Chapter 1

Introduction

In this book we deal with minimization of real-valued functions of real variables over convex domains. This is a broad class of problems that admits many interesting subclasses, as well as many instances of pathological structure. Our focus is on the types of optimization formulations that arise in data analysis and machine learning. We discuss the properties of specific problems in this area later, but let us focus first on important features that they all share.

- They can be formulated naturally as functions of *real variables*, which we typically arrange in a vector $\mathbb{R}^n$. (We do not consider explicitly the important class of data analysis problems in which the unknowns are matrices.)

- We do not consider problems in which the variables are restricted to integer or binary (zero/one) values. The number of possible values for the full set of variables is usually very large in these problems, and they are typically "hard" in a computational-complexity sense, as well as in practice. A great deal of research has been done on heuristics to find solutions of these problems.

- We often assume that the functions are continuous and even smooth, and that when nonsmoothness appears in the formulation, it does so in a structured way that can be exploited by the algorithm. Smoothness properties allow an algorithm to make good inferences about the behavior of the function on the basis of knowledge gained at points that have been visited already.

- The objective is often made up in part of a summation of many terms, where each term depends on just a subset of the data. A common structure for the objective is a sum of two terms: a "loss term" (sometimes arising from a maximum likelihood expression for some statistical model) and a "regularization term" whose purpose is to impose structure and "generalizability" on the recovered model, avoiding overfitting to the data available. (The data is viewed as a sample of some underlying infinite data set, so the function we are minimizing is an empirical estimate of some unknown, fundamental underlying function.)

This book will some several fundamental approaches to nonlinear optimization problems with the properties above, focusing on the methods and their convergence properties.

1.1 Examples

Sketch several examples from ML applications, definitely including primal SVM classification, logistic regression, least squares and robust regression. Include ℓ_1 regularization and possibly other regularizers. Probably also include kernel SVM, multiclss logistic regression. Include graph structures in the formulations.

We could just sketch these problems here and briefly outline their relevance. Perhaps could develop them further in "topics" chapters at the end.

1.2 Basic Concepts

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

$$\min_{x \in \Omega} f(x), \tag{1.1}$$

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 (1.1) 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 (1.1) if $f(x) \geq f(x^*)$ for all $x \in \Omega$.

1.3 Convexity

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

$$x, y \in \Omega \quad \Rightarrow \quad (1-\alpha)x + \alpha y \in \Omega \;\text{ for all } \alpha \in [0,1]. \tag{1.2}$$

Need a figure. A *supporting hyperplane* for the set Ω at a point in the set $\bar{x} \in \Omega$ is defined by a nonzero vector $g \in \mathbb{R}^n$ with the property that

$$g^T(x - \bar{x}) \leq 0, \quad \text{for all } x \in \Omega.$$

The convex sets that we consider in this book are usually *closed.* We have the following definition or normal cones, which is key to recognizing optimality.

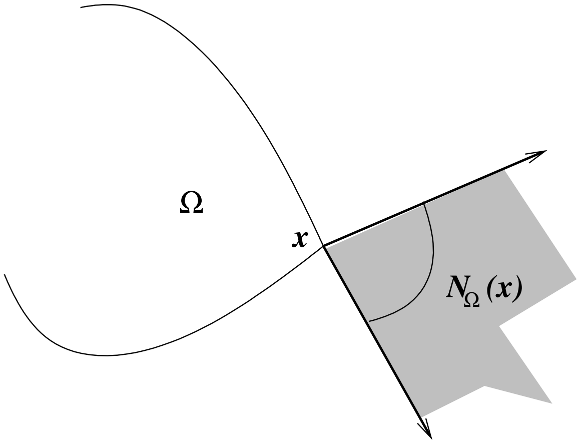

Figure 1.1: Normal Cone

Definition 1.1. *Let $\Omega \subset \mathbb{R}^n$ be a convex set. At any $x \in \Omega$ we the* normal cone $N_\Omega(x)$ *is:*

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

(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 1.1 for an example.

In optimization, we often deal with sets that are intersections of closed convex sets. We have the following result for normal cones of such sets.

Theorem 1.2. *Let Ω_i, $i = 1, 2, \ldots, m$ be convex sets and let $\Omega = \cap_{i=1,2,\ldots,m} \Omega_i$. Then for $x \in \Omega$, we have*

$$N_\Omega(x) \supset N_{\Omega_1}(x) + N_{\Omega_2}(x) + \ldots + N_{\Omega_m}(x). \tag{1.3}$$

Proof. Consider vectors $v_i \in N_{\Omega_i}(x)$ for all $i = 1, 2, \ldots, m$, and define $v := \sum_{i=1}^m v_i$. Let z be any point in the intersection $\Omega = \cap_{i=1}^m \Omega_i$. Since $z \in \Omega_i$, we have $v_i^T(z-x) \le 0$ for all $i = 1, 2, \ldots, m$, so that $v^T(z-x) = (\sum_{i=1}^m v_i)^T(z-x) \le 0$, and thus $v \in N_\Omega(x)$. □

The following is an example for which strict inclusion holds in (1.3). Define the following two convex subsets of $\mathbb{R}^2$:

$$\Omega_1 := \{x \in \mathbb{R}^2 \,:\, x_1 \le 0\}, \quad \Omega_2 := \{x \in \mathbb{R}^2 \,:\, (x_1-1)^2 + x_2^2 \le 1\}, \tag{1.4}$$

for which clearly $\Omega_1 \cap \Omega_2 = \{0\}$. The normal cones at the interesting point 0 are

$$N_{\Omega_1}(0) = \left\{ \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \,:\, v_1 \ge 0 \right\}, \quad N_{\Omega_2}(0) = \left\{ \begin{bmatrix} v_1 \\ 0 \end{bmatrix} \,:\, v_1 \le 0 \right\}, \quad N_{\Omega_1 \cap \Omega_2}(0) = \mathbb{R}^2.$$

Since $N_{\Omega_1}(0) + N_{\Omega_1}(0) = \mathbb{R} \times \{0\}$, strict inclusion holds. See Figure 1.2.

Additional conditions are sometimes assumed to ensure that equality holds in (1.3); these conditions are called *constraint qualifications.* Some constraint qualifications are expressed in terms of the geometry of the sets while others focus on their algebraic descriptions. One common theme

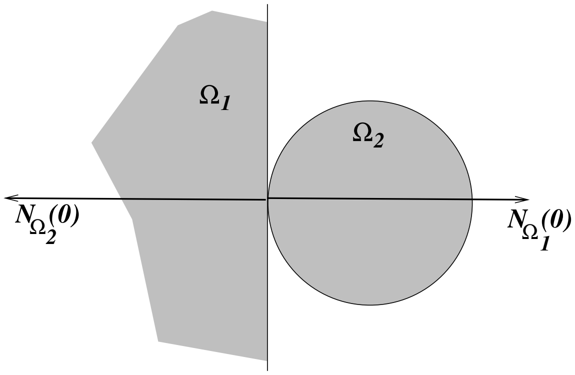

Figure 1.2: Example for which strict inclusion holds in (1.3).

among constraint qualifications is that a linear approximation of the sets near the point in question needs to capture the essential geometry of the set itself in a neighborhood of the point. This is not true of the example above, where the tangents (linear approximations) to Ω_1 and Ω_2 are the vertical axis (so the intersection of their linear approximations is also the vertical axis), while the intersection of the two sets is the single point $\{0\}$.

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.

A convex function $\phi : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ maps $\mathbb{R}^n$ to the extended reals, which is the set of real numbers augmented by $+\infty$ and $-\infty$, denoted (unsurprisingly) by $\mathbb{R} \cup \{\pm\infty\}$. The defining property of a convex function is that

$$\phi((1-\alpha)x + \alpha y) \le (1-\alpha)\phi(x) + \alpha\phi(y), \quad \text{for all } x, y \in \mathbb{R}^n \text{ and all } \alpha \in [0,1]. \tag{1.5}$$

The following definitions are useful.

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

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

  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 $x \in \Omega$ and $\phi(x) > -\infty$ for all $x \in \Omega$. This class encompasses almost all convex functions of practical interest.
- ϕ is a *closed proper* convex function if it is a proper convex function and the set $\{x \in \Omega : \phi(x) \le \bar{t}\}$ 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 simplier than for the general case. In particular, the distinction between "local" and "global" solutions goes away, as we show now.

Theorem 1.3. *Suppose that in* (1.1)*, the function f is convex and the set Ω is closed and convex. We have the following.*

(a) Any local solution of (1.1) *is also a global solution.*

(b) The set of global solutions of (1.1) *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 $\mu > 0$ such that

$$\phi((1-\alpha)x + \alpha y) \leq (1-\alpha)\phi(x) + \alpha\phi(y) - \frac{1}{2}\mu\alpha(1-\alpha)\|x - y\|_2^2 \tag{1.6}$$

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

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

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

Indicator functions are useful devices for deriving optimality conditions for constrained problems, and even for developing algorithms. The constrained optimization problem (1.1) can be restated equivalently as follows:

$$\min f(x) + I_\Omega(x). \tag{1.7}$$

1.4 Subgradients

We turn now to *subgradients*, which generalize the concept of a gradient to a (possibly nonsmooth) convex function, and which are instrumental in deriving first-order methods. We assume throughout this section that f is a closed, proper, convex function.

The subgradient and subdifferential of are defined as follows.

Definition 1.4. *A vector $v \in \mathbb{R}^n$ is a* subgradient *of f at a point x if*

$$f(x+d) \geq f(x) + v^T d. \quad \textit{for all } d \in \mathbb{R}^n.$$

The subdifferential*, denoted $\partial f(x)$, is the set all subgradients of f at x.*

Each subgradient can be identified with a supporting hyperplane to the epigraph of f. We have the following result.

Theorem 1.5. *$g \in \partial f(x)$ if and only if $(g, -1)$ is a supporting hyperplane to* epi *f at the point $(x, f(x))$.*

Proof. Given a supporting hyperplane defined by $(g, -1)$ at $(x, f(x))$, we have for any y that $(y, f(y)) \in \operatorname{epi} f$ and therefore

$$g^T(y-x) - (f(y) - f(x)) \leq 0 \;\Leftrightarrow\; f(y) \geq f(x) + g^T(y-x),$$

which implies that $g \in \partial f(x)$. The converse follows by reversing the argument. □

We can easily characterize a minimum in terms of the subdifferential.

Theorem 1.6. *The point x^* is the minimizer of a convex function f if and only if $0 \in \partial f(x^*)$.*

Proof. Suppose that $0 \in \partial f(x^*)$, we have by substituting $x = x^*$ and $v = 0$ into Definition 1.4 that $f(x^* + d) \geq f(x^*)$ for all $d \in \mathbb{R}^n$, which implies that x^* is a minimizer of f. The converse follows trivially by showing that $v = 0$ satisfies Definition 1.4 when x^* is a minimizer. □

Subdifferentials satisfy a *monotonicity* property, as shown in the following result.

Lemma 1.7. *If $a \in \partial f(x)$ and $b \in \partial f(y)$, we have $(a-b)^T(x-y) \geq 0$.*

Proof. From convexity of f and the definitions of a and b, we have $f(y) \geq f(x) + a^T(y-x)$ and $f(x) \geq f(y) + b^T(x-y)$. The result follows by adding these two inequalities. □

The subdifferential generalizes the concept of derivative of a smooth function.

Theorem 1.8. *If f is convex and differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. Differentiability of f implies that for all unit vectors d in $\mathbb{R}^n$ (that is, $\|d\| = 1$) and all scalars t, we have $f(x+td) = f(x) + t\nabla f(x)^T d + o(|t|)$. For all $v \in \partial f(x)$, we have from this fact and Definition 1.4 that for all unit vectors d and for $t > 0$, we have

$$\begin{aligned} f(x+td) &= f(x) + t\nabla f(x)^T d + o(t) \geq f(x) + tv^T d \\ f(x-td) &= f(x) - t\nabla f(x)^T d + o(t) \geq f(x) - tv^T d. \end{aligned}$$

By combining these expressions and dividing by t, we have

$$v^T d + o(t)/t \leq \nabla f(x)^T d \leq v^T d + o(t)/t, \quad \text{for all unit vectors } d,$$

which, by taking $t \downarrow 0$ and using the fact that d is arbitrary, implies that $\nabla f(x) = v$. □

A converse of this result is also true. Specifically, if the subdifferential of a convex function f at x contains a single subgradient, then f is differentiable with gradient equal to this subgradient (see [21, Theorem 25.1]).

Theorem 1.9. *For a convex set $\Omega \subset \mathbb{R}^n$, we have that $N_\Omega(x) = \partial I_\Omega(x)$ for all $x \in \Omega$.*

Proof. Given $v \in N_\Omega(x)$, we have

$$I_\Omega(y) - I_\Omega(x) = 0 - 0 = 0 \geq v^T(y-x), \quad \text{for all } y \in \Omega,$$

and

$$I_\Omega(y) - I_\Omega(x) = \infty - 0 = \infty \geq v^T(y-x), \quad \text{for all } y \notin \Omega.$$

It follows from Definition 1.4 that $v \in \partial I_\Omega(x)$. Supposing now that $v \in \partial I_\Omega(x)$, we have

$$0 = I_\Omega(y) \geq I_\Omega(x) + v^T(y-x) = v^T(y-x), \quad \text{for all } y \in \Omega,$$

which implies that $v \in N_\Omega(x)$, completing the proof. □

Some basic rules of calculus for subdifferentials are easily proved. Supposing that f and g are convex functions and α is a positive scalar, the following are true.

$$\partial(f_1 + f_2)(x) \supset \partial f_1(x) + \partial f_2(x), \tag{1.8}$$

$$\partial(\alpha f)(x) = \alpha \partial f(x). \tag{1.9}$$

The relationship in (1.8) is not an equality in general. In fact, we already saw an example in which the inclusion is strict in the discussion following Theorem 1.2; consider $f_1(x) = I_{\Omega_1}(x)$ and $f_2(x) = I_{\Omega_2}(x)$ for the closed convex sets Ω_1 and Ω_2 defined in (1.4). Additional conditions (discussed later in ???) are needed to ensure equality in (1.8). *Need a Slater-type condition. Also see paper [6].*

Prove that Subdifferential is nonempty and bounded. (Uses supporting hyperplane theorem - should this go in the appendix?)

1.5 Proximal Operators and the Moreau Envelope

For a closed proper convex function h and a positive scalar λ, we define the *Moreau envelope* as

$$M_{\lambda,h}(x) := \inf_u \left\{ h(u) + \frac{1}{2\lambda}\|u-x\|^2 \right\} = \frac{1}{\lambda} \inf_u \left\{ \lambda h(u) + \frac{1}{2}\|u-x\|^2 \right\}. \tag{1.10}$$

The prox-operator of the function λh is the value of u that achieves the infimum in (1.10), that is,

$$\text{prox}_{\lambda h}(x) := \arg\min_u \left\{ \lambda h(u) + \frac{1}{2}\|u-x\|^2 \right\}. \tag{1.11}$$

Note that from optimality properties, we have from (1.11) that

$$0 \in \lambda \partial h(\text{prox}_{\lambda h}(x)) + (\text{prox}_{\lambda h}(x) - x). \tag{1.12}$$

The Moreau envelope can be seen as a smoothing of regularization of the function h. It has a finite valute for all x, even when h take on infinite values for some $x \in \mathbb{R}^n$. In fact, it is differentiable everywhere: Its gradient is

$$\nabla M_{\lambda,h}(x) = \frac{1}{\lambda}(x - \text{prox}_{\lambda h}(x)).$$

It is easy to check moreover that x^* is a minimizer of h if and only if it is a minimizer of $M_{\lambda,h}$.

The proximal operator satisfies a nonexpansiveness property. From the optimality conditions (1.12) at two points x and y, we have

$$x - \text{prox}_{\lambda h}(x) \in \lambda\partial(\text{prox}_{\lambda h}(x)), \quad y - \text{prox}_{\lambda h}(y) \in \lambda\partial(\text{prox}_{\lambda h}(y)).$$

By applying monotonicity (Lemma 1.7), we have

$$(1/\lambda)\big((x - \text{prox}_{\lambda h}(x)) - (y - \text{prox}_{\lambda h}(y))\big)^T(\text{prox}_{\lambda h}(x) - \text{prox}_{\lambda h}(y)) \geq 0,$$

which by rearrangement and application of the Cauchy-Schwartz inequality yields

$$\|\text{prox}_{\lambda h}(x) - \text{prox}_{\lambda h}(y)\|^2 \leq (x-y)^T(\text{prox}_{\lambda h}(x) - \text{prox}_{\lambda h}(y)) \leq \|x - y\| \, \|\text{prox}_{\lambda h}(x) - \text{prox}_{\lambda h}(y)\|,$$

from which we obtain $\|\text{prox}_{\lambda h}(x) - \text{prox}_{\lambda h}(y)\| \leq \|x - y\|$, as claimed.

We note several special cases of the prox operator which are useful in later chapters.

- $h(x) = 0$ for all x, for which we have $\text{prox}_{\lambda h}(x) = 0$. (Though trivial, this observation is useful in proxing that the prox-gradient method of Chapter 8 reduces to the familiar steepest descent method when the objective contains no regularization term.)

- $h(x) = I_\Omega(x)$, the indicator function for a closed convex set Ω. In this case, we have for any $\lambda > 0$ that

$$\text{prox}_{\lambda I_\Omega}(x) = \arg\min_u \left\{ \lambda I_\Omega(u) + \frac{1}{2}\|u - x\|^2 \right\} = \arg\min_{u \in \Omega} \frac{1}{2}\|u - x\|^2,$$

  which is simply the projection of x onto the set Ω.

- $h(x) = \|x\|_1$. By substituting into definition (1.11) we see that the minimization separates into its n separate components, and that the ith component of $\text{prox}_{\lambda\|\cdot\|_1}$ is

$$[\text{prox}_{\lambda\|\cdot\|_1}]_i = \arg\min_{u_i} \left\{ \lambda|u_i| + \frac{1}{2}(u_i - x_i)^2 \right\}.$$

  It is not hard to verify that

$$[\text{prox}_{\lambda\|\cdot\|_1}(x)]_i = \begin{cases} x_i - \lambda & \text{if } x_i > \lambda; \\ 0 & \text{if } x_i \in [-\lambda, \lambda]; \\ x_i + \lambda & \text{if } x_i < -\lambda, \end{cases} \tag{1.13}$$

  an operator that is known as *soft-thresholding.*

- $h(x) = \|x\|_0$, where $\|x\|_0$ denotes the *cardinality* of the vector x, its number of nonzero components. Although this h is not a convex function (as we can see by considering convex combinations of the vectors $(0,1)^T$ and $(1,0)^T$ in $\mathbb{R}^2$), its prox-operator is well defined to be the *hard thresholding* operation:

$$[\mathrm{prox}_{\lambda\|\cdot\|_0}(x)]_i = \begin{cases} x_i & \text{if } |x_i| \geq \sqrt{2\lambda}; \\ 0 & \text{if } |x_i| < \sqrt{2\lambda}. \end{cases}$$

For the cardinality function, the definition (1.11) separates into n individual components, and the fixed price of λ for allowing u_i to be nonzero is not worth paying unless $|x_i| \geq \sqrt{2\lambda}$.

1.6 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 1.10. *Given a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, and given $x, p \in \mathbb{R}^n$, we have that*

$$f(x+p) = f(x) + \int_0^1 \nabla f(x+\gamma p)^T p \, d\gamma, \tag{1.14}$$

$$f(x+p) = f(x) + \nabla f(x+\gamma p)^T p, \quad \text{some } \gamma \in (0,1). \tag{1.15}$$

If f is twice continuously differentiable, we have

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+\gamma p) p \, d\gamma, \tag{1.16}$$

$$f(x+p) = f(x) + \nabla f(x)^T p + \frac{1}{2} p^T \nabla^2 f(x+\gamma p) p, \quad \text{some } \gamma \in (0,1). \tag{1.17}$$

(We sometimes call the relation (1.14) the "integral form" and (A.6) the "mean-value form" of Taylor's theorem.)

For the remainder of this section, we assume that f is continuously differentiable and also *convex*. The definition of convexity (1.5) and the fact that $\partial f(x) = \{\nabla f(x)\}$ implies that

$$f(y) \geq f(x) + \nabla f(x)^T (y-x), \quad \text{for any } x, y \in \mathrm{dom}(f). \tag{1.18}$$

We defined *strong convexity with modulus μ* in (1.6). When f is differentiable, we have the following equivalent definition:

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2, \tag{1.19}$$

Another crucial quantity is the Lipschitz constant for the gradient of f, which is the nonnegative number L such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \quad \text{for all } x, y \in \mathrm{dom}(f). \tag{1.20}$$

From (1.14), 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 (1.20), we have

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

and therefore

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

By combining this expression with (1.19), we have proved the following result.

Lemma 1.11. *Given convex f satisfying (1.6), with ∇f uniformly Lipschitz continuous with constant L, we have for any $x, y \in \mathrm{dom}(f)$ that*

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

When f is *twice* continuously differentiable, we can chancterize the constants μ and L defined above as bounds on the eigenvalues of the Hessian $\nabla f(x)$, that is,

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \tag{1.21}$$

We have from (A.7) that

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

thus verifying that the definition of L in (1.21) is consistent with (1.20).

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,

$$P_S(x) := \arg\min_z \frac{1}{2}\|z-x\|_2^2. \tag{1.22}$$

Lemma 1.12. *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 [13, Theorems 2.1.5 and 2.1.12]):*

$$f(x) + \nabla f(x)^T(y-x) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le f(y), \tag{1.23}$$

$$\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 \le (\nabla f(x) - \nabla f(y))^T(x-y) \le L\|x-y\|^2. \tag{1.24}$$

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

$$f(y) - f(x) \ge -\frac{1}{2\mu}\|\nabla f(x)\|^2. \tag{1.25}$$

Proof. For (1.23), 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 1.11 to ϕ, we have

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

By substituting the definition of ϕ into this inequality, we obtain the result (1.23).

We obtain the left inequality in (1.24) by adding two copies of (1.23) with x and y interchanged. The right inequality in (1.24) follows from L being a Lipschitz constant for ∇f.

For (1.25), we have from Lemma 1.11 that

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

□

The condition (1.25) 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 (1.25) that

$$\|\nabla f(x)\|^2 \ge 2\mu[f(x) - f^*]^2, \quad \text{for some } \mu > 0. \tag{1.26}$$

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 null(A) is the nullspace of A and x^* is a particular solution. When the rank of A is $r \le n$, 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

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

so (1.26) holds for $\mu = \lambda_r$. (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 $r \le n$ and orthonormal columns.)

We conclude by noting that when f is strongly convex and twice continuously differentiable, (A.8) implies the following, when x^* is the minimizer:

$$f(x) - f(x^*) = \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|^2). \tag{1.27}$$

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 propoerties of algorithms just by studying convex quadratic functions.

1.7 First-Order Optimality Conditions

We now consider first-order optimality conditions for functions that are the sum of a smooth function and a convex, possibly nonsmooth function. This is a type of objective that we encounter often in later chapters, and in many machine learning applications.

Consider the composite function

$$\phi(x) := f(x) + \psi(x), \tag{1.28}$$

where f is differentiable and ψ is convex. We deal first with the case in which f is convex too.

Theorem 1.13. *When f is convex and differentiable and ψ is convex, the point x^* is a minimizer of ϕ defined in* (1.28) *if and only if* $0 \in \nabla f(x^*) + \partial\phi(x^*)$.

Proof. By Theorem 1.8, we have that

$$\partial\phi(x) = \partial f(x) + \partial\psi(x) = \nabla f(x) + \partial\psi(x).$$

The result follows immediately from Theorem 1.6. □

Corollary 1.14. *Consider the constrained optimization problem* (1.1)*, where $\Omega \subset \mathbb{R}^n$ is closed and convex and f is convex and differentiable. Then $x^* \in \Omega$ is a solution of* (1.1) *if and only if $-\nabla f(x^*) \in N_\Omega(x^*)$.*

Proof. We define

$$\phi(x) := f(x) + I_\Omega(x),$$

which has the form (1.28). By combining Theorem 1.13 with the characterization of ∂I_Ω in Theorem 1.9, we write the optimality condition as

$$0 \in \partial\phi(x^*) = \nabla f(x^*) + \partial I_\Omega(x^*) = \nabla f(x^*) + N_\Omega(x^*),$$

proving the claim. □

When f is strongly convex, the problem (1.28) has a minimizer and it is unique.

Theorem 1.15. *Suppose that the conditions of Theorem 1.13 hold, and in addition that f is strongly convex. Then the problem* (1.28) *has a unique solution.*

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

$$f(x^\ell) + \psi(x^\ell) \leq f(x^0) + \psi(x^0). \tag{1.29}$$

By convexity of ψ, we have that $\psi(x^\ell) \geq \psi(x^0) + g^T(x^\ell - x^0)$ for some g. By strong convexity of f, we have for some $\mu > 0$ that

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

By substituting these relationships in (1.29), and rearranging slghtly, we obtain

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

By dividing both sides by $(\mu/2)\|x^\ell - x^0\|$, we obtain $\|x^\ell - x^0\| \leq (2/\mu)\|\nabla f(x^0) + g\|$ for all ℓ, which contradicts unboundedness of $\{x^\ell\}$. Thus, the level set is bounded.

Since ϕ is continuous, it attains its minimum on the level set, which is also the solution of $\min_x \phi(x)$, and we denote it by x^*. By Theorem 1.13, we have that there is $g \in \partial\phi(x^*)$ such that $0 = \nabla f(x^*) + g = 0$. By strong convexity of f, we have for any $x \neq x^*$ that

$$f(x) + \psi(x) \leq f(x^*) + \psi(x^*) + (\nabla f(x^*) + g)^T(x - x^*) + \frac{\mu}{2}\|x - x^*\|^2 > f(x^*),$$

proving that x^* is the *unique* minimizer. □

For the more general case in which f is possibly nonconvex, we have a first-order necessary condition.

Theorem 1.16. *Suppose that f is differentiable and ψ is convex, and let ψ be defined by* (1.28). *Then if x^* is a local minimizer of ψ, we have that $0 \in \nabla f(x^*) + \partial\psi(x^*)$.*

Proof. Supposing that $0 \notin \nabla f(x^*) + \partial\psi(x^*)$, we show that x^* cannot be a local minimizer. We define the following convex function approximation to $\phi(x + d)$:

$$\bar{\phi}(d) := f(x^*) + \nabla f(x^*)^T d + \psi(x^* + d),$$

By differentiability of f we have that for all $\alpha \in [0, 1]$ and for any d that $\bar{\phi}(\alpha d) = \phi(x + \alpha d) + o(\alpha|d|)$. Since by assumption $0 \notin \partial\bar{\phi}(0) = \nabla f(x^*) + \partial\psi(x^*)$, we have from Theorem 1.6 that 0 is not a minimizer of $\bar{\phi}(d)$. Hence there exists $\bar{d}$ with $\bar{\phi}(\bar{d}) < \bar{\phi}(0)$, so that the quantity $c := \bar{\phi}(0) - \bar{\phi}(\bar{d})$ is strictly positive. By convexity of $\bar{\phi}$, we have for all $\alpha \in [0, 1]$ that

$$\bar{\phi}(\alpha\bar{d}) \leq \bar{\phi}(0) - \alpha(\bar{\phi}(0) - \bar{\phi}(\bar{d})) = \phi(x^*) - \alpha c,$$

and therefore

$$\phi(x^* + \alpha\bar{d}) \leq \phi(x^*) - \alpha c + o(\alpha|d|).$$

Therefore $\phi(x^* + \alpha\bar{d}) < \phi(x^*)$ for all $\alpha > 0$ sufficiently small, so x^* is not a local minimizer of ϕ. □

1.8 Outline

In Chapters 2 and 3, we consider minimization of a smooth, convex function f, via methods that require evaluation of a gradient $\nabla f(x)$ (for some x) at each iteration. We assume that the gradient can be evaluated exactly, to within the limits of floating-point computation. The algorithms we describe in these chapters can in many cases be extended to more general problems, for example,

- Objectives consisting of a smooth convex term plus a nonconvex regularization term
- Minimization of smooth functions over simple constraint sets, such as bounds on the components of x;
- Functions for which f of ∇f cannot be evaluated exactly without a complete sweep through the data set, but unbiased estimates of ∇f can be obtained easily.
- Situations in which it is much cheaper to evaluate an individual component or a subvector of ∇f than the full gradient vector.
- Smooth but nonconvex f.

Insert more as chapter frameworks are added.

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 $\eta_k \le M\zeta_k$ 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 $\zeta_k = O(\eta_k)$.
- $S\mathbb{R}^{n\times n}$ denotes the set of symmetric positive definite real $n \times n$ matrices.

Sources and Further Reading

Further background on Moreau envelopes and the proximal mapping is given in [16].

Exercises

1. Prove that the effective domain of a convex function is a convex set.

2. Show that I_Ω is a convex function if and only if Ω is a convex set.

3. Show that Ω is a nonempty closed convex set if and only if $I_\Omega(x)$ is a closed proper convex function.

4. Prove the rules of calculus (1.8) and (1.9) for convex functions.

5. Show that the subdifferential $\partial f(x)$ of a convex function f is a closed convex set, for all x.

6. Suppose that f is defined as a maximum of m convex functions, that is, $f(x) := \max_{i=1,2,\ldots,m} f_i(x)$, where each f_i is convex. Show that
$$\partial f(x) = \left\{ \sum_{i\,:\,f_i(x)=f(x)} \lambda_i v_i \; : \; v_i \in \partial f_i(x), \; \lambda_i \geq 0, \sum_{i\,:\,f_i(x)=f(x)} \lambda_i = 1 \right\}.$$

7. Show that a closed proper convex function h and its Moreau envelope $M_{\lambda,h}$ have identical minimizers.

8. Calculate $\mathrm{prox}_{\lambda h}(x)$ and $M_{\lambda,h}(x)$ for $h(x) = \frac{1}{2}\|x\|_2^2$.