Chapter 2

Foundations

We outline in this chapter the foundations of the algorithms and theory discussed in later chapters. These foundations include the elements of convex analysis, optimality conditions for convex problems, Taylor's theorem (which is the basis of much of smooth nonlinear optimization), and proximal operators (the basis of most algorithms for regularized optimization).

2.1 A Taxonomy of Solutions

Suppose that f is a function mapping some domain $\mathcal{D} \subset \mathbb{R}^n$ to the real line $\mathbb{R}$. We have the following definitions.

- $x^* \in \mathcal{D}$ is a *local minimizer* of f if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a *global minimizer* of f if $f(x) \geq f(x^*)$ for all $x \in \mathcal{D}$.
- $x^* \in \mathcal{D}$ is a *strict local minimizer* if it is a local minimizer and in addition $f(x) > f(x^*)$ for all $x \in \mathcal{N}$ with $x \neq x^*$.
- x^* is an *isolated local minimizer* if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \mathcal{D}$ and in addition, $\mathcal{N}$ contains no local minimizers other than x^*.

For the constrained optimization problem

$$\min_{x \in \Omega} f(x), \tag{2.1}$$

where $\Omega \subset \mathcal{D} \subset \mathbb{R}^n$ is a closed set, we modify the terminology slightly to use the word "solution" rather than "minimizer." That is, we have the following definitions.

- $x^* \in \Omega$ is a *local solution* of (2.1) if there is a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N} \cap \Omega$.
- $x^* \in \Omega$ is a *global solution* of (2.1) if $f(x) \geq f(x^*)$ for all $x \in \Omega$.

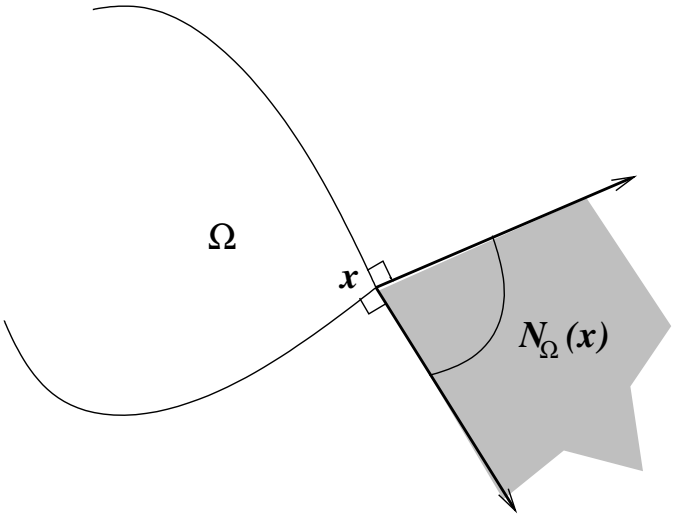

Figure 2.1: Normal Cone

2.2 Convex Sets and Functions

A convex set $\Omega \subset \mathbb{R}^n$ has the property that

$$x, y \in \Omega \;\; \Rightarrow \;\; (1-\alpha)x + \alpha y \in \Omega \;\text{ for all } \alpha \in [0,1]. \tag{2.2}$$

The convex sets that we consider in this book are usually *closed*.

We have the following definition of normal cones, which is key to recognizing optimality in problems (2.1), as we see in Chapter 7.

Definition 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a convex set. At any $x \in \Omega$ the* normal cone $N_\Omega(x)$ *is defined as*

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

(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 2.1 for an example.

Given a closed convex set $\Omega \subset \mathbb{R}^n$, the projection operator $P : \mathbb{R}^n \to \Omega$ is defined as follows:

$$P(y) = \arg\min_{z \in \Omega} \|z - y\|_2.$$

That is, $P(y)$ is the point in Ω that is closest to y in the sense of the Euclidean norm. This operator is useful both in defining optimality conditions and in defining algorithms. We prove several useful results about this operator in Section A.6.

We consider convex functions that map ϕ defined on all of $\mathbb{R}^n$, but possibly taking the value ∞ at some points. We say that $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ maps $\mathbb{R}^n$ to the extended reals, or that it is an *extended-value function.* The defining property of a convex function is the following inequality:

$$\phi((1-\alpha)x + \alpha y) \le (1-\alpha)\phi(x) + \alpha\phi(y), \quad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \alpha \in [0,1]. \tag{2.3}$$

The following definitions are useful.

- The *effective domain* of ϕ is the set of points $x \in \mathbb{R}^n$ such that $\phi(x) < +\infty$.
- The *epigraph* of ϕ, denoted by $\operatorname{epi}\phi$, is the following subset of $\mathbb{R}^{n+1}$:

$$\operatorname{epi}\phi := \{(x,t) \in \Omega \times \mathbb{R} \,:\, t \geq \phi(x)\}.$$

  The effective domain is therefore the set of points x such that $(x,t) \in \operatorname{epi}\phi$ for some $t \in \mathbb{R}$.
- ϕ is a *proper* convex function if $\phi(x) < +\infty$ for some $x \in \mathbb{R}^n$ and $\phi(x) > -\infty$ for all $x \in \mathbb{R}^n$. All convex functions of practical interest are proper.
- ϕ is a *closed proper* convex function if it is a proper convex function and the set $\{x \in \mathbb{R}^n \,:\, \phi(x) \leq \bar{t}\}$ is a closed set for all $\bar{t} \in \mathbb{R}$.

The concepts of "minimizer" and "solution" for the case of convex objective function and constraint set are simpler than for the general case. In particular, the distinction between "local" and "global" solutions goes away, as we show now.

Theorem 2.2. *Suppose that in* (2.1)*, the function f is convex and the set Ω is closed and convex. We have the following.*

(a) Any local solution of (2.1) *is also a global solution.*

(b) The set of global solutions of (2.1) *is a convex set.*

Proof. For (a), suppose for contradiction that $x^* \in \Omega$ is a local solution but not a global solution, so there exists a point $\bar{x} \in \Omega$ such that $f(\bar{x}) < f(x^*)$. Then by convexity we have for any $\alpha \in (0,1)$ that

$$f(x^* + \alpha(\bar{x} - x^*)) \leq (1-\alpha)f(x^*) + \alpha f(\bar{x}) < f(x^*).$$

But for any neighborhood $\mathcal{N}$, we have for sufficiently small $\alpha > 0$ that $x^* + \alpha(\bar{x} - x^*)) \in \mathcal{N} \cap \Omega$ and $f(x^* + \alpha(\bar{x} - x^*)) < f(x^*)$, contradicting the definition of a local minimizer.

For (b), we simply apply the definition of convexity for both sets and functions. Given any global solutions x^* and $\bar{x}$, we have $f(\bar{x}) = f(x^*)$, so for any $\alpha \in [0,1]$ we have

$$f(x^* + \alpha(\bar{x} - x^*)) \leq (1-\alpha)f(x^*) + \alpha f(\bar{x}) = f(x^*).$$

We have also that $f(x^* + \alpha(\bar{x} - x^*)) \geq f(x^*)$, since $x^* + \alpha(\bar{x} - x^*) \in \Omega$ and x^* is a global minimizer. It follows from these two inequalities that $f(x^* + \alpha(\bar{x} - x^*)) = f(x^*)$, so that $x^* + \alpha(\bar{x} - x^*)$ is also a global minimizer. □

If there exists a value $m > 0$ such that

$$\phi((1-\alpha)x + \alpha y) \leq (1-\alpha)\phi(x) + \alpha\phi(y) - \frac{1}{2}m\alpha(1-\alpha)\|x - y\|_2^2 \tag{2.4}$$

for all x and y in the domain of ϕ, we say that ϕ is *strongly convex with modulus of convexity m.*

For any set $\Omega \subset \mathbb{R}^n$ we define the *indicator function* $I_\Omega(x)$ as follows:

$$I_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{otherwise.} \end{cases}$$

When Ω is a convex set, then I_Ω is a convex function.

Indicator functions are useful devices for deriving optimality conditions for constrained problems, and even for developing algorithms. The constrained optimization problem (2.1) can be restated equivalently as follows:

$$\min_{x \in \mathbb{R}^n} \; f(x) + I_\Omega(x). \tag{2.5}$$

2.3 Taylor's Theorem and Convexity

The foundational result for many algorithms in smooth nonlinear optimization is Taylor's theorem. This result shows how smooth functions can be approximated locally by low-order (linear or quadratic) functions. Note that this result does not require f to be a convex function!

Theorem 2.3. *Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, and given $x, p \in \mathbb{R}^n$, we have that*

$$f(x+p) = f(x) + \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma, \tag{2.6}$$

$$f(x+p) = f(x) + \nabla f(x+\gamma p)^T p, \quad \textit{some } \gamma \in (0,1). \tag{2.7}$$

If f is twice continuously differentiable, we have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p) p \, d\gamma, \tag{2.8}$$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+\gamma p) p, \quad \textit{some } \gamma \in (0,1). \tag{2.9}$$

(We sometimes call the relation (2.6) the "integral form" and (2.7) the "mean-value form" of Taylor's theorem.)

A consequence (2.7) is that for f continuously differentiable at x, we have

$$f(x+p) = f(x) + \nabla f(x)^T p + o(\|p\|). \tag{2.10}$$

We prove this claim by manipulating (2.7) as follows:

$$\begin{aligned} f(x+p) &= f(x) + \nabla f(x+\gamma p)^T p \\ &= f(x) + \nabla f(x)^T p + (\nabla f(x+\gamma p) - \nabla f(x))^T p \\ &= f(x) + \nabla f(x)^T p + O(\|\nabla f(x+\gamma p) - \nabla f(x)\| \|p\|) \\ &= f(x) + \nabla f(x)^T p + o(\|p\|), \end{aligned}$$

where the last step follows from continuity: $\nabla f(x+\gamma p) - \nabla f(x) \to 0$ as $p \to 0$, for all $\gamma \in (0,1)$.

For the remainder of this section, we assume that f is continuously differentiable and also *convex*. By applying Taylor's theorem to the left-hand side of the definition of convexity (2.3), we obtain

$$f(x) + \alpha \nabla f(x)^T (y-x) + O(\alpha^2) \leq (1-\alpha) f(x) + \alpha f(y).$$

By cancelling the $f(x)$ term, rearranging, and dividing by α, we obtain

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + O(\alpha),$$

so by letting $\alpha \downarrow 0$, we obtain

$$f(y) \geq f(x) + \nabla f(x)^T (y - x), \quad \text{for any } x, y \in \text{dom}(f), \tag{2.11}$$

which is a fundamental characterization of convexity of a smooth function. We defined "strong convexity with modulus m" in (2.4). When f is differentiable, we have the following equivalent definition, obtained by working on (2.4) with a similar argument to the one above:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}\|y - x\|^2. \tag{2.12}$$

Another crucial quantity is the Lipschitz constant L for the gradient of f, which is defined to satisfy

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \text{for all } x, y \in \text{dom}(f). \tag{2.13}$$

From (2.6), we have

$$f(y) - f(x) - \nabla f(x)^T (y - x) = \int_0^1 [\nabla f(x + \gamma(y - x)) - \nabla f(x)]^T (y - x)\, d\gamma.$$

By using (2.13), we have

$$[\nabla f(x + \gamma(y - x)) - \nabla f(x)]^T (y - x) \leq \|\nabla f(x + \gamma(y - x)) - \nabla f(x)\|\|y - x\| \leq L\gamma\|y - x\|^2.$$

By substituting this bound into the previous integral, we obtain

$$f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2}\|y - x\|^2. \tag{2.14}$$

By combining this expression with (2.12), we have proved the following result.

Lemma 2.4. *Given convex f satisfying (2.4), with ∇f uniformly Lipschitz continuous with constant L, we have for any $x, y \in \text{dom}(f)$ that*

$$\frac{m}{2}\|y - x\|^2 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{L}{2}\|y - x\|^2.$$

When f is *twice* continuously differentiable, we can characterize the constants m and L in terms of the eigenvalues of the Hessian $\nabla f(x)$. Specifically, we have

$$mI \preceq \nabla^2 f(x) \preceq LI, \quad \text{for all } x \tag{2.15}$$

as the following result proves.

Lemma 2.5. *Suppose f is twice continuously differentiable on $\mathbb{R}^n$. Then*

(a) f is strongly convex with modulus of convexity m if and only if $\nabla^2 f(x) \succeq mI$ for all x.

(b) ∇f is Lipschitz continuous with Lipschitz constant L if and only if $\nabla^2 f(x) \preceq LI$ for all x.

Proof. We first prove (a). m. Then for any $x, u \in \mathbb{R}^n$ and $\alpha > 0$, we have from Taylor's theorem that

$$f(x + \alpha u) = f(x) + \alpha \nabla f(x)^T + \frac{1}{2}\alpha^2 u^T \nabla^2 f(x + t\alpha u)u, \quad \text{for some } t \in (0, 1).$$

From the strong convexity property, we have

$$f(x + \alpha u) \geq f(x) + \alpha \nabla f(x)^T + \frac{m}{2}\alpha^2 \|u\|^2.$$

By comparing these two expressions, cancelling terms, and dividing by α^2, we obtain

$$u^T \nabla^2 f(x + t\alpha u)u \geq m\|u\|^2.$$

By taking $\alpha \downarrow 0$, we obtain $u^T \nabla^2 f(x)u \geq m\|u\|^2$, thus proving that $\nabla^2 f(x) \succeq mI$.

For the converse, suppose that $\nabla^2 f(x) \succeq mI$ for all x. Using the same form of Taylor's theorem as above, we obtain

$$f(z) = f(x) + \nabla f(x)^T (z - x) + \frac{1}{2}(z - x)^T \nabla^2 f(x + t(z - x))(z - x), \quad \text{for some } t \in (0, 1).$$

We obtain the strong convexity expression when we bound the last term as follows:

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

completing the proof of (a).

For (b), we assume first that ∇f is Lipschitz continuous with Lipschitz constant L. From (2.14), we have by setting $y = x + \alpha p$ for some $\alpha > 0$ that

$$f(x + \alpha p) - f(x) - \alpha \nabla f(x)^T p \leq \frac{L}{2}\alpha^2 \|p\|^2.$$

From formula (2.9) from Taylor's theorem, we have

$$f(x + \alpha p) - f(x) - \alpha \nabla f(x)^T p = \frac{1}{2}\alpha^2 p^T \nabla^2 f(x + \gamma\alpha p)p.$$

By comparing these two expressions, we obtain

$$p^T \nabla^2 f(x + \gamma\alpha p)p \leq L\|p\|^2.$$

By letting $\alpha \downarrow 0$, we have that all eigenvalues of $\nabla^2 f(x)$ are bounded by L, so that $\nabla^2 f(x) \preceq LI$, as claimed.

Suppose now that $\nabla^2 f(x) \preceq LI$. We have from (2.8) that

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| &= \left\| \int_{t=0}^{1} \nabla^2 f(x + t(y - x))(y - x)\, dt \right\| \\
&\leq \int_{t=0}^{1} \|\nabla^2 f(x + t(y - x))\| \|y - x\|\, dt \\
&\leq \int_{t=0}^{1} L\|y - x\|\, dt = L\|y - x\|,
\end{aligned}$$

as required. This completes the proof of (b). □

Strongly convex functions have unique minimizers, as we now show.

Theorem 2.6. *Let f be differentiable and strongly convex with modulus $m > 0$. Then the minimizer x^* of f exists and is unique.*

Proof. We show first that for any point x^0, the level set $\{x \mid f(x) \leq f(x^0)\}$ is closed and bounded, and hence compact. Suppose for contradiction that there is a sequence $\{x^\ell\}$ such that $\|x^\ell\| \to \infty$ and

$$f(x^\ell) \leq f(x^0). \tag{2.16}$$

By strong convexity of f, we have for some $m > 0$ that

$$f(x^\ell) \geq f(x^0) + \nabla f(x^0)^T(x^\ell - x^0) + \frac{m}{2}\|x^\ell - x^0\|^2.$$

By rearranging slghtly, and using (2.16), we obtain

$$\frac{m}{2}\|x^\ell - x^0\|^2 \leq -\nabla f(x^0)^T(x^\ell - x^0) \leq \|\nabla f(x^0)\|\|x^\ell - x^0\|.$$

By dividing both sides by $(m/2)\|x^\ell - x^0\|$, we obtain $\|x^\ell - x^0\| \leq (2/m)\|\nabla f(x^0)\|$ for all ℓ, which contradicts unboundedness of $\{x^\ell\}$. Thus, the level set is bounded. Since it is also closed (by continuity of f), it is compact.

Since f is continuous, it attains its minimum on the compact level set, which is also the solution of $\min_x f(x)$, and we denote it by x^*. Suppose for contradiction that the minimizer is not unique, so that we have two points x_1^* and x_2^* that minimize f. Obviously, these points must attain equal objective values, so that $f(x_1^*) = f(x_2^*) = f^*$. By taking (2.4) and setting $\phi = f$, $x = x_1^*$, $y = x_2^*$, and $\alpha = 1/2$, we obtain

$$f((x_1^* + x_2^*)/2) \leq \frac{1}{2}(f(x_1^*) + f(x_2^*)) - \frac{1}{8}m\|x_1^* - x_2^*\|^2 < f^*,$$

so the point $(x_1^* + x_2^*)/2$ has a smaller function value than both x_1^* and x_2^*, contradicting our assumption that x_1^* and x_2^* are both minimizers. Hence, the minimizer x^* is unique. □

We now prove several other (slightly trickier) technical results that are useful in subsequent analysis. We recall the definition of S to be the set of minimizers of the function f, and define P_S to be the Euclidean projection operator of a vector x onto this set, that is,

$$P_S(x) := \arg\min_z \frac{1}{2}\|z - x\|_2^2. \tag{2.17}$$

Lemma 2.7. *Given convex, uniformly Lipschitz continuously differentiable f (with Lipschitz constant L for ∇f), we have for any $x, y \in \mathrm{dom}(f)$ that the following bounds hold (see [29, Theorems 2.1.5 and 2.1.12]):*

$$f(x) + \nabla f(x)^T(y - x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y), \tag{2.18}$$

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))^T(x - y) \leq L\|x - y\|^2. \tag{2.19}$$

If, in addition, f is strongly convex with modulus m and unique minimizer x^*, we have for all $x, y \in \mathrm{dom}(f)$ that

$$f(y) - f(x) \geq -\frac{1}{2m}\|\nabla f(x)\|^2. \tag{2.20}$$

Proof. For (2.18), we define

$$\phi(y) := f(y) - \nabla f(x)^T y.$$

Note that ϕ is convex with $\nabla\phi(y) = \nabla f(y) - \nabla f(x)$, and that $\nabla\phi(x) = \nabla f(x) - \nabla f(x) = 0$, so that x is a minimizer of ϕ. By using the latter fact, and applying Lemma 2.4 to ϕ, we have

$$\begin{aligned}\phi(x) \leq \phi(y - (1/L)\nabla\phi(y)) &\leq \phi(y) + \nabla\phi(y)^T[(-1/L)\nabla\phi(y)] + \frac{L}{2}\|(-1/L)\nabla\phi(y)\|^2 \\ &= \phi(y) - \frac{1}{2L}\|\nabla\phi(y)\|^2.\end{aligned}$$

By substituting the definition of ϕ into this inequality, we obtain the result (2.18).

We obtain the left inequality in (2.19) by adding two copies of (2.18) with x and y interchanged. The right inequality in (2.19) follows from L being a Lipschitz constant for ∇f.

For (2.20), we have from Lemma 2.4 that

$$\begin{aligned}f(y) - f(x) &\geq \nabla f(x)^T(y-x) + \frac{m}{2}\|y-x\|^2 \\ &= \frac{m}{2}\left\|y - x + \frac{1}{m}\nabla f(x)\right\|^2 - \frac{1}{2m}\|\nabla f(x)\|^2 \\ &\geq -\frac{1}{2m}\|\nabla f(x)\|^2.\end{aligned}$$

□

The condition (2.20) plays an important role in the analysis of many methods in this book. By choosing y to be any solution of the problem $\min_x f(x)$, and defining f^* to be the optimal objective value for this problem, we have from (2.20) that

$$\|\nabla f(x)\|^2 \geq 2m[f(x) - f^*], \quad \text{for some } m > 0. \tag{2.21}$$

We call this condition the *generalized strong convexity* condition, and note that it holds in situations other than when f is strongly convex. One such situation is when f is the convex quadratic function

$$f(x) := \frac{1}{2}x^T A x - b^T x,$$

where A is a symmetric positive semidefinite matrix. The minimizers x^* of f satisfy the condition $\nabla f(x^*) = Ax^* - b = 0$, so when A is rank deficient, the solution set is either empty or else is the affine space $S = x^* + \text{null}(A)$, where $\text{null}(A)$ is the nullspace of A and x^* is a particular solution. When the rank of A is $r \leq n$, we can write the eigenvalue decomposition of A as $A = U\Lambda U^T$, where U is an $n \times r$ matrix with orthonormal solutions and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ contains the positive eigenvalues of A arranged in decreasing order. (In particular, $\lambda_r > 0$.) For any x, we have noting that $Ax^* = b$ for any solution x^* that

$$\begin{aligned}\|\nabla f(x)\|^2 = \|Ax - b\|^2 &= \|A(x - x^*)\|^2 \\ &= \|U\Lambda U^T(x - x^*)\|^2 \\ &= \|\Lambda U^T(x - x^*)\|^2 \\ &\geq \lambda_r\|\Lambda^{1/2}U^T(x - x^*)\|^2 \\ &= \lambda_r(x - x^*)^T U^T \Lambda U^T(x - x^*) \\ &= \lambda_r(x - x^*)^T A(x - x^*) = 2\lambda_r(f(x) - f^*),\end{aligned}$$

so (2.21) holds for $m = \lambda_r$. (Note that in the fourth equality we used the fact that $\|Uz\| = \|z\|$ for all z, where U is an $n \times r$ matrix with $r \leq n$ and orthonormal columns.)

We conclude by noting that when f is strongly convex and twice continuously differentiable, (2.9) implies the following, when x^* is the minimizer:

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|^2). \tag{2.22}$$

Thus, f behaves like a strongly convex *quadratic* function in a neighborhood of x^*. It follows that we can learn a lot about local convergence properties of algorithms just by studying convex quadratic functions.

2.4 Characterizing Optimality for Smooth Unconstrained Problems

The results of the previous section give us the tools needed to characterize solutions of the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{2.23}$$

where f is a smooth function. We refer here to the taxonomy of solution types developed in Section 2.1, and we will consider both convex and nonconvex functions f.

We start with *necessary* conditions, which give properties of the derivatives of f that are satisfied when x^* is a local solution. We have the following result.

Theorem 2.8 (Necessary Conditions for Smooth Unconstrained Optimization)**.**

(a) Suppose that f is continuously differentiable. Then if x^* is a local minimizer of (2.23)*, then* $\nabla f(x^*) = 0$.

(b) Suppose that f is twice continuously differentiable. Then if x^* is a local minimizer of (2.23)*, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite.*

Proof. We start by proving (a). Suppose for contradiction that $\nabla f(x^*) \neq 0$, and consider a step $-\alpha \nabla f(x^*)$ away from x^*, where α is a small positive number. By setting $p = -\alpha \nabla f(x^*)$ in formula (2.7) from Theorem 2.3, we have

$$f(x^* - \alpha \nabla f(x^*)) = f(x^*) - \alpha \nabla f \left(x^* - \gamma \alpha \nabla f(x^*)\right)^T \nabla f(x^*), \quad \text{for some } \gamma \in (0, 1). \tag{2.24}$$

Since ∇f is continuous, we have that

$$\nabla f \left(x^* - \gamma \alpha \nabla f(x^*)\right)^T \nabla f(x^*) \geq \frac{1}{2} \|\nabla f(x^*)\|^2,$$

for all α sufficiently small, and any $\gamma \in (0, 1)$. Thus by substituting into (2.24), we have that

$$f(x^* - \alpha \nabla f(x^*)) = f(x^*) - \frac{1}{2} \alpha \|\nabla f(x^*)\|^2 < f(x^*),$$

for all positive and sufficiently small α. This it is impossible to choose a neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$, so x^* is not a local minimizer.

We now prove (b). It follows immediately from (a) that $\nabla f(x^*) = 0$, so we need to prove only positive semidefiniteness of $\nabla^2 f(x^*)$. Suppose for contradiction that $\nabla^2 f(x^*)$ has a negative eigenvalue, so there exists a vector $v \in \mathbb{R}^n$ and a positive scalar λ such that $v^T \nabla^2 f(x^*) v \leq -\lambda$. We set $x = x^*$ and $p = \alpha v$ in formula (2.9) from Theorem 2.3, where α is a small positive constant, to obtain

$$f(x^+ \alpha v) = f(x^*) + \alpha \nabla f(x^*)^T v + \frac{1}{2}\alpha^2 v^T \nabla^2 f(x^* + \gamma \alpha v) v, \quad \text{for some } \gamma \in (0,1). \tag{2.25}$$

For all α sufficiently small, we have for λ defined above that $v^T \nabla^2 f(x^* + \gamma \alpha v) v \leq -\lambda/2$, for all $\gamma \in (0,1)$. By substituting this bound together with $\nabla f(x^*) = 0$ into (2.25), we obtain

$$f(x^+ \alpha v) = f(x^*) - \frac{1}{4}\alpha^2 \lambda < f(x^*),$$

for all sufficiently small, positive values of α. Thus there is no neighborhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ for all $x \in \mathcal{N}$, so x^* is not a local minimizer. Thus we have proved by contradiction that $\nabla^2 f(x^*)$ is positive semidefinite. □

Condition (a) in Theorem 2.8 is called the *first-order necessary condition*, because it involves the first-order derivatives of f. For obviou reasons, condition (b) is called the *second-order necessary condition*.

Theorem 2.8 holds for nonconvex functions f. When f is convex, the first-order necessary condition is actually a *sufficient* condition, as the following theorem shows.

Theorem 2.9. *Suppose that f is continuously differentiable and convex. Then if $\nabla f(x^*) = 0$, then x^* is a global minimizer of* (2.23)*. When, in addition, f is* strongly *convex, then x^* is the* unique *global minimizer.*

Proof. The proof of the first part follows immediately from condition (2.11), if we set $x = x^*$. Using this inequality together with $\nabla f(x^*) = 0$, we have for any y that

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

so that x^* is a global minimizer. For the second claim, we use (2.12) □

Returining to nonconvex f, we have the following *second-order sufficient condition.*

Theorem 2.10 (Sufficient Conditions for Smooth Unconstrained Optimization)**.** *Suppose that f is twice continuously differentiable and that for some x^*, we have $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite. Then x^* is a strict local minimizer of* (2.23)*.*

Proof. We use formula (2.9) from Taylor's theorem. Define a radius ρ sufficiently small and positive such that the eigenvalues of $\nabla^2 f(x^* + \gamma p)$ are bounded below by some positive number ϵ, for all $p \in \mathbb{R}^n$ with $\|p\| \leq \rho$, and all $\gamma \in (0,1)$. (Because $\nabla^2 f$ is positive definite at x^* and continuous, and because the eigenvalues of a matrix are continuous functions of the elements of a matrix, it is possible to choose $\rho > 0$ and $\epsilon > 0$ with these properties.) By setting $x = x^*$ in (2.9), we have

$$f(x^* + p) = f(x^*) + \nabla f(x^*)^T p + \frac{1}{2} p^T \nabla^2 f(x^* + \gamma p) p \geq f(x^*) + \frac{1}{2}\epsilon \|p\|^2, \quad \text{for all } p \text{ with } \|p\| \leq \rho.$$

thus by setting $\mathcal{N} = \{x^* + p \,|\, \|p\| < \rho\}$, we have found a neighborhood of x^* such that $f(x) > f(x^*)$ for all $x \in \mathcal{N}$ with $x \neq x^*$, thus satisfying the conditions for a strict local minimizer. □

Notation

We list key notational conventions that are used in the rest of the book.

- We use $\|\cdot\|$ to denote the Euclidean norm $\|\cdot\|_2$ of a vector in $\mathbb{R}^n$. Other norms, such as $\|\cdot\|_1$ and $\|\cdot\|_\infty$, will be denoted explicitly.
- Given two sequences of nonnegative scalars $\{\eta_k\}$ and $\{\zeta_k\}$, with $\zeta_k \to \infty$, we write $\eta_k = O(\zeta_k)$ if there exists a constant M such that $\eta_k \leq M\zeta_k$ for all k sufficiently large. The same definition holds if $\zeta_k \to 0$.
- For sequences $\{\eta_k\}$ and $\{\zeta_k\}$ as above, we write $\eta_k = o(\zeta_k)$ if $\eta)k/\zeta_k \to 0$ as $k \to \infty$. We write $\eta_k = \Omega(\zeta_k)$ if both $\eta_k = O(\zeta_k)$ and $\zeta_k = O(\eta_k)$.

Sources and Further Reading

Further background on Moreau envelopes and the proximal mapping is given in [33].

Exercises

1. Prove that the effective domain of a convex function is a convex set.
2. Prove that epi f is a convex subset of $\mathbb{R}^{n+1}$ for any convex function f.
3. Show that I_Ω is a convex function if and only if Ω is a convex set.
4. Show that Ω is a nonempty closed convex set if and only if $I_\Omega(x)$ is a closed proper convex function.
5. Show rigorously how (2.12) is derived from (2.4) when f is continuously differentiable.
6. Prove Theorem 8.16 by doing a careful generalization of the proof of Theorem 2.6.