Chapter 7

First-Order Methods for Constrained Optimization

In constrained optimization, we aim to find a point x which achieves the smallest value of some function f subject to the requirement that x lives in some specified set Ω.

Constrained optimization lets us design considerably more rich and complex optimization problems. The constraints could simply be bounds on the values of the variables, but could model temporal dependencies, resource constraints, or statistical models. In this chapter, we will focus on case when Ω is a simple convex set. We will move to more complicated scenarios in the later chapters.

7.1 Optimality Conditions

We consider the problem (2.1), restate here as

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

where Ω is closed and convex and f is smooth (at least differentiable). We recall material from Chapter 2, including the definitions of (local and global) solutions from Section 2.1 and the definitions of convexity and normal cones from Section 2.4.

In order to characterize optimality for minimizing a smooth function f over a closed convex set Ω, we need a bit of additional machinery. Suppose all of the minimizers of f lie outside Ω. Then the gradient of f will not vanish in Ω. Instead, the function will be decreasing as it pushes against the boundary.

To make this notion rigorous, we first introduce the notion of a *normal cone*. We define the *normal cone* to Ω at a point $x \in \Omega$ as follows.

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*

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

(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.)

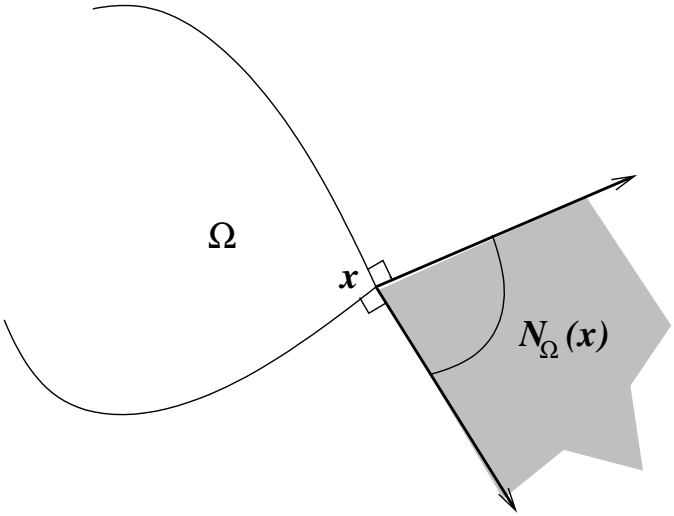

Figure 7.1: Normal Cone

Theorem 7.2. *Consider the constrained optimization problem* (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 converse holds.*

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

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

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 expression above, 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. □

When f is *strongly* convex, the 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. The same technique as in the proof of Theorem 2.8 can be used to show existence of a solution x^*. Theorem 7.2 tells us that $-\nabla f(x^*) \in N_\Omega(x^*)$. Letting z be any other point in Ω, we

have from the characterization (2.18) of strong convexity that

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

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

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. Denote this point by $\Pi_\Omega(x)$. Note that $\Pi_\Omega(x)$ is the solution of a constrained optimization problem:

$$\Pi_\Omega(x) = \arg\min\{\|z - x\| \ : \ z \in \Omega\}$$

That is, $\Pi_\Omega(x)$ is the solution to the optimization problem

$$\begin{array}{ll} \text{minimize}_z & \frac{1}{2}\|z - x\|^2 \\ \text{subject to} & z \in \Omega \end{array}.$$

Since the cost function of this problem is strongly convex, this proves that $\Pi_\Omega(x)$ is unique for all x.

Using the minimum principle, we can compute a variety of projections onto simple sets.

Example 1: The nonnegative orthant The nonnegative orthant is the set of vectors which are nonnegative in all coordinates.

$$\Omega = \{x \ : \ x_i \geq 0 \ \forall \ i = 1, \ldots, d\}$$

Note that Ω is a closed, convex cone.

Unpacking the condition $\langle \Pi_\Omega(x) - x, z - \Pi_\Omega(x)\rangle \geq 0$, we must have that

$$[\Pi_\Omega(x) - x]_i \geq 0$$

for all coordinates. Note that simply setting

$$[\Pi_\Omega(x)]_i = \begin{cases} x_i & x_i \geq 0 \\ 0 & x_i < 0 \end{cases}$$

satisfies the Minimum Principle.

Example 2: Unit norm ball Let

$$\Omega = \{x \ : \ \|x\| \leq 1\}$$

To compute $\Pi_\Omega(x)$, note that we require $\langle \Pi_\Omega(x) - x, z - \Pi_\Omega(x)\rangle \geq 0$. for all $z \in \Omega$. One can readily check that

$$\Pi_\Omega(x) = \frac{x}{\|x\|}$$

satisfies the Minimum Principle.

One of the most useful properties of the Euclidean projection is the fact that projections are *nonexpansive* in the following sense:

$$\|\Pi_\Omega(x) - \Pi_\Omega(y)\| \leq \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n. \tag{7.2}$$

(See Proposition A.8 for a proof.)

7.3 The projected gradient algorithm

The projected gradient algorithm combines a projection step with a gradient step. This lets us solve a variety of constrained optimization problems with simple constraints.

We will aim to solve the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in \Omega \end{array} \tag{7.3}$$

where f is smooth and Ω is convex. Let us assume as usual that ∇f is Lipschitz so that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.

Let us define a projected gradient scheme to solve this problem. Let $\alpha_0, \ldots, \alpha_T, \ldots$, be a sequence of positive step sizes. Choose $x_0 \in X$, and iterate

$$x_{k+1} = \Pi_\Omega(x_k - \alpha_k \nabla f(x_k))\,. \tag{7.4}$$

The algorithm simply alternates between taking gradient steps and then taking projection steps.

The key idea behind this algorithm is summed up by the following proposition

Proposition 7.4. *Let f be differentiable and convex and let Ω be convex. $x_\star$ is an optimal solution of (7.3) if and only if $x_\star = \Pi_\Omega(x_\star - \alpha\nabla f(x_\star))$ for all $\alpha > 0$.*

Proof. $x_\star$ is an optimal solution if and only if $\langle \nabla f(x_\star), x - x_\star \rangle \geq 0$ for all $x \in \Omega$. This is equivalent to

$$\langle x_\star - (x_\star - \alpha\nabla f(x_\star)), x - x_\star \rangle \geq 0\,,$$

which, by the Minimum Principle is equivalent to $x_\star$ being the Euclidean projection of $x_\star - \alpha\nabla f(x_\star)$ onto Ω. In other words, $x_\star = \Pi_\Omega(x_\star - \alpha\nabla f(x_\star))$. □

For non-convex f, we see that a fixed point of the projected gradient iteration is a stationary point of h. We first analyze the convergence of this projected gradient method for arbitrary smooth f, and then focus on strongly convex f.

7.3.1 General Case

Let f_* denote the optimal value of (7.3). Suppose we set $\alpha_k = 1/M$ for all k with $M \geq L$. Then we have

$$\min_{k \leq T} \|x_{k+1} - x_k\| \leq \sqrt{\frac{2(f(x_0) - f_\star)}{M(T+1)}}\,. \tag{7.5}$$

This expression confirms that we will find a point x where

$$\|\Pi_\Omega(x - \alpha\nabla f(x)) - x\| \leq \epsilon\,.$$

To verify this inequality, note that for any x, y,

$$f(x) = f(x) \leq f(y) + \nabla f(y)^T(x-y) + \frac{M}{2}\|x-y\|^2 =: \ell(x;y)$$

for any $M \geq L$. This is just Taylor's series. Note that the minimizer of $\ell(x;y)$ (with respect to x) over Ω is equal to

$$\Pi_\Omega(y - 1/M\nabla f(y))\,.$$

and also note that $\ell(x;y)$ is strongly convex with parameter M.

Now we have the chain of inequalities

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq f(x_k) - \ell(x_{k+1};x_k) \\ &= \ell(x_k;x_k) - \ell(x_{k+1};x_k) \\ &\geq \frac{M}{2}\|x_{k+1}-x_k\|^2 \end{aligned}$$

Summing these inequalities up for $k = 1, \ldots, T$, we have

$$\sum_{k=0}^{T} \|x_{k+1}-x_k\|^2 \leq \frac{2}{M}(f(x_0) - f_\star)$$

and the conclusion follows.

7.3.2 Strongly Convex Case

Let's now assume that f is strongly convex with strong convexity parameter m:

$$f(z) \geq f(x) + \nabla f(x)^T(z-x) + \frac{m}{2}\|z-x\|^2\,. \tag{7.6}$$

Let $x_\star$ denote the optimal solution of (7.3). $x_\star$ is unique because of strong convexity. Observe that

$$\begin{aligned} \|x_{k+1} - x_\star\| &= \|\Pi_\Omega(x_k - \alpha_k\nabla f(x_k) - \Pi_\Omega(x_\star - \alpha_k\nabla f(x_\star))\| &(7.7)\\ &\leq \|x_k - \alpha_k\nabla f(x_k) - x_\star + \alpha_k\nabla f(x_\star)\| &(7.8)\end{aligned}$$

Here, the first equality follows by the definition of x_{k+1} and because $x_\star$ is optimal (see Proposition 7.4). (7.8) follows from Proposition A.8.

Since f is strongly convex and has a Lipschitz continuous gradient, it follows that for all vectors x and y and all positive scalars η

$$\|x - \eta\nabla f(x) - (y - \eta\nabla f(y))\| \leq \max\{|1-\eta L|, |1-\eta m|\}\|x-y\|\,. \tag{7.9}$$

To see this, note that there exists a $\hat{t} \in [0,1]$ such that

$$\|x - \eta\nabla f(x) - (y - \eta\nabla f(y))\| = \left\|\left(I - \eta\nabla^2 f(x + \hat{t}(y-x))\right)(y-x)dt\right\|\,. \tag{7.10}$$

From this, it follows that

$$\|x - \eta\nabla f(x) - (y - \eta\nabla f(y))\| \leq \sup_z \|I - \eta\nabla^2 f(z)\|\|y-z\|\,. \tag{7.11}$$

Note that the minimum eigenvalue of $\nabla^2 f(z)$ is at least m and the maximum eigenvalue is at least L. Therefore the eigenvalues of $I - \eta\nabla^2 f(z)$ are at most $\max(1-\eta L, 1-\eta m)$ and at least $\min(1-\eta L, 1-\eta m)$. Therefore, $\|I - \eta\nabla^2 f(z)\| \leq \max(|1-\eta L|, |1-\eta m|)$.

Using the upper bound (7.9), we have

$$\|x_{k+1} - x_\star\| \leq \max\{|1-\alpha_k L|, |1-\alpha_k m|\}\|x - y\|\,. \tag{7.12}$$

Note that $\alpha_k = \frac{2}{L+m}$ minimizes the right hand side for all k. Setting α_k to this value, we find that

$$\|x_{k+1} - x_\star\| \leq \left(\frac{L-m}{L+m}\right)\|x_k - x_\star\| \tag{7.13}$$

or, denoting $\kappa = \frac{L}{m}$ and $D_0 = \|x_0 - x_\star\|$,

$$\|x_k - x_\star\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k D_0 \tag{7.14}$$

That is, for strongly convex f and arbitrary Ω, the projected gradient algorithm converges at a linear rate under a constant step-size policy.

For the case where we lack strong convexity, we defer the analysis to Chapter 9 where we provide a more general algorithm for nonsmooth optimization that reduces to projected gradient descent as a special case.

7.3.3 Nesterov Iteration

We can even define an *accelerated* version of the projected gradient method. Iterations take the form:

$$\begin{aligned} \xi_{k+1} &= \Pi_\Omega\left(y_k - \alpha\nabla f(y_k)\right) \\ y_k &= \xi_k + \beta(\xi_k - \xi_{k-1}) \end{aligned} \tag{7.15}$$

Note that when $\Pi_\Omega = I$, we recover the standard Nesterov algorithm. When $\beta = 0$, we recover the proximal gradient method. This method will converge in

$$O\left(\sqrt{\frac{L}{m}}\log(1/\epsilon)\right)$$

iterations for strongly convex functions. We leave a proof of this convergence rate as an exercise.

7.4 The Conditional Gradient ("Frank-Wolfe") Method

Often times, the computation of the projection onto the set Ω is a very expensive operation. Moreover, for many sets that arise in optimization, it is often considerably simpler to minimize a linear objective over Ω than it is to project onto this set. For example, minimizing a linear objective over the simplex simply requires extracting a maximum, whereas the Euclidean projection naively requires sorting a list of numbers. The conditional gradient method, the first variant of which was proposed by Frank and Wolfe [13], provides an effective algorithm for constrained optimization that requires only linear minimization rather than Euclidean projection.

Conditional gradient method replaces the objective in (7.3) by a linear Taylor-series approximation around the current iterate x^k, and solves the following subproblem:

$$\bar{x}^k := \arg\min_{\bar{x}\in\Omega} f(x^k) + \nabla f(x^k)^T(\bar{x} - x^k) = \arg\min_{\bar{x}\in\Omega} \nabla f(x^k)^T\bar{x}. \tag{7.16}$$

Note that the constraint set Ω is unchanged. The next iterate is obtained by steeping toward $\bar{x}^k$ from x^k, as follows

$$x^{k+1} = x^k + \alpha_k(\bar{x}^k - x^k), \quad \text{for some } \alpha_k \in (0, 1]. \tag{7.17}$$

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 $\bar{x}^k$, $k = 0, 1, 2 \ldots$.

This method is practical when the linearized subproblem (7.16) is much easier to solve than the original problem (7.3). As we have discussed, this is the case in many applications of interest.

The original Frank-Wolfe approach made the particular choice of step length $\alpha_k = 2/(k + 2)$, $k = 0, 1, 2, \ldots$. 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:

$$D := \max_{x,y\in\Omega} \|x - y\|. \tag{7.18}$$

Theorem 7.5. *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 that solution x^* to (7.3) exists. Then if algorithm (7.16)-(7.17) is applied from some $x^0 \in \Omega$ with steplength $\alpha_k = 2/(k + 2)$, we have*

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

Proof. Since f has L-Lipschitz gradients, we have

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \alpha_k \nabla f(x^k)^T(\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 L\|\bar{x}^k - x^k\|^2 \\ &\leq f(x^k) + \alpha_k \nabla f(x^k)^T(\bar{x}^k - x^k) + \frac{1}{2}\alpha_k^2 LD^2, \end{aligned} \tag{7.19}$$

where the second inequality comes from the definition of D. For the first-order term, we have by definition of $\bar{x}^k$ in (7.16) 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 both sides of (7.19) and subtracting $f(x^*)$ from both sides, we have

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

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

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

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

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

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.

Notes and References

Pointer to proof of Nesterov's method with projection.

Homework: projection and linear optimization on the simplex

Exercises

1. Exercise