Demystifying Duality

Shaddin Dughmi
Department of Computer Science
University of Southern California
shaddin@usc.edu

May 4, 2024

Preface

This text concerns a central phenomenon in mathematical optimization known as *duality*. The discovery of duality was arguably one of the most influential mathematical ideas to come out of the twentieth century, and its impact can be felt in a multitude of fields such as computer science, economics, operations research, signal processing, machine learning, physics, and engineering. Despite its importance, we have found that an intuitive and deep understanding of duality can be elusive, often requiring years of experience with the topic, if it is ever attained at all.

Most texts on optimization treat duality largely as a symbolic transformation. Comparatively little emphasis is placed on the geometry of duality, or on its various pedagogical interpretations. Through teaching the topic over many years, the author has found that it is those interpretations that are most effective at instilling an intuitive sense of duality. This book distills this particular version of the story. Moreover, instead of a general treatment of linear or convex optimization, we instead focus entirely on duality, and present topics in optimization only as they are needed to understand the phenomenon.

The material in this book is adapted from portions of a course developed and taught by the author at USC between 2013 and 2024, titled "Convex and Combinatorial Optimization". We describe the duality of linear and convex optimization problems in terms of three interpretations: one economic, one physical, and one as an algebraic proof system. We hope that, between these three perspectives, most readers will find something familiar with which to anchor their intuition. In addition to the traditional algebraic form of duality, known as Lagrangian duality, we also describe its equivalent geometric form known as polar duality. This geometric perspective enables a common understanding of the duality of optimization problems, sets, and functions.

The audience for this book are students looking to further their understanding of duality, beyond that encountered in most standard courses or texts on optimization. In an instructional setting, we envision this book being used to supplement one of the many excellent introductory texts on optimization, or as a stand-alone text for a followup course focused on duality.

Contents

Chapter 1

A Brief Review of Linear Programming

In this first chapter, we briefly review the basics of linear programming. We present just enough detail to enable our exploration of duality in subsequent chapters. We refer the reader to more comprehensive texts on linear programming and optimization for a more thorough treatment.

We emphasize the geometric and economic perspectives of linear programming. This comes at the expense of the algebraic perspective, often rooted in the simplex method, which is predominant in most other texts on optimization. While the algebraic perspective is valuable for experts and practitioners, we find our approach to be more accessible, and also particularly suitable as a foundation for an intuitive grasp of duality — the goal of this book.

1.1 Linear Programming and Standard Forms

A *linear programming problem* is concerned with optimizing a linear function over a region of Euclidean space defined by a set of linear equalities and inequalities. Figure 1.1 illustrates the most general form of a linear program (henceforth, LP). Here, $x \in \mathbb{R}^n$ is a set of n real-valued *decision variables*, $c \in \mathbb{R}^n$ is a fixed set of coefficients describing the linear *objective function* $\langle c, x\rangle \triangleq \sum_{j=1}^n c_j x_j$, and $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ are given real numbers describing the ith linear *constraint* comparing $\langle a_i, x\rangle \triangleq \sum_{j=1}^n a_{ij} x_j$ to b_i. Each constraint can be an inequality $\leq$ or $\geq$, or an equality $=$, and we use $C^{\leq}$, $C^{\geq}$,$C^{=}$ to denote the corresponding sets of indices. Note that the parameters c, $\{a_i\}$, $\{b_i\}$, choice of linear relationships ($\leq$, $\geq$, or $=$ for each i), and the choice of whether to maximize or minimize, all form the *inputs* to the linear program. The decision variables x are the *outputs*. We refer to a vector x satisfying the constraints as a *feasible solution* of the LP (or *solution* for short). The family of all feasible solutions forms the *feasible region* of the LP. If x is a feasible solution and satisfies $\langle c, x\rangle \geq \langle c, y\rangle$ (or $\langle c, x\rangle \leq \langle c, y\rangle$ in the case of minimization) for every other feasible solution y, we say x is an *optimal solution* to the LP, and refer to the real

$$\begin{array}{lll}
\text{maximize (or minimize)} & \sum_{j=1}^n c_j x_j & \\
\text{subject to} & \sum_{j=1}^n a_{ij} x_j \leq b_i, & \text{for } i \in \mathcal{C}^{\leq}. \\
& \sum_{j=1}^n a_{ij} x_j \geq b_i, & \text{for } i \in \mathcal{C}^{\geq}. \\
& \sum_{j=1}^n a_{ij} x_j = b_i, & \text{for } i \in \mathcal{C}^{=}.
\end{array}$$

Figure 1.1: General Form of Linear Programming

$$\begin{array}{lll} \text{maximize} & \langle c, x\rangle & \\ \text{subject to} & \langle a_i, x\rangle \le b_i, & \text{for } i = 1, \dots, m. \\ & x_j \ge 0, & \text{for } j = 1, \dots, n. \end{array}$$

(a) Maximization

$$\begin{array}{lll} \text{minimize} & \langle c, x\rangle & \\ \text{subject to} & \langle a_i, x\rangle \ge b_i, & \text{for } i = 1, \dots, m. \\ & x_j \ge 0, & \text{for } j = 1, \dots, n. \end{array}$$

(b) Minimization

Figure 1.2: The Standard Forms of Linear Programming

$$\begin{array}{ll} \text{maximize} & x_1 + x_2 \\ \text{subject to} & x_1 + 2x_2 \le 2 \\ & 2x_1 + x_2 \le 2 \\ & x_1, x_2 \ge 0 \end{array}$$

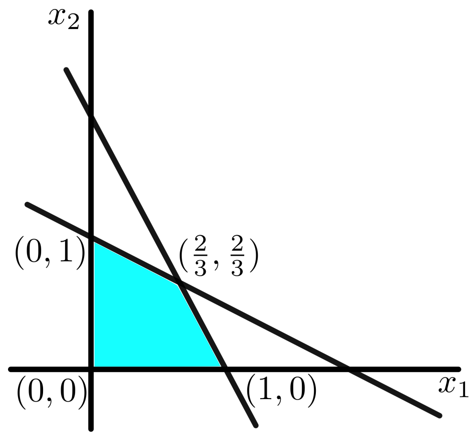

Figure 1.3: A 2-D example of LP

number $\langle c, x\rangle$ as the *optimal value* of the LP. If a constraint is satisfied with equality at a solution x (i.e. $\langle a_i, x\rangle = b_i$), then we say the constraint is *binding* or *tight* at x.

In general, the feasible region of an LP can be equivalently described using only $\le$ or only $\ge$ constraints. To see this, observe that every equality constraint can be replaced by a pair of inequalities, and an inequality can be reversed by multiplying both sides by -1. Moreover, we can turn a maximization problem into a minimization problem, or vice versa, by negating the objective: minimizing $\langle c, x\rangle$ is equivalent to maximizing $-\langle c, x\rangle$. Note that these modifications are merely syntactic, and do not change the geometry of the feasible set nor the semantics of the objective function. In order to further simplify our study of linear programming, we will often also restrict our attention to LPs where variables are constrained to be nonnegative. This restriction is without loss of semantic expressivity, since each real valued variable x_i can be replaced with two nonnegative variables x_i^+ and x_i^- plus the constraint $x_i = x_i^+ - x_i^-$.[1] Thus, we can write any LP either in *maximization standard form* or *minimization standard form*, both of which are illustrated in Figure 1.2.[2]

As shorthand, we often express the feasible region of an LP in maximization [minimization] standard form using the vector inequalities $Ax \preceq b$ [$Ax \succeq b$] and $x \succeq 0$, where A is the matrix with ith row a_i. We term the former *generic constraints* and the latter *nonnegativity constraints*. Here, the symbol $\preceq$ refers to entry-wise comparison between vectors, where $u \preceq v$ if and only if $u_j \le v_j$ for every coordinate j.

[1] That being said, this transformation does change the geometry of the feasible set.

[2] Note that there are many different definitions of the "standard form" in different communities and different textbooks; all serve the same purpose.

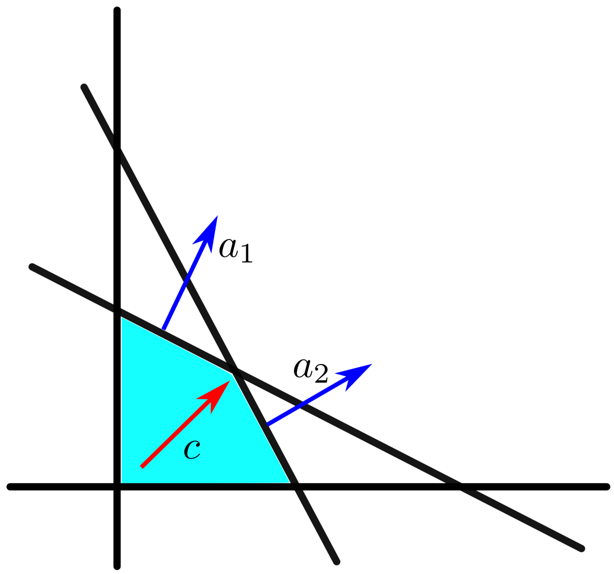

Figure 1.4: Physical Interpretation of LP

A Simple Example Consider the 2-variable linear program in Figure 1.3. The optimal value of this LP is $\frac{4}{3}$, and the optimal solution is $(\frac{2}{3}, \frac{2}{3})$. Note that the optimal solution is the point at which the top two constraints are binding — specifically, it is the solution of the linear system given by $x_1 + 2x_2 = 2$ and $2x_1 + x_2 = 2$.

1.2 Two Interpretations

We now present two interpretations of linear programming, one economic and the other physical/geometric. These interpretations will serve as an indispensable lens on duality in later chapters. We restrict attention, without loss of generality, to maximization standard form.

Economic Interpretation. Consider a facility which produces n different products — say, household chemicals — from m different raw materials. The coefficient c_j denotes the profit from selling each kilogram of product j, a_{ij} denotes the amount (in kilograms, say) of raw material i needed to produce a single kilogram of product j, and b_i denotes the amount of raw material i available. The linear program in Figure 1.2 can then be interpreted as solving for an amount x_i of each product; the objective is to maximize profit, and the constraints are imposed by the limited amount of raw materials available.

Physical Interpretation. As another interpretation of linear programming, consider a room in n dimensional space delimited by a number of walls, one for each inequality constraint. The sign of a constraint determines on which side of the wall the room lies. The vector of objective coefficients c described a direction, and the goal to travel as far as possible in the direction c while staying in the room. This is illustrated in Figure 1.4.

1.3 Basic Facts and Terminology

Before delving deeper into the properties of linear programs, we establish some basic terminology. A *hyperplane* is a region of Euclidean described by a single linear equality of the form $\langle a, x\rangle = b$. A *closed halfspace* (or *halfspace* for short) is the region defined by a linear inequality of the form $\langle a, x\rangle \leq b$ or $\langle a, x\rangle \geq b$. An *open halfspace* is the region defined by a strict linear inequality of the form $\langle a, x\rangle < b$ or $\langle a, x\rangle > b$. A *polyhedron* is the intersection of a finite number of closed halfspaces; note that the feasible region of linear program is a polyhedron. A *polytope* is a bounded polyhedron — i.e., one which does not go on forever in any direction. A *vertex* (a.k.a. *corner* or *extreme point*) of a polyhedron P is a point $x \in P$ with the property that there is no vector $y \neq 0$ with $x + y \in P$ and $x - y \in P$ — i.e., P does not include a nontrivial line segment through x. A *face* of a polyhedron P is the intersection with P of a hyperplane H disjoint from the interior of P. A face can be anywhere from zero-dimensional (a vertex) to $n-1$ dimensional; for example, the faces of a two dimensional polyhedron are a collection of points and lines.

We briefly mention some basic properties of polyhedrons and linear programs which we will need in this section. We leave the proofs as an exercise to the reader. Some of these properties will be revisited more thoroughly when we discuss convex sets.

Fact 1.3.1. *Every polyhedron P is a* convex set*; i.e., the line segment between any two points in P is also in P.*

When P is the feasible region of an LP with n variables, say one with constraints $Ax \preceq b$ for some matrix $A \in \mathbb{R}^{m\times n}$, right hand side vector $b \in \mathbb{R}^m$, and objective coefficients $c \in \mathbb{R}^n$, we mention three notable facts.

Fact 1.3.2. *The family of optimal solutions of the LP is a convex subset of P of dimension at most $n-1$; in fact, it is a face of P.*

To see this, observe that the family of optimal solutions is the intersection of P with the hyperplane $\langle c, x\rangle = OPT$, where OPT is the optimal value of the linear program.

Fact 1.3.3. *At each vertex x of P, at least n constraints of the linear program are binding; moreover, there exist n binding constraints at x which, when written as equalities, form a non-singular system of linear equations.*

In other words, there is an $n \times n$ non-singular submatrix A' of A, and corresponding subvector b' of b, satisfying $A'x = b'$.

Fact 1.3.4. *An LP falls into one of these three categories:*

- *The LP is* bounded*: There exists an optimal solution.*
- *The LP is* unbounded*: There exist arbitrarily good solutions of ever-increasing objective values.*
- *The LP is* infeasible*: there exist no feasible solutions to the LP, i.e. $P = \emptyset$.*

We have already seen an example of a bounded LP in Figure 1.3. As an example of an unbounded LP, consider the feasible region on the left side of Figure 1.5 paired with any objective pointing in the northeast direction. The right side of Figure 1.5 is an empty polyhedron — i.e., the LP is infeasible. Note the terminological distinction between a bounded feasible set and a bounded LP: Even though the feasible region on the left side of Figure 1.5 is unbounded, the corresponding LP with an objective pointing south-east would be bounded.

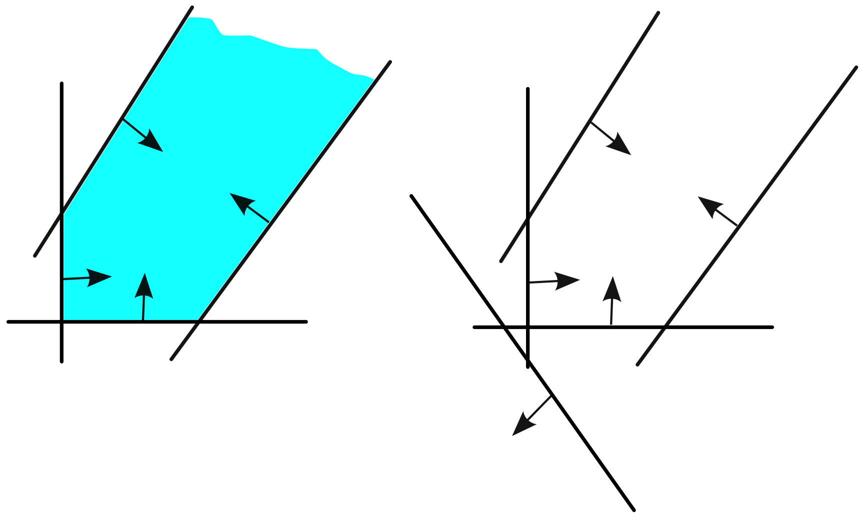

Figure 1.5: Unbounded and Infeasible LPs: arrows point to the interior of each halfspace

1.4 The Fundamental Theorem of LP

An important property of linear programs is that they typically admit "simple" optimal solutions. This is captured by the so-called "fundamental theorem" of linear programming, which in fact comes in a number of different guises. The most general variant is below.

Theorem 1.4.1. *If the feasible region P of an LP includes no lines (i.e., for every $y, d \in \mathbb{R}^n$, there exists $\alpha \in \mathbb{R}$ such that $y + \alpha d \notin P$), and moreover the LP admits an optimal solution, then it also admits an optimal solution at a vertex of P.*

More useful variants of this theorem are below.

Theorem 1.4.2. *If an LP has a bounded feasible region P (i.e., its feasible region is a polytope), then it admits an optimal solution at a vertex of P.*

Theorem 1.4.3. *If an LP in standard form admits an optimal solution, then it admits an optimal solution at a vertex of P.*

Next, we present a proof of Theorem 1.4.1. Theorems 1.4.2 and 1.4.3 follow immediately from Theorem 1.4.1 and the fact that neither polytopes nor the nonnegative orthant include a line.

Proof of Theorem 1.4.1. Let $P = \{x \in \mathbb{R}^n : Ax \preceq b\}$ for some matrix A and vector b — recall that any polyhedron can be written in this form. Let $c \in \mathbb{R}^n$ be the coefficients of the linear objective function. Let $\widehat{x} \in P$ be an optimal solution to the LP with the maximum number of binding constraints; let A' be the submatrix of A corresponding to the binding constraints at $\widehat{x}$.

Suppose for a contradiction that $\widehat{x}$ is not a vertex of P. By the definition of a vertex, this implies the existence of $y \neq \mathbf{0}$ such that $\widehat{x} + y \in P$ and $\widehat{x} - y \in P$. It must be that $\langle c, y \rangle = 0$, since otherwise either $\widehat{x} + y$ or $\widehat{x} - y$ has better objective value than x. Similarly, it must be that $A'y = \mathbf{0}$, since otherwise one of $\widehat{x} + y$ or $\widehat{x} - y$ must be infeasible.

Consider the line $L = \{\widehat{x} + \alpha y : \alpha \in \mathbb{R}\}$. By the previous discussion, every point on L has the same objective value as $\widehat{x}$, and all the binding constraints at $\widehat{x}$ are also binding everywhere on L. Nevertheless, by assumption $L \not\subseteq P$. Convexity implies that $L' = L \cap P$ is either a line segment with two endpoints, or a ray with a single endpoint. Let $\widetilde{x}$ be an endpoint of L'. An additional constraint (beyond the constraints in A') binds at $\widetilde{x}$, since otherwise we could "go further" along the line L. Therefore $\widetilde{x}$ is an optimal solution to the LP with strictly more binding constraints than $\widehat{x}$, a contradiction.

□

As a corollary, Theorem 1.4.2 implies the following immensely-useful structural property of linear programs in standard form, useful for LPs with more variables than constraints (excluding nonnegativity constraints).

Corollary 1.4.4. *Consider an LP in standard form with n variables and m generic constraints. If the LP admits an optimal solution, then it also admits an optimal solution with at most m non-zero variables.*

Proof. By Theorem 1.4.2, there is an optimal solution at a vertex x^* of the feasible region. Since x^* is a vertex, at least n constraints bind at x^*. Of those, at most m can be generic constraints, meaning that at least $n-m$ nonnegativity constraints bind at x^*. This implies that at least $n-m$ of the variables are zero at x^*. □

Application to Optimal Production. To appreciate the utility of Corollary 1.4.4, consider the optimal production interpretation of linear programs in standard form. The Corollary implies that there exists an optimal production plan which produces no more products than the number of different raw materials.

Chapter 2

Lagrangian Duality of Linear Programs, and Three Interpretations

In this chapter, we present linear programming duality in its traditional algebraic form: as a syntactic transformation of LPs. This is also often referred to as Lagrangian duality. While such a mechanical algebraic definition might appear dry and devoid of intuition at first glance, we endeavor to make it more accessible and intuitive by leaning on three equivalent interpretations: one economic, another physical/geometric, and a third based in logical proofs.

In Chapter 9, we will complement this algebraic form with its instructive, and arguably most enlightening, geometric analogue: polar duality.

2.1 The LP Duality Transformation

Linear programs come in pairs, where each maximization problem is associated with a *dual* minimization problem and vice versa. Moreover, the dual of the dual of an LP is the LP itself — i.e., duality, viewed as a relation on the space of linear programs, is an *involution.* When starting with an LP which most naturally encodes the task at hand, it is customary to refer to that original LP as *the primal LP* and to its dual as *the dual LP*, although in principle this nomenclature could be flipped. The dual of an LP provides a different, and often enlightening and useful, way of looking at the same optimization problem.

We begin by defining the syntactic transformation which derives the dual of an LP. Whereas this transformation might appear arbitrary to the unfamiliar reader, its genesis will become more clear over the course of this Chapter as well as Chapter 8. Starting with a primal LP in maximization standard form, as in Figure 2.1a, its dual is the LP in minimization standard form shown in Figure 2.1b. It is useful to view these programs in the shorthand form shown in Figure 2.2. Here, the

$$\begin{array}{lll} \text{maximize} & \sum_{j=1}^{n} c_j x_j & \\ \text{subject to} & \sum_{j=1}^{n} a_{ij} x_j \leq b_i, & \text{for } i = 1, \ldots, m. \\ & x_j \geq 0, & \text{for } j = 1, \ldots, n. \end{array}$$

(a) Maximization

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{m} b_i y_i & \\ \text{subject to} & \sum_{i=1}^{m} a_{ij} y_i \geq c_i, & \text{for } j = 1, \ldots, n. \\ & y_i \geq 0, & \text{for } i = 1, \ldots, m. \end{array}$$

(b) Minimization

Figure 2.1: The Primal and Dual of an LP in Standard Form

$$\begin{array}{ll} \text{maximize} & \langle c, x \rangle \\ \text{subject to} & Ax \preceq b \\ & x \succeq 0 \end{array}$$

(a) Maximization

$$\begin{array}{ll} \text{minimize} & \langle b, y \rangle \\ \text{subject to} & A^\intercal y \succeq c \\ & y \succeq 0 \end{array}$$

(b) Minimization

Figure 2.2: The Primal and Dual of an LP in Standard Form (shorthand)

	x_1	x_2	x_3	x_4	
y_1	a_{11}	a_{12}	a_{13}	a_{14}	b_1
y_2	a_{21}	a_{22}	a_{23}	a_{24}	b_2
y_3	a_{31}	a_{32}	a_{33}	a_{34}	b_3
	c_1	c_2	c_3	c_4	

Figure 2.3: Visualizing Variable/Constraint Correspondence

(given) parameters of each LP are the matrix $A \in \mathbb{R}^{m \times n}$ and vectors $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$, whereas $x \in \mathbb{R}^n$ is the vector of *primal variables* and $y \in \mathbb{R}^m$ is the vector of *dual variables*. Conversely, if we start with an LP in minimization standard form, as in Figure 2.1b, then its dual is the LP in maximization standard form shown in Figure 2.1a; in this case, we term y the primal variables and x the dual variables instead.

Note that generic constraints of the standard-form primal are in one-to-one correspondence with variables of its dual, and similarly for variables of the primal and constraints of the dual. We say that y_i is the dual variable corresponding to the primal constraint $\sum_{j=1}^n a_{ij} x_j \leq b_i$, and that $\sum_{i=1}^m a_{ij} y_i \geq c_i$ is the dual constraint corresponding to primal variable x_j. Figure 2.3 can serve as an aid to visualizing this correspondence: the rows give the primal constraints and corresponding dual variables, and the columns give the dual constraints and corresponding primal variables. This correspondence between variables and constraints is not a coincidence: each variable of the dual measures the "importance" of the corresponding primal constraint at the optimal solution, and vice versa. We elaborate on this in Section 3.3.

When faced with an LP in general form, we could derive a dual by first converting it to a standard form, then applying the transformation of Figure 2.1. There are multiple equivalent ways of doing this, all of which yield LPs which are essentially equivalent, but some of which are simpler to describe than others. We fix a standardized and convenient way of deriving the dual, shown in Figure 2.4: The dual of a maximization problem as in 2.4a is the minimization problem in 2.4b, and vice versa. Here, the (given) parameters of each LP are the matrix $A \in \mathbb{R}^{m \times n}$, vectors $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$, disjoint index sets $\mathcal{C}_1, \mathcal{C}_2 \subseteq [m]$, and disjoint index sets $\mathcal{D}_1, \mathcal{D}_2 \subseteq [n]$. As shorthand, we use a_i to denote the ith row of A and $\overline{a}_j$ to denote the jth column of A. Note that, as in the case of a primal/dual pair in standard form, there is still a one-to-one correspondence between primal constraints and dual variables, as well as primal variables and dual constraints. A useful rule of thumb for remembering the transformation of Figure 2.4 is the following: A "loose" (i.e., unconstrained) variable is dual to a "tight" (i.e., equality) constraint, and a "tight" (i.e., nonnegative or nonpositive) variable is dual to a "loose" (i.e., inequality) constraint.

It is simple to verify that the duality transformation of Figure 2.1 is a special case of that in Figure 2.4. The reader is also invited to verify that deriving the dual using the transformation of Figure 2.4 is equivalent — up to simple syntactic transformations and a change of variables — to first converting to a standard form then applying the transformation of Figure 2.1. We note that

$$
\begin{array}{lll}
\text{maximize} & \langle c, x\rangle & \\
\text{subject to} & \langle a_i, x\rangle \le b_i, & \text{for } i \in \mathcal{C}_1. \\
& \langle a_i, x\rangle \ge b_i, & \text{for } i \in \mathcal{C}_2. \\
& \langle a_i, x\rangle = b_i, & \text{for } i \in [m] \setminus (\mathcal{C}_1 \cup \mathcal{C}_2). \\
& x_j \ge 0, & \text{for } j \in \mathcal{D}_1. \\
& x_j \le 0, & \text{for } j \in \mathcal{D}_2.
\end{array}
$$

(a) Maximization

$$
\begin{array}{lll}
\text{minimize} & \langle b, y\rangle & \\
\text{subject to} & \langle \overline{a}_j, y\rangle \ge c_j, & \text{for } j \in \mathcal{D}_1. \\
& \langle \overline{a}_j, y\rangle \le c_j, & \text{for } j \in \mathcal{D}_2. \\
& \langle \overline{a}_j, y\rangle = c_j, & \text{for } j \in [n] \setminus (\mathcal{D}_1 \cup \mathcal{D}_2). \\
& y_i \ge 0, & \text{for } i \in \mathcal{C}_1. \\
& y_i \le 0, & \text{for } i \in \mathcal{C}_2.
\end{array}
$$

(b) Minimization

Figure 2.4: The Primal and Dual of a General LP

we distinguish variable nonnegativity and nonpositivity constraints from generic constraints when deriving the dual; this is merely for convenience, as an equivalent — though less elegant — dual could be derived by treating all constraints generically.

2.2 Three Interpretations of LP Duality

There is both a natural and formal relationship between an LP and its dual. Before presenting the formal relationship, captured by the theorems of weak and strong LP duality, we build intuition through three different interpretations: one economic, one logical, and one physical. We will reflect back on these three interpretations when we present the duality theorems in Section 2.3. The reader can choose whichever of these interpretations best develops their intuition.

Economic Interpretation Consider an LP in maximization standard form as in Figure 2.1a, interpreted as an optimal production problem faced by a production facility as described in Sec 1.2. The dual, shown in Figure 2.1b admits a natural interpretation as a problem facing a buyer of raw materials. In particular, the buyer must choose a price y_i to offer the facility per unit of the ith raw material. The prices are constrained so that the facility has incentive to sell all its raw materials rather than engage in any production. In particular, the jth generic constraint of the dual requires that the materials involved in producing a single unit of product j would fetch at least as much from being sold in raw form as would a unit of product j. Subject to incentivizing the facility to sell, the buyer looks to minimize the total price paid for all raw materials, as captured by the objective.

Logical Interpretation The dual of a maximization (minimization) LP can be interpreted as searching for a proof of the best possible upper bound (lower bound) on the optimal value which can be obtained by algebraically combining the constraints. This is best illustrated via an example. Consider the maximization LP in Figure 1.3, with optimal value $4/3$. Observe that the constraint $x_1 + 2x_2 \le 2$, along with the fact that variables are nonnegative, implies that $x_1 + x_2 \le x_1 + 2x_2 \le 2$. This proves an upper bound of 2 on the optimal value. We could obtain a tight upper bound by forming an appropriate linear combination of the two generic constraints:

$$\begin{array}{rl} 1/3\times & (x_1+2x_2\le 2) \\ & + \\ 1/3\times & (2x_1+x_2\le 2) \\ \hline & x_1+x_2\le 4/3 \end{array}$$

More generally, consider an LP in maximization standard form as in Figure 2.2a. Its dual, shown in Figure 2.2b, searches for nonnegative coefficients $y_1,\dots,y_m$ for each of the m generic constraints. Multiplying the ith constraint by y_i and adding them up yields the inequality $(y^\intercal A)x \le y^\intercal b$. When $A^\intercal y \succeq c$, as required by the dual constraints, we conclude an upper bound of $\langle b, y\rangle$ on the primal optimal value:

$$\langle c, x\rangle \le \langle A^\intercal y, x\rangle = y^\intercal Ax \le y^\intercal b = \langle b, y\rangle.$$

The objective of the dual is to minimize $\langle b, y\rangle$, in order to get the tightest upper bound possible.

A similar exercise shows that dual of a minimization LP searches for the best possible lower bound of the optimal value of the primal.

Physical Interpretation Recall that any LP with variables $x \in \mathbb{R}^n$ can be written in the following form, without altering the geometry of the feasible set nor the semantics of the objective function.

$$\begin{array}{ll} \text{maximize} & \langle c, x\rangle \\ \text{subject to} & \langle a_i, x\rangle \le b_i, \quad \text{for } i = 1,\dots,m. \end{array}$$

As described in Section 1.2, and illustrated in Figure 1.4, this can be interpreted as physical optimization problem set in an n-dimensional room. The room has m walls, with the ith wall described by the equation $\langle a_i, x\rangle = b_i$. The vector a_i is normal to the wall, and determines the "side" of the wall containing the room: a position vector a_i with its tail at the wall points "outside" the room. The objective is to travel as far as possible in the direction c without leaving the room.

We interpret this (primal) LP as describing the outcome of the following physical process. There is a tiny ball which is initially at an arbitrary starting point inside the room. Without loss of generality by a basic shifting argument, we may assume that this arbitrary starting point is the origin $\mathbf{0} \in \mathbb{R}^n$. A force field then applies a force vector c to the ball, and maintains this force indefinitely regardless of the position of the ball. Assuming frictionless walls, and a room which is "bounded" in the direction c, the ball eventually comes to rest at the furthest point in the room in the direction c; i.e., at the optimal solution of the primal LP. When the ball comes to rest, the walls adjacent to the ball collectively counteract the force vector c, with each wall producing a force perpendicular to its surface. This process is illustrated in Figure 2.5. Another interpretation of the objective function $\langle c, x\rangle$ is the amount of work done on the ball, by the force field, in traveling from the origin to the point x; equivalently, the potential energy difference between the origin and x.

Using the rules of Figure 2.4 we obtain the following dual, with the equality constraints combined into a single vector equality for convenience.

$$\begin{array}{lll} \text{minimize} & \langle b, y\rangle & \\ \text{subject to} & \sum_i y_i a_i = c & \\ & y_i \ge 0, & \text{for } i = 1,\dots,m. \end{array}$$

The dual can be interpreted as solving for the forces exerted on the ball at its resting place x^*. The dual variable $y_i \ge 0$ determines the magnitude $y_i||a_i||$ of the force exerted on the ball by the

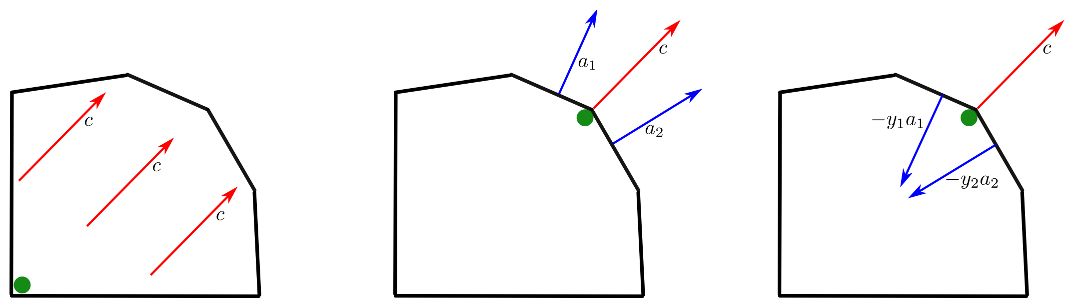

Figure 2.5: Physical Interpretation of LP

ith wall in direction $-\frac{a_i}{||a_i||}$, and the dual equality constraints guarantee that net force on the ball $c - \sum_i y_i a_i$ is zero.

To appreciate the dual objective function, imagine shrinking the room to a point (the origin) by slowly shifting each wall i a distance of $\frac{b_i}{||a_i||}$ in the direction $-\frac{a_i}{||a_i||}$. This takes the ball from its resting place x^* back to the origin, doing work on the ball in the process. Specifically, wall i would do a total work of $y_i||a_i|| \cdot \frac{b_i}{||a_i||} = y_i b_i$ on the ball. The dual objective, therefore, is the total work done on the ball in the process of moving it from its resting place back to the origin. Underlying the dual is the fact that the forces on the ball at its resting place can return the ball to the origin with the minimum amount of work; by strong duality (Section 2.3.2), an amount of work equal to the potential energy difference (due to the force field) between the resting position and the origin of the ball.

2.3 The LP Duality Theorems

2.3.1 Weak Duality

Weak duality states that the dual of a maximization problem bounds the value of the primal from above, and the dual of a minimization problem bounds its value from below.

Theorem 2.3.1 (Weak Duality)**.** *Consider a primal LP with variables $x \in \mathbb{R}^n$ and maximization objective $\langle c, x\rangle$, and its dual with variables $y \in \mathbb{R}^m$ and minimization objective $\langle b, y\rangle$. For every primal feasible x and dual feasible y, we have $\langle c, x\rangle \leq \langle b, y\rangle$.*

Proof. From the discussion of Section 2.1, it is sufficient to prove this for a primal and dual in standard form as in Figure 2.2. Let $x \succeq 0$ and $y \succeq 0$ be such that $Ax \preceq b$ and $A^\intercal y \succeq c$. We have

$$\begin{aligned} \langle c, x\rangle &\leq \langle A^\intercal y, x\rangle && \text{(By } x \succeq 0 \text{ and } c \preceq A^\intercal y) \\ &= y^\intercal A x \\ &= \langle y, Ax\rangle \\ &\leq \langle y, b\rangle && \text{(By } y \succeq 0 \text{ and } Ax \preceq b) \end{aligned}$$

□

The following corollary lists some immediate consequences of Theorem 2.3.1.

Corollary 2.3.2. *Consider a primal LP with variables $x \in \mathbb{R}^n$ and maximization objective $\langle c, x\rangle$, and its dual with variables $y \in \mathbb{R}^m$ and minimization objective $\langle b, y\rangle$. Let OPT_p and OPT_d denote the optimal values of the primal and dual, respectively, in the event that the corresponding program is feasible and bounded. The following hold:*

- *If both the primal and dual are feasible and bounded, then* $OPT_p \leq OPT_d$.
- *If* y *is dual feasible, and the primal program is feasible, then the primal is bounded and* $OPT_p \leq \langle b, y \rangle$.
- *If* x *is primal feasible, and the dual program is feasible, then the dual is bounded and* $OPT_d \geq \langle c, x \rangle$.
- *Suppose that* x *is primal feasible,* y *is dual feasible, and* $\langle c, x \rangle = \langle b, y \rangle$*. It follows that* x *and* y *are optimal solutions of the primal and dual, respectively, and moreover that* $OPT_p = \langle c, x \rangle = \langle b, y \rangle = OPT_d$.
- *If the primal is unbounded, then the dual is infeasible. Conversely, if the dual is unbounded then the primal is infeasible.*

Whereas the proof of Theorem 2.3.1 was somewhat algebraic and mechanical, much intuition can be gleaned from reflecting on weak duality in the context of our three interpretations of Section 2.2. Below, we restate weak duality in the language of each interpretation.

Economic Interpretation If prices on raw materials are set so that the facility (weakly) prefers to sell the raw materials over producing any individual product, then selling all raw materials is at least as profitable than any production schedule.

Logical Interpretation Algebraically combining the constraints, in the manner suggested by the dual and described in Section 2.2, yields a sound proof system for establishing bounds on the optimal value of the primal.

Physical Interpretation Consider a ball at position x in the room. Shrinking the room to a point as described in Section 2.2, and in the process returning the ball from x to the origin, must do at least as much work on the ball as the potential energy difference (due to the force field) between the origin and x.

2.3.2 Strong Duality

Strong duality states that the bounds implied by weak duality are in fact tight. In other words, the primal and dual are two different perspectives on what is essentially the same optimization problem.

Theorem 2.3.3 (Strong Duality)**.** *If either a linear program or its dual is feasible and bounded, then both programs are feasible, bounded, and have the same optimal value.*

The formal proof of strong LP duality is somewhat technical, so we defer it to Appendix A. Instead, we include an informal proof at the end of this section, which we hope is useful for building the reader's intuition. But first, as we did for weak duality, we restate strong duality in the language of each of our three interpretations.

Economic Interpretation One can set prices on raw materials so that, in addition to weakly preferring to sell raw materials over producing any individual product, the facility is indifferent between selling all raw materials and engaging in optimal production.

Logical Interpretation Algebraically combining the constraints, in the manner suggested by the dual and described in Section 2.2, yields a complete proof system for establishing bounds on the optimal value of the primal.

Physical Interpretation Consider a ball at its final resting position x against some walls of the room. Shrinking the room to a point as described in Section 2.2 can return the ball to the origin without wasting any energy. In other words, the amount of work done on the ball by the walls can — through a proper assignment of forces to the walls — be made to equal the potential energy difference (due to the force field) between the origin and x.

Our informal proof of strong duality hinges on the physical interpretation above.

Informal Proof of Theorem 2.3.3. Consider a pair of dual linear programs, and assume one of them is feasible and bounded — without loss of generality, the primal. Also without loss of generality, we assume that the primal is written as a maximization problem, and the dual is therefore a minimization problem, and both are written in the forms used for our physical interpretation of Section 2.2. By weak duality (Theorem 2.3.1), it suffices to exhibit primal and dual solutions with the same objective value.

Consider the physical interpretation of the primal, and let $x \in \mathbb{R}^n$ be the final resting position of the ball; i.e., x is an optimal solution to the primal, with optimal value $\langle c, x \rangle$. Since the ball is stationary at x, the force field c will be neutralized by the forces exerted on the ball by the walls. In particular, there exist force multipliers $y \succeq 0$ such that $\sum_{i=1}^{m} y_i a_i = c$. Therefore y is a feasible solution to the dual.

We now make a somewhat informal leap based on our intuition regarding this physical system. In particular, when the ball is stationary we expect that only the walls adjacent to the ball exert any force at all on the ball. Formally, this means that $y_i > 0$ only if $\langle a_i, x \rangle = b_i$, or equivalently that $y_i(b_i - \langle a_i, x \rangle) = 0$. This allows us the complete the proof.

$$\begin{aligned}
\langle b, y \rangle - \langle c, x \rangle &= \langle b, y \rangle - \langle A^\intercal y, x \rangle \\
&= \langle b, y \rangle - x^\intercal A^\intercal y \\
&= \langle b, y \rangle - \langle Ax, y \rangle \\
&= \langle y, b - Ax \rangle \\
&= \sum_i y_i (b_i - \langle a_i, x \rangle) \\
&= 0
\end{aligned}$$

□

We find that this informal proof is all one needs to appreciate the mathematics of LP duality. That said, a formal proof is presented in Appendix A for the interested reader.

Chapter 3

Some Consequences of LP Duality

In this chapter, we present some important consequences of the duality theorems of linear programming, viewed through the lens of our three interpretations.

3.1 Complementary Slackness

Complementary slackness describes a relationship which holds between every pair of primal and dual optimal solutions. This relationship is useful for a number of reasons, as it often reveals structure in the underlying linear programs, and moreover enables recovering a primal solution from a dual solution and vice versa.

Consider a primal LP and its dual in general form as in Figure 2.4. Let x and y be feasible solutions to the primal and dual, respectively. The pair (x, y) is said to satisfy *complementary slackness* if and only if the following hold:

- $x_j(\langle \overline{a}_j, y\rangle - c_j) = 0$ for all j.
- $y_i(b_i - \langle a_i, x\rangle) = 0$ for all i.

In other words, complementary slackness requires that each primal (dual) variable is nonzero only if its corresponding dual (primal) constraint is binding. Equivalently, each primal (dual) constraint is slack only if its corresponding dual (primal) variable is zero. We can show that complementary slackness holds at optimality for every pair of dual linear programs.

Theorem 3.1.1 (Complementary Slackness)**.** *Consider a primal LP and its dual in general form as in Figure 2.4. Let x and y be feasible solutions to the primal and dual, respectively. x and y are both optimal for their respective programs if and only if complementary slackness holds for (x, y).*

Proof. We begin by proving this for linear programs in standard form, as in Figure 2.1. Let x and y be primal and dual feasible solutions: $x \succeq 0$, $Ax \preceq b$, $y \succeq 0$, $A^\intercal y \succeq c$. Let $s_i = b_i - \langle a_i, x\rangle$ be the ith *primal slack variable*, and let $t_j = \langle \overline{a}_j, y\rangle - c_j$ denote the jth *dual slack variable*. Note that $s, t \succeq 0$. We can write the gap between the value of the dual and the value of the primal as follows.

$$\begin{aligned}\langle y, b\rangle - \langle c, x\rangle &= y^\intercal(Ax + s) - (A^\intercal y - t)^\intercal x \\ &= y^\intercal Ax + y^\intercal s - y^\intercal Ax + t^\intercal x \\ &= \langle y, s\rangle + \langle x, t\rangle\end{aligned}$$

Strong and weak duality imply that x and y are both optimal if and only if the above gap is zero. Since $y, x, s, t \succeq 0$, the expression $\langle y, s\rangle + \langle x, t\rangle$ is zero if and only if complementary slackness holds for (x, y).

To generalize this proof to LPs in general form, we first convert to a pair of dual LPs in standard form as in Figure 2.1. For each pair of solutions to the original pair of linear programs, we convert to a pair of solutions to the standard form LPs, deduce complementary slackness there, then convert back. The details are straightforward, and left as an exercise for the reader. □

To build intuition, we now restate Theorem 3.1.1 in the language of two of our interpretations from Section 2.2

Economic Interpretation Given a production schedule x and dual offer prices y for the raw materials, both x and y are optimal if and only if

- Facility only manufactures products for which it is indifferent between production and sale.
- Only raw materials which are binding constraints on production command a non-zero price.

Physical Interpretation Let x be a position for the ball, and let y be forces exerted on the ball by the walls. The ball is at rest if and only if only the walls adjacent to the ball at x push on the ball with a non-zero amount of force.

3.2 Computational Equivalence of Primal and Dual

We will encounter linear programs where it is easier to solve a linear program's dual than to solve the primal directly (or vice versa). Typically in such scenarios, complementary slackness permits recovering the solution of the primal from that of the dual (or vice versa) through solving a linear system. The general situation is somewhat intricate, so we first present it under the assumption that both LPs are in standard form, and that neither exhibits *degeneracy.*

A linear program is said to be *degenerate* if there exists an optimal vertex where the number of tight constraints exceeds the number of the variables of the LP. This can occur only when the constraints of the LP are *redundant*: some of the constraints can be removed without affecting the feasible region of the LP. An example of a degenerate LP is shown in Figure 3.1.

Theorem 3.2.1. *Consider a primal/dual pair of linear programs in standard form, and let n and m denote the number of variables and generic constraints of the primal, respectively. Suppose both LPs are non-degenerate. Given a vertex optimal solution to one of the LPs, a vertex optimal solution to the other can be computed by solving a non-singular* $\min(n, m) \times \min(n, m)$ *system of linear equations.*

Proof. We show how to recover the solution of a primal LP in maximization standard form from a solution to its dual; the other direction is essentially identical. Let the primal and dual be as in Figure 2.1, and let $y \in \mathbb{R}^m$ be a given vertex optimal solution to the dual. The primal admits at least one vertex optimal solution, by Theorem 2.3.3.

Note that the dual feasible region is described by n generic constraints and m nonnegativity constraints, for a total of $m+n$ constraints. Fact 1.3.3 and the assumption of non-degeneracy imply that exactly m dual constraints bind at y, leaving n loose dual constraints (some generic, some nonnegativity). Complementary slackness (Theorem 3.1.1) then helps us identify n corresponding constraints of the primal (some nonnegativity, some generic) which must be tight at every primal

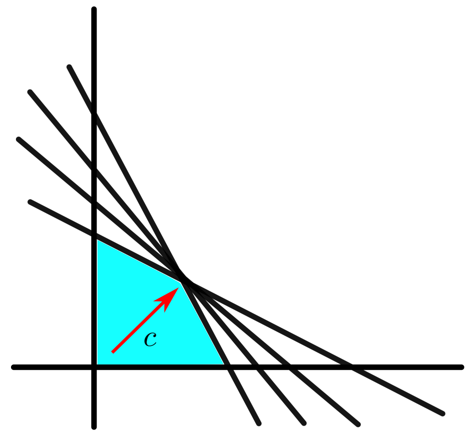

Figure 3.1: A Degenerate LP

optimal solution. This yields an $n \times n$ system of linear equations which is satisfied by every primal optimal solution. Non-degeneracy implies that no other constraints can be tight at a vertex optimal solution to the primal. Invoking Fact 1.3.3, this in turn implies that our $n \times n$ system of equations is non-singular, and its solution is the unique optimal solution to the primal.

When $n \leq m$, our proof is complete. However, when $n > m$, we observe that our non-singular $n \times n$ system of equations includes at most m generic primal constraints, with the remaining constraints setting at least $n - m$ primal variables to zero. Discarding $n - m$ variables which are set to zero, we obtain a non-singular $m \times m$ system of linear equations whose solution sets the remaining m variables. □

Our assumption that our linear programs are in standard form is typically without loss, since in most cases we may convert the LPs to standard form without significant computational overhead. Our assumption of non-degeneracy simplified our proof. However, this assumption tends to be unnecessary in practice, since most algorithms for solving a linear program produce, as a byproduct, a non-singular system of tight constraints describing the solution. In particular, alongside the dual optimal solution y, we can identify the tight dual constraints at y which are non-redundant. The other tight constraints at y are redundant and may be discarded or, equivalently, treated as if they were loose by setting the corresponding primal variables to zero.

3.3 Sensitivity Analysis

It is often useful to analyze how small changes in the inputs to a linear program affect its outputs. Consider a linear program in (maximization or minimization) standard form as in Figure 2.2. Its optimal value OPT is a function of its input parameters A, b, and c. This function is continuous, but is in general non-differentiable. Nevertheless, where it is differentiable its partial derivatives yield useful information about the "importance" of each input in determining the optimal value. In our economic interpretation, for example, one may ask how a small increase in one or more raw materials affects profits. We summarize some basic results on sensitivity in the following theorem.

Theorem 3.3.1. *Consider a primal-dual pair of linear programs in standard form in Figure 2.1, with input parameters $\{a_{ij}\}$, $\{b_i\}$, and $\{c_j\}$. The following holds for the optimal value OPT as a function of the input parameters.*

- *If there is a unique primal optimal* x^**, then* OPT *is differentiable with respect to* c *at* (A, b, c) *with* $\frac{\partial OPT}{\partial c_j} = x_j^*$.
- *If there is a unique dual optimal* y^**, then* OPT *is differentiable with respect to* b *at* (A, b, c) *with* $\frac{\partial OPT}{\partial b_i} = y_i^*$.

Proof. Consider the primal linear program, and suppose x^* is the unique primal optimal. It follows from Theorem 1.4.3 that x^* is a vertex and that every other vertex is strictly suboptimal. Consider adding δ to c_j, where $\delta \neq 0$ is sufficiently small to preserve optimality of x^* among the finite set of vertices (and therefore overall by Theorem 1.4.3). The optimal value increases by δx_j^*. Taking the limit as $\delta \to 0$ (whether from above or from below), it follows that $\frac{\partial OPT}{\partial c_j} = x_j^*$.

A symmetric argument applied to the dual linear program, with unique optimal y^*, shows that $\frac{\partial OPT}{\partial b_i} = y_i^*$.

□

Chapter 4

Examples of LP Duality Relationships

This chapter presents three illustrative examples of duality relationships, which will help build an intuitive grasp of the concept. In each case, the dual lends an alternative perspective to the optimization problem at hand.

4.1 The Shortest Path Problem

In the shortest path problem, we are given a directed graph (a.k.a. a network) $G = (V, E)$, where V is a finite set of *nodes* and $E \subseteq V \times V$ is a set of *edges* (a.k.a. links). Each edge $e \in E$ is labeled with its *length* $\ell_e \in \mathbb{R}$. We are given two nodes $s, t \in V$, and the goal is to find the shortest path from s to t in the graph. An example is shown in Figure 4.1.

The problem of searching for the shortest path can be encoded as a linear program as follows. We designate a nonnegative variable x_e for each edge e, serving as an indicator as to whether e is part of the shortest path. We envision that our variables will be set to 0 or 1, though do not impose those constraints directly (in fact, such an integer constraint is impossible in linear programming). The constraints we do impose, however, require that the path exits s one more time than it enters, enters t one more time that it exits, and for every $v \in V \setminus \{s, t\}$ the path enters v as many times as it exits. The LP objective minimizes the length of the path. The resulting LP is shown in Figure 4.2a, where $E^{\mathtt{in}}(v)$ and $E^{\mathtt{out}}(v)$ denote the edges into and out of $v \in V$, respectively. It can be formally shown that vertices of this LP just so happen to set each variable to 0 or 1, rendering

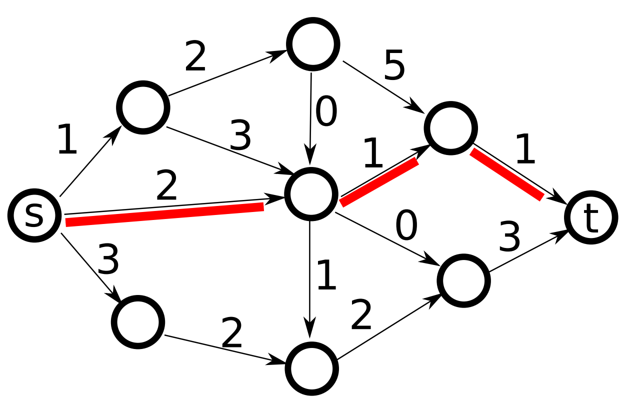

Figure 4.1: The Shortest Path Problem

$$
\begin{array}{lll}
\text{minimize} & \sum_{e\in E} \ell_e x_e & \\
\text{subject to} & \sum\limits_{v\in E^{\mathtt{in}}(s)} x_e - \sum\limits_{e\in E^{\mathtt{out}}(s)} x_e = -1 & \\
& \sum\limits_{v\in E^{\mathtt{in}}(t)} x_e - \sum\limits_{e\in E^{\mathtt{out}}(t)} x_e = 1 & \\
& \sum\limits_{v\in E^{\mathtt{in}}(v)} x_e - \sum\limits_{e\in E^{\mathtt{out}}(v)} x_e = 0, & \text{for } v \in V \setminus \{s,t\}\,. \\
& x_e \geq 0, & \text{for } e \in E.
\end{array}
$$

(a) Primal

$$
\begin{array}{lll}
\text{maximize} & y_t - y_s & \\
\text{subject to} & y_v - y_u \leq \ell_e, & \text{for } (u,v) \in E.
\end{array}
$$

(b) Dual

Figure 4.2: The Shortest Path LP and its Dual

integer constraints unnecessary; we say the LP is *integral*, or an *(exact) LP formulation* for the shortest path problem.

Applying the duality transformation of Section 2.1, we derive the dual LP shown in Figure 4.2b. The dual LP can be interpreted as solving a natural maximization problem, as follows. Suppose that each edge $e = (u, v)$ is a rope of length ℓ_e connecting nodes u and v. Holding node s in one hand and t in the other, how far apart can you pull s and t without severing any of the ropes? This question is modeled by the dual LP as follows: We interpret y_v as the "height" of node v. The LP constraints require, for each edge $e = (u, v)$, that v not be so high relative to u so as to result in severing e. Subject to these constraints, the objective maximizes the difference in heights between t and s.

Some thought reveals that the shortest path between s and t serves as the limit on the difference in heights between t and s. Hence the two problems are equivalent or, more precisely, *dual* to each other.

4.2 Maximum Weight Matching

In the maximum-weight matching problem, there is an undirected graph $G = (V, E)$, where V is a set of nodes and $E \subseteq \binom{V}{2}$ is the set of (undirected) edges. Moreover, for each $e \in E$ we are given a *weight* $w_e \in \mathbb{R}$. A matching is a set of edges $M \subseteq E$ such that each node appears at most once in M. The objective is to find a matching M of maximum total weight $w(M) = \sum_{e\in M} w_e$.

Consider the following linear program for the maximum weight matching problem. We designate a nonnegative variable x_e for each pair $e \in E$, serving as an indicator as to whether $e \in M$. As in the shortest path problem, we envision that our variables will be set to 0 or 1, though do not impose this integer constraint directly (which, again, is impossible in a linear program). The constraints we do impose restrict each node to appearing at most once. The LP objective maximizes the total weight of the matching. The resulting LP is shown in Figure 4.3a, where $E(v) \subseteq E$ denotes the edges incident on node v.

It is worth noting that, unlike in the shortest path problem, we can not guarantee that solving this LP yields an integer solution in general. In other words, for some graphs we might obtain a *fractional* matching. This linear program is therefore not an exact LP formulation of maximum weight matching in general graphs, but rather what is known as a linear programming *relaxation* of the problem. As we will see later in this text, a relaxation of a combinatorial problem is often of great utility in the design of approximation algorithms and heuristics, and in identifying tractable special cases. In the case of maximum weight matching, it can be shown that our LP is indeed an

maximize	$\sum_{e\in E} w_e x_e$		minimize	$\sum_{v\in V} p_v$	
subject to	$\sum_{e\in E(v)} x_e \leq 1,$	for $v \in V$.	subject to	$p_u + p_v \geq w_{(u,v)},$	for $(u,v) \in E$.
	$x_e \geq 0,$	for $e \in E$.		$p_v \geq 0,$	for $v \in V$.

(a) Primal

(b) Dual

Figure 4.3: An LP for Max-Weight Matching and its Dual

exact formulation when the graph is *bipartite*: when vertices can be partitioned into two "sides" such that each edge is incident to a node from each side. The reader may assume that the graph is bipartite, and hence the LP integral, for the rest of this section.

Applying the duality transformation of Section 2.1, we derive the dual LP shown in Figure 4.3b. In order to appreciate the relationship between the primal and dual, consider the following interpretation of both programs. We interpret each node $v \in V$ as a production resource, and each edge $e = (u, v)$ as a potential investment project which requires resources u and v and generates profit w_e. The primal problem captures the task of an investor looking to select a profit-maximizing set of projects, subject to the constraint that a resource can be used by at most a single project at a time. The dual problem, on the other hand, captures the task of a buyer looking to purchase all the resources by offering a price p_v for each resource v. For the offer to be accepted, it must be more profitable for the investor to sell any pair of resources (u, v) than to undertake the investment $e = (u, v)$. The dual LP solves the buyer's problem of minimizing the total amount paid, subject to the offer being acceptable to the investor.

Given this interpretation, weak duality implies that acceptable offers have the buyer paying at least as much as the investor's profit from any feasible set of projects. Strong duality implies that, at optimality, the buyer's cost equals the investor's profit.

4.3 Zero Sum Games

One of the early motivating applications of linear programming is in the context of *game theory*. This is the field which studies scenarios, known as *games*, where strategic and self-interested agents interact. Such interactions may involve competition, cooperation, or elements of both. Linear programming is very closely related to games of pure competition between two players, known as *two-player zero-sum games*. Despite the restriction to two players and pure competition, the influence that the study of these games has had on mathematics, economics, and science is difficult to overstate.

We model a two-player zero-sum game using a matrix A, where rows index the actions of the *row player*, and columns index the strategies of the *column player*. Each cell of the matrix describes the utility to each player from the corresponding pair of strategies. Being a model of pure competition, one player's gain is the other's loss in zero-sum games. A single real number a_{ij} then suffices for each entry of the matrix corresponding to row i and column j, which by convention we will fix to be the utility of the column player. The utility of the row player is therefore presumed to be $-a_{ij}$.[1] We restrict attention to games with finitely many actions, and use m to n to denote the number of strategies for the row and column player respectively; in other words, A is an $m \times n$ matrix. A popular example of zero-sum games is the game of Rock-Paper-Scissors, shown in Figure 4.4.

[1] Zero-sum games are equivalent to constant sum games, where the sum of players' utilities is fixed regardless of their strategies. For simplicity, we will stick to zero-sum games.

	R	P	S
R	0	1	-1
P	-1	0	1
S	1	-1	0

Figure 4.4: Rock-Paper-Scissors as a Zero-Sum Game

When the row player plays i and the column player plays j, the column player's utility is a_{ij} (and the row player's utility is therefore $-a_{ij}$). We also allow randomized play, where each player chooses their strategy randomly. In particular, we allow the column player selects column j with probability $x_j \geq 0$, where $\sum_{j=1}^{n} x_j = 1$. Similarly, the row player selects row i with probability $y_i \geq 0$, where $\sum_{i=1}^{m} y_i = 1$. We refer to the vectors x and y as *mixed strategies*. The original (deterministic) strategies, indexed by rows and column, are contrastingly referred to as *pure strategies* to avoid ambiguity. We assume that the players do not coordinate their play, in the sense that they draw their random strategies independently of each other. It follows therefore that with probability $y_i x_j$, the row player plays i and the column player plays j, yielding utility a_{ij} to the column player. We can therefore write the column player's expected utility as follows:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} y_i x_j a_{ij} = y^T A x.$$

Naturally, the row player's expected utility is the negation of this quantity.

To fully specify the game, we need to fix an order of events. We examine two possible setups, as understanding the relationships between them is key to appreciating the connection between zero-sum games and linear programming duality. The first, and perhaps most familiar, is the *simultaneous move game* in which players select their strategies simultaneously. The second is the *Stackelberg game*, where one player moves first by selecting his strategy, and the second responds by selecting her own. When the first mover chooses a mixed strategy, we assume the second-mover only learns its distribution but not its realization. Whereas in the simultaneous move game neither player has an advantage over the other beyond that baked into the game matrix, the asymmetry of the Stackelberg setting can result in a first mover or second mover advantage in general games.

A primary concern of game theory is to predict the outcome of games, when players act strategically to maximize their utilities. The key concept here is that of an *equilibrium*: a choice of strategy by each player so that neither player can gain by unilaterally deviating. In the simultaneous move setting, where the row player selects y^* and the column player selects x^*, this means that each player's strategy is a *best response* to the other's strategy as follows:

- y^* is a best response by the row player to x^*, in that it minimizes $y^T A x^*$ over all y.
- x^* is a best response by the column player to y^*, in that it maximizes $(y^*)^T A x$ over all x.

When y^* and x^* are pure strategies, meaning that $y^* = e_i$ and $x^* = e_j$ for a row i and column j, we refer to (x^*, y^*) as a *pure equilibrium*. Otherwise, we refer to it as a *mixed equilibrium*. The only simultaneous-move equilibrium of the Rock-Paper-Scissors game is that where each player uniformly randomizes over his strategies, assigning equal probability of $1/3$ to each of Rock, Paper, and Scissors. Each player has an equal probability of a win, lose, and tie, and therefore has expected utility zero in this equilibrium. There is no pure equilibrium of Rock-Paper-Scissors.

In the Stackelberg setting, the situation is somewhat more nuanced. The player who moves second best responds as in the simultaneous move game. The player moving first, however, cannot

$$\begin{array}{ll} \text{maximize} & u \\ \text{subject to} & Ax \succeq u\vec{1} \\ & \sum_j x_j = 1 \\ & x \succeq 0 \end{array}$$

(a) Column Player Moves First (Maximin)

$$\begin{array}{ll} \text{minimize} & v \\ \text{subject to} & A^T y \preceq v\vec{1} \\ & \sum_i y_i = 1 \\ & y \succeq 0 \end{array}$$

(b) Column Player Moves Second (Minimax)

Figure 4.5: Stackelberg Equilibria of Zero-Sum Game as Linear Programs

best respond in the same manner: he must plan ahead, knowing that his opponent will base their strategy on his own. When the column player moves first, the equilibrium conditions are as follows:

- x^* maximizes the column player's expected utility, given the row player will best respond to x^*. In particular, x^* maximizes $\min_y y^T Ax$ over all x.
- y^* is a best response by the row player to x^*, in that it minimizes $y^T Ax^*$ over all y. (as in the simultaneous move game)

We note the player who moves second need not randomize. Consider the (second mover) row player's best response problem: $y^T Ax^*$ is minimized at a pure strategy $y^* = e_i$, where i is any index of a smallest entry of the vector Ax^*. Typically there are many such minimal entries i, and any y^* which randomizes between them is a best response. In the simultaneous-move setting, the requirement that x^* be a best response to y^* further constrains y^*, typically entailing randomization. When the column player moves first, however, his best response constraint makes no explicit mention of y^*, leaving the row player free to choose any minimizer of $y^T Ax^*$, including a pure strategy.

In summary, when the column player moves first his utility is

$$\max_x \min_y y^T Ax = \max_x \min_i (Ax)_i.$$

We refer to this as the *column player's maximin utility*, and to its negation as the *row player's minimax utility*. We refer to x^* attaining this maximum as the *column player's maximin strategy*. When the column player moves second, the situation is reversed and his utility is

$$\min_y \max_x y^T Ax = \min_y \max_j (y^T A)_j.$$

We refer to this as the *column player's minimax utility*, and to its negation as the *row player's maximin utility*. We refer to y^* attaining this minimum as the row player's *maximin strategy*. In an equilibrium of the Stackelberg Rock-Paper-Scissors game, the first mover uniformly randomizes his strategy, and the second mover's mixed strategy is completely arbitrary — all pure and mixed strategies are best responses. The expected utility of both players is zero, as in the the simultaneous move setting. This is no coincidence.

Theorem 4.3.1 (Minimax Theorem)**.** *Let A be utility matrix of a zero-sum game. There is no first or second mover advantage in the Stackelberg game since a player's minimax and maximin utilities are equal:*

$$\max_x \min_i (Ax)_i = \min_y \max_j (y^T A)_j.$$

Moreover, the maximin strategies x^* and y^* of the two players form an equilibrium of the simultaneous move game. Each player obtains exactly their maximin (or equivalently, minimax) utility at this equilibrium.

Proof. The column player's first-mover (maximin) and second-mover (minimax) utilities are the optimal values of the pair of linear programs shown in Figure 4.5. Using the rules of duality in Figure 2.4, it is easy to show that these two LPs are in fact duals. Weak duality implies that the maximin utility is no more than the minimax utility, ruling out a first-mover advantage. Strong duality implies they are equal, ruling out a second-mover advantage.

Let $x^* \in \operatorname{argmax}_x \min_i (Ax)_i$ and $y^* \in \operatorname{argmin}_y \max_j (y^T A)_j$ the maximin strategies of the column and row player respectively. Let u^* be the maximin utility, or equivalently the minimax utility, of the column player. Suppose the column player plays x^* and the row player plays y^*. By playing x^*, the column player guarantees himself at least u^* regardless of the row player's strategy. By playing y^*, the row player guarantees the column player's utility is no more than u^* regardless of the column player's strategy. Therefore, the column player obtains utility exactly u^*, and moreover cannot improve his utility by deviating. Similarly, the row player obtains utility exactly $-u^*$, and cannot improve her utility by deviating. It follows that (x^*, y^*) is an equilibrium, and each player obtains their maximin (or equivalently, minimax) utility. □

Since the maximin, minimax, and simultaneous-move utility are all equal, we singularly refer to them as *the value of the game*.

We close with some observations. First, we needed randomization to eliminate the second-mover advantage in zero-sum games. Indeed, if player's can't randomize in Rock-Paper-Scissors, and one moves first, then the second-mover always wins! Second, complementary slackness implies that $x_j^* = 0$ unless the j'th entry of $y^{*T} A$ is of maximum value, and $y_i^* = 0$ unless the ith entry of Ax is of minimum value. This implies that, at equilibrium, the column player randomizes over only the pure best responses to the row player's mixed strategy, and vice versa. This stands to reason, as any random strategy involving a non-best-response can be improved, and is therefore not itself a best response.

Chapter 5

A Brief Review of Convex Optimization

FORTHCOMING

Chapter 6

Extending Lagrangian Duality to Convex Programs

FORTHCOMING

Chapter 7

Some Consequences of CP Duality

FORTHCOMING

Chapter 8

Examples of CP Duality

FORTHCOMING

Chapter 9

Polar Duality: A Geometric Analogue of Lagrangian Duality

FORTHCOMING

Chapter 10

Consequences of Polar Duality

FORTHCOMING

Chapter 11

A Unified view of Duality: Sets, Functions, and Optimization Problems

FORTHCOMING

Appendix A

A Formal Proof of Strong LP Duality

FORTHCOMING