7

First-Order Methods for Constrained Optimization

In constrained optimization, we seek the point x^* in a specified set Ω that attains the smallest value of the objective function f in Ω. The set Ω is called the *feasible set*, and it is often defined via a number of algebraic equalities and inequalities, called *constraints*. The constraints can simply be bounds on the values of the variables, or they can be more complex formulas that capture temporal dependencies, resource usage, or statistical models. In this chapter, we focus on case in which Ω is a simple closed convex set. Later chapters consider setups in which the feasible set is more complicated.

7.1 Optimality Conditions

We consider problem (2.1), restated here as

$$\min_{x \in \Omega} f(x), \tag{7.1}$$

where $\Omega \subset \mathbb{R}^n$ is closed and convex and f is smooth (at least differentiable). We refer to earlier definitions of local and global solutions in Section 2.1 and convexity of sets and functions in Sections 2.4 and 2.5.

To characterize optimality for minimizing a smooth function f over a closed convex set Ω, we need to generalize beyond the optimality theory of Section 2.3, which was for unconstrained optimization. Typically, the unconstrained first-order conditions $\nabla f(x) = 0$ are *not* satisfied at the solution of (7.1). To define optimality conditions for this constrained problem, we need the notion of a *normal cone* to a closed convex set Ω at a point $x \in \Omega$.

Definition 7.1 Let $\Omega \subset \mathbb{R}^n$ be a closed convex set. At any $x \in \Omega$, the *normal cone* $N_\Omega(x)$ is defined as

$$N_\Omega(x) = \{d \in \mathbb{R}^n : \; d^T(y - x) \leq 0 \;\; \text{for all } y \in \Omega\}.$$

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

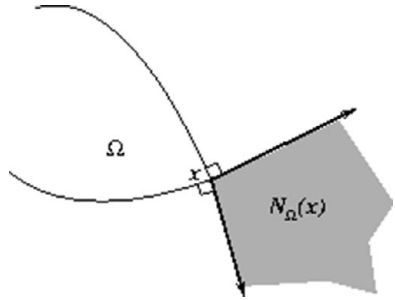

Figure 7.1 Normal Cone

(Note that $N_\Omega(x)$ satisfies trivially the definition of a *cone* $C \in \mathbb{R}^n$, which is that $z \in C \Rightarrow tz \in C$ for all $t > 0$.) See Figure 7.1 for an example of a normal cone.

The following result is a first-order necessary condition for x^* to be a solution of (7.1). When f is convex, the condition is also sufficient.

Theorem 7.2 *Consider* (7.1)*, where* $\Omega \subset \mathbb{R}^n$ *is closed and convex and* f *is continuously differentiable. If* $x^* \in \Omega$ *is a local solution of* (7.1)*, then* $-\nabla f(x^*) \in N_\Omega(x^*)$*. If* f *is also convex, then the condition* $-\nabla f(x^*) \in N_\Omega(x^*)$ *implies that* x^* *is a global solution of* (7.1)*.*

Proof Suppose that x^* is a local solution, and let z be any point in Ω. We have that $x^* + \alpha(z - x^*) \in \Omega$ for all $\alpha \in [0, 1]$, and, by Taylor's theorem (specifically (2.3)), we have

$$\begin{aligned} f(x^* + \alpha(z - x^*)) &= f(x^*) + \alpha \nabla f(x^*)^T (z - x^*) \\ &\quad + \alpha \left[\nabla f(x^* + \gamma_\alpha \alpha (z - x^*)) - \nabla f(x^*) \right]^T (z - x^*) \\ &= f(x^*) + \alpha \nabla f(x^*)^T (z - x^*) + o(\alpha) \end{aligned}$$

for some $\gamma_\alpha \in (0, 1)$. Since x^* is a local solution, we have that $f(x^* + \alpha(z - x^*)) \geq f(x^*)$ for all $\alpha > 0$ sufficiently small. By substituting this inequality into the previous expression and letting $\alpha \downarrow 0$, we have that $-\nabla f(x^*)^T (z - x^*) \leq 0$. Since the choice of $z \in \Omega$ was arbitrary, we conclude that $-\nabla f(x^*) \in N_\Omega(x^*)$, as required.

Suppose now that f is also convex, and that $-\nabla f(x^*) \in N_\Omega(x^*)$. Then $-\nabla f(x^*)^T (z - x^*) \leq 0$ for all $z \in \Omega$. By convexity of f, we have

$$f(z) \geq f(x^*) + \nabla f(x^*)^T (z - x^*) \geq f(x^*),$$

verifying that x^* minimizes f over Ω, proving the second claim. □

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

When f is *strongly* convex (see (2.19)), problem (7.1) has a unique solution.

Theorem 7.3 *Suppose that in the problem* (7.1)*, f is differentiable and strongly convex, while Ω is closed, convex, and nonempty. Then* (7.1) *has a unique solution x^*, characterized by $-\nabla f(x^*) \in N_\Omega(x^*)$.*

Proof Given any $z \in \Omega$, it follows immediately from (2.19) that f is *globally* bounded below by a quadratic function – that is,

$$f(x) \geq f(z) + \nabla f(z)^T(x - z) + \frac{m}{2}\|x - z\|^2,$$

with $m > 0$. Thus, the set $\Omega \cap \{x \mid f(x) \leq f(z)\}$ is closed and bounded, hence compact, so f attains its minimum value on this set at some point x^*, which is thus a solution of (7.1).

For uniqueness of this solution x^*, we note that, for any point $x \in \Omega$, we have, from (2.19) again, together with the property $-\nabla f(x^*) \in N_\Omega(x^*)$ from Theorem 7.2, that

$$f(x) \geq f(x^*) + \nabla f(x^*)^T(x - x^*) + \frac{m}{2}\|x - x^*\|^2 > f(x^*),$$

since $\nabla f(x^*)^T(x - x^*) \geq 0$, $m > 0$, and $x \neq x^*$. □

7.2 Euclidean Projection

Let Ω be a closed, convex set. The *Euclidean projection* of a point x onto Ω is the closest point in Ω to x, measured by the Euclidean norm (which we denote by $\|\cdot\|$). Denoting this point by $P_\Omega(x)$, we see that it solves the following constrained optimization problem:

$$P_\Omega(x) = \arg\min\{\|z - x\| \mid z \in \Omega\},$$

or, equivalently,

$$P_\Omega(x) = \arg\min_{z \in \Omega} \tfrac{1}{2}\|z - x\|_2^2. \tag{7.2}$$

Since the cost function of this problem is strongly convex, Theorem 7.3 tells us that $P_\Omega(x)$ exists and is unique, so well defined. The same theorem gives us the following characterization of $P_\Omega(x)$:

$$x - P_\Omega(x) \in N_\Omega(P_\Omega(x));$$

that is, from the Definition 7.1,

$$(x - P_\Omega(x))^T(z - P_\Omega(x)) \leq 0, \quad \text{for all } z \in \Omega. \tag{7.3}$$

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

In fact, this inequality characterizes $P_\Omega(x)$; there is no other point $\bar{x} \in \Omega$ such that $(x - \bar{x})^T(z - \bar{x}) \leq 0$ for all $z \in \Omega$, since if such a point existed, it would also be a solution of the projection subproblem.

We refer to (7.3) as a *minimum principle*. We can use it to compute a variety of projections onto simple sets Ω.

Example 7.4 (Nonnegative Orthant) Consider the set of vectors whose components are all nonnegative: $\Omega = \{x \mid x_i \geq 0,\ i = 1, 2, \ldots, n\}$. Note that Ω is a closed, convex cone. We have

$$P_\Omega(x) = \max(x, 0);$$

that is, the ith component of $P_\Omega(x)$ is x_i if $x_i \geq 0$, and 0 otherwise. We prove this claim by referring to the minimum principle (7.3). We have

$$\begin{aligned}
&(x - P_\Omega(x))^T(z - P_\Omega(x)) \\
&\quad = \sum_{x_i<0} (x_i - [P_\Omega(x)]_i)(z_i - [P_\Omega(x)]_i) + \sum_{x_i \geq 0} (x_i - [P_\Omega(x)]_i)(z_i - [P_\Omega(x)]_i) \\
&\quad = \sum_{x_i<0} x_i z_i \leq 0,
\end{aligned}$$

since $z_i \geq 0$ for all i.

Example 7.5 (Unit Norm Ball) Defining $\Omega = \{x \mid \|x\| \leq 1\}$, we have

$$P_\Omega(x) = \begin{cases} x & \text{if } \|x\| \leq 1, \\ x/\|x\| & \text{otherwise.} \end{cases}$$

We leave the proof as an Exercise.

The following result is an immediate consequence of (7.3).

Lemma 7.6 *Let Ω be closed and convex. Then $(P_\Omega(y) - z)^T(y - z) \geq 0$ for all $z \in \Omega$, with equality if and only if $z = P_\Omega(y)$.*

Proof

$$\begin{aligned}
(P_\Omega(y) - z)^T(y - z) &= (P_\Omega(y) - z)^T(y - P_\Omega(y) + P_\Omega(y) - z) \\
&= (P_\Omega(y) - z)^T(y - P_\Omega(y)) + \|P_\Omega(y) - z\|^2 \\
&\geq (P_\Omega(y) - z)^T(y - P_\Omega(y)) \geq 0,
\end{aligned}$$

where the final inequality follows from (7.3). When $(P_\Omega(y) - z)^T(y - z) = 0$, we have from the same reasoning that $\|P_\Omega(y) - z\| = 0$, proving the final claim. □

Euclidean projections are *nonexpansive* operators, as we show now.

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

Proposition 7.7 *Let Ω be a closed convex set. Then $P_\Omega(\cdot)$ is a nonexpansive operator – that is,*

$$\|P_\Omega(x) - P_\Omega(y)\| \leq \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n.$$

Proof We have

$$\begin{aligned}
&\|x - y\|^2 \\
&\quad = \|(x - P_\Omega(x)) - (y - P_\Omega(y)) + P_\Omega(x) - P_\Omega(y)\|^2 \\
&\quad = \left\|(x - P_\Omega(x)) - (y - P_\Omega(y))\right\|^2 + \|P_\Omega(x) - P_\Omega(y)\|^2 \\
&\qquad - 2\,[x - P_\Omega(x)]^T \left[P_\Omega(y) - P_\Omega(x)\right] - 2\left[y - P_\Omega(y)\right]^T \left[P_\Omega(x) - P_\Omega(y)\right] \\
&\quad \geq \|(x - P_\Omega(x)) - (y - P_\Omega(y))\|^2 + \|P_\Omega(x) - P_\Omega(y)\|^2 \\
&\quad \geq \|P_\Omega(x) - P_\Omega(y)\|^2,
\end{aligned}$$

where the first inequality follows from (7.3). □

7.3 The Projected Gradient Algorithm

We consider (7.1) in which f is Lipschitz continuously differentiable with constant L (see (2.7)) and Ω is closed and convex. Iteration k of the projected gradient algorithm consists of a step along the negative gradient direction $-\nabla f(x^k)$, followed by projection onto the feasible set Ω. The steplength is chosen to ensure descent in f at each iteration. This approach is most useful when the projection operation $P_\Omega(\cdot)$ is inexpensive to compute, no greater than the same order as the cost of evaluating a gradient ∇f.

Given a feasible starting point $x^0 \in \Omega$, the projected gradient algorithm is defined by the formula

$$x^{k+1} = P_\Omega\left(x^k - \alpha_k \nabla f(x^k)\right), \tag{7.4}$$

where $\alpha_k > 0$ is a steplength. Figure 7.2 shows the path traced by $P_\Omega(x - tg)$ for given $x, g \in \mathbb{R}^n$ and scalar $t > 0$ for a box-shaped set Ω. In this case, the path is piecewise linear.

The following proposition shows that if x^k is a point satisfying first-order conditions (see Theorem 7.2), then the projected gradient algorithm will not move away from x^k – that is, $x^{k+1} = x^k$, regardless of the value $\alpha_k > 0$ chosen for the steplength.

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

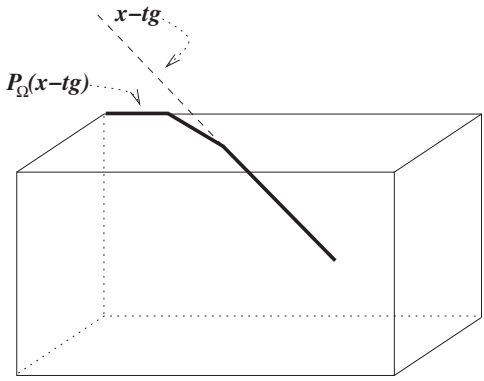

Figure 7.2 Projection of path $x - tg$ for $t \geq 0$ onto the feasible set Ω is piecewise linear.

Proposition 7.8 *Suppose that f is smooth and Ω is closed and convex. Then the point $x^* \in \Omega$ satisfies the first-order condition $-\nabla f(x^*) \in N_\Omega(x^*)$ if and only if $x^* = P_\Omega(x^* - \alpha \nabla f(x^*))$ for all $\alpha > 0$.*

Proof Suppose that x^* satisfies the first-order condition. Then for any $\alpha > 0$, we have

$$0 \geq -\alpha \nabla f(x^*)^T (z - x^*) = [(x^* - \alpha \nabla f(x^*)) - x^*]^T (z - x^*), \quad \text{for all } z \in \Omega,$$

so that, by (7.3), we must have $x^* = P_\Omega(x^* - \alpha \nabla f(x^*))$. Conversely, if $x^* = P_\Omega(x^* - \alpha \nabla f(x^*))$, the same inequality shows that the first-order condition is satisfied. □

7.3.1 General Case: A Short-Step Approach

We first examine the case in which f satisfies (2.7) but may be nonconvex, and set $\alpha_k \equiv 1/L$ in (7.4), where L is the Lipschitz constant for ∇f:

$$x^{k+1} = P_\Omega \left(x^k - (1/L) \nabla f(x^k) \right). \tag{7.5}$$

Then for any $T > 0$, and denoting by $\bar{f}$ a value such that $f(x) \geq \bar{f}$ for all $x \in \Omega$, we have the sublinear convergence bound

$$\min_{0 \leq k \leq T-1} \|x^{k+1} - x^k\| \leq \sqrt{\frac{2(f(x^0) - \bar{f})}{LT}}. \tag{7.6}$$

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

This expression confirms that within the first T iterations, we will find a point x such that

$$\|P_\Omega(x - (1/L)\nabla f(x)) - x\| \leq \epsilon.$$

To verify the bound (7.6), we have from Lemma 2.2 that for any $x \in \Omega$,

$$f(x) \leq q_k(x) := f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{L}{2}\|x - x^k\|^2. \tag{7.7}$$

The minimizer of $q_k(x)$ over $x \in \Omega$ is simply $P_\Omega(x^k - (1/L)\nabla f(x^k))$ (see the Exercises), which is x^{k+1} by (7.5). We thus have, from Theorem 7.2 applied to $\min_{x\in\Omega} q_k(x)$, that

$$-\nabla q_k(x^{k+1}) = -\nabla f(x^k) - L(x^{k+1} - x^k) \in N_\Omega(x^{k+1}).$$

Thus, by Definition 7.1, it follows that

$$\begin{aligned}&[-\nabla f(x^k) - L(x^{k+1} - x^k)]^T(x^k - x^{k+1}) \leq 0\\ &\quad\Longrightarrow\ \nabla f(x^k)^T(x^k - x^{k+1}) \geq L\|x^k - x^{k+1}\|^2.\end{aligned}$$

Since $f(x^k) = q_k(x^k)$ and $f(x^{k+1}) \leq q_k(x^{k+1})$, we have

$$\begin{aligned}f(x^k) - f(x^{k+1}) &\geq q_k(x^k) - q_k(x^{k+1})\\ &= -\nabla f(x^k)^T(x^{k+1} - x^k) - \frac{L}{2}\|x^{k+1} - x^k\|^2\\ &\geq \frac{L}{2}\|x^{k+1} - x^k\|^2.\end{aligned}$$

By summing these inequalities up for $k = 0, 1, \ldots, T-1$, we have

$$\sum_{k=0}^{T-1}\|x^{k+1} - x^k\|^2 \leq \frac{2}{L}(f(x^0) - f(x^T)) \leq \frac{2}{L}(f(x^0) - \bar{f}),$$

from which the result follows, in a similar fashion to Section 3.2.1.

7.3.2 General Case: Backtracking

We now describe a backtracking version of the projected gradient method, which does not require knowledge of the Lipschitz constant L. We follow the backtracking approach for unconstrained optimization described in Section 3.5, but include the projection operator to ensure that all iterates x^k are feasible.

The scheme is shown in Algorithm 7.1. At each iteration, we choose some initial guess of the steplength $\bar{\alpha}_k > 0$. (This could be either some constant, such as $\bar{\alpha}_k = 1$ for all k, or a slight increase on the successful steplength

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

Algorithm 7.1 Projected Gradient with Backtracking

Given $0 < c_1 < \frac{1}{2}$, $\beta \in (0,1)$; Choose x^0;
for $k = 0, 1, 2, \ldots$ **do**
  Set $\alpha_k = \bar{\alpha}_k$, for some initial guess of steplength $\bar{\alpha}_k > 0$;
  while $f(P_\Omega(x^k - \alpha_k \nabla f(x^k))) > f(x^k) + c_1 \nabla f(x^k)^T (P_\Omega(x^k - \alpha_k \nabla f(x^k)) - x^k)$
  do
    $\alpha_k \leftarrow \beta \alpha_k$;
  end while
  Set $x^{k+1} = P_\Omega(x^k - \alpha_k \nabla f(x^k))$;
end for

from the previous iteration, such as $\bar{\alpha}_k = 1.2\alpha_{k-1}$.) We then test a sufficient decrease condition, similar to (3.26a). This condition asks whether the actual improvement in f obtained with this value of α_k is at least a fraction c_1 of the improvement expected from the first-order Taylor series expansion of f around the current iterate x^k. If this condition is not satisfied, we decrease α_k by a factor $\beta \in (0,1)$, repeating the process until the sufficient decrease condition holds.

Provided the initial guess $\bar{\alpha}_k$ is chosen larger than $1/L$, the steps that are accepted by this backtracking approach are typically larger than the $1/L$ steps of the previous section, and convergence is often faster in practice. We derive convergence results for Algorithm 7.1 in the Exercises.

7.3.3 Smooth Strongly Convex Case

We now consider f that is strongly convex with modulus of convexity m (see (2.19)), as well as having L-Lipschitz gradients (2.7). Moreover, we assume that f is twice continuously differentiable so that (2.4) from Theorem 2.1 applies. We have from the latter result that for any $y, z \in \mathbb{R}^n$ and any $\alpha \geq 0$,

$$\begin{aligned}
&\|(y - \alpha \nabla f(y)) - (z - \alpha \nabla f(z))\| \\
&\quad = \left\| \int_0^1 \left[I - \alpha \nabla^2 f(z + t(y - z)) \right] (y - z)\, dt \right\| \\
&\quad \leq \int_0^1 \left\| I - \alpha \nabla^2 f(z + t(y - z)) \right\| dt\, \|y - z\| \\
&\quad \leq \sup_{t \in [0,1]} \left\| I - \alpha \nabla^2 f(z + t(y - z)) \right\| \|y - z\| \\
&\quad \leq \max(|1 - \alpha m|, |1 - \alpha L|)\, \|y - z\|, \qquad (7.8)
\end{aligned}$$

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

where the second inequality follows from the fact that the spectrum of $\nabla^2 f(\cdot)$ is contained in the interval $[m, L]$. The right-hand side is minimized by setting $\alpha = 2/(L+m)$ (see the Exercises), for which value we have

$$\alpha = \frac{2}{L+m} \quad \Longrightarrow \quad \|(y - \alpha\nabla f(y)) - (z - \nabla f(z))\| \leq \frac{L-m}{L+m}\|y - z\|.$$

We set $y = x^k$, $z = x^*$, and $\alpha_k \equiv 2/(L+m)$ in (7.8) and use the characterization of x^* in Proposition 7.8 and the nonexpansive property (Proposition 7.7) to obtain

$$\begin{aligned}\|x^{k+1} - x^*\| &= \left\| P_\Omega(x^k - \alpha_k \nabla f(x^k)) - P_\Omega(x^* - \alpha_k \nabla f(x^*)) \right\| \\ &\leq \|(x^k - x^*) - \alpha_k(\nabla f(x^k) - \nabla f(x^*))\| \\ &\leq \frac{L-m}{L+m}\|x^k - x^*\|,\end{aligned}$$

which indicates linear convergence of $\{x^k\}$ to the optimal x^* for a fixed-steplength version of projected gradient. Note that when $0 < m \ll L$, the linear rate constant is approximately $(1 - 2m/L)$.

The projected gradient method analyzed here is a special case of the proximal-gradient algorithm described in Section 9.3. We refer to that section for analysis of cases other than those analyzed here – for example, the case in which f is convex but not strongly convex.

7.3.4 Momentum Variants

There are versions of the projected gradient method that make use of the momentum ideas of Chapter 4. Following (4.7), Nesterov's method can be adapted to (7.1) as follows:

$$y^k = x^k + \beta_k(x^k - x^{k-1}) \tag{7.9a}$$

$$x^{k+1} = P_\Omega(y^k - \alpha_k \nabla f(y^k)), \tag{7.9b}$$

where we define $x^{-1} = x^0$ as before, so that $y^0 = x^0$. (When $\beta_k \equiv 0$, we recover the projected gradient method (7.4).) Note that the sequence $\{x^k\}$ is feasible, whereas the y^k are not necessarily feasible. With appropriate choices of α_k and β_k, and when applied to strongly convex f, the iterations (7.9) will converge at an approximate linear rate of $(1 - \sqrt{m/L})$.

7.3.5 Alternative Search Directions

Recall that in Section 3.1, we show that search directions d^k other than the negative gradient could be used in conjunction with line searches in algorithms

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

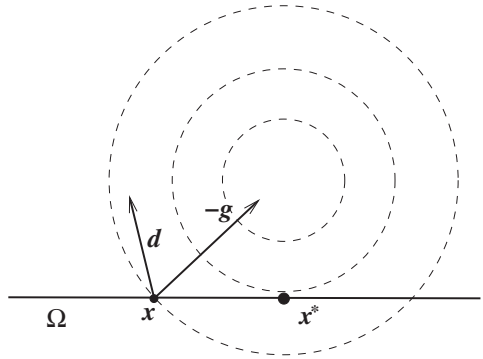

Figure 7.3 Searching along a direction $d \neq -g = -\nabla f(x)$ and projecting onto Ω fails to yield descent in f, even for small steplengths α.

for smooth unconstrained optimization. We ask here whether such general choices of d^k can be used in the projected gradient method for solving the *constrained* problem (7.1). That is, can we define steps of the form $x^{k+1} = P_\Omega(x^k + \alpha_k d^k)$ for d^k satisfying conditions like those of (3.22)? The answer is *no*, in general. It is sufficient to illustrate with a picture; see Figure 7.3. Here we show minimization of a quadratic function whose contours are shown, subject to $x \in \Omega$ where Ω is the half-space below the line. The solution is shown at x^*. From the point x, we show the direction $-g = -\nabla f(x)$, the negative gradient direction, which is orthogonal to the contours. Clearly, if we take steps of size $\alpha > 0$ along this direction and project onto Ω, we are moving toward x^* and decreasing the function (provided α is not too large). Consider now the direction d, which satisfies conditions like (3.22) – it makes an angle of significantly less than $\pi/2$ radians with $-\nabla f(x)$ and is similar in length. Although we can decrease f by moving along d with steplength $\alpha > 0$, the same does not hold when we project $x + \alpha d$ onto Ω. In fact, the function *increases* along the path defined by $P_\Omega(x + \alpha d)$ for $\alpha > 0$.

7.4 The Conditional Gradient (Frank–Wolfe) Method

For some feasible sets Ω, the projection operator P_Ω can be expensive to compute, whereas minimization of a linear objective over this same sets is relatively inexpensive. For example, minimizing a linear objective over the simplex $\left\{x \in \mathbb{R}^n \mid x \geq 0,\ \sum_{i=1}^n x_i = 1\right\}$ simply requires finding the minimum element of the gradient, whereas projection of an arbitrary vector y onto this

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

set requires (naively) a sorting of the elements of y. The conditional gradient method, the first variant of which was proposed by Frank and Wolfe (1956), provides an effective algorithm for constrained optimization that requires only linear minimization rather than Euclidean projection.

The conditional gradient method replaces the objective in (7.1) by a linear Taylor series approximation around the current iterate x^k and solves the following subproblem:

$$\bar{x}^k := \arg\min_{\bar{x}\in\Omega} f(x^k) + \nabla f(x^k)^T(\bar{x} - x^k) = \arg\min_{\bar{x}\in\Omega} \nabla f(x^k)^T \bar{x}. \quad (7.10)$$

The next iterate is obtained by stepping toward $\bar{x}^k$ from x^k as follows:

$$x^{k+1} = x^k + \alpha_k(\bar{x}^k - x^k), \quad \text{for some } \alpha_k \in (0, 1]. \quad (7.11)$$

Note that if the initial iterate x^0 is feasible (that is, $x^0 \in \Omega$), all subsequent iterates x^k, $k = 1, 2, \ldots$ are also feasible, as are all the subproblem solutions $\bar{x}^k$, $k = 0, 1, 2 \ldots$. The method is usually applied only when Ω is compact (that is, closed and bounded) and convex, so that $\bar{x}^k$ in (7.10) is well defined for all k. The conditional gradient method is practical only when the linearized subproblem (7.10) is much easier to solve than the original problem (7.1). As we have discussed, such is the case for various interesting choices of Ω.

The original approach of Frank and Wolfe makes the particular choice of steplength $\alpha_k = 2/(k + 2)$, $k = 0, 1, 2, \ldots$. The resulting method converges at a sublinear rate, as we show now. Again assume that $\Omega \subset \mathbb{R}^n$ is a closed, bounded convex set and f is a smooth convex function. We define the *diameter* D of Ω as follows:

$$D := \max_{x, y\in\Omega} \|x - y\|. \quad (7.12)$$

We have the following result.

Theorem 7.9 *Suppose that f is a convex function whose gradient is Lipschitz continuously differentiable with constant L on an open neighborhood of Ω, where Ω is a closed bounded convex set with diameter D, and let x^* be the solution to (7.1). Then if algorithm* (7.10)–(7.11) *is applied from some $x^0 \in \Omega$ with steplength $\alpha_k = 2/(k + 2)$, we have*

$$f(x^k) - f(x^*) \le \frac{2LD^2}{k+2}, \quad k = 1, 2, \ldots .$$

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

Proof Since f has L-Lipschitz gradients, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 L \|\bar{x}^k - x^k\|^2 \\ &\leq f(x^k) + \alpha_k \nabla f(x^k)^T (\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 L D^2, \end{aligned} \tag{7.13}$$

where the second inequality comes from the definition of D. For the first-order term, we have by definition of $\bar{x}^k$ in (7.10) and feasibility of x^* that

$$\nabla f(x^k)^T (\bar{x}^k - x^k) \leq \nabla f(x^k)^T (x^* - x^k) \leq f(x^*) - f(x^k).$$

By substituting this bound into (7.13) and subtracting $f(x^*)$ from both sides, we have

$$f(x^{k+1}) - f(x^*) \leq (1 - \alpha_k)[f(x^k) - f(x^*)] + \frac{1}{2}\alpha_k^2 L D^2.$$

We now demonstrate the required bound by induction. By setting $k = 0$ and substituting $\alpha_0 = 1$, we have

$$f(x^1) - f(x^*) \leq \frac{1}{2} L D^2 < \frac{2}{3} L D^2,$$

as required. For the inductive step, we suppose that the claim holds for some k, and demonstrate that it still holds for $k + 1$. We have

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq \left(1 - \frac{2}{k+2}\right)[f(x^k) - f(x^*)] + \frac{1}{2}\frac{4}{(k+2)^2} L D^2 \\ &= L D^2 \left[\frac{2k}{(k+2)^2} + \frac{2}{(k+2)^2}\right] \\ &= 2 L D^2 \frac{(k+1)}{(k+2)^2} \\ &= 2 L D^2 \frac{k+1}{k+2} \frac{1}{k+2} \\ &\leq 2 L D^2 \frac{k+2}{k+3} \frac{1}{k+2} = \frac{2 L D^2}{k+3}, \end{aligned}$$

as required. □

Note that the same result holds if we choose α_k to exactly minimize f along the line from x^k to $\bar{x}^k$; only minimal changes to the proof are needed.

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

Notes and References

The projected gradient method originated with Goldstein (1964) and Levitin and Polyak (1966). Goldstein proposed the steplength acceptance condition used in Algorithm 7.1 in Goldstein, 1974. Convergence properties of projected gradient were developed further by Bertsekas (1976) and Dunn (1981).

The conditional gradient approach was described first for the case of convex quadratic programming by Frank and Wolfe (1956). Extensions to more general problems of the type (7.1) are described by Dem'yanov and Rubinov (1967) (which is difficult to read) and in Dem'yanov and Rubinov (1970). Dunn (1980) presents comprehensive results for various line-search procedures, including linear convergence results for problems that satisfy a condition akin to second-order sufficiency, and results for nonconvex problems. The revival of interest in the conditional gradient approach in the machine learning community is due largely to Jaggi (2013).

Exercises

1. Prove that the formula for $P_\Omega(x)$ in Example 7.5 is correct.
2. Prove that (7.8) is minimized by setting $\alpha = 2/(L+m)$, when $0 < m < L$. Prove that the alternative choice of steplength $\alpha = 1/L$ leads to a linear convergence rate of $(1 - m/L)$ in $\|x^k - x^*\|$ (similar to the rate obtained for the unconstrained case in Section 3.2.3). How do these two different choices compare in terms of the number of iterations T required to guarantee $\|x^T - x^*\| \leq \epsilon$ for some tolerance $\epsilon > 0$?
3. By adapting the analysis of Section 4.3 to the projected version of Nesterov's method for the constrained case (7.9) and for the choice of parameters α_k and β_k shown in (4.23), prove linear convergence of this method, and find the constant for the linear rate.
4. Find the minimizer of $c^T x$ (for $c \in \mathbb{R}^n$, a constant vector, and for $x \in \mathbb{R}^n$, a variable) over Ω, where Ω is each of the following sets:
   (a) The unit ball: $\{x \mid \|x\|_2 \leq 1\}$
   (b) The unit simplex: $\left\{x \in \mathbb{R}^n \mid x \geq 0,\ \sum_{i=1}^n x_i = 1\right\}$
   (c) A box: $\{x \mid 0 \leq x_i \leq 1,\ i = 1, 2, \ldots, n\}$
5. Show that Theorem 7.9 continues to hold if α_k is chosen in (7.11) to minimize $f(x^k + \alpha_k(\bar{x}^k - x^k))$ for $\alpha_k \in [0, 1]$, rather than from the formula $\alpha_k = 2/(k+2)$.

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press

6. Prove that for any $\alpha_k > 0$ and for x^{k+1} defined by (7.4), we have

$$x^{k+1} = \arg\min_{x\in\Omega} f(x^k) + \nabla f(x^k)^T(x - x^k) + \frac{1}{2\alpha_k}\|x - x^k\|^2$$

and

$$\|P_\Omega(x^k - \alpha_k \nabla f(x^k)) - x^k\|^2 \le \alpha_k \nabla f(x^k)^T[x^k - P_\Omega(x^k - \alpha_k \nabla f(x^k))].$$

(Note that with $\alpha_k = 1/L$, it follows that, for q_k defined in (7.7), we have $x^{k+1} = \min_{x\in\Omega} q_k(x)$.)

7. Show by using arguments similar to those of Section 7.3.1 that when f is L-smooth, the sufficient decrease condition in Algorithm 7.1 will be satisfied whenever $\alpha_k \le 1/L$ – that is,

$$f(P_\Omega(x^k - \alpha_k \nabla f(x^k))) \le f(x^k) + c_1 \nabla f(x^k)^T (P_\Omega(x^k - \alpha_k \nabla f(x^k)) - x^k), \tag{7.14}$$

where $c_1 \in (0, 1/2)$. Deduce that, provided $\bar{\alpha}_k \ge 1/L$, the inner loop in Algorithm 7.1 terminates with $\alpha_k \ge \beta/L$.

8. Show by combining with the results of the previous two questions that for *any* $\alpha_k > 0$ such that (7.14) is satisfied, we have, using (7.4), that

$$f(x^{k+1}) \le f(x^k) - c_1 \frac{1}{\alpha_k}\|x^{k+1} - x^k\|^2 \le f(x^k) - c_1 \frac{1}{\bar{\alpha}_k}\|x^{k+1} - x^k\|^2.$$

Hence, taking $\bar{\alpha}_k = 1/M$ for some $M > 0$ and all k, derive a convergence bound similar to (7.6) for Algorithm 7.1.

https://doi.org/10.1017/9781009004282.008 Published online by Cambridge University Press