10
Duality and Algorithms

To this point, we have considered optimization over simple sets – sets over which it is easy to minimize a linear objective or to compute a Euclidean projection. The methods we described have strong theory and often good performance, but in many cases, they do not extend well to cases in which the feasible set has more complicated structure – for example, when it is defined as the intersection of several sets or implicitly via algebraic equalities or inequalities. In this chapter, we explore the use of *duality* to obtain a different class of optimization methods that may perform better in such cases. For any constrained optimization problem, duality defines an associated concave maximization problem – the *dual problem* – whose solutions lower-bound the optimal value of the original problem. In fact, under mild assumptions, we can solve the original problem (also referred to as the *primal* problem in this context) by first solving the dual problem. While there is a vast literature on general techniques for constrained optimization, we highlight a few methods that exploit duality and build on the algorithms studied in earlier chapters.

We begin by discussing how duality arises in problems in which the feasible set Ω is the intersection of a hyperplane and a closed convex set $\mathcal{X}$. We introduce the *Lagrangian* function and discuss optimality conditions for constrained problems of this form. We then present two methods based on the Lagrangian function for problems of this type. Finally, we mention several interesting problems to which these algorithms are particularly well suited.

10.1 Quadratic Penalty Function

Consider the following formulation for an optimization problem with both a set inclusion constraint $x \in \mathcal{X}$ and a linear equality constraint $Ax = b$:

$$\min_x \ f(x) \quad \text{subject to } Ax = b, x \in \mathcal{X}. \tag{10.1}$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

Here, $\mathcal{X}$ is a closed convex set, $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable, and $A \in \mathbb{R}^{m\times n}$ has full row rank m (thus, $m \leq n$). We described in Chapter 7 first-order methods for the case in which only the set inclusion constraint is present. The addition of an equality constraint complicates matters.

One approach to dealing with the equality constraint is to move it into the objective function via a *penalty*. That is, we add a positive term to the objective when the constraint is violated, with larger penalties being incurred for larger violations. One simple type of penalty is a *quadratic penalty*, which leads to the following approximation to (10.1):

$$\min_{x\in\mathcal{X}} f(x) + \frac{1}{2\alpha}\|Ax - b\|^2, \tag{10.2}$$

where $\alpha > 0$ is the *penalty parameter*. As α tends to zero, the penalty for violating the constraint $Ax = b$ become more severe, so the solution of (10.2) will more nearly satisfy this constraint.

An intuitive approach to solving (10.1) would be to solve (10.2) with a large value of α to yield a minimizer $x^*(\alpha)$. Then decrease the value of α (by a factor of 2 or 5, say) and solve (10.2) again, "warm-starting" from the solution obtained at the previous value of α. Generally, we have that $Ax^*(\alpha) - b \to 0$ as $\alpha \downarrow 0$. In the limit, as $\alpha \downarrow 0$, we hope that $x(\alpha)$ approaches the solution of (10.1).

We can make the relationship between (10.2) and (10.1) more crisp by considering the following penalized min-max problem (also known as a saddle point problem)

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^m} f(x) - \lambda^T(Ax - b) - \frac{\alpha}{2}\|\lambda\|^2. \tag{10.3}$$

To see that this problem is equivalent to (10.2), note that we can carry out the maximization with respect to λ explicitly, because the function is strongly concave in λ with a simple Hessian. The optimal value is $\lambda = -(Ax - b)/\alpha$. By substituting this value into (10.3), we obtain (10.2).

Note too that (10.3) is well defined even for $\alpha = 0$. In this case, we have

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^m} f(x) - \lambda^T(Ax - b), \tag{10.4}$$

and this problem is *equivalent* to (10.1). To see this, note that if $Ax \neq b$, then the maximization with respect to λ is infinite. On the other hand, if $Ax = b$, then $f(x) - \lambda^T(Ax - b) = f(x)$ for all values of λ, so inner maximization with respect to λ in (10.4) yields $f(x)$ in this case. Hence, the outer minimization in (10.4) considers only points in $\mathcal{X}$ with $Ax = b$, and it minimizes f over the set of such points. The problem (10.4) is the starting point for our discussion of duality.

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

10.2 Lagrangians and Duality

The function $\mathcal{L}\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined by

$$\mathcal{L}(x,\lambda) := f(x) - \lambda^T(Ax - b) \tag{10.5}$$

is called the *Lagrangian function* (often abbreviated to simply *Lagrangian*) associated with the constrained optimization problem (10.1). This function appears frequently in theory and algorithms for constrained optimization, both convex and nonconvex. The vector λ is known as a *Lagrange multiplier*, specifically, the Lagrange multiplier associated with the constraint $Ax = b$. As we saw in (10.4), the problem

$$\min_{x\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n} \mathcal{L}(x,\lambda) \tag{10.6}$$

is equivalent to (10.1). When we switch the order of the minimization and maximization, we obtain the following *dual problem* associated with (10.1):

$$\max_{\lambda\in\mathbb{R}^n} q(\lambda), \quad \text{where } q(\lambda) := \min_{x\in\mathcal{X}} \mathcal{L}(x,\lambda) \tag{10.7}$$

In discussing duality, we often refer to the original formulation (10.1) as the *primal problem*.

Note that the function $q(\lambda)$ defined in (10.7) is always a concave function, as can be proved from first principles. Thus, the dual problem is a concave maximization problem, regardless of whether f is a convex function and whether $\mathcal{X}$ is a convex set. We now show that the solution of (10.7) always lower-bounds the optimal objective of the primal problem (10.1).

Proposition 10.1 *For any function $\varphi(x,z)$, we have*

$$\min_x \max_z \varphi(x,z) \geq \max_z \min_x \varphi(x,z). \tag{10.8}$$

Proof The proof is essentially tautological. Note that we always have

$$\varphi(x,z) \geq \min_x \varphi(x,z).$$

By taking the maximization with respect to the second argument, we obtain

$$\max_z \varphi(x,z) \geq \max_z \min_x \varphi(x,z) \quad \text{for all } x.$$

Minimizing the left-hand side of this expression with respect to x yields our assertion (10.8). □

When applied to (10.6) and (10.7), Proposition 10.1 yields a result known as *weak duality*: The maximum value of q gives a lower bound on the optimal objective value from (10.1). (The gap between the these two values is known

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

as a *duality gap*.) This result would be especially useful if the inequality (10.8) were to be replaced by an equality – that is, the duality gap is zero. In this case, knowledge of the dual maximum value would tell us the *optimal* value of the primal (10.1), so we would know when to terminate an algorithm for solving the latter problem. However, the inequality in (10.8) can be strict, as the following example shows.

Example 10.2 (Bertsekas et al., 2003, p. 203) For $x \in \mathbb{R}^2$ and $z \in \mathbb{R}$, define

$$\varphi(x,z) := \exp(-\sqrt{x_1 x_2}) + z x_1 + I_X(x) + I_Z(z),$$

where I_X and I_Z are the indicator functions for the sets X and Z (respectively) defined by $X = \{x \in \mathbb{R}^2 \mid x \geq 0\}$ and $Z = \{z \in \mathbb{R} \mid z \geq 0\}$. We have that

$$1 = \min_x \max_z \varphi(x,z) > \max_z \min_x \varphi(x,z) = 0. \tag{10.9}$$

We will see in the sequel that if the minimization problem is convex, the primal and dual problems usually attain *equal optimal values* (that is, the duality gap is zero), and we are able to reconstruct minimizers of the primal problem from the solution of the dual problem. Even in the convex case, though, there are exceptions: The inequality can still be strict.

Example 10.3 (Todd, 2001) In semidefinite programming, we work with matrix variables that are required to be symmetric positive semidefinite. We also work with an inner product operation $\langle \cdot, \cdot \rangle$ defined on two $n \times n$ symmetric matrices X and Y as follows: $\langle X, Y \rangle = \sum_{i=1}^n \sum_{j=1}^n X_{ij} Y_{ij}$. Consider the following Lagrangian for a semidefinite program:

$$\varphi(X,\lambda) = \langle C, X \rangle - \lambda_1(\langle A_1, X \rangle - b_1) - \lambda_2(\langle A_2, X \rangle - b_2) + I_{X \succeq 0}, \tag{10.10}$$

where

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{12} & X_{22} & X_{23} \\ X_{13} & X_{23} & X_{33} \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

and $b_1 = 0$, $b_2 = 2$, where $X \in \mathbb{R}^{3\times 3}$ and $\lambda \in \mathbb{R}^2$. The last term in (10.10) is an indicator function for the positive semidefinite cone; that is, it is zero when X is positive semidefinite and ∞ otherwise. By substituting the definitions of C, A_1, etc., into (10.10), we obtain

$$\varphi(X,\lambda) = X_{33} - \lambda_1 X_{11} - \lambda_2(2X_{12} + 2X_{33} - 2) + I_{X \succeq 0}. \tag{10.11}$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

In considering $\max_\lambda \varphi(X,\lambda)$, we note that this value will be infinite if the coefficients of λ_1 or λ_2 are nonero. (If, for example $X_{11} < 0$, we can drive λ_1 to $+\infty$ to make $\max_\lambda \varphi(X,\lambda) = \infty$.) Thus, in seeking (X,λ) that achieve finite values of $\varphi(X,\lambda)$, we need only consider X for which $X_{11} = 0$ and $X_{12} + X_{33} = 1$, and also X positive semidefinite. These conditions on X are satisfied only when $X_{11} = X_{12} = X_{13} = 0$ and $X_{33} = 1$. Thus, we have that $\min_X \max_\lambda \varphi(X,\lambda) = 1$.

In considering $\max_\lambda \min_X \varphi(X,\lambda)$, we rewrite (10.11) as

$$\varphi(X,\lambda) = \langle X, S\rangle + I_{X\succeq 0}, \quad \text{where } S = \begin{bmatrix} -\lambda_1 & -\lambda_2 & 0 \\ -\lambda_2 & 0 & 0 \\ 0 & 0 & 1-2\lambda_2 \end{bmatrix}.$$

If S were to have a negative eigenvalue μ with corresponding eigenvector v, we have $\langle vv^T, S\rangle = \mu\|v\|^2$, so by setting $X = \beta vv^T$ for $\beta > 0$, we have that $\varphi(\beta vv^T,\lambda) = \mu\beta\|v\|^2 \downarrow -\infty$ as $\beta \uparrow \infty$. Thus, the maximum with respect to λ of $\min_X \varphi(X,\lambda)$ cannot be attained by any λ for which S has a negative eigenvalue. We therefore have

$$S = \begin{bmatrix} -\lambda_1 & -\lambda_2 & 0 \\ -\lambda_2 & 0 & 0 \\ 0 & 0 & 1-2\lambda_2 \end{bmatrix} \succeq 0,$$

which is satisfied only when $\lambda_2 = 0$ and $\lambda_1 \le 0$, for which values we have

$$\varphi(X,\lambda) = X \bullet S + I_{X\succeq 0} = -\lambda_1 X_{11} + X_{33} + I_{X\succeq 0}.$$

The minimum over X is achieved at $X = 0$, so we have $\max_\lambda \min_X \varphi(X,\lambda) = 0$.

In conclusion, we have

$$1 = \min_X \max_\lambda \varphi(X,\lambda) > \max_\lambda \min_X \varphi(X,\lambda) = 0,$$

so for this choice of φ, the inequality in Proposition 10.1 is strict.

In the next section, we identify conditions under which (10.8) holds with *equality*, when φ obtained from the constrained optimization problem (10.1).

10.3 First-Order Optimality Conditions

In this section, we describe algebraic and geometric conditions that are satisfied by the solutions of constrained optimization problems of the form (10.1). Such problems admit "checkable" conditions that allow us to recognize solutions as being solutions and allow practical algorithms to be constructed.

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

These conditions are related to stationary points of the Lagrangian (10.5). In the next section, we describe algorithms that seek points at which these optimality conditions are satisfied.

We will build on fundamental first-order optimality conditions, like the one proved in Theorem 7.2 for the problem $\min_{x\in\Omega} f(x)$, for the case of Ω closed and convex – namely, that $-\nabla f(x^*) \in N_\Omega(x^*)$. The normal cone has a particular structure for the case in which $\Omega = \mathcal{X} \cap \{x \mid Ax = b\}$ (as in (10.1)), which, when characterized, yields the optimality conditions. This characterization is described in the following result, which uses the definition of the relative interior of a set C (denoted by ri (C)) from (A.3).

Theorem 10.4 *Suppose that $\mathcal{X} \in \mathbb{R}^n$ is a closed convex set and that $\mathcal{A} := \{x \mid Ax = b\}$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, and define $\Omega := \mathcal{X} \cap \mathcal{A}$. Then for any $x \in \Omega$, we have*

$$N_\Omega(x) \supset N_{\mathcal{X}}(x) + \{A^T\lambda \mid \lambda \in \mathbb{R}^m\}. \tag{10.12}$$

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

$$N_\Omega(x) = N_{\mathcal{X}}(x) + \{A^T\lambda \mid \lambda \in \mathbb{R}^m\}. \tag{10.13}$$

This result is proved in the Appendix (see Theorem A.18). We demonstrate the need for the assumption ri $(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$ for the "$\subset$" inclusion in (10.13) with an example. Consider

$$\mathcal{X} = \{x \in \mathbb{R}^2 \mid \|x\|_2 \leq 1\}, \quad A = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad b = [1],$$

for which ri $(\mathcal{X}) \cap \mathcal{A} = \emptyset$ and $\Omega = \mathcal{X} \cap \mathcal{A} = (0,1)^T$. We have that $N_\Omega((0,1)^T) = \mathbb{R}^2$, whereas

$$N_{\mathcal{X}}((0,1)^T) + \{A^T\lambda \mid \lambda \in \mathbb{R}\} = \{(0,\tau)^T \mid \tau \in \mathbb{R}\},$$

so the left-hand set in (10.13) is a superset of the right-hand set.

The condition ri $(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$ is an example of a *constraint qualification*. These conditions appear often in the theory of constrained optimization, particularly in the definition of optimality conditions. Broadly speaking, constraint qualifications are conditions under which the local geometry of a set – in particular, its normal cone at a point – is captured accurately by some alternative representation, usually more convenient and more "arithmetic" than geometric. In the case of the set Ω defined before, the representation of the normal cone on the right-hand side of (10.13) can be much easier to use when determining membership of $N_\Omega(x)$ than when directly checking this condition.

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

Using Theorem 10.4, we can now write the first-order optimality conditions for (10.1) as follows.

Theorem 10.5 *Consider the problem* (10.1) *in which f is continuously differentiable and $\mathcal{X}$ is a closed convex set, with* $\text{ri}\,(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$, *where* $\mathcal{A} = \{x \mid Ax = b\}$. *If x^* is a local solution of* (10.1), *then there exists $\lambda^* \in \mathbb{R}^m$ such that*

$$x^* \in \Omega = \mathcal{X} \cap \mathcal{A}, \quad -\nabla f(x^*) + A^T\lambda^* \in N_{\mathcal{X}}(x^*). \tag{10.14}$$

Proof The proof follows immediately by combining Theorems 7.2 and 10.4, noting that Ω is a closed convex set. □

We next show that a converse of this result holds when we assume additionally that f is convex. Note that the assumption $\text{ri}\,(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$ is not needed for this result.

Theorem 10.6 *Consider the problem* (10.1) *in which f is continuously differentiable and convex, and $\mathcal{X}$ is a closed convex set. If there exists $\lambda^* \in \mathbb{R}^m$ such that the conditions* (10.14) *are satisfied at some x^*, then x^* is a solution of* (10.1).

Proof Note from the first part of Theorem 10.4 that (10.14) implies that $-\nabla f(x^*) \in N_\Omega(x^*)$. The second part of Theorem 7.2 can then be applied to obtain the result. □

The following is an immediate corollary of the last two results, which applies to constrained convex optimization problems of the form (10.1).

Corollary 10.7 *Consider the problem* (10.1) *in which f is continuously differentiable and convex, and $\mathcal{X}$ is a closed convex set, with* $\text{ri}\,(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$, *where* $\mathcal{A} = \{x \mid Ax = b\}$. *Then the conditions* (10.14) *are necessary and sufficient for x^* to be a solution of* (10.1).

Example 10.8 Consider the problem

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^n x_i \quad \text{subject to } \|x\|_2 \leq 1, x_1 = 1/2, x_2 = 1/2, \tag{10.15}$$

for some $n \geq 3$. Note that we can eliminate the variables x_1 and x_2, and write the problem equivalently as

$$\min_{x_3, x_4, \ldots, x_n} \sum_{i=3}^n x_i \quad \text{subject to} \quad \sqrt{\sum_{i=3}^n x_i^2} \leq \frac{1}{\sqrt{2}}. \tag{10.16}$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

By using Theorem 7.2, we can check that the point

$$(x_3, x_4, \ldots, x_n)^T = \frac{-1}{\sqrt{2(n-2)}}(1, 1, \ldots, 1)^T \tag{10.17}$$

is the global solution of (10.16). It follows that the solution of (10.15) is

$$x^* = \left(\frac{1}{2}, \frac{1}{2}, \frac{-1}{\sqrt{2(n-2)}}, \frac{-1}{\sqrt{2(n-2)}}, \ldots, \frac{-1}{\sqrt{2(n-2)}}\right). \tag{10.18}$$

We can use Corollary 10.7 to verify optimality of this point directly, by noting that (10.15) has the form of (10.1) with $\mathcal{X} = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ and

$$A = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}.$$

Note that the condition $\text{ri}\,(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$ is satisfied, because $\text{ri}\,(\mathcal{X}) = \{x \in \mathbb{R}^n \mid \|x\|_2 < 1\}$, and we have, for example, that $(1/2, 1/2, 0, 0, \ldots, 0)^T \in \text{ri}\,(\mathcal{X}) \cap \mathcal{A}$. For any x with $\|x\| = 1$, we have that $N_{\mathcal{X}}(x) = \alpha x$ for any $\alpha \geq 0$. Thus, the optimality condition (10.14) at x^* defined by (10.18) is

$$-\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \alpha \begin{bmatrix} 1/2 \\ 1/2 \\ \frac{-1}{\sqrt{2(n-2)}} \\ \vdots \frac{-1}{\sqrt{2(n-2)}} \end{bmatrix},$$

for some $\alpha \geq 0$, $\lambda_1 \in \mathbb{R}$, and $\lambda_2 \in \mathbb{R}$. It is easy to check that this equality holds when we set

$$\alpha = \sqrt{2(n-2)}, \quad \lambda_1 = \lambda_2 = 1 + \sqrt{\frac{n-2}{2}}.$$

Example 10.9 Consider the following problem, which has a combination of nonnegativity bound constraints and equality constraints:

$$\min\ f(x) \quad \text{subject to } Ax = b,\ x \geq 0.$$

By defining $\mathcal{X} := \{x \mid x \geq 0\}$, we have for any $x \in \mathcal{X}$ that

$$N_{\mathcal{X}}(x) = \{v \mid v_i \in (-\infty, 0] \text{ if } x_i = 0,\ \ v_i = 0 \text{ if } x_i > 0\}.$$

Thus, the first-order optimality condition (10.14) becomes that there exists $\lambda^* \in \mathbb{R}^m$ such that $Ax^* = b$, $x^* \geq 0$, and

$$\left[-\nabla f(x^*) + A^T\lambda^*\right]_i \leq 0 \text{ when } x_i^* = 0, \quad \left[-\nabla f(x^*) + A^T\lambda^*\right]_i = 0 \text{ when } x_i^* > 0.$$

Note that since $\text{ri}\,(\mathcal{X}) = \{x \mid x > 0\}$, the constraint qualification $\text{ri}\,(\mathcal{X}) \cap \mathcal{A}$ requires existence of and x with $x > 0$ (all positive components) for which

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

$Ax = b$. In fact, it can be shown that in this particular case, the characterization (10.13) holds even when this condition does not hold, because all the constraints (equalities and inequalities) are linear functions of x.

10.4 Strong Duality

Having characterized optimality conditions, we now return to proving that the primal problem (10.1) and the dual problem (10.7) attain the same optimal objective values for many convex optimization problems. The following theorem also shows that if we know a solution to the dual problem, we can extract a solution to the primal via a simpler optimization problem.

Theorem 10.10 (Strong Duality) *Suppose that f in* (10.1) *is continuously differentiable and convex, that $\mathcal{X}$ is closed and convex, and that the condition* $\text{ri}\,(\mathcal{X}) \cap \mathcal{A} \neq \emptyset$ *holds, where $\mathcal{A} = \{x \mid Ax = b\}$. We then have the following.*

1. *If* (10.1) *has a solution x^*, then the dual problem* (10.7) *also has an optimal solution λ^*, and the primal and dual optimal objective values are equal.*
2. *For x^* to be optimal for the primal and λ^* optimal for the dual, it is necessary and sufficient that $Ax^* = b$, $x^* \in \mathcal{X}$, and*

$$x^* \in \arg\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^*) = f(x) - (\lambda^*)^T (Ax - b).$$

Proof For all $\lambda \in \mathbb{R}^n$ and all x feasible for (10.1), we have, from (10.7), that

$$q(\lambda) \leq f(x) - \lambda^T (Ax - b) = f(x),$$

where the equality holds because $Ax = b$. By Corollary 10.7, $x^* \in \Omega$ is optimal if and only if there exists a $\lambda^* \in \mathbb{R}^m$ such that

$$(\nabla f(x^*) - A^T \lambda^*)^T (x - x^*) \geq 0, \quad \text{for all } x \in \mathcal{X}. \tag{10.19}$$

But since $\mathcal{L}(\cdot, \lambda^*)$ is convex as a function of its first argument, and $\nabla_x \mathcal{L}(x, \lambda^*) = \nabla f(x) - A^T \lambda^*$, condition (10.19) shows that x^* minimizes $\mathcal{L}(x, \lambda^*)$ over $x \in \mathcal{X}$. It now follows that

$$q(\lambda^*) = \inf_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^*) = \mathcal{L}(x^*, \lambda^*) = f(x^*) - (\lambda^*)^T (Ax^* - b) = f(x^*),$$

completing the proof of Part 1. The proof of Part 2 is left as an Exercise. □

Note that even if λ is only *approximately* dual optimal, minimizing the Lagrangian with respect to x gives a reasonable approximation to the original optimization problem. This claim follows from the calculation

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

$$f(x^*) = q(\lambda^*) \le q(\lambda) + \epsilon = \inf_{x \in \Omega} \mathcal{L}(x, \lambda) + \epsilon \le \mathcal{L}(x, \lambda) + \epsilon$$
$$= f(x) - \lambda^T (Ax - b) + \epsilon.$$

Hence, if $\|Ax - b\|$ is small and our dual optimal value λ is accurate to within an objective margin of ϵ, then $f(x)$ is a reasonable approximation to the optimal function value $f(x^*) = q(\lambda^*)$.

10.5 Dual Algorithms

Though the dual objective function q is concave, it is typically nonsmooth, so minimization may not be a straightforward operation. In this section, we review how the algorithms derived earlier for nonsmooth optimization can be leveraged to solve dual problems.

10.5.1 Dual Subgradient

Since the concave dual objective q defined by (10.7) is a minimum of linear functions parametrized by the primal variable x, we can compute a subgradient by finding the minimizing x and then applying Danskin's theorem (Theorem 8.13). Since $-q$ is a convex function, we have

$$\partial(-q)(\lambda) := \left\{ Az - b \mid z \in \arg\min_{x \in \mathcal{X}} \{ f(x) - \lambda^T (Ax - b) \} \right\}.$$

Starting from some initial guess λ^1 of the optimum, step k of the subgradient method of Section 9.2 applied to $-q$ thus has the form

$$x^k \leftarrow \arg\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^k), \quad \lambda^{k+1} \leftarrow \lambda^k - s_k (Ax^k - b),$$

where $s_k \in \mathbb{R}_+$ is a steplength. To analyze this method, note that the maximum norm of any subgradient of $-q$ is bounded by the maximal infeasibility of the equality constraints over the set $\mathcal{X}$. If we set

$$M = \sup_{x \in \mathcal{X}} \|Ax - b\|,$$

we can apply our analysis of the subgradient method from Section 9.2 to obtain

$$q\left(\frac{1}{\sum_{k=1}^T s_k} \sum_{k=1}^T s_k \lambda^k \right) - q^* \ge - \frac{\|\lambda^1 - \lambda^*\|^2 + M^2 \sum_{k=1}^T s_k^2}{2 \sum_{k=1}^T s_k}.$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

Hence, a convergence rate of $O(T^{-1/2})$ is attainable, for the choices of steplength s_k discussed in Section 9.2. We can achieve a faster rate of convergence by appealing to the proximal point method rather than the subgradient method, as we show next.

10.5.2 Augmented Lagrangian Method

Application of the proximal point method of Section 9.5 to the problem of maximizing the dual objective $q(\lambda)$ leads to the following iteration:

$$\begin{aligned}\lambda^{k+1} &\leftarrow \arg\max_{\lambda}\ q(\lambda) - \frac{1}{2\alpha_k}\|\lambda - \lambda^k\|^2 \\ &= \arg\max_{\lambda}\ \inf_{x\in\mathcal{X}} \left\{ f(x) - \lambda^T(Ax-b) - \frac{1}{2\alpha_k}\|\lambda-\lambda^k\|^2 \right\},\end{aligned}$$

where α_k is the proximality parameter. This is a *saddle point problem* in (x,λ). Since the objective is convex in x and strongly convex in λ, we can swap the infimum and supremum by Sion's minimax theorem (Sion, 1958) to obtain the equivalent problem

$$\inf_{x\in\mathcal{X}} \left\{ \max_{\lambda}\ f(x) - \lambda^T(Ax-b) - \frac{1}{2\alpha_k}\|\lambda-\lambda^k\|^2 \right\}. \tag{10.20}$$

The inner problem is quadratic in λ and has the trivial solution $\lambda = \lambda^k - \alpha_k(Ax-b)$, which we can substitute into (10.20), to obtain

$$\min_{x\in\mathcal{X}}\ f(x) - (\lambda^k)^T(Ax-b) + \frac{\alpha_k}{2}\|Ax-b\|^2 =: \mathcal{L}_{\alpha_k}(x,\lambda^k).$$

The function $\mathcal{L}_\alpha(x,\lambda)$ is called the *augmented Lagrangian*. It consists of the ordinary Lagrangian function added to a quadratic penalty term that penalizes violation of the equality constraint $Ax = b$. Iteration k of this overall approach can be summarized as follows:

$$x^k \leftarrow \arg\min_{x\in\mathcal{X}} \mathcal{L}_{\alpha_k}(x,\lambda^k), \quad \lambda^{k+1} \leftarrow \lambda^k - \alpha_k(Ax^k - b)\,.$$

This algorithm was historically referred to as the *method of multipliers* in the optimization literature but more recently has been known as the *augmented Lagrangian method*.

For a fixed parameter α_k (that is, $\alpha_k \equiv \alpha$), we have, from the convergence rate of the proximal point method (Theorem 9.8), that

$$q^* - q(\lambda^T) \le \frac{\|\lambda^* - \lambda^1\|^2}{2\alpha T}, \quad T = 1,2,\ldots;$$

that is, the dual objective converges at a rate of $O(1/T)$.

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

The only difference between the augmented Lagrangian approach and the dual subgradient method is that we have to minimize the *augmented* Lagrangian for the x-step instead of the original Lagrangian. This may add algorithmic difficulty, but in many cases, it does not; the augmented Lagrangian can be as inexpensive to minimize as its non-augmented counterpart. We give several examples in what follows.

Although the proximal point method is guaranteed to converge even for a constant step size α_k, the use of some heuristics frequently improves its practical performance. In particular, the following approach is suggested by Conn et al. (1992).

Algorithm 10.1 Augmented Lagrangian

Choose initial point λ^1, initial parameter $\alpha_1 > 0$, $\delta_1 = \infty$, and parameters $\eta \in (0,1)$ and $\gamma > 1$;
for $k = 1, 2, \ldots$ **do**
  Set $x^k = \arg\min_{x \in \mathcal{X}} \mathcal{L}_{\alpha_k}(x, \lambda^k)$;
  Set $\delta = \|Ax^k - b\|^2$;
  if $\delta < \eta\delta_k$ **then**
    $\lambda^{k+1} \leftarrow \lambda^k - \alpha_k(Ax^k - b)$; $\alpha_{k+1} \leftarrow \alpha_k$; $\delta_{k+1} \leftarrow \delta$; {Improvement in feasibility of x is acceptable; take step in λ.}
  else
    $\lambda^{k+1} \leftarrow \lambda^k$; $\alpha_{k+1} \leftarrow \gamma\alpha_k$; $\delta_{k+1} \leftarrow \delta_k$; {Insufficient improvement in feasibilty; don't update λ but increase penalty parameter α for next iteration.}
  end if
end for

Typical values of the parameters are $\eta = 1/4$ and $\gamma = 10$.

10.5.3 Alternating Direction Method of Multipliers

The alternating direction method of multipliers (ADMM) is a powerful extension of the method of multipliers that is well suited to a variety of interesting problems in data analysis and elsewhere. ADMM is targeted to problems of the form

$$\min_{x,z} \; f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = c, \;\; x \in \mathcal{X}, \; z \in \mathcal{Z}, \tag{10.21}$$

where $\mathcal{X}$ and $\mathcal{Z}$ are closed convex sets. The augmented Lagrangian for this problem is

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

$$\mathcal{L}_\alpha(x,z,\lambda) = f(x) + g(z) - \lambda^T(Ax + Bz - c) + \frac{\alpha}{2}\|Ax + Bz - c\|^2.$$

ADMM essentially performs one step of block coordinate descent on the primal problem and then updates the Lagrange multiplier, as follows:

$$x^k = \arg\min_{x\in\mathcal{X}} \mathcal{L}_{\alpha_k}(x, z^{k-1}, \lambda^k) \tag{10.22a}$$

$$z^k = \arg\min_{z\in\mathcal{Z}} \mathcal{L}_{\alpha_k}(x^k, z, \lambda^k) \tag{10.22b}$$

$$\lambda^{k+1} = \lambda^k - \alpha_k(Ax^k + Bz^k - c). \tag{10.22c}$$

Note that if we were to loop on the first two update steps until $\mathcal{L}_{\alpha_k}(x,z,\lambda^k)$ were minimized with respect to the primal variables (x,z), this approach would become a particular implementation of the ordinary method of multipliers. But the fact that only one round of block coordinate descent steps is taken before updating λ is what distinguishes ADMM. In practice, taking multiple coordinate descent steps may be advantageous in some contexts, but traditional convergence proofs for ADMM have an "operator splitting" character that does not exploit their relationship to the augmented Lagrangian method. The paper of Eckstein and Yao (2015) explores this point and also gives computational comparisons of ADMM with variants that more closely approximate the augmented Lagrangian method. A proof of convergence of ADMM (10.22) for the case of convex f and g is given in (Boyd et al., 2011, section 3.2 and appendix A).

10.6 Some Applications of Dual Algorithms

Here we describe several applications for which the duality-based methods of this chapter may be a good fit.

10.6.1 Consensus Optimization

Let $G = (V,E)$ be a graph with vertex set V and edge set E. Consider the following optimization problem in unknowns $[x_v]_{v\in V}$, where each $x_v \in \mathbb{R}^{n_v}$ and the functions $f_v\colon \mathbb{R}^{n_v} \to \mathbb{R}$ are convex:

$$\min_x \sum_{v\in V} f_v(x_v) \quad \text{subject to } x_u = x_v \text{ for all } (u,v) \in E. \tag{10.23}$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

The Lagrangian for this problem is

$$\begin{aligned}\mathcal{L}(x,\lambda) &= \sum_{v\in V} f_v(x_v) - \sum_{(u,v)\in E} \lambda_{u,v}^T (x_u - x_v) \\ &= \sum_{v\in V} \left\{ f_v(x_v) - \left(\sum_{(v,w)\in E} \lambda_{v,w} - \sum_{(u,v)\in E} \lambda_{u,v} \right)^T x_v \right\}.\end{aligned}$$

Note that this function is separable in the components of x, so we can minimize with respect to each x_v, $v \in V$ separately, even in a distributed fashion. The λ-step of the dual subgradient method is

$$\lambda_{u,v}^{k+1} = \lambda_{u,v}^k - s_k(x_u^k - x_v^k), \quad \text{for all } (u,v) \in E.$$

Many problems can be stated in the form (10.23). For instance, the case in which we wish to minimizing a finite-sum objective with a shared variable:

$$\min_x \sum_{i=1}^m f_i(x), \tag{10.24}$$

(where some of the f_i may even be indicator functions for convex sets) can be stated in the form (10.23) by defining $V := \{1, 2, \ldots, m\}$, giving each node its own version of the variable x and defining an edge set E so that the graph $G = (V, E)$ is completely connected.

The augmented Lagrangian for (10.23) does not yield a problem that is separable in the x_v, because the quadratic penalty term couples the x_v at different nodes. We can, however, devise an equivalent formulation that enables a convenient splitting with ADMM. Introducing new "edge variables" $z_{u,v}$ for all $(u,v) \in E$, we rewrite (10.23) as follows:

$$\min \sum_{v\in V} f_v(x_v) \quad \text{subject to } x_u = z_{u,v},\ x_v = z_{u,v}, \ \text{ for all } (u,v) \in E. \tag{10.25}$$

The augmented Lagrangian for this formulation is

$$\begin{aligned}\mathcal{L}_\alpha(x,z,\lambda,\beta) = \sum_{v\in V} f_v(x_v) &- \sum_{(u,v)\in E} \lambda_{u,v}^T (x_u - z_{u,v}) - \sum_{(u,v)\in E} \beta_{u,v}^T (x_v - z_{u,v}) \\ &+ \sum_{(u,v)\in E} \frac{\alpha}{2} (x_u - z_{u,v})^2 + \sum_{(u,v)\in E} \frac{\alpha}{2} (x_v - z_{u,v})^2 .\end{aligned}$$

This function is separable in the components x_v, $v \in V$, so the x-update step in ADMM can be performed in a separated manner, possibly on a distributed computational platform. Similarly, it is separable in the $z_{u,v}$ variables, and

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

also in the dual variables $\lambda_{u,v}$ and $\beta_{u,v}$. Note that distributed implementations would require information to be passed between nodes, or to a central server, between updates of the various components.

A particular method for the finite-sum problem (10.24) is to allow each function f_i to have its own variable x_i and then define a "master variable" x and constraints that ensure that all x_i are identical to x. We thus obtain the following formulation, equivalent to (10.24):

$$\min_{x,x_1,x_2,\ldots,x_m} \sum_{i=1}^{m} f_i(x_i) \quad \text{subject to } x_i = x,\ i = 1,2,\ldots,m. \tag{10.26}$$

The augmented Lagrangian for this problem is

$$\mathcal{L}_\alpha(x,z,\lambda) = \sum_{i=1}^{m} f_i(x_i) - \sum_{i=1}^{m} \lambda_i^T(x_i - x) + \frac{\alpha}{2}\sum_{i=1}^{m} \|x_i - x\|^2,$$

where we defined $z := (x_1,x_2,\ldots,x_m)$. The z-update step in ADMM is separable in the replicates x_i, $i = 1,2,\ldots,m$; the step (10.22b) can be performed as m separate optimization problems of the form

$$x_i^k = \arg\min_{x_i} f_i(x_i) - (\lambda_i^k)^T x_i + \frac{\alpha}{2}\|x_i - x^k\|^2.$$

The x-update step (10.22a) can be performed explicitly, since the augmented Lagrangian is a simple convex quadratic in x. We have

$$x^k = \frac{1}{m}\sum_{i=1}^{m}\left(x_i^{k-1} - \frac{1}{\alpha_k}\lambda_i^k\right).$$

This example illustrates the flexibility that is possible with dual algorithms. Different problem formulations play to the strengths of different algorithms, and sometimes the algorithms with better worst-case complexities are not the most appropriate, due to issues surrounding overhead and communication in distributed implementations.

10.6.2 Utility Maximization

The general utility maximization problem is

$$\max \sum_{i=1}^{n} U_i(x_i) \quad \text{subject to } Rx \le c,$$

where R is a $p \times n$ matrix. Each utility function U_i represents some measure of well-being for the ith agent as a function of the amount of resource x_i available to it. The inequalities are resource constraints, coupling the amount of utility

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

available to each user. Rewriting in our form (10.1), using minimization and slack variables s, we have

$$\min_{(x,s)} -\sum_{i=1}^{n} U_i(x_i) \quad \text{subject to} \quad -Rx + c - s = 0, \;\; s \geq 0.$$

The Lagrangian is

$$\mathcal{L}(x, s, \lambda) = \sum_{i=1}^{n} -U_i(x_i) - \lambda^T(-Rx + c - s).$$

The dual subgradient method requires us to minimize this function over (x, s) for $s \geq 0$. Note that this minimization is unbounded below if any components of λ are negative: If $\lambda_i < 0$, we can drive s_i to $+\infty$ to force $\mathcal{L}(x, s, \lambda)$ to $-\infty$. Thus, the dual problem (10.7) is equivalent to

$$\begin{aligned} &\max_{\lambda \geq 0} \min_{(x,s)\,:\,s \geq 0} \sum_{i=1}^{n} -U_i(x_i) - \lambda^T(-Rx + c - s) \\ &\quad = \max_{\lambda \geq 0} \min_{x} \sum_{i=1}^{n} -U_i(x_i) - \lambda^T(-Rx + c), \end{aligned}$$

where we can eliminate s because when $\lambda \geq 0$, the optimal value of s is clearly $s = 0$. The x-step of the dual subgradient method is separable; agent i maximizes

$$U(x_i) - \left[\sum_{j=1}^{p} R_{ji}\lambda_j^k\right] x_i.$$

The λ-update step is the projection of a subgradient step onto the nonnegative orthant defined by $\lambda \geq 0$; that is,

$$\lambda^{k+1} \leftarrow \left[\lambda^k - \alpha_k(-Rx^k + c)\right]_+ .$$

The dynamics of this model are interesting. Component j of λ can be interpreted as a *price* for the resources represented by jth row of R and c. If the prices are high, users incur a negative cost for acquiring more of their quantities x. When the resource constraints are loose, the prices go down. When they are violated, the prices go up.

10.6.3 Linear and Quadratic Programming

Consider the bound-constrained convex quadratic program,

$$\min_{x} \; c^T x + \tfrac{1}{2} x^T Q x, \quad \text{subject to} \;\; Ax = b, \;\; \ell \leq x \leq u, \tag{10.27}$$

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

where $Q \succeq 0$ and ℓ and u represent vectors of lower and upper bounds on the components of x, respectively. (Some or all components of ℓ and u may be infinite.) When $\ell = 0$, the components of u are all $+\infty$, and $Q = 0$, then (10.27) is a linear program – the fundamental problem in constrained optimization. The augmented Lagrangian for (10.27) is

$$\mathcal{L}_{\alpha_k}(x, \lambda) = c^T x + \tfrac{1}{2} x^T Q x - \lambda^T (Ax - b) + \frac{\alpha_k}{2} \|Ax - b\|^2,$$

so the x-step of the augmented Lagrangian method reduces to the following bound-constrained quadratic problem:

$$x^k = \min_{\ell \le x \le u} c^T x + \tfrac{1}{2} x^T Q x - (\lambda^k)^T (Ax - b) + \frac{\alpha_k}{2} \|Ax - b\|^2.$$

This problem can be solved via first-order methods, such as the projected gradient or conditional gradient methods of Chapter 7.

To apply ADMM to this problem, we could formulate (10.27) equivalently as

$$\min_{(x,z)} c^T x + \frac{1}{2} x^T Q x \quad \text{subject to } Ax = b, \ \ell \le z \le u, \ z = x. \tag{10.28}$$

The augmented Lagrangian for this problem is

$$\mathcal{L}_\alpha(x, z, \lambda) = c^T x + \tfrac{1}{2} x^T Q x - \lambda^T (x - z) + \frac{\alpha}{2} \|z - x\|^2,$$

where we choose to enforce the constraints $Ax = b$ and $\ell \le z \le u$ explicitly. The ADMM updates are therefore

$$x^{k+1} = \arg\min_x \mathcal{L}_{\alpha_k}(x, z^k, \lambda^k) \qquad \text{subject to } Ax = b, \tag{10.29a}$$

$$z^{k+1} = \arg\min_z \mathcal{L}_{\alpha_k}(x^{k+1}, z, \lambda^k) \quad \text{subject to } \ell \le z \le u, \tag{10.29b}$$

$$\lambda^{k+1} = \lambda^k - \alpha_k (x^{k+1} - z^{k+1}). \tag{10.29c}$$

The x update can be solved by solving an equality constrained quadratic program, which reduces to solving a system of linear equations, as follows:

$$\begin{bmatrix} Q + \alpha_k I & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \nu \end{bmatrix} = \begin{bmatrix} -c + \lambda^k + \alpha_k z^k \\ b \end{bmatrix}.$$

Note that if α_k is constant, only the right-hand side changes from iteration to iteration, so a factorization of the left-hand side can be precomputed and reused at every iteration. A closed-form solution is available for the z update (see the Exercises). This strategy for solving QPs is the main algorithmic idea behind the OSQP quadratic programming solver (Stellato et al., 2020).

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press

Notes and References

Several further examples of duality gaps (gaps between the primal and dual optimal objective values) in convex problems appear in (Luo et al., 2000; Vandenberghe and Boyd, 1996).

The method of multipliers (a.k.a. the augmented Lagrangian method) was invented in the late 1960s by Hestenes (1969) and Powell (1969). It was developed further by Rockafellar (1973, 1976a) and Bertsekas (1982) and made into the practical general software package Lancelot for nonlinear programming by Conn et al. (1992).

The alternating direction method of multipliers is described in the classic review paper of Boyd et al. (2011). The approach was first proposed in the 1970s in Glowinski and Marrocco (1975) and Gabay and Mercier (1976), while Eckstein and Bertsekas (1992) is an important early reference.

Exercises

1. Consider minimization of a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ over the polyhedral set defined by a combination of linear equalities and inequalities as follows:

$$\{x \mid Ex = g,\ Cx \geq d\},$$

where $E \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{p\times n}$. Show that the first-order necessary conditions for x^* to be a solution of this problem are that there exist vectors $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^p$ such that

$$\nabla f(x^*) - E^T\lambda - C^T\mu = 0, \quad Ex^* = g, \quad 0 \leq \mu \perp Cx^* - d \geq 0,$$

where $0 \leq u \perp v \geq 0$ for two vectors $u, v \in \mathbb{R}^p$ indicates that for all $i = 1, 2, \ldots, p$, we have $u_i \geq 0$, $v_i \geq 0$, and $u_i v_i = 0$. (Hint: Introduce slack variables $s \in \mathbb{R}^p$, and reformulate the problem equivalently as follows:

$$\min_{(x,s)\in\mathbb{R}^{n+p}} f(x) \quad \text{s.t. } Ex = g,\ Cx - s = d,\ s \geq 0.$$

Now, by defining $\mathcal{X}$, A, and b appropriately, use Theorem 10.5 to find optimality conditions for the reformulated problem; then eliminate s to obtain the aforementioned conditions.)
2. Prove the strict duality gap (10.9) for the function in Example 10.2.
3. Prove Part 2 of Theorem 10.10.
4. Verify by checking the condition $-\nabla f(x^*) \in N_\Omega(x^*)$ that the point x^* defined by (10.17) is the solution of (10.16).
5. Write down a closed-form solution for the step (10.29b).

https://doi.org/10.1017/9781009004282.011 Published online by Cambridge University Press