
This forces each coordinate to be either 0 or linked by $x = z = -y$. Checking the resulting cases under $x^2+y^2+z^2 = 1$ shows that the minimum value is -1, attained for instance at $(x,y,z) = (-1,0,0)$, $(0,1,0)$, or $(0,0,-1)$.

Solution 2.6

Back to Exercise

1. Assume $a_j \le 0$ for some j.
If $a_i \le 0$ for all i, then the constraint $\mathbf{a}\cdot\mathbf{x} = 1$ is impossible with $\mathbf{x} > 0$, so the problem is infeasible and the optimal value is $-\infty$ by convention.
Otherwise, there exists k with $a_k > 0$. Then the optimal value is $+\infty$: one can send $x_j \to +\infty$ and adjust x_k to keep $\mathbf{a}\cdot\mathbf{x} = 1$ while maintaining all other coordinates positive. Since $s_j \ln(x_j) \to +\infty$, the objective is unbounded above.
2. If $a_i > 0$ for all i, then for any feasible $\mathbf{x}$ we have $x_i \le 1/a_i$. Moreover, for any finite value K achieved by some feasible point, any optimizer must lie in a compact box $[\varepsilon_1, 1/a_1] \times \cdots \times [\varepsilon_n, 1/a_n]$ for suitable $\varepsilon_i > 0$, because $\ln(x_i) \to -\infty$ as $x_i \downarrow 0$. Therefore an optimal solution exists by compactness.
3. KKT conditions apply (affine equality constraint, $\mathbf{x} > 0$). There exists $\lambda \in \mathbb{R}$ such that for every i,
$$\frac{s_i}{x_i^\sharp} + \lambda a_i = 0 \quad \Rightarrow \quad x_i^\sharp = -\frac{s_i}{\lambda a_i}.$$
Using $\mathbf{a}\cdot\mathbf{x}^\sharp = 1$ gives $\lambda = -\sum_{j=1}^n s_j$, hence
$$x_i^\sharp = \frac{s_i}{a_i \sum_{j=1}^n s_j} \qquad \forall i \in [n].$$

Solution 2.7

Back to Exercise

1. The feasible set $\{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$ is compact and nonempty, and the objective is continuous, hence a minimizer $\mathbf{x}^\sharp$ exists.
2. Squaring the constraint does not change feasibility and simplifies the derivatives. KKT conditions yield the existence of $\lambda \in \mathbb{R}$ such that
$$(M^\top + M)\mathbf{x}^\sharp + 2\lambda \mathbf{x}^\sharp = \mathbf{0},$$
i.e. $\mathbf{x}^\sharp$ is an eigenvector of $M^\top + M$. The optimal value is the smallest eigenvalue of the symmetric part of M.
3. In particular, $M^\top + M$ always has an eigenvector; when M is symmetric, this is an eigenvector of M.

Solution 2.8

Back to Exercise

The objective is coercive and the feasible set is nonempty and closed, so an optimal solution exists. Constraints are affine, hence qualified, and we can use KKT conditions.

Let $\mu_1, \mu_2 \geq 0$ be the multipliers. Stationarity gives

$$2x_1 - 2\mu_1 - \mu_2 = 0, \qquad 2x_2 - \mu_1 = 0,$$

and complementary slackness gives

$$\mu_1(-2x_1 - x_2 + 10) = 0, \qquad \mu_2 x_1 = 0.$$

If $\mu_1 = 0$ then $x_1 = x_2 = 0$, which is infeasible. Hence $\mu_1 > 0$ and $2x_1 + x_2 = 10$.

If additionally $\mu_2 = 0$, then $2x_2 = \mu_1$ and $2x_1 = 2\mu_1$ give $x_1 = 2x_2$. Together with $2x_1 + x_2 = 10$ we get $(x_1, x_2) = (4, 2)$.

If $\mu_2 > 0$, then $x_1 = 0$ and $2x_1 + x_2 = 10$ gives $(x_1, x_2) = (0, 10)$.

Comparing objective values, $(4, 2)$ is optimal.

Solution 2.9

Back to Exercise

The objective is coercive and the feasible set is nonempty and closed (e.g. $(4, 3)$ is feasible), so an optimal solution exists.

First consider the case $x_2 = 0$. Feasibility requires $x_1^2 \geq 25$ and $x_1 \leq 4$, hence $x_1 \leq -5$. The best choice is then $x_1 = -5$, with objective value 125.

Now assume $x_2 \neq 0$. Then the gradients of the two constraints, $(1, 0)^\top$ and $-2(x_1, x_2)^\top$, are linearly independent, so KKT applies. Let $\mu_1, \mu_2 \geq 0$ be the multipliers. Stationarity gives

$$10x_1 + \mu_1 - 2\mu_2 x_1 = 0, \qquad 12x_2 - 2\mu_2 x_2 = 0,$$

and complementary slackness gives

$$\mu_1(x_1 - 4) = 0, \qquad \mu_2(25 - x_1^2 - x_2^2) = 0.$$

Since $x_2 \neq 0$, the second stationarity equation forces $\mu_2 = 6$, hence the circle constraint is active: $x_1^2 + x_2^2 = 25$. If $\mu_1 = 0$ then $x_1 = 0$ and $x_2 = \pm 5$. If $\mu_1 > 0$ then $x_1 = 4$ and $x_2 = \pm 3$.

Comparing all candidates $(-5, 0)$, $(0, \pm 5)$ and $(4, \pm 3)$ shows that $(-5, 0)$ is optimal.

Solution 2.10

Back to Exercise

At $(0, 0)$ the feasible set is given by $x_2^2 \leq x_1^3$, hence necessarily $x_1 \geq 0$. One finds

$$T_X(0, 0) = \mathbb{R}_+ \times \{0\},$$

since directions with positive x_1 component and zero x_2 component are admissible, while any direction with nonzero x_2 component cannot be approximated by feasible curves (a picture suffices).

The linearized cone uses the gradient of the active constraint at $(0, 0)$:

$$\nabla(-x_1^3 + x_2^2)\big|_{(0,0)} = (0, 0),$$

so the linearized constraint imposes nothing and

$$T_X^\ell(0,0) = \mathbb{R}^2.$$

Thus constraint qualification fails.

Moreover, feasibility implies $x_1 \geq 0$, so the minimum value is clearly 1, attained at $(0,0)$. If KKT applied, differentiating with respect to x_1 would give a multiplier $\mu \geq 0$ such that $2(0+1) - (3\mu)\cdot 0^2 = 0$, a contradiction.

Solution 2.11

Back to Exercise

1. The feasible set is nonempty and compact (closed and bounded by $x_i \geq 1$ and $\sum_i x_i = n^2$), and the objective is continuous on it, so an optimum exists.
2. The feasible point $x_i = n$ for all i yields a strictly positive value. If some coordinate were equal to 1, then $\ln(1) = 0$ and the product would be 0, so such a point cannot be optimal.
3. Since the inequality constraints are inactive at the optimum (by (2)), applying KKT conditions to the equality constraint gives a multiplier $\lambda \in \mathbb{R}$ such that, for each i,

$$\frac{\partial}{\partial x_i}\left(\prod_{j=1}^{n} \ln(x_j)\right) + \lambda = 0.$$

   This simplifies to the stated relation.
4. The right-hand side is nonzero, so $\lambda \neq 0$. Hence all $x_i^\sharp \ln(x_i^\sharp)$ are equal. Since $x \mapsto x \ln x$ is strictly increasing on $[1,\infty)$, all $x_i^\sharp$ are equal. Using $\sum_i x_i^\sharp = n^2$ yields $x_i^\sharp = n$.
5. The optimum is unique by symmetry and the strict concavity of $\sum_i \ln(\ln(x_i))$ on $(1,\infty)^n$ (equivalently, of the objective up to a monotone transformation).

Solution 2.12

Back to Exercise

We show that the optimal value is 0. Let $\mathbf{x}^\sharp$ be an optimal solution (it exists by compactness). If some $x_i^\sharp = 0$ then $\prod_i x_i^\sharp = 0$ and the objective is ≥ 0.

Assume now that all $x_i^\sharp > 0$. Then the inequality constraints are inactive, so KKT conditions apply with only the equality constraint $\|\mathbf{x}\|^2 = 1$. There exists $\lambda \in \mathbb{R}$ such that for every j,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^\sharp\right)^{n-1} - \prod_{i\neq j} x_i^\sharp + \lambda x_j^\sharp = 0.$$

Multiplying by $x_j^\sharp$ yields a quadratic equation

$$\lambda (x_j^\sharp)^2 + \left(\frac{s}{n}\right)^{n-1} x_j^\sharp - p = 0,$$

where $s := \sum_i x_i^\sharp$ and $p := \prod_i x_i^\sharp$. Summing over j gives

$$\frac{\lambda}{n} + \left(\frac{s}{n}\right)^{n} - p = 0.$$

If $\lambda \leq 0$, then the objective $(s/n)^n - p$ is nonnegative. If $\lambda > 0$, the quadratic equation has at most one nonnegative root, hence all $x_j^\sharp$ are equal and the objective value is 0.

Therefore, for $\|\mathbf{x}\| = 1$ and $\mathbf{x} \geq 0$, we have $(\frac{1}{n}\sum_i x_i)^n \geq \prod_i x_i$. By homogeneity, this gives AM–GM for all positive vectors.

Solution 2.13

Back to Exercise

One convenient form of the dual is

$$\begin{array}{ll} \text{Max} & \mathbf{b}\cdot\mathbf{y} \\ \text{s.t.} & A^\top \mathbf{y} = \mathbf{c} \\ & \mathbf{y} \leq \mathbf{0}. \end{array}$$

Note that it is of the same nature as the primal (affine objective and affine constraints).

Solution 2.14

Back to Exercise

1. Write each constraint in ≤ 0 form:

$$19 - x - y + z \le 0, \quad 13 + y - z \le 0, \quad 24 - x - y \le 0, \quad -x \le 0,\ -y \le 0,\ -z \le 0.$$

With multipliers $l, m, o, p, q, r \ge 0$, the Lagrangian is

$$\mathcal{L} = 5x + 11y + z + l(19 - x - y + z) + m(13 + y - z) + o(24 - x - y) - px - qy - rz.$$

The dual is

$$\begin{array}{ll} \text{Max} & 19l + 13m + 24o \\ \text{s.t.} & l + o \le 5 \\ & l - m + o \le 11 \\ & -l + m \le 1 \\ & l, m, o \ge 0. \end{array}$$

2. One checks that $(32, 0, 13)$ is primal-feasible and $(5, 6, 0)$ is dual-feasible, and that they give the same objective value. By weak duality, both are optimal.

Solution 2.15

Back to Exercise

Let $\boldsymbol{\mu} \ge \mathbf{0}$ be the multiplier for $A\mathbf{x} \le \mathbf{b}$. The dual is

$$\max_{\boldsymbol{\mu} \ge \mathbf{0}} \inf_{\mathbf{x} \in \mathbb{R}^n} \left(\mathbf{x}^\top M \mathbf{x} + \mathbf{c} \cdot \mathbf{x} + \boldsymbol{\mu}^\top (A\mathbf{x} - \mathbf{b}) \right).$$

The inner problem is an unconstrained strictly convex quadratic minimization. The minimizer satisfies

$$2M\mathbf{x} + \mathbf{c} + A^\top \boldsymbol{\mu} = \mathbf{0} \quad \Rightarrow \quad \mathbf{x}(\boldsymbol{\mu}) = -\frac{1}{2} M^{-1}(\mathbf{c} + A^\top \boldsymbol{\mu}).$$

Substituting back gives the dual objective

$$-\frac{1}{4}(\mathbf{c} + A^\top \boldsymbol{\mu})^\top M^{-1}(\mathbf{c} + A^\top \boldsymbol{\mu}) - \mathbf{b} \cdot \boldsymbol{\mu}.$$

Hence the dual problem is

$$\begin{array}{ll} \text{Max} & -\frac{1}{4}(\mathbf{c} + A^\top \boldsymbol{\mu})^\top M^{-1}(\mathbf{c} + A^\top \boldsymbol{\mu}) - \mathbf{b} \cdot \boldsymbol{\mu} \\ \text{s.t.} & \boldsymbol{\mu} \ge \mathbf{0}. \end{array}$$

Solution 2.16

Back to Exercise

1. The objective depends only on x_2, and feasibility requires $x_2 \ge 0$. Choosing $x_2 = 0$ (and any $x_1 \le 0$) gives value 1, which is optimal since $e^{x_2} \ge 1$ for $x_2 \ge 0$.
2. Use multipliers $\mu_1 \ge 0$ for $x_1 - x_2 \le 0$ and $\mu_2 \ge 0$ for $-x_2 \le 0$. The Lagrangian is $e^{x_2} + \mu_1(x_1 - x_2) - \mu_2 x_2$. Minimizing over x_1 forces $\mu_1 = 0$, otherwise the infimum is $-\infty$. With $\mu_1 = 0$, the dual function is

$$\inf_{x_2 \in \mathbb{R}} (e^{x_2} - \mu_2 x_2) = \mu_2(1 - \ln \mu_2),$$

with the convention $\mu_2 \ln \mu_2 = 0$ at $\mu_2 = 0$. Thus the dual is

$$\begin{array}{ll} \text{Max} & y(1 - \ln y) \\ \text{s.t.} & y \ge 0, \end{array}$$

whose optimum is 1 (attained at $y = 1$).

3. Yes: primal and dual optimal values coincide.

Solution 2.17

Back to Exercise

1. The optimal value is 1 (attained at $(0,0)$).
2. The Lagrangian is
$$(x_1+1)^2 + x_2^2 + \mu(-x_1^3 + x_2^2), \qquad \mu \geq 0.$$
If $\mu > 0$, the infimum over x_1 is $-\infty$ (the cubic term dominates as $x_1 \to +\infty$). Hence a finite dual value requires $\mu = 0$, in which case the infimum is 0 (attained at $x_1 = -1, x_2 = 0$). Therefore the dual is
$$\begin{array}{ll} \text{Max} & 0 \\ \text{s.t.} & \mu = 0, \end{array}$$
with optimal value 0.
3. No: there is a duality gap ($1 \neq 0$).

Solution 3.1

Back to Exercise

Let $S := \frac{1}{2}(M + M^\top)$ be the symmetric part of M. Since $\mathbf{x}^\top M \mathbf{x} = \mathbf{x}^\top S \mathbf{x}$, we can assume M is symmetric.

Then $\nabla^2(\mathbf{x}^\top M \mathbf{x}) = M + M^\top = 2M$, hence the function is convex if and only if M is positive semidefinite, i.e. $M + M^\top \succeq 0$ in general.

Solution 3.2

Back to Exercise

1. We take
$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{pmatrix}, \qquad \theta = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 15 \end{pmatrix}.$$
Then $F(a,b) = \|A\theta - \mathbf{y}\|_2^2$.
2. The Hessian is $2A^\top A \succeq 0$, so the objective is convex. The columns of A are linearly independent; hence $A^\top A$ is positive definite and the minimizer is unique.
3. The normal equations are
$$A^\top A \theta = A^\top \mathbf{y}, \qquad \begin{pmatrix} 30 & 10 \\ 10 & 5 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 94 \\ 31 \end{pmatrix}.$$
4. Solving gives $a = 16/5$ and $b = -1/5$, so
$$\widehat{y} = \frac{16}{5}x - \frac{1}{5}.$$
5. The residuals $\widehat{y}_i - y_i$ are
$$-\frac{6}{5}, \quad 0, \quad \frac{6}{5}, \quad \frac{12}{5}, \quad -\frac{12}{5},$$
and their squared sum is $72/5 = 14.4$. The final observation pulls the slope away from the line $y = 2x + 1$ through the first four observations. The same data are revisited with ℓ_1 and ℓ_∞ criteria in Exercise 4.5.

Solution 3.3

Back to Exercise

Assume first that f is convex. For any $t, s \in [0,1]$ and $\lambda \in [0,1]$, since C is convex we have $\mathbf{x} + (\lambda t + (1-\lambda)s)\mathbf{d} \in C$ and

$$\begin{aligned} g(\lambda t + (1-\lambda)s) &= f\big(\mathbf{x} + (\lambda t + (1-\lambda)s)\mathbf{d}\big) \\ &= f\big(\lambda(\mathbf{x} + t\mathbf{d}) + (1-\lambda)(\mathbf{x} + s\mathbf{d})\big) \\ &\leq \lambda f(\mathbf{x} + t\mathbf{d}) + (1-\lambda) f(\mathbf{x} + s\mathbf{d}) = \lambda g(t) + (1-\lambda) g(s), \end{aligned}$$

so g is convex.

Conversely, assume that for every $\mathbf{x}, \mathbf{d}$ the corresponding g is convex. Take any $\mathbf{y}, \mathbf{z} \in C$ and $\lambda \in [0,1]$. Let $\mathbf{x} := \mathbf{z}$ and $\mathbf{d} := \mathbf{y} - \mathbf{z}$. Then $\mathbf{x} + \mathbf{d} = \mathbf{y} \in C$ and

$$f(\lambda \mathbf{y} + (1-\lambda)\mathbf{z}) = f(\mathbf{z} + \lambda(\mathbf{y} - \mathbf{z})) = g(\lambda).$$

By convexity of g,

$$g(\lambda) \le \lambda g(1) + (1-\lambda) g(0) = \lambda f(\mathbf{y}) + (1-\lambda) f(\mathbf{z}),$$

which is exactly the convexity of f.

Solution 3.4

Back to Exercise

Let $x_1, \dots, x_n > 0$. Since ln is concave, Jensen's inequality gives, with $\lambda_i = 1/n$,

$$\ln\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \ge \frac{1}{n}\sum_{i=1}^{n} \ln(x_i).$$

Exponentiating both sides yields

$$\frac{1}{n}\sum_{i=1}^{n} x_i \ge \left(\prod_{i=1}^{n} x_i\right)^{1/n}.$$

Solution 3.5

Back to Exercise

Let $Z := \sum_{k=1}^{n} e^{x_k}$ and define $p_i := e^{x_i}/Z$. Then $p_i \ge 0$ and $\sum_i p_i = 1$.
We have $\nabla f(\mathbf{x}) = (p_1, \dots, p_n)^\top$. Differentiating again gives the Hessian

$$\nabla^2 f(\mathbf{x}) = \operatorname{diag}(p) - pp^\top.$$

For any $\mathbf{u} \in \mathbb{R}^n$,

$$\mathbf{u}^\top(\operatorname{diag}(p) - pp^\top)\mathbf{u} = \sum_{i=1}^{n} p_i u_i^2 - \left(\sum_{i=1}^{n} p_i u_i\right)^2 = \operatorname{Var}(U) \ge 0,$$

where U is a random variable taking value u_i with probability p_i. Therefore $\nabla^2 f(\mathbf{x}) \succeq 0$ and f is convex.

Solution 3.6

Back to Exercise

The objective is convex and all constraints are affine, so the KKT conditions characterize the optimum. Let $\mu_1, \mu_2 \ge 0$ be the multipliers. Stationarity gives

$$2x_1 - 6 + 2\mu_1 = 0, \qquad 2x_2 - 2 + \mu_1 + \mu_2 = 0,$$

and complementary slackness gives

$$\mu_1(2x_1 + x_2 - 2) = 0, \qquad \mu_2(x_2 - 1) = 0.$$

If $x_2 = 1$ then μ_2 can be nonzero, but stationarity would force $\mu_1 + \mu_2 = 0$ hence $\mu_1 = \mu_2 = 0$ and $x_1 = 3$, which violates $2x_1 + x_2 \le 2$. Therefore $x_2 \ne 1$ and $\mu_2 = 0$.

If $\mu_1 = 0$ then $x_1 = 3$ and $x_2 = 1$, impossible. Hence $\mu_1 > 0$ and the first constraint is active: $2x_1 + x_2 = 2$. Solving

$$2x_1 - 6 + 2\mu_1 = 0, \qquad 2x_2 - 2 + \mu_1 = 0, \qquad 2x_1 + x_2 = 2$$

gives $\mu_1 = 2$, $x_2 = 0$, $x_1 = 1$.

Thus the unique optimal solution is $(x_1, x_2) = (1, 0)$.

Solution 3.7

Back to Exercise

The objective and inequality constraints are convex, and Slater's condition holds (for example at $(0,1)$), so the KKT conditions characterize the optimum.

Let $\mu_1, \mu_2 \ge 0$ be the multipliers for $x_1^2 - x_2 \le 0$ and $x_2 - 4 \le 0$. Stationarity gives

$$-2 + 2\mu_1 x_1 = 0 \quad \Rightarrow \quad x_1\mu_1 = 1, \qquad 1 - \mu_1 + \mu_2 = 0,$$

and complementary slackness gives

$$\mu_1(x_1^2 - x_2) = 0, \qquad \mu_2(x_2 - 4) = 0.$$

From $x_1\mu_1 = 1$ we get $\mu_1 \neq 0$ and $x_1 > 0$, hence $x_2 = x_1^2$ (first constraint active). If $\mu_2 \neq 0$ then $x_2 = 4$, $x_1 = 2$, and $1 - \mu_1 + \mu_2 = 0$ would imply $\mu_2 = \mu_1 - 1 = 1/2 - 1 < 0$, impossible. Therefore $\mu_2 = 0$, so $\mu_1 = 1$ and $x_1 = 1$, $x_2 = 1$.

Thus the unique optimal solution is $(x_1, x_2) = (1, 1)$.

Solution 3.8

Back to Exercise

If $\mathbf{c} = \mathbf{0}$, then $\mathbf{x} = \mathbf{a}$ is feasible and optimal. Assume $\mathbf{c} \neq \mathbf{0}$.

The constraint is active at the optimum, and KKT conditions give the existence of a multiplier $\mu > 0$ such that

$$\mathbf{c} + 2\mu M(\mathbf{x} - \mathbf{a}) = 0, \qquad (\mathbf{x} - \mathbf{a})^\top M(\mathbf{x} - \mathbf{a}) = 1.$$

From the first equation,

$$\mathbf{x} = \mathbf{a} - \frac{1}{2\mu} M^{-1}\mathbf{c}.$$

Plugging into the active constraint yields

$$\frac{1}{4\mu^2}\mathbf{c}^\top M^{-1}\mathbf{c} = 1 \quad \Rightarrow \quad \mu = \frac{1}{2}\sqrt{\mathbf{c}^\top M^{-1}\mathbf{c}}.$$

Therefore the (unique) optimal solution is

$$\mathbf{x}^\sharp = \mathbf{a} - \frac{M^{-1}\mathbf{c}}{\sqrt{\mathbf{c}^\top M^{-1}\mathbf{c}}}.$$

Solution 3.9

Back to Exercise

1. The delivery time for resource i from region j is $c_j x_{ij}^2$. To have all resources available, we must wait for the slowest delivery, hence the objective $\max_{i,j} c_j x_{ij}^2$. The constraints enforce that we meet each demand a_i and that quantities are nonnegative.

2. Introduce a variable t that upper bounds the maximum:

$$\begin{array}{lll} \text{Min} & t & \\ \text{s.t.} & c_j x_{ij}^2 \le t & \forall i \in [m],\ \forall j \in [n] \\ & \sum_{j=1}^n x_{ij} = a_i & \forall i \in [m] \\ & x_{ij} \ge 0 & \forall i \in [m],\ \forall j \in [n]\,. \end{array}$$

3. The Lagrangian reads

$$\mathcal{L} = \Big(1 - \sum_{i,j} \mu_{ij}\Big)t + \sum_{i,j} c_j \mu_{ij} x_{ij}^2 - \sum_{i,j} (\omega_{ij} + \lambda_i) x_{ij} + \sum_i \lambda_i a_i,$$

with $\mu_{ij} \ge 0$, $\omega_{ij} \ge 0$ and $\lambda_i \in \mathbb{R}$. Minimizing over t forces $\sum_{i,j} \mu_{ij} = 1$.
Minimizing over each x_{ij} yields the value $h(\omega_{ij} + \lambda_i, c_j\mu_{ij})$, where

$$h(\nu, \nu') = \begin{cases} -\dfrac{\nu^2}{4\nu'} & \text{if } \nu' > 0, \\ 0 & \text{if } \nu' = 0 \text{ and } \nu = 0, \\ -\infty & \text{if } \nu' = 0 \text{ and } \nu \neq 0. \end{cases}$$

Hence the dual is

$$\begin{array}{lll} \text{Max} & \displaystyle\sum_{i=1}^m a_i\lambda_i + \sum_{i=1}^m \sum_{j=1}^n h(\omega_{ij} + \lambda_i, c_j\mu_{ij}) & \\ \text{s.t.} & \sum_{i=1}^m \sum_{j=1}^n \mu_{ij} = 1 & \\ & \omega_{ij} \ge 0,\ \mu_{ij} \ge 0 & \forall i \in [m],\ \forall j \in [n]. \end{array}$$

4. With $\mu_{ij} = 1/(mn)$ and $\omega_{ij} = 0$, we have $h(\lambda_i, c_j/(mn)) = -(mn)\lambda_i^2/(4c_j)$. The dual objective becomes

$$\sum_{i=1}^{m} a_i\lambda_i - \sum_{i=1}^{m}\sum_{j=1}^{n} mn\frac{\lambda_i^2}{4c_j}.$$

Maximizing independently over each λ_i gives the value $\sum_i a_i^2/(mn\sum_j 1/c_j)$, which is therefore a valid lower bound on the primal optimal value.

Solution 3.10

Back to Exercise

1. With $\overline{\mathbf{p}} = (4, 6, 10)$, the problem is

$$\begin{aligned} \text{Min} \quad & \sum_{i=1}^{3} C_i(p_i) \\ \text{s.t.} \quad & D - \sum_{i=1}^{3} p_i = 0, \\ & 0 \le p_i \le \overline{p}_i \qquad i = 1, 2, 3. \end{aligned}$$

2. The objective is strictly convex, since each cost has a positive quadratic coefficient, and the feasible set is convex. It is nonempty exactly when $0 \le D \le 20$ and is then compact. Thus an optimum exists, and strict convexity makes it unique.
3. Write the bounds as $-p_i \le 0$ and $p_i - \overline{p}_i \le 0$, with multipliers $\alpha_i \ge 0$ and $\beta_i \ge 0$, respectively. With $\lambda \in \mathbb{R}$ for the balance equation, the Lagrangian is

$$\mathcal{L}(\mathbf{p}, \lambda, \alpha, \beta) = \sum_i C_i(p_i) + \lambda\Big(D - \sum_i p_i\Big) - \sum_i \alpha_i p_i + \sum_i \beta_i(p_i - \overline{p}_i).$$

The KKT conditions are primal feasibility, $\alpha_i, \beta_i \ge 0$, and

$$C_i'(p_i) - \lambda - \alpha_i + \beta_i = 0, \qquad \alpha_i p_i = 0, \qquad \beta_i(p_i - \overline{p}_i) = 0 \quad (i = 1, 2, 3).$$

4. For $D = 6$, generators 1 and 2 are strictly inside their bounds, so

$$C_1'(p_1) = p_1 + 2 = \lambda, \qquad C_2'(p_2) = \frac{1}{2}p_2 + 4 = \lambda.$$

Together with $p_1 + p_2 = 6$, this gives

$$\mathbf{p}^\sharp = \left(\frac{10}{3}, \frac{8}{3}, 0\right).$$

5. For $D = 10$, generator 1 is at capacity, while generators 2 and 3 are strictly inside their bounds. Their common marginal cost gives

$$\frac{1}{2}p_2 + 4 = \frac{1}{3}p_3 + 6 = \lambda, \qquad p_2 + p_3 = 6.$$

Hence

$$\mathbf{p}^\sharp = \left(4, \frac{24}{5}, \frac{6}{5}\right).$$

6. For $D = 6$, generators 1 and 2 are strictly inside their bounds, generator 3 is off, and $\lambda = 16/3$. The bound multipliers are

$$(\alpha_1, \alpha_2, \alpha_3) = \left(0, 0, \frac{2}{3}\right), \qquad (\beta_1, \beta_2, \beta_3) = (0, 0, 0),$$

since $C_3'(0) - \lambda = 6 - 16/3 = 2/3$.

For $D = 10$, generator 1 is at capacity, generators 2 and 3 are strictly inside their bounds, and $\lambda = 32/5$. Here

$$(\alpha_1, \alpha_2, \alpha_3) = (0, 0, 0), \qquad (\beta_1, \beta_2, \beta_3) = \left(\frac{2}{5}, 0, 0\right),$$

because $C_1'(4) - \lambda + \beta_1 = 6 - 32/5 + 2/5 = 0$. In both cases the displayed multipliers satisfy dual feasibility, stationarity, and complementary slackness, so the KKT conditions certify the solutions.

7. With the chosen sign convention, the derivative of the Lagrangian with respect to the demand parameter is $\partial\mathcal{L}/\partial D = \lambda$. The marginal-value result therefore gives $v'(D) = \lambda$ whenever the derivative exists and the active set is unchanged.
8. Economically, λ is the marginal cost, or marginal electricity price, of supplying one additional unit of demand. The change between the two cases illustrates how this price and the marginal generators change when a capacity bound becomes active.

Solution 3.11

Back to Exercise

1. Substituting the conductances and boundary voltages gives
$$E(v_1, v_2) = \frac{1}{2}\left((v_1 - 10)^2 + \frac{1}{2}(v_2 - 10)^2 + (v_1 - v_2)^2 + \frac{1}{2}v_1^2 + v_2^2\right).$$
2. Its Hessian is
$$\nabla^2 E = \begin{pmatrix} 5/2 & -1 \\ -1 & 5/2 \end{pmatrix}.$$
Its first leading principal minor is positive and its determinant is $21/4 > 0$, so it is positive definite and E is strictly convex.
3. The positive-definite quadratic term also makes E coercive. Hence it has a unique minimizer (see also Section 3.2).
4. The stationarity equations are
$$\frac{5}{2}v_1 - v_2 = 10, \qquad -v_1 + \frac{5}{2}v_2 = 5.$$
5. More generally, differentiating the energy terms incident to an internal node i gives exactly
$$\sum_{j:\{i,j\}\text{ is an edge}} \sigma_{ij}(v_i - v_j) = 0.$$
6. Since $\sigma_{ij}(v_i - v_j) = I_{ij}$, this equation says that the algebraic sum of the currents leaving each internal node is zero. This is Kirchhoff's current law.
7. Solving the two stationarity equations gives
$$v_1 = \frac{40}{7}, \qquad v_2 = \frac{30}{7}.$$
In the requested directions, the currents are
$$I_{01} = \frac{30}{7}, \qquad I_{02} = \frac{20}{7}, \qquad I_{12} = \frac{10}{7}, \qquad I_{13} = \frac{20}{7}, \qquad I_{23} = \frac{30}{7}.$$
At node 1, $I_{01} = I_{12} + I_{13} = 30/7$. At node 2, $I_{02} + I_{12} = I_{23} = 30/7$. Thus current is conserved at both internal nodes.

Solution 3.12

Back to Exercise

This is a convex problem on the domain $\{x_k > 0\}$, and Slater's condition holds (*e.g.,*take $x_k = N/(2n)$). Let $\mu \geq 0$ be the multiplier for $\sum_k x_k \leq N$. KKT conditions give, for an optimal solution $x^\sharp$,
$$a - \frac{b_k}{(x_k^\sharp)^2} + \mu = 0 \quad \forall k, \qquad \mu\Big(\sum_{k=1}^n x_k^\sharp - N\Big) = 0, \qquad \sum_{k=1}^n x_k^\sharp \leq N.$$
Thus $x_k^\sharp = \sqrt{\frac{b_k}{a+\mu}}$ for all k.

If $\frac{1}{\sqrt{a}}\sum_{k=1}^n \sqrt{b_k} \leq N$, then $\mu = 0$ and
$$x_k^\sharp = \sqrt{\frac{b_k}{a}}.$$
Otherwise the constraint is active and μ is chosen so that $\sum_k x_k^\sharp = N$, which yields
$$\mu = \left(\frac{\sum_{k=1}^n \sqrt{b_k}}{N}\right)^2 - a, \qquad x_k^\sharp = \frac{N\sqrt{b_k}}{\sum_{j=1}^n \sqrt{b_j}}.$$

The solution is unique since the objective is strictly convex on $\{x_k > 0\}$.

Solution 3.13

Back to Exercise

1. Let $\mathbf{x}$ be feasible for $\mathbf{u}$ and $\mathbf{x}'$ feasible for $\mathbf{u}'$. By convexity of each h_j,

$$h_j(t\mathbf{x} + (1-t)\mathbf{x}') \le t h_j(\mathbf{x}) + (1-t) h_j(\mathbf{x}') \le t u_j + (1-t) u_j',$$

so $t\mathbf{x} + (1-t)\mathbf{x}'$ is feasible for $t\mathbf{u} + (1-t)\mathbf{u}'$. By convexity of f,

$$v(t\mathbf{u} + (1-t)\mathbf{u}') \le f(t\mathbf{x} + (1-t)\mathbf{x}') \le t f(\mathbf{x}) + (1-t) f(\mathbf{x}').$$

Minimizing over feasible $\mathbf{x}$ and $\mathbf{x}'$ gives $v(t\mathbf{u} + (1-t)\mathbf{u}') \le t v(\mathbf{u}) + (1-t) v(\mathbf{u}')$, hence v is convex.

2. (a) Consider the Lagrangian for $\mathbf{u} = \mathbf{0}$:

$$\mathcal{L}_0(\mathbf{x}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{j=1}^{q} \mu_j h_j(\mathbf{x}).$$

Since $(\mathbf{x}(\mathbf{0}), \boldsymbol{\mu}^0)$ is a saddle point, we have $\mathcal{L}_0(\mathbf{x}, \boldsymbol{\mu}^0) \ge \mathcal{L}_0(\mathbf{x}(\mathbf{0}), \boldsymbol{\mu}^0) = v(\mathbf{0})$ for all $\mathbf{x}$. In particular, for an optimal solution $\mathbf{x}(\mathbf{u})$ at parameter $\mathbf{u}$,

$$v(\mathbf{0}) \le \mathcal{L}_0(\mathbf{x}(\mathbf{u}), \boldsymbol{\mu}^0) = v(\mathbf{u}) + \sum_{j=1}^{q} \mu_j^0 h_j(\mathbf{x}(\mathbf{u})) \le v(\mathbf{u}) + \boldsymbol{\mu}^0 \cdot \mathbf{u},$$

hence $v(\mathbf{u}) \ge v(\mathbf{0}) - \boldsymbol{\mu}^0 \cdot \mathbf{u}$.

(b) If v is differentiable at $\mathbf{0}$ then $v(\mathbf{u}) = v(\mathbf{0}) + \nabla v(\mathbf{0}) \cdot \mathbf{u} + o(\|\mathbf{u}\|)$. Combining with the inequality from (a) gives

$$\nabla v(\mathbf{0}) \cdot \mathbf{u} + o(\|\mathbf{u}\|) \ge -\boldsymbol{\mu}^0 \cdot \mathbf{u}.$$

Taking $\mathbf{u} = t\,\mathbf{e}_i$ and letting $t \to 0^+$ and $t \to 0^-$ yields $\nabla v(\mathbf{0}) = -\boldsymbol{\mu}^0$.

Solution 3.14

Back to Exercise

We have $\nabla f(\mathbf{x}) = (x_1, \gamma x_2)^\top$. With step size $\ell > 0$,

$$x_1^{k+1} = (1-\ell) x_1^k, \qquad x_2^{k+1} = (1-\ell\gamma) x_2^k.$$

Therefore

$$\mathbf{x}^k = \begin{pmatrix} (1-\ell)^k \gamma \\ (1-\ell\gamma)^k \end{pmatrix}.$$

We have $\mathbf{x}^k \to \mathbf{0}$ (the unique minimizer) if and only if $|1-\ell| < 1$ and $|1-\ell\gamma| < 1$, i.e.

$$0 < \ell < 2\min(1, 1/\gamma).$$

Solution 4.1

Back to Exercise

From inequality form to standard form: add slack variables and replace a free variable $\mathbf{x}$ by $\mathbf{u} - \mathbf{v}$ with $\mathbf{u}, \mathbf{v} \ge \mathbf{0}$.

From standard form to inequality form: replace an equality constraint by two inequalities $\le$ and $\ge$, and view $\mathbf{x} \ge \mathbf{0}$ as a family of inequality constraints.

Solution 4.2

Back to Exercise

If there is an optimal solution, then there exists a basic optimal solution. Let B be the corresponding basis. The basic solution is $\mathbf{x}_B = A_B^{-1}\mathbf{b}$, which has rational entries since A_B and $\mathbf{b}$ are rational.

Solution 4.3

Back to Exercise

Let x, y, z be the weekly production (in liters) of XXX, YYY, and ZZZ. The linear program is

$$\begin{array}{ll} \text{Max} & 1.55x + 3.05y + 4.80z \\ \text{s.t.} & x \geq 1020, \quad y \geq 1450, \quad z \geq 750 \\ & \frac{1}{20}(2x + 3y + 6z) \leq 708 \\ & \frac{1}{20}(4x + 5.5y + 7.5z) \leq 932 \\ & \frac{1}{20}(2x + y + 3.5z) \leq 342. \end{array}$$

Solution 4.4

Back to Exercise

Let x_1, x_2, x_3 be the amounts of components 1,2,3 used in gasoline A, and y_1, y_2, y_3 the amounts used in gasoline B. The profit-maximization LP is

$$\begin{array}{ll} \text{Max} & 5.5(x_1 + x_2 + x_3) + 4.5(y_1 + y_2 + y_3) - 3(x_1 + y_1) - 6(x_2 + y_2) - 4(x_3 + y_3) \\ \text{s.t.} & x_1 + y_1 \leq 3000, \quad x_2 + y_2 \leq 2000, \quad x_3 + y_3 \leq 4000 \\ & x_1 \leq 0.3(x_1 + x_2 + x_3), \quad x_2 \geq 0.4(x_1 + x_2 + x_3), \quad x_3 \leq 0.5(x_1 + x_2 + x_3) \\ & y_1 \leq 0.5(y_1 + y_2 + y_3), \quad y_2 \geq 0.1(y_1 + y_2 + y_3) \\ & x_1, x_2, x_3, y_1, y_2, y_3 \geq 0. \end{array}$$

Solution 4.5

Back to Exercise

1. The decision variables are a and b. Introduce auxiliary variables to model the supremum or absolute values. For the ℓ_∞ fit:

$$\begin{array}{lll} \text{Min} & t & \\ \text{s.t.} & t \geq ax_i + b - y_i & \forall i \\ & t \geq y_i - ax_i - b & \forall i \\ & a, b, t \in \mathbb{R}. & \end{array}$$

   For the ℓ_1 fit:

$$\begin{array}{lll} \text{Min} & \sum_{i=1}^{n} t_i & \\ \text{s.t.} & t_i \geq ax_i + b - y_i & \forall i \\ & t_i \geq y_i - ax_i - b & \forall i \\ & a, b, t_1, \ldots, t_n \in \mathbb{R}. & \end{array}$$

2. Let $r_i = ax_i + b - y_i$, with the data ordered as in the question. For the ℓ_1 problem, take

$$\mathbf{q} = \left(-\frac{3}{5}, \frac{1}{10}, \frac{3}{5}, \frac{9}{10}, -1\right).$$

   Its components belong to $[-1, 1]$ and satisfy $\sum_i q_i = \sum_i q_i x_i = 0$. Therefore every affine line satisfies

$$\sum_i |r_i| \geq \sum_i q_i r_i = -\sum_i q_i y_i = 6.$$

   The line $y = 2x + 1$ has residuals $(0, 0, 0, 0, -6)$, so it attains this lower bound. Moreover, $|q_i| < 1$ for the first four observations, so equality in the displayed bound requires their four residuals to vanish. Hence this fit is unique and its total absolute residual is 6.

   For the ℓ_∞ problem, use

$$\mathbf{q} = \left(-\frac{1}{8}, 0, 0, \frac{1}{2}, -\frac{3}{8}\right).$$

   Here $\sum_i q_i = \sum_i q_i x_i = 0$ and $\sum_i |q_i| = 1$. Thus, for $t = \max_i |r_i|$,

$$t \geq \sum_i q_i r_i = -\sum_i q_i y_i = \frac{9}{4}.$$

   The line

$$y = \frac{7}{2}x - \frac{5}{4}$$

   has residuals

$$-\frac{9}{4}, \quad -\frac{3}{4}, \quad \frac{3}{4}, \quad \frac{9}{4}, \quad -\frac{9}{4},$$

so it attains the bound. Equality in the preceding bound forces the three residuals with nonzero q_i, at the observations with $x = 0, 3, 4$, to attain $-t, t, -t$, respectively; these equations determine a, b, and t uniquely.

By contrast, Exercise 3.2 gives the least-squares line $y = \frac{16}{5}x - \frac{1}{5}$. In this example, the ℓ_2 criterion spreads the outlier's influence across the fit, the ℓ_1 criterion preserves the line through the first four observations, and the ℓ_∞ criterion balances the largest residuals.

Solution 4.6

Back to Exercise

1. Among all subsets of $[n]$ containing k indices, the largest sum is obtained by selecting the k largest components (with any choice among tied components). Hence

$$\sum_{i=1}^{k} x^{[i]} = \max_{\substack{S \subseteq [n] \\ |S|=k}} \sum_{i \in S} x_i.$$

Introducing an epigraph variable z gives the LP

$$\begin{array}{ll} \text{Min} & z \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & z \geq \displaystyle\sum_{i \in S} x_i \qquad \forall S \subseteq [n],\ |S| = k \\ & \mathbf{x} \geq 0. \end{array}$$

For fixed $\mathbf{x}$, minimizing z makes it equal to the maximum subset sum and therefore to $\sum_{i=1}^{k} x^{[i]}$. This formulation has $\binom{n}{k}$ subset inequalities.

2. For the compact formulation, use the identity

$$\sum_{i=1}^{k} x^{[i]} = \min_{t \in \mathbb{R}} \left(kt + \sum_{j=1}^{n} \max(0, x_j - t) \right).$$

Indeed, for every $t \in \mathbb{R}$,

$$kt + \sum_{j=1}^{n} (x_j - t)_+ \geq kt + \sum_{i=1}^{k} (x^{[i]} - t) = \sum_{i=1}^{k} x^{[i]}.$$

Equality holds for $t \in [x^{[k+1]}, x^{[k]}]$ when $k < n$: the first k ordered components are at least t and all the others are at most t. When $k = n$, equality holds for every $t \leq x^{[n]}$. This also covers ties between components.

Introduce variables $t \in \mathbb{R}$ and $u_i \geq 0$ with $u_i \geq x_i - t$. The compact LP is

$$\begin{array}{ll} \text{Min} & kt + \sum_{i=1}^{n} u_i \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & u_i \geq x_i - t \qquad \forall i \in [n] \\ & \mathbf{u}, \mathbf{x} \geq 0, \quad t \in \mathbb{R}. \end{array}$$

For fixed $(\mathbf{x}, t)$, minimizing over $\mathbf{u}$ gives $u_i = (x_i - t)_+$, so the identity proves equivalence with the original problem. Ignoring sign constraints, this formulation has m equality constraints from $A\mathbf{x} = \mathbf{b}$ and n inequalities, hence $m + n$ constraints.

Solution 4.7

Back to Exercise

The distance from a point $\mathbf{x}$ to the hyperplane $H = \{\mathbf{y} : \mathbf{a} \cdot \mathbf{y} + \beta = 0\}$ is

$$\frac{|\mathbf{a} \cdot \mathbf{x} + \beta|}{\|\mathbf{a}\|}.$$

Consider a polyhedron defined by inequalities $\mathbf{a}^i \cdot \mathbf{y} + \beta_i \geq 0$ for $i = 1, \ldots, m$. A ball of radius r centered at $\mathbf{x}$ is inside the polyhedron if the center is at distance at least r from every supporting hyperplane, i.e.

$$\mathbf{a}^i \cdot \mathbf{x} + \beta_i \geq \|\mathbf{a}^i\|\, r \qquad \forall i.$$

Thus the largest inscribed ball is obtained by the LP:

$$\begin{array}{ll} \text{Max} & r \\ \text{s.t.} & \mathbf{a}^i \cdot \mathbf{x} + \beta_i \geq \|\mathbf{a}^i\| \, r \qquad \forall i \\ & \mathbf{x} \in \mathbb{R}^n, \quad r \in \mathbb{R}_+. \end{array}$$

Solution 4.8

Back to Exercise

1. (a) Introduce slack variables $y_i \geq 0$ with $x_i + y_i = 1$:

$$\begin{array}{lll} \text{Min} & \mathbf{c} \cdot \mathbf{x} & \\ \text{s.t.} & \sum_{i=1}^n a_i x_i = b & \\ & x_i + y_i = 1 & i = 1, \ldots, n \\ & x_i, y_i \geq 0. & \end{array}$$

   The feasible set is bounded, hence the objective is bounded below, and an optimal basic feasible solution exists.

   (b) There are $n+1$ independent equality constraints, so any basis has cardinality $n+1$. For each i, at least one of x_i and y_i must be basic, hence at most one index $\bar{\imath}$ can have both $x_{\bar{\imath}}$ and $y_{\bar{\imath}}$ basic, i.e. at most one x_i can be fractional.

   (c) Suppose $c_k/a_k < c_j/a_j$ but $x_k^\sharp < x_j^\sharp$. For small $\varepsilon > 0$, define a new feasible point by

$$\bar{x}_i = x_i^\sharp \ (i \notin \{j,k\}), \quad \bar{x}_k = x_k^\sharp + \frac{\varepsilon}{a_k}, \quad \bar{x}_j = x_j^\sharp - \frac{\varepsilon}{a_j}.$$

   Feasibility holds and the objective decreases by $\left(\frac{c_k}{a_k} - \frac{c_j}{a_j}\right)\varepsilon < 0$, a contradiction. Hence $x_k^\sharp \geq x_j^\sharp$.

   (d) Sort items so that $c_1/a_1 \leq \cdots \leq c_n/a_n$. Then in an optimal solution one fills items with smallest c_i/a_i first: set $x_i = 1$ as long as possible until the equality constraint would be violated, then set one fractional variable

$$x_{\ell+1} = \frac{b - \sum_{i=1}^{\ell} a_i}{a_{\ell+1}},$$

   and set $x_i = 0$ for $i > \ell + 1$.

2. (a) A dual formulation is

$$\begin{array}{ll} \text{Max} & b\lambda - \sum_{i=1}^n \mu_i \\ \text{s.t.} & a_i \lambda - c_i \leq \mu_i \\ & \mu_i \geq 0 \qquad i = 1, \ldots, n. \end{array}$$

   (b) Eliminating μ_i gives the concave function

$$d(\lambda) = b\lambda - \sum_{i=1}^{n} \max(0, a_i\lambda - c_i),$$

   and the dual amounts to maximizing $d(\lambda)$ over $\lambda \in \mathbb{R}$.

   (c) (*) The maximum is attained at a breakpoint $\lambda^\sharp = c_{\ell+1}/a_{\ell+1}$ where ℓ is the largest index such that $\sum_{i=1}^{\ell} a_i \leq b$.

Solution 4.9

Back to Exercise

Let $\mathbf{x}$ be any feasible solution. We can write $\mathbf{x}_B = A_B^{-1}\mathbf{b} - A_B^{-1}A_N\mathbf{x}_N$, hence

$$\begin{aligned} \mathbf{c} \cdot \mathbf{x} &= \mathbf{c}_B^\top A_B^{-1}\mathbf{b} + (\mathbf{c}_N^\top - \mathbf{c}_B^\top A_B^{-1}A_N)\mathbf{x}_N \\ &= \mathbf{c}_B^\top A_B^{-1}\mathbf{b} + \mathbf{r} \cdot \mathbf{x}_N \\ &\geq \mathbf{c}_B^\top A_B^{-1}\mathbf{b} + (\min_j r_j) \sum_{i \in N} x_i. \end{aligned}$$

Since $\sum_i x_i = 1$ and $\mathbf{x}_B \geq 0$, we have $\sum_{i \in N} x_i \leq 1$. Moreover $\mathbf{r} \not\geq \mathbf{0}$ implies $\min_j r_j \leq 0$, hence $(\min_j r_j)\sum_{i \in N} x_i \geq \min_j r_j$. Therefore

$$\mathbf{c} \cdot \mathbf{x} \geq \mathbf{c}_B^\top A_B^{-1}\mathbf{b} + \min_j r_j,$$

and in particular $v \geq \mathbf{c}_B^\top A_B^{-1}\mathbf{b} + \min_j r_j$.

For the upper bound, take the basic feasible solution $\mathbf{y}$ with $\mathbf{y}_B = A_B^{-1}\mathbf{b}$ and $\mathbf{y}_N = 0$, which is feasible and has objective value $\mathbf{c}_B^\top A_B^{-1}\mathbf{b}$. Hence $v \leq \mathbf{c}_B^\top A_B^{-1}\mathbf{b}$.

Solution 4.10

Back to Exercise

We prove strong duality for

$$\begin{array}{ll} \text{Min} & \mathbf{c}\cdot\mathbf{x} \\ \text{s.t.} & A\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}. \end{array}$$

If the primal is unbounded below, weak duality already implies the equality of optimal values (both are $-\infty$ or the dual is infeasible).

Assume the objective is bounded below. Then the simplex method terminates at an optimal basis $B^\sharp$, for which the reduced costs satisfy $\mathbf{r} \geq \mathbf{0}$. This means $\mathbf{c}_{N^\sharp}^\top - \mathbf{c}_{B^\sharp}^\top A_{B^\sharp}^{-1} A_{N^\sharp} \geq \mathbf{0}^\top$, and the optimal primal value is $\mathbf{c}_{B^\sharp}^\top A_{B^\sharp}^{-1}\mathbf{b}$.

Define $(\mathbf{y}^\sharp)^\top := \mathbf{c}_{B^\sharp}^\top A_{B^\sharp}^{-1}$. Then

$$(\mathbf{y}^\sharp)^\top A = (\mathbf{y}^\sharp)^\top (A_{B^\sharp}, A_{N^\sharp}) = (\mathbf{c}_{B^\sharp}^\top,\ \mathbf{c}_{B^\sharp}^\top A_{B^\sharp}^{-1} A_{N^\sharp}) \leq (\mathbf{c}_{B^\sharp}^\top,\ \mathbf{c}_{N^\sharp}^\top) = \mathbf{c}^\top,$$

so $\mathbf{y}^\sharp$ is dual-feasible. Its dual objective value is $(\mathbf{y}^\sharp)^\top\mathbf{b} = \mathbf{c}_{B^\sharp}^\top A_{B^\sharp}^{-1}\mathbf{b}$, which matches the primal optimal value. By weak duality, this proves strong duality.

Solution 4.11

Back to Exercise

The dual (D^I) is obtained from (D) by removing some constraints, hence its feasible set is larger and its optimal value is an upper bound on the optimal value of (D).

If $\tilde{\mathbf{y}}$ is feasible for (D), then it is feasible for (D) and optimal for (D^I), so it must be optimal for (D) as well (otherwise (D^I) could not have a larger value). Therefore the optimal values of (D) and (D^I) coincide. By strong duality, the optimal values of (P) and (P^I) coincide.

Solution 5.1

Back to Exercise

Variables. $x_{uv} \in \{0,1\}$ equals 1 if agent u is assigned to task v, and 0 otherwise.

Formulation.

$$\begin{array}{lll} \text{Min} & \displaystyle\sum_{(u,v)\in U\times V} c_{uv}x_{uv} & \\ \text{s.t.} & \displaystyle\sum_{v\in V} x_{uv} = 1 & u \in U \\ & \displaystyle\sum_{u\in U} x_{uv} = 1 & v \in V \\ & x_{uv} \in \{0,1\} & (u,v) \in U \times V. \end{array}$$

Forbidden pairs. If $I \subseteq U \times V$ is the set of forbidden pairs, add constraints $x_{uv} = 0$ for all $(u,v) \in I$.

Solution 5.2

Back to Exercise

Variables. $x_{uv} \in \{0,1\}$ equals 1 if arc (u,v) is in the path, and 0 otherwise.

Formulation.

$$\begin{array}{lll} \text{Min} & \displaystyle\sum_{(u,v)\in E} c_{uv}x_{uv} & \\ \text{s.t.} & \displaystyle\sum_{v:(u,v)\in E} x_{uv} - \sum_{v:(v,u)\in E} x_{vu} = b_u & u \in V \\ & x_{uv} \in \{0,1\} & (u,v) \in E, \end{array}$$

with $b_s = 1$, $b_t = -1$, and $b_u = 0$ otherwise (flow conservation).

Validity. An integral unit s–t flow can consist of an s–t path together with directed cycles. Since all arc costs are nonnegative, these cycles can be removed without increasing the objective, so there exists an optimal

solution corresponding to an s–t path. The LP relaxation is integral because the node–arc incidence matrix is totally unimodular and the right-hand side is integral.

Solution 5.3

Back to Exercise

Variables. $x_{uv} \in \{0,1\}$ indicates whether arc (u,v) is used. Introduce auxiliary variables $z_i \in \mathbb{R}$ to eliminate subtours (MTZ).

$$
\begin{aligned}
\text{Min} \quad & \sum_{(u,v)\in E} c_{uv} x_{uv} && \\
\text{s.t.} \quad & \sum_{v:(u,v)\in E} x_{uv} = 1 && u \in V \\
& \sum_{v:(v,u)\in E} x_{vu} = 1 && u \in V \\
& z_u - z_v + |V|\, x_{uv} \le |V| - 1 && (u,v) \in E,\ u \neq v,\ u, v \neq 1 \\
& x_{uv} \in \{0,1\},\ z_u \ge 0 && (u,v) \in E,\ u \neq 1 \\
& z_1 = 0. &&
\end{aligned}
$$

The first two families ensure exactly one arc enters and leaves each node; the MTZ constraints prevent subtours. This formulation can be compared with the subtour-elimination formulations in Example 1.3 and Example 5.5; all three describe the same tours, although their LP relaxations need not have the same strength.

Solution 5.4

Back to Exercise

Variables. $x_{ij}^k \in \{0,1\}$ equals 1 if vehicle k travels from i to j (with $i, j \in N \cup \{0\}$). $y_k \in \{0,1\}$ equals 1 if vehicle k is used. $u_i^k \ge 0$ denotes the load of vehicle k after serving customer i.

$$
\begin{aligned}
\text{Min} \quad & \sum_{k\in K} \sum_{i,j\in N\cup\{0\}} c_{ij} x_{ij}^k + f \sum_{k\in K} y_k && \\
\text{s.t.} \quad & \sum_{j\in N\cup\{0\}} x_{ij}^k = \sum_{j\in N\cup\{0\}} x_{ji}^k && i \in N,\ k \in K \quad \text{(flow conservation)} \\
& \sum_{k\in K} \sum_{j\in N\cup\{0\}} x_{ij}^k = 1 && i \in N \quad \text{(serve each customer)} \\
& \sum_{j\in N} x_{0j}^k = \sum_{i\in N} x_{i0}^k = y_k && k \in K \quad \text{(start/end at depot)} \\
& u_0^k = 0 && k \in K \\
& u_j^k \ge u_i^k + d_j - Q(1 - x_{ij}^k) && i \in N\cup\{0\},\ j \in N,\ k \in K \quad \text{(capacity)} \\
& d_i \le u_i^k \le Q && i \in N,\ k \in K \\
& x_{ij}^k \in \{0,1\},\ y_k \in \{0,1\},\ u_i^k \ge 0 && (i,j) \in (N\cup\{0\})^2,\ k \in K.
\end{aligned}
$$

Solution 5.5

Back to Exercise

Variables. $y_j \in \{0,1\}$ indicates whether facility j is opened; $x_{ij} \ge 0$ is the shipped flow from j to i.

$$
\begin{aligned}
\text{Min} \quad & \sum_{j\in J} f_j y_j + \sum_{i\in I} \sum_{j\in J} c_{ij} x_{ij} && \\
\text{s.t.} \quad & \sum_{j\in J} x_{ij} = d_i && i \in I \\
& \sum_{i\in I} x_{ij} \le Q_j y_j && j \in J \\
& x_{ij} \ge 0,\ y_j \in \{0,1\}. &&
\end{aligned}
$$

Strengthening. For each (i, j), the individual linking inequality $x_{ij} \le d_i\, y_j$ is valid: demand satisfaction and nonnegativity imply $x_{ij} \le d_i$, while $y_j = 0$ forces $x_{ij} = 0$ through the aggregate capacity constraint. It can strengthen the LP relaxation when $d_i < Q_j$, because the aggregate capacity constraint alone only implies $x_{ij} \le Q_j y_j$.

Solution 5.6

Back to Exercise

Variables. For each $i \in I$, $x_i \in \{0, 1\}$ equals 1 if item i is selected, and 0 otherwise.
Formulation.

$$\begin{aligned} \text{Max} \quad & \sum_{i \in I} b_i x_i \\ \text{s.t.} \quad & \sum_{i \in I} w_i x_i \le W \\ & x_i \in \{0, 1\} \qquad i \in I. \end{aligned}$$

Valid inequality (cover). For any subset $S \subseteq I$ such that $\sum_{i \in S} w_i > W$ (a *cover*), we have

$$\sum_{i \in S} x_i \le |S| - 1.$$

In particular, taking S as a set of the heaviest items whose total weight exceeds W yields a simple strengthening cut.

Solution 5.7

Back to Exercise

Variables. $u_t \ge 0$ (production), $s_t \ge 0$ (end-of-period inventory), $y_t \in \{0, 1\}$ (setup), for $t = 1, \ldots, T$.

$$\begin{aligned} \text{Min} \quad & \sum_{t=1}^{T} \big(c_t u_t + h_t s_t + f_t y_t\big) & & \\ \text{s.t.} \quad & s_t = s_{t-1} + u_t - d_t & t = 1, \ldots, T \quad & \text{(balance)} \\ & u_t \le M_t y_t & t = 1, \ldots, T \quad & \text{(link production–setup)} \\ & u_t, s_t \ge 0,\ y_t \in \{0, 1\}. & & \end{aligned}$$

Optimality-preserving strengthening. For the original formulation with free terminal stock, production and inventory have no finite globally valid upper bounds. Under the nonnegative-cost assumptions, however, excess production cannot improve the objective. Define the unavoidable terminal inventory

$$\bar{s}_T := \max\left\{0,\ s_0 - \sum_{\tau=1}^{T} d_\tau\right\}.$$

There exists an optimal solution satisfying $s_T = \bar{s}_T$: if terminal stock exceeds $\bar{s}_T$, unnecessary production can be reduced while preserving feasibility, without increasing production, setup, or holding costs.

We may therefore impose $s_T = \bar{s}_T$ as a without-loss-of-optimality restriction. Summing the balances from t through T gives

$$\sum_{\tau=t}^{T} u_\tau = \bar{s}_T + \sum_{\tau=t}^{T} d_\tau - s_{t-1},$$

and hence

$$u_t \le \bar{s}_T + \sum_{\tau=t}^{T} d_\tau.$$

Thus an optimality-preserving choice is

$$M_t = \bar{s}_T + \sum_{\tau=t}^{T} d_\tau.$$

Similarly, summing from $t + 1$ through T yields the strengthening

$$s_t = \bar{s}_T + \sum_{\tau=t+1}^{T} d_\tau - \sum_{\tau=t+1}^{T} u_\tau \le \bar{s}_T + \sum_{\tau=t+1}^{T} d_\tau.$$

These inequalities are without-loss-of-optimality restrictions, not valid inequalities for every feasible solution of the original free-terminal-stock formulation. If $s_0 \le \sum_{\tau=1}^{T} d_\tau$, then $\bar{s}_T = 0$ and the bounds reduce to the simpler formulas without the $\bar{s}_T$ term.

Solution 5.8

Back to Exercise

Variables. For $j \in J$: start time $D_j \ge 0$, finish time $F_j \ge 0$, tardiness $\eta_j \ge 0$. For $i \ne j$: $y_{ij} \in \{0,1\}$ equals 1 if job i is scheduled before job j.

Objective. Min $\sum_{j\in J} \eta_j$.

Define the horizon

$$H := \max_{j\in J} r_j + \sum_{j\in J} \tau_j.$$

For any fixed job order, shifting jobs left whenever possible cannot increase any completion time or the total tardiness, and the resulting schedule finishes by H. We may therefore impose $F_j \le H$ for every j without loss of optimality and use $M = H$.

Constraints. For all $j \in J$ and $i \ne j$:

$F_j - D_j = \tau_j$	(processing time)
$D_j \ge r_j$	(release date)
$F_j \le H$	(optimality-preserving horizon)
$\eta_j \ge F_j - d_j, \quad \eta_j \ge 0$	(tardiness definition)
$D_j \ge D_i + \tau_i - H(1 - y_{ij})$	(if i before j)
$D_i \ge D_j + \tau_j - H y_{ij}$	(otherwise j before i)
$y_{ij} + y_{ji} = 1$	$(i < j)$
$y_{ij} \in \{0,1\}$.	

When the first disjunct is inactive, its right-hand side is $F_i - H \le 0$ and $D_j \ge 0$ makes it redundant. Symmetrically, the inactive second disjunct has right-hand side $F_j - H \le 0$ and is redundant because $D_i \ge 0$.

Solution 5.9

Back to Exercise

(1) Bilinear model.

$$\begin{aligned} \text{Max} \quad & p\,q - c\,q \\ \text{s.t.} \quad & q = A - Bp, \\ & 0 \le q \le Q^{\max}, \\ & \underline{p} \le p \le \overline{p}, \\ & \underline{q} \le q \le \overline{q}. \end{aligned}$$

The nonlinearity comes from the bilinear term $p\,q$ in the objective.

(2) McCormick relaxation. Introduce a revenue variable w intended to represent $p\,q$. McCormick envelopes for $p \in [\underline{p}, \overline{p}]$ and $q \in [\underline{q}, \overline{q}]$ give:

$$\begin{aligned} w &\ge \underline{q}\,p + \underline{p}\,q - \underline{p}\,\underline{q}, \quad & w &\ge \overline{q}\,p + \overline{p}\,q - \overline{p}\,\overline{q}, \\ w &\le \overline{q}\,p + \underline{p}\,q - \underline{p}\,\overline{q}, \quad & w &\le \underline{q}\,p + \overline{p}\,q - \overline{p}\,\underline{q}. \end{aligned}$$

The resulting LP relaxation is

$$\begin{aligned} \text{Max} \quad & w - c\,q \\ \text{s.t.} \quad & q = A - Bp, \\ & 0 \le q \le Q^{\max},\ \underline{p} \le p \le \overline{p},\ \underline{q} \le q \le \overline{q}, \\ & \text{(the four McCormick constraints above).} \end{aligned}$$

Because this system enlarges the feasible set in (p, q, w) and the problem is a maximization problem, its optimal value is an upper bound on the true optimal profit.

(3) The relaxation need not be exact. Consider

$$A = B = 1, \qquad \underline{p} = \underline{q} = 0, \qquad \overline{p} = \overline{q} = 1, \qquad Q^{\max} = 1, \qquad c = 0.$$

Then $q = 1 - p$, and the true problem is

$$\max_{0 \le p \le 1} p(1-p),$$

whose optimal value is $1/4$ at $p = q = 1/2$. On $[0,1]^2$, the McCormick inequalities reduce to

$$w \ge 0, \qquad w \ge p + q - 1, \qquad w \le p, \qquad w \le q.$$

At $p = q = 1/2$, all four inequalities allow $w = 1/2$, so the LP relaxation has value at least $1/2$, strictly above the true value $1/4$.

Strengthening hint. Tighten bounds before building McCormick envelopes: for instance, $q \in \big[\max\{\underline{q}, A - B\overline{p}\},\ \min\{\overline{q}, A - B\underline{p}\}\big]$.

Bibliography

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] S. Bubeck. *Convex Optimization: Algorithms and Complexity*. now publishers inc., 2015.

[3] G. B. Dantzig. Maximization of a linear function of variables subject to linear inequalities. In Tj. C. Koopmans, editor, *Activity Analysis of Production and Allocation*. Wiley, 1951.

[4] G. B. Dantzig, A. Orden, and P. Wolfe. The generalized simplex method for minimizing a linear form under linear inequality constraints. *Pacific Journal of Mathematics*, 1955.

[5] J. Matoušek and B. Gärtner. *Understanding and Using Linear Programming*. Springer, 2007.

[6] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004.