Chapter 2

Smooth minimization

In this chapter, we consider the problem of minimizing a smooth function on a Euclidean space $\mathbf{E}$. Such problems are ubiquitous in computation mathematics and applied sciences. Before we delve into the formal development, it is instructive to look at some specific and typical examples that will motivate much of the discussion. We will often refer to these examples in latter parts of the chapter to illustrate the theory and techniques.

Example 2.1 (Linear Regression)**.** Suppose we wish to predict an output $b \in \mathbf{R}$ of a certain system on an input $a \in \mathbf{R}^n$. Let us also make the following two assumptions: (i) the relationship between the input a and the output b is fairly simple and (ii) we have available examples $a_i \in \mathbf{R}^n$ together with inexactly observed responses $b_i \in \mathbf{R}$ for $i = 1, \ldots, m$. Taken together, $\{(b, a_i)\}_{i=1}^m$ is called the training data.

Linear regression is an important example, where we postulate a linear relationship between the examples a and the response b. Trying to learn such a relationship from the training data amounts to finding a weight vector $x \in \mathbb{R}^{n+1}$ satisfying

$$b_i \approx x_0 + \langle a_i, x \rangle \qquad \text{for each } i = 1, \ldots, m.$$

To simplify notation, we may assume that the examples a_i lie in $\mathbf{R}^{n+1}$ with the first coordinate of a_i equal to one, so that we can simply write $b_i \approx \langle a_i, x \rangle$. The linear regression problem then takes the form

$$\min_x \ \sum_{i=1}^n \tfrac{1}{2} |\langle a_i, x \rangle - b_i|^2 = \tfrac{1}{2} \|Ax - b\|^2,$$

where $A \in \mathbb{R}^{m \times (n+1)}$ is a matrix whose rows are the examples a_i and b is the vector of responses. The use of the squared l_2-norm as a measure of misfit is a choice here. Other measures of misfit are often more advantageous from a modeling viewpoint – more on this later.

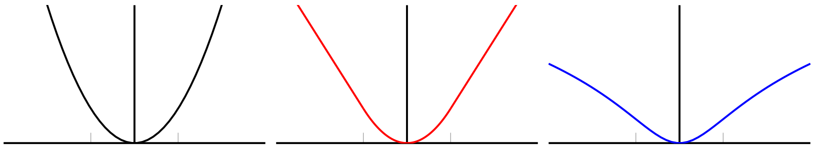

Figure 2.1: Least squares, Huber, and Student's t penalties. All three take (nearly) identical values for small inputs, but Huber and Student's t penalize larger inputs less harshly than least squares.

Example 2.2 (Ridge regularization)**.** The linear regression problem always has a solution, but the solution is only unique if A has a trivial null-space. To obtain unique solutions, as well as to avoid "over-fitting", *regularization* is often incorporated in learning problems. The simplest kind of regularization is called Tikhonov regularization or ridge regression:

$$\min_x \ \tfrac{1}{2}\|Ax - b\|^2 + \lambda\|x - x_0\|^2,$$

where $\lambda > 0$ is a regularization parameter that must be chosen, and x_0 is most often taken to be the zero vector.

Example 2.3 (Robust Regression)**.** While least squares is a good criterion in many cases, it is known to be vulnerable to outliers in the data. Therefore other smooth criteria ρ can be used to measure the discrepancy between b_i and $\langle a_i, x\rangle$:

$$\min_x \ \sum_{i=1}^{m} \rho(\langle a_i, x\rangle - b_i).$$

Two common examples of robust penalties are:

- Huber: $\rho_\kappa(z) = \begin{cases} \frac{1}{2}\|z\|^2 & |z| \le \kappa \\ \kappa|z| - \frac{1}{2}\kappa^2 & |z| > \kappa. \end{cases}$
- Student's t: $\rho_\nu(z) = \log(\nu + z^2)$.

Note that both (nearly) agree with $\frac{1}{2}\|z\|^2$ for small values of z, but penalize larger z less harshly, see Figure 2.1.

Example 2.4 (General Linear Models)**.** The use of the squared l_2-norm in linear regression (Example 2.1) was completely ad hoc. Let us see now how statistical modeling dictates this choice and leads to important extensions. Suppose that the observed response b_i is a realization of a random variable $\mathbf{b}_i$, which is indeed linear in the input vector up to an additive statistical error. That is, assume that there is a vector $x \in \mathbf{R}^{n+1}$ satisfying

$$\mathbf{b}_i = \langle a_i, x\rangle + \boldsymbol{\varepsilon}_i \qquad \text{for } i = 1, \ldots, m,$$

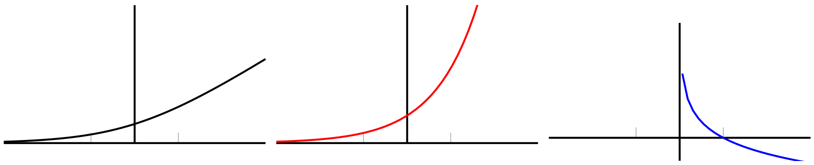

Figure 2.2: Penalties $\log(1+\exp(\cdot))$, $\exp(\cdot)$, and $-\log(\cdot)$ are used to model binary observations, count observations, and non-negative observations.

where ε_i is a normally distributed random variable with zero mean and variance σ^2. Thus $\mathbf{b}_i$ is normally distributed with mean $\mu_i := \langle a_i, x\rangle$ and variance σ^2. Assuming that the responses are independent, the *likelihood* of observing $\mathbf{b}_i = b_i$ for $i = 1, \ldots, m$ is given by

$$L(\{b_i|\mu_i, \sigma^2\}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{m}\frac{1}{2}(b_i - \mu_i)^2\right).$$

To find a good choice of x, we maximize this likelihood with respect to x, or equivalently, minimize its negative logarithm:

$$\begin{aligned} \min_x \quad -\log(L\{b_i|\mu_i(x), \sigma^2\}) &= \min_x \ \tfrac{1}{2}\sum_{i=1}^{m}(b_i - \langle a_i, x\rangle)^2 \\ &= \min_x \ \tfrac{1}{2}\|Ax - b\|^2. \end{aligned}$$

This is exactly the linear regression problem in Example 2.1

The assumption that the $\mathbf{b_i}$ are normally distributed limits the systems one can model. More generally, responses $\mathbf{b}_i$ can have special restrictions. They may be count data (number of trees that fall over in a storm), indicate class membership (outcome of a medical study, hand-written digit recognition), or be non-negative (concentration of sugar in the blood). These problems and many others can be modeled using *general linear models (GLMs)*. Suppose the distribution of $\mathbf{b}_i$ is parametrized by (μ_i, σ^2):

$$L(b_i|\mu_i, \sigma^2) = g_1(b_i, \sigma^2)\exp\left(\frac{b_i\mu_i - g_2(\mu_i)}{g_3(\sigma^2)}\right).$$

To obtain the GLM, set $\mu_i := \langle a_i, x\rangle$, and minimize the negative log-likelihood:

$$\min_x \ \sum_{i=1}^{m} g_2(\langle a_i, x\rangle) - b_i\langle a_i, x\rangle,$$

ignoring g_1 and g_3 as they do not depend on x. This problem is smooth exactly when g_2 is smooth. Important examples are shown in Figure 2.2:

- Linear regression ($b_i \in \mathbf{R}$): $\qquad g_2(z) = \frac{1}{2}\|z\|^2$.

- Binary classification ($b_i \in \{0,1\}$): $g_2(z) = \log(1+\exp(z))$.
- Poisson regression ($b_i \in \mathbb{Z}_+$): $g_2(z) = \exp(z)$.
- Exponential regression ($b_i \geq 0$): $g_2(z) = -\ln(z)$.

Example 2.5 (Nonlinar inverse problems)**.** Suppose that we are given multivariate responses $b_i \in \mathbf{R}^k$, along with functions $f_i : \mathbf{R}^n \to \mathbf{R}^k$. Our task is to find the best weights $x \in \mathbf{R}^n$ to describe b_i. This gives rise to *nonlinear* least squares:

$$\min_x \ \sum_{i=1}^m \frac{1}{2}\|f_i(x) - b_i\|^2$$

For example, global seismologists image the subsurface of the earth using earthquake data. In this case, x encodes density of subterranean layers and initial conditions at the start the earthquake i (e.g. 'slip' of a tectonic plate), b_i are echograms collected during earthquake i, and f_i is a (smooth) function of layer density and initial conditions that predicts b_i.

Example 2.6 (Low-rank factorization)**.** Suppose that we can observe some entries a_{ij} of a large matrix $A \in \mathbf{R}^{m\times n}$, with ij ranging over some small index set $\mathcal{I}$. The goal in many applications is to recover A (i.e. fill in the missing entries) from the partially observed information and an a priori upper bound k on the rank of A. One approach is to determine a factorization $A = LR^T$, for some matrices $L \in \mathbf{R}^{m\times k}$ and $R \in \mathbf{R}^{n\times k}$. This approach leads to the problem

$$\min_{L,R} \ \tfrac{1}{2}\sum_{ij\in\mathcal{I}} \|(LR^T)_{ij} - a_{ij}\|^2 + g(L,R),$$

where g is a smooth regularization function. Such formulations were successfully used, for exampe, to 'fill in' the Netflix Prize dataset, where only about 1% of the data (ratings of 15000 movies by 500,000 users) was present.

2.1 Optimality conditions: Smooth Unconstrained

We begin the formal development with a classical discussion of optimality conditions. To this end, consider the problem

$$\min_{x\in\mathbf{E}} \ f(x)$$

where $f\colon \mathbf{E} \to \mathbf{R}$ is a C^1-smooth function. Without any additional assumptions on f, finding a global minimizer of the problem is a hopeless task. Instead, we focus on finding a *local minimizer*: a point x for which there exists a neighborhood U of x such that $f(x) \leq f(y)$ for all $y \in U$. After all, gradients and Hessians provide only local information on the function.

When encountering an optimization problem, such as above, one faces two immediate tasks. First, design an algorithm that solves the problem. That is, develop a rule for going from one point x_k to the next x_{k+1} by using computable quantities (e.g. function values, gradients, Hessians) so that the limit points of the iterates solve the problem. The second task is easier: given a test point x, either verify that x solves the problem or exhibit a direction along which points with strictly better function value can be found. Though the verification goal seems modest at first, it always serves as the starting point for algorithm design.

Observe that naively checking if x is a local minimizer of f from the very definition requires evaluation of f at every point near x, an impossible task. We now derive a *verifiable necessary condition* for local optimality.

Theorem 2.7. *(First-order necessary conditions) Suppose that x is a local minimizer of a function $f : U \to \mathbf{R}$. If f is differentiable at x, then equality $\nabla f(x) = 0$ holds.*

Proof. Set $v := -\nabla f(x)$. Then for all small $t > 0$, we deduce from the definition of derivative

$$0 \leq \frac{f(x+tv) - f(x)}{t} = -\|\nabla f(x)\|^2 + \frac{o(t)}{t}.$$

Letting t tend to zero, we obtain $\nabla f(x) = 0$, as claimed. □

A point $x \in U$ is a *critical point* for a C^1-smooth function $f\colon U \to \mathbf{R}$ if equality $\nabla f(x) = 0$ holds. Theorem 2.7 shows that all local minimizers of f are critical points. In general, even finding local minimizers is too ambitious, and we will for the most part settle for critical points.

To obtain *verifiable sufficient conditions* for optimality, higher order derivatives are required.

Theorem 2.8. *(Second-order conditions)*
Consider a C^2-smooth function $f\colon U \to \mathbf{R}$ and fix a point $x \in U$. Then the following are true.

1. *(Necessary conditions) If $x \in U$ is a local minimizer of f, then*

$$\nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succeq 0.$$

2. *(Sufficient conditions) If the relations*

$$\nabla f(x) = 0 \quad \text{and} \quad \nabla^2 f(x) \succ 0$$

hold, then x is a local minimizer of f. More precisely,

$$\liminf_{y \to x} \frac{f(y) - f(x)}{\frac{1}{2}\|y - x\|^2} \geq \lambda_{\min}(\nabla^2 f(x)).$$

Proof. Suppose first that x is a local minimizer of f. Then Theorem 2.7 guarantees $\nabla f(x) = 0$. Consider an arbitrary vector $v \in \mathbf{E}$. Then for all $t > 0$, we deduce from a second-order expansion

$$0 \leq \frac{f(x+tv) - f(x)}{\frac{1}{2}t^2} = \langle \nabla^2 f(x)v, v\rangle + \frac{o(t^2)}{t^2}.$$

Letting t tend to zero, we conclude $\langle \nabla^2 f(x)v, v\rangle \geq 0$ for all $v \in \mathbf{E}$, as claimed.

Suppose $\nabla f(x) = 0$ and $\nabla^2 f(x) \succ 0$. Let $\epsilon > 0$ be such that $B_\epsilon(x) \subset U$. Then for points $y \to x$, we have from a second-order expansion

$$\begin{aligned} \frac{f(y) - f(x)}{\frac{1}{2}\|y-x\|^2} &= \left\langle \nabla^2 f(x)\left(\frac{y-x}{\|y-x\|}\right), \frac{y-x}{\|y-x\|}\right\rangle + \frac{o(\|y-x\|^2)}{\|y-x\|^2} \\ &\geq \lambda_{\min}(\nabla^2 f(x)) + \frac{o(\|y-x\|^2)}{\|y-x\|^2}. \end{aligned}$$

Letting y tend to x, the result follows. □

The reader may be misled into believing that the role of the necessary conditions and the sufficient conditions for optimality (Theorem 2.8) is merely to determine whether a putative point x is a local minimizer of a smooth function f. Such a viewpoint is far too limited.

Necessary conditions serve as the basis for algorithm design. If necessary conditions for optimality fail at a point, then there must be some point nearby with a strictly smaller objective value. A method for discovering such a point is a first step for designing algorithms.

Sufficient conditions play an entirely different role. In Section 2.3, we will see that sufficient conditions for optimality at a point x guarantee that the function f is *strongly convex* on a neighborhood of x. Strong convexity, in turn, is essential for establishing rapid convergence of numerical methods.

2.2 Optimality Conditions: Smooth Constrained

By using the tangent cones in Definition 1.17, as well as Exercises 1.20 and 1.21, simple optimality conditions for C^1-smooth convexly constrained problems are easily obtained.

Theorem 2.9. *(First-order necessary conditions) Suppose U is an open set in $\mathbf{E}$ and that $\bar{x}$ is a local minimizer of a function $f : U \to \mathbf{R}$ over the nonempty closed set $\Omega \subset U$. If f is differentiable at $\bar{x}$, then $\langle \nabla f(\bar{x}), v\rangle \geq 0$ for all $v \in T(\bar{x} \,|\, \Omega)$.*

Proof. Let $v \in T(\bar{x} \,|\, S)$. With no loss in generality, we may assume that $v \neq 0$. Then, by Exercise 1.19, there is a $t > 0$ and a sequence $\{x^k\} \subset \Omega \setminus \{\bar{x}\}$

such that $x^k \to \bar{x}$, $\left\|x^k - \bar{x}\right\|^{-1}(x^k - \bar{x}) \to u$, and $v = tu$. Since $\bar{x}$ is a local minimizer of f on Ω, we may assume that $f(\bar{x}) \le f(x^k)$ for all k. Then, for all k,

$$f(\bar{x}) \le f(\bar{x}) + \left\langle \nabla f(\bar{x}),\, x^k - \bar{x} \right\rangle + o\left(\left\|x^k - \bar{x}\right\|\right),$$

and so

$$0 \le \left\langle \nabla f(\bar{x}),\, x^k - \bar{x} \right\rangle + o\left(\left\|x^k - \bar{x}\right\|\right) \qquad \forall\, k.$$

Dividing through by $\left\|x^k - \bar{x}\right\|$ and taking the limit in k yields the result. □

Notice that the first-order necessary condition above works for arbitrary nonempty closed sets Ω. However, to obtain a second-order conditions, we assume that Ω is convex, and make use of the tangent cone characterizations in Exercises 1.20 and 1.21.

Theorem 2.10. *(Second-order conditions)*
Consider a C^2-smooth function $f : U \to \mathbf{R}$, where $U \subset \mathbf{E}$ is open. Fix a point $\bar{x} \in \Omega \subset U$, where Ω is a nonempty close convex set. Then the following are true.

1. *(Necessary conditions) Assume that Ω is a convex polyhedron. If $\bar{x}$ is a local minimizer of f on Ω, then*

$$\langle \nabla f(\bar{x}),\, v\rangle \ge 0 \qquad \forall\ v \in T\left(\bar{x}\,|\,\Omega\right)$$

and

$$v^T \nabla^2 f(x) v \ge 0 \qquad \forall\ v \in T\left(\bar{x}\,|\,\Omega\right) \cap \operatorname{span}\left(\nabla f(\bar{x})\right)^{\perp}.$$

2. *(Sufficient conditions) If the relations*

$$\langle \nabla f(\bar{x}),\, v\rangle \ge 0 \qquad \forall\ v \in T\left(\bar{x}\,|\,\Omega\right)$$

and

$$v^T \nabla^2 f(\bar{x}) v > 0 \qquad \forall\ v \in \left(T\left(\bar{x}\,|\,\Omega\right) \cap \operatorname{span}\left(\nabla f(\bar{x})\right)^{\perp}\right) \setminus \{0\}.$$

hold, then there is an $\epsilon > 0$ and $\beta > 0$ such that

$$f(x) \ge f(\bar{x}) + \frac{\beta}{2}\left\|x - \bar{x}\right\|^2 \qquad \forall\ x \in B_\epsilon(\bar{x}) \cap \Omega. \tag{2.1}$$

Proof. Theorem 2.9 tells us that $\langle \nabla f(\bar{x}),\, v\rangle \ge 0$ for all $v \in T\left(\bar{x}\,|\,\Omega\right)$. Next let $v \in T\left(\bar{x}\,|\,\Omega\right) \cap \operatorname{span}\left(\nabla f(\bar{x})\right)^{\perp}$. With no loss in generality, we may assume that $\|v\| = 1$. By Exercise 1.21 there exists $\bar{t} > 0$ such that $\bar{x} + tv \in \Omega$ for all $t \in (0, \bar{t})$. Since $\bar{x}$ is a local solution, we may take $\bar{t}$ so small that that $f(\bar{x}) \le f(\bar{x} + tv)$ for all $t \in (0, \bar{t})$. Then, for all $t \in (0, \bar{t})$,

$$f(\bar{x}) \le f(\bar{x}) + t\langle \nabla f(\bar{x}),\, v\rangle + \frac{t^2}{2}\langle \nabla f(\bar{x})v,\, v\rangle + o\left(t^2\right)$$

and so

$$0 \le \frac{1}{2}\langle \nabla f(\bar{x})v,\, v\rangle + \frac{o(t^2)}{t^2} \quad \forall\, t \in (0, \bar{t}).$$

Letting $t \to 0$ yields the second-order necessary condition.

To see the second-order sufficient condition, we suppose that the result is false so that there exists a sequences $\beta_k \downarrow 0$ and $x^k \to \bar{x}$ such that

$$f(x^k) < f(\bar{x}) + \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k,$$

or equivalently,

$$\begin{aligned} f(\bar{x}) + \left\langle \nabla f(\bar{x}),\, x^k - \bar{x}\right\rangle + \frac{1}{2}\left\langle \nabla^2 f(\bar{x})(x^k - \bar{x}),\, (x^k - \bar{x})\right\rangle + o(\left\|x^k - \bar{x}\right\|^2) \\ \le f(\bar{x}) + \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k. \end{aligned} \tag{2.2}$$

With no loss in generality, we may assume that there is a unit vector u such that $\left\|x^k - \bar{x}\right\|^{-1}(x^k - \bar{x}) \to u \in T(\bar{x}\,|\,\Omega)$. Dividing by $\left\|x^k - \bar{x}\right\|$ and letting $k \uparrow \infty$ yields $0 \le \langle \nabla f(\bar{x}),\, u\rangle \le 0$ so that $u \in (T(\bar{x}\,|\,\Omega) \cap \mathrm{span}(\nabla f(\bar{x}))^{\perp}) \backslash \{0\}$. Further note that by Exercise 1.20, $(x^k - \bar{x}) \in T(\bar{x}\,|\,\Omega)$ for all k so that $\left\langle \nabla f(\bar{x}),\, x^k - \bar{x}\right\rangle \ge 0$ for all k. Hence, (2.2) tells us that

$$\frac{1}{2}\left\langle \nabla^2 f(\bar{x})(x^k - \bar{x}),\, (x^k - \bar{x})\right\rangle + o(\left\|x^k - \bar{x}\right\|^2) \le \frac{\beta_k}{2}\left\|x^k - \bar{x}\right\|^2 \qquad \forall\, k.$$

Dividing by $\left\|x^k - \bar{x}\right\|^2$ and taking the limit as $k \uparrow \infty$ gives the contradiction $\left\langle \nabla^2 f(\bar{x})u,\, u\right\rangle \le 0$ which proves the result. □

In later sections we will improve on the second-order conditions in this theorem by delving deeper into the curvature properties of the set Ω. These later results will not only allow us to remove the convexity hypotheses, but will also be stronger even in the convex case. As a first illustration of the limitations of Theorem 2.10, the following example shows that the polyhedrality hypothesis used in the necessary condition cannot be weakened.

Example 2.11. Consider the problem

$$\begin{array}{ll} \min & \frac{1}{2}(x_2 - x_1^2) \\ \text{subject to} & 0 \le x_2,\ x_1^3 \le x_2^2. \end{array}$$

Observe that the constraint region in this problem can be written as $\Omega := \{(x_1, x_2) : |x_1|^{\frac{3}{2}} \le x_2\}$, therefore

$$\begin{aligned} f(x) &= \frac{1}{2}(x_2 - x_1^2) \\ &\ge \frac{1}{2}(|x_1|^{\frac{3}{2}} - |x_1|^2) \\ &= \frac{1}{2}|x_1|^{\frac{3}{2}}(1 - |x_1|^{\frac{1}{2}}) > 0 \end{aligned}$$

whenever $0 < |x_1| \leq 1$. Consequently, the origin is a strict local solution for this problem. Nonetheless,

$$T\,(0\,|\,\Omega) \cap [\nabla f(0)]^{\perp} = \{(\delta, 0)\,:\,\delta \in \mathbf{R}\},$$

while

$$\nabla^2 f(0) = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}.$$

That is, even though the origin is a strict local solution, the Hessian of f is negative definite on $T\,(0\,|\,\Omega) \cap [\nabla f(0)]^{\perp}$.

The second-order sufficiency condition in Theorem 2.10 is also lacking since, as is shown in the next example, the quadratic growth condition (2.1) can be satisfied even if the hessian is not positive definite on the set $(T\,(\bar{x}\,|\,\Omega) \cap \operatorname{span}(\nabla f(\bar{x}))^{\perp}) \setminus \{0\}$.

Example 2.12. Consider the problem

$$\begin{array}{ll} \min & x_2 \\ \text{subject to} & x_1^2 \leq x_2. \end{array}$$

Clearly, $\bar{x} = 0$ is the unique global solution to this convex program. Moreover,

$$\begin{aligned} f(\bar{x}) + \frac{1}{2}\|x - \bar{x}\|^2 &= \frac{1}{2}(x_1^2 + x_2^2) \\ &\leq \frac{1}{2}(x_2 + x_2^2) \\ &\leq x_2 = f_0(x) \end{aligned}$$

for all x in the constraint region Ω with $\|x - \bar{x}\| \leq 1$. However, $\nabla^2 f(\bar{x}) = 0$.

2.3 Convexity, a first look

Finding a global minimizer of a general smooth function $f\colon \mathbf{E} \to \mathbf{R}$ is a hopeless task, and one must settle for local minimizers or even critical points. This is quite natural since gradients and Hessians only provide local information on the function. However, there is a class of smooth functions, prevalent in applications, whose gradients provide *global information.* This is the class of convex functions – the main setting for the book. This section provides a short, and limited, introduction to the topic to facilitate algorithmic discussion. Later sections of the book explore convexity in much greater detail.

Definition 2.13 (Convexity)**.** A function $f\colon U \to (-\infty, +\infty]$ is *convex* if the inequality

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

holds for all points $x, y \in U$ and real numbers $\lambda \in [0, 1]$.

In other words, a function f is convex if any secant line joining two point in the graph of the function lies above the graph. This is the content of the following exercise.

Exercise 2.14. Show that a function $f\colon U \to (-\infty, +\infty]$ is convex if and only if the *epigraph*

$$\operatorname{epi} f := \{(x, r) \in U \times \mathbf{R} : f(x) \leq r\}$$

is a convex subset of $\mathbf{E} \times \mathbf{R}$.

Convexity is preserved under a variety of operations. Point-wise maximum is an important example.

Exercise 2.15. Consider an arbitrary set T and a family of convex functions $f_t\colon U \to (-\infty, +\infty]$ for $t \in T$. Show that the function $f(x) := \sup_{t\in T} f_t(x)$ is convex.

Convexity of smooth functions can be characterized entirely in terms of derivatives.

Theorem 2.16 (Differential characterizations of convexity)**.** *The following are equivalent for a C^1-smooth function $f\colon U \to \mathbf{R}$.*

(a) **(convexity)** *f is convex.*

(b) **(gradient inequality)** *$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ for all $x, y \in U$.*

(c) **(monotonicity)** *$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq 0$ for all $x, y \in U$.*

If f is C^2-smooth, then the following property can be added to the list:

(d) The relation $\nabla^2 f(x) \succeq 0$ holds for all $x \in U$.

Proof. Assume (a) holds, and fix two points x and y. For any $t \in (0, 1)$, convexity implies

$$f(x + t(y - x)) = f(ty + (1 - t)x) \leq tf(y) + (1 - t)f(x),$$

while the definition of the derivative yields

$$f(x + t(y - x)) = f(x) + t\langle \nabla f(x), y - x\rangle + o(t).$$

Combining the two expressions, canceling $f(x)$ from both sides, and dividing by t yields the relation

$$f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + o(t)/t.$$

Letting t tend to zero, we obtain property (b).

Suppose now that (b) holds. Then for any $x, y \in U$, appealing to the gradient inequality, we deduce

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

and

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle.$$

Adding the two inequalities yields (c).

Finally, suppose (c) holds. Define the function $\varphi(t) := f(x + t(y - x))$ and set $x_t := x + t(y - x)$. Then monotonicity shows that for any real numbers $t, s \in [0, 1]$ with $t > s$ the inequality holds:

$$\begin{aligned}\varphi'(t) - \varphi'(s) &= \langle \nabla f(x_t), y - x \rangle - \langle \nabla f(x_s), y - x \rangle \\ &= \frac{1}{t - s} \langle \nabla f(x_t) - \nabla f(x_s), x_t - x_s \rangle \geq 0.\end{aligned}$$

Thus the derivative φ' is nondecreasing, and hence for any $x, y \in U$, we have

$$f(y) = \varphi(1) = \varphi(0) + \int_0^1 \varphi'(r)\, dr \geq \varphi(0) + \varphi'(0) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Some thought now shows that f admits the representation

$$f(y) = \sup_{x \in U} \{ f(x) + \langle \nabla f(x), y - x \rangle \}$$

for any $y \in U$. Since a pointwise supremum of an arbitrary collection of convex functions is convex (Excercise 2.15), we deduce that f is convex, establishing (a).

Suppose now that f is C^2-smooth. Then for any fixed $x \in U$ and $h \in \mathbf{E}$, and all small $t > 0$, property (b) implies

$$f(x) + t\langle \nabla f(x), h \rangle \leq f(x + th) = f(x) + t\langle \nabla f(x), h \rangle + \frac{t^2}{2} \langle \nabla^2 f(x)h, h \rangle + o(t^2).$$

Canceling out like terms, dividing by t^2, and letting t tend to zero we deduce $\langle \nabla^2 f(x)h, h \rangle \geq 0$ for all $h \in \mathbf{E}$. Hence (d) holds. Conversely, suppose (d) holds. Then Corollary 1.13 immediately implies for all $x, y \in \mathbf{E}$ the inequality

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle = \int_0^1 \int_0^t \langle \nabla^2 f(x + s(y - x))(y - x), y - x \rangle\, ds\, dt \geq 0.$$

Hence (b) holds, and the proof is complete. □

Exercise 2.17. Show that the functions f and F in Exercise 1.10 are convex.

Exercise 2.18. Consider a C^1-smooth function $f\colon \mathbf{R}^n \to \mathbf{R}$. Prove that each condition below holding for all points $x, y \in \mathbf{R}^n$ is equivalent to f being β-smooth and convex.

1. $f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq f(y)$
2. $\frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle$
3. $0 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle \leq \beta\|x - y\|^2$

Global minimality, local minimality, and criticality are equivalent notions for smooth convex functions.

Corollary 2.19 (Minimizers of convex functions)**.** *For any C^1-smooth convex function $f\colon U \to \mathbf{R}$ and a point $x \in U$, the following are equivalent.*

(a) x is a global minimizer of f,

(b) x is a local minimizer of f,

(c) x is a critical point of f.

Proof. The implications $(a) \Rightarrow (b) \Rightarrow (c)$ are immediate. The implication $(c) \Rightarrow (a)$ follows from the gradient inequality in Theorem 2.16. □

Exercise 2.20. Consider a C^1-smooth convex function $f\colon \mathbf{E} \to \mathbf{R}$. Fix a linear subspace $\mathcal{L} \subset \mathbf{E}$ and a point $x_0 \in \mathbf{E}$. Show that $x \in \mathcal{L}$ minimizes the restriction $f_{\mathcal{L}}\colon \mathcal{L} \to \mathbf{R}$ if and only if the gradient $\nabla f(x)$ is orthogonal to $\mathcal{L}$.

Strengthening the gradient inequality in Theorem 2.16 in a natural ways yields an important subclass of convex functions. These are the functions for which numerical methods have a chance of converging at least linearly.

Definition 2.21 (Strong convexity)**.** We say that a C^1-smooth function $f\colon U \to \mathbf{R}$ is *α-strongly convex* (with $\alpha \geq 0$) if the inequality

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2 \quad \text{holds for all } x, y \in U.$$

Figure 2.3 illustrates geometrically a β-smooth and α-convex function.

In particular, a very useful property to remember is that if x is a minimizer of an α-strongly convex C^1-smooth function f, then for all y it holds:

$$f(y) \geq f(x) + \frac{\alpha}{2}\|y - x\|^2.$$

Exercise 2.22. Show that a C^1-smooth function $f\colon U \to \mathbf{R}$ is α-strongly convex if and only if the function $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.

The following is an analogue of Theorem 2.16 for strongly convex functions.

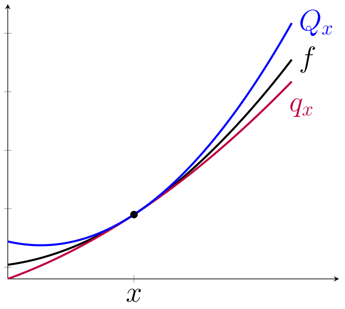

Figure 2.3: Illustration of a β-smooth and α-strongly convex function f, where $Q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2$ is an upper models based at x and $q_x(y) := f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2$ is a lower model based at x. The fraction $Q := \beta/\alpha$ is often called the *condition number* of f.

Theorem 2.23 (Characterization of strong convexity)**.** *The following properties are equivalent for any C^1-smooth function $f : U \to \mathbf{R}$ and any constant $\alpha \geq 0$.*

(a) f is α-convex.

(b) The inequality $\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \alpha\|y - x\|^2$ holds for all $x, y \in U$.

If f is C^2-smooth, then the following property can be added to the list:

(c) The relation $\nabla^2 f(x) \succeq \alpha I$ holds for all $x \in U$.

Proof. By Excercise 2.22, property (a) holds if and only if $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex, which by Theorem 2.16, is equivalent to (b). Suppose now that f is C^2-smooth. Theorem 2.16 then shows that $f - \frac{\alpha}{2}\|\cdot\|^2$ is convex if and only if (c) holds. □

2.4 Rates of convergence

In the next section, we will begin discussing algorithms. A theoretically sound comparison of numerical methods relies on precise rates of progress in the iterates. For example, we will predominantly be interested in how fast the quantities $f(x_k) - \inf f$, $\nabla f(x_k)$, or $\|x_k - x^*\|$ tend to zero as a function of the counter k. In this section, we review three types of convergence rates that we will encounter.

Fix a sequence of real numbers $a_k > 0$ with $a_k \to 0$.

1. We will say that a_k converges *sublinearly* if there exist constants $c, q > 0$ satisfying
$$a_k \leq \frac{c}{k^q} \qquad \text{for all } k.$$

Larger q and smaller c indicates faster rates of convergence. In particular, given a target precision $\varepsilon > 0$, the inequality $a_k \leq \varepsilon$ holds for every $k \geq (\frac{c}{\varepsilon})^{1/q}$. The importance of the value of c should not be discounted; the convergence guarantee depends strongly on this value.

2. The sequence a_k is said to *converge linearly* if there exist constants $c > 0$ and $q \in (0, 1]$ satisfying

$$a_k \leq c \cdot (1-q)^k \qquad \text{for all } k.$$

In this case, we call $1 - q$ the *linear rate of convergence.* Fix a target accuracy $\varepsilon > 0$, and let us see how large k needs to be to ensure $a_k \leq \varepsilon$. To this end, taking logs we get

$$c \cdot (1-q)^k \leq \varepsilon \quad \iff \quad k \geq \frac{-1}{\ln(1-q)} \ln\left(\frac{c}{\varepsilon}\right).$$

Taking into account the inequality $\ln(1-q) \leq -q$, we deduce that the inequality $a_k \leq \varepsilon$ holds for every $k \geq \frac{1}{q}\ln(\frac{c}{\varepsilon})$. The dependence on q is strong, while the dependence on c is very weak, since the latter appears inside a log.

3. The sequence a_k is said to *converge quadratically* if there is a constant c satisfying

$$a_{k+1} \leq c \cdot a_k^2 \qquad \text{for all } k.$$

Observe then unrolling the recurrence yields

$$a_{k+1} \leq \frac{1}{c}(ca_0)^{2^{k+1}}.$$

The only role of the constant c is to ensure the starting moment of convergence. In particular, if $ca_0 < 1$, then the inequality $a_k \leq \varepsilon$ holds for all $k \geq \log_2 \ln(\frac{1}{c\varepsilon}) - \log_2(-\ln(ca_0))$. The dependence on c is negligible.

2.5 Two basic methods

This section presents two classical minimization algorithms: gradient descent and Newton's method. It is crucial for the reader to keep in mind how the convergence guarantees are amplified when (strong) convexity is present.

2.5.1 Majorization view of gradient descent

Consider the optimization problem

$$\min_{x \in \mathbf{E}} f(x),$$

where f is a β-smooth function. Our goal is to design an iterative algorithm that generates iterates x_k, such that any limit point of the sequence $\{x_k\}$ is critical for f. It is quite natural, at least at first, to seek an algorithm that is monotone, meaning that the sequence of function values $\{f(x_k)\}$ is decreasing. Let us see one way this can be achieved, using the idea of *majorization*. In each iteration, we will define a simple function m_k (the "upper model") agreeing with f at x_k, and majorizing f globally, meaning that the inequality $m_k(x) \geq f(x)$ holds for all $x \in \mathbf{E}$. Defining x_{k+1} to be the global minimizer of m_k, we immediately deduce

$$f(x_{k+1}) \leq m_k(x_{k+1}) \leq m_k(x_k) = f(x_k).$$

Thus function values decrease along the iterates generated by the scheme, as was desired.

An immediate question now is where such upper models m_k can come from. Here's one example of a quadratic upper model:

$$m_k(x) := f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{\beta}{2}\|x - x_k\|^2. \tag{2.3}$$

Clearly m_k agrees with f at x_k, while Corollary 1.14 shows that the inequality $m_k(x) \geq f(x)$ holds for all $x \in \mathbf{E}$, as required. It is precisely this ability to find quadratic upper models of the objective function f that separates minimization of smooth functions from those that are non-smooth.

Notice that m_k has a unique critical point, which must therefore equal x_{k+1} by first-order optimality conditions, and therefore we deduce

$$x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k).$$

This algorithm, likely familiar to the reader, is called *gradient descent*. Let us now see what can be said about limit points of the iterates x_k. Appealing to Corollary 1.14, we obtain the descent guarantee

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \langle \nabla f(x_k), \beta^{-1}\nabla f(x_k)\rangle + \frac{\beta}{2}\|\beta^{-1}\nabla f(x_k)\|^2 \\ &= f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2. \end{aligned} \tag{2.4}$$

Rearranging, and summing over the iterates, we deduce

$$\sum_{i=1}^{k} \|\nabla f(x_i)\|^2 \leq 2\beta\big(f(x_1) - f(x_{k+1})\big).$$

Thus either the function values $f(x_k)$ tend to $-\infty$, or the sequence $\{\|\nabla f(x_i)\|^2\}$ is summable and therefore every limit point of the iterates x_k is a critical points of f, as desired. Moreover, setting $f^* := \lim_{k\to\infty} f(x_k)$, we deduce the precise rate at which the gradients tend to zero:

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{1}{k}\sum_{i=1}^{k} \|\nabla f(x_i)\|^2 \leq \frac{2\beta\big(f(x_1) - f^*\big)}{k}.$$

We have thus established the following result.

Theorem 2.24 (Gradient descent). *Consider a β-smooth function $f\colon \mathbf{E} \to \mathbf{R}$. Then the iterates generated by the gradient descent method satisfy*

$$\min_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 \leq \frac{2\beta\big(f(x_1) - f^*\big)}{k}.$$

Convergence guarantees improve dramatically when f is convex. Henceforth let x^* be a minimizer of f and set $f^* = f(x^*)$.

Theorem 2.25 (Gradient descent and convexity). *Suppose that $f\colon \mathbf{E} \to \mathbf{R}$ is convex and β-smooth. Then the iterates generated by the gradient descent method satisfy*

$$f(x_k) - f^* \leq \frac{\beta\|x_0 - x^*\|^2}{2k}$$

and

$$\min_{i=1,\ldots k} \|\nabla f(x_i)\| \leq \frac{2\beta\|x_0 - x^*\|}{k}.$$

Proof. Since x_{k+1} is the minimizer of the β-strongly convex quadratic $m_k(\cdot)$ in (2.3), we deduce

$$f(x_{k+1}) \leq m_k(x_{k+1}) \leq m_k(x^*) - \frac{\beta}{2}\|x_{k+1} - x^*\|^2.$$

We conclude

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x^* - x^k\rangle + \frac{\beta}{2}(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2) \\ &\leq f^* + \frac{\beta}{2}(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2). \end{aligned}$$

Summing for $i = 1, \ldots, k+1$ yields the inequality

$$\sum_{i=1}^{k}(f(x_i) - f^*) \leq \frac{\beta}{2}\|x_0 - x^*\|^2,$$

and therefore

$$f(x_k) - f^* \leq \frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f^*) \leq \frac{\beta\|x_0 - x^*\|^2}{2k},$$

as claimed. Next, summing the basic descent inequality

$$\frac{1}{2\beta}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

for $k = m, \ldots, 2m-1$, we obtain

$$\frac{1}{2\beta}\sum_{i=m}^{2m-1}\|\nabla f(x_i)\|^2 \leq f(x_m) - f^* \leq \frac{\beta\|x_0 - x^*\|^2}{2m},$$

Taking into account the inequality

$$\frac{1}{2\beta}\sum_{i=m}^{2m-1}\|\nabla f(x_i)\|^2 \geq \frac{m}{2\beta}\cdot \min_{i=1,\ldots 2m}\|\nabla f(x_i)\|^2,$$

we deduce

$$\min_{i=1,\ldots 2m}\|\nabla f(x_i)\| \leq \frac{2\beta\|x_0 - x^*\|}{2m}$$

as claimed. □

Thus when the gradient method is applied to a potentially nonconvex β-smooth function, the gradients $\|\nabla f(x_k)\|$ decay as $\frac{\beta\|x_1-x^*\|}{\sqrt{k}}$, while for convex functions the estimate significantly improves to $\frac{\beta\|x_1-x^*\|}{k}$.

Better *linear rates* on gradient, functional, and iterate convergence is possible when the objective function is strongly convex.

Theorem 2.26 (Gradient descent and strong convexity).
Suppose that $f\colon \mathbf{E} \to \mathbf{R}$ is α-strongly convex and β-smooth. Then the iterates generated by the gradient descent method satisfy

$$\|x_k - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^k \|x_0 - x^*\|^2,$$

where $Q := \beta/\alpha$ is the condition number of f.

Proof. Appealing to strong convexity, we have

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \beta^{-1}\nabla f(x_k)\|^2 \\ &= \|x_k - x^*\|^2 + \frac{2}{\beta}\langle \nabla f(x_k), x^* - x_k\rangle + \frac{1}{\beta^2}\|\nabla f(x_k)\|^2 \\ &\leq \|x_k - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_k) - \frac{\alpha}{2}\|x_k - x^*\|^2\Big) + \frac{1}{\beta^2}\|\nabla f(x_k)\|^2 \\ &= \left(1 - \frac{\alpha}{\beta}\right)\|x_k - x^*\|^2 + \frac{2}{\beta}\Big(f^* - f(x_k) + \frac{1}{2\beta}\|\nabla f(x_k)\|^2\Big).\end{aligned}$$

Seeking to bound the second summand, observe the inequalities

$$f^* + \frac{\alpha}{2}\|x_{k+1} - x^*\|^2 \leq f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2.$$

Thus we deduce

$$\|x_{k+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_k - x^*\|^2 - \frac{\alpha}{\beta}\|x_{k+1} - x^*\|^2.$$

Rearranging yields

$$\|x_{k+1} - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)\|x_k - x^*\|^2 \leq \left(\frac{Q-1}{Q+1}\right)^{k+1}\|x_0 - x^*\|^2,$$

as claimed. □

Thus for gradient descent, the quantities $\|x_k - x^*\|^2$ converge to zero at a linear rate $\frac{Q-1}{Q+1} = 1 - \frac{2}{Q+1}$. We will often instead use the simple upper bound, $1 - \frac{2}{Q+1} \leq 1 - Q^{-1}$, to simplify notation. Analogous linear rates for $\|\nabla f(x_k)\|$ and $f(x_k) - f^*$ follow immediately from β-smoothness and strong convexity. In particular, in light of Section 2.4, we can be sure that the inequality $\|x_k - x^*\|^2 \leq \varepsilon$ holds after $k \geq \frac{Q+1}{2}\ln\left(\frac{\|x_0 - x^*\|^2}{\varepsilon}\right)$ iterations.

Example 2.27 (Linear Regression)**.** Consider a linear regression problem as in Example 2.1:

$$\min_x \ \frac{1}{2}\|Ax - b\|^2. \tag{2.5}$$

This problem has a unique solution only if A is injective, and in this case, the solution is

$$\bar{x} = (A^T A)^{-1}A^T b.$$

When the solution is not unique, for example if $m < n$, it is common to *regularize* the problem by adding a strongly-convex quadratic perturbation:

$$\min_x \ \frac{1}{2}\|Ax - b\|^2 + \frac{\eta}{2}\|x\|^2. \tag{2.6}$$

This strategy is called *ridge regression* (Example 2.2). This problem always has a closed form solution, regardless of properties of A:

$$\bar{x}_\eta = (A^T A + \eta I)^{-1}A^T b,$$

In this example, we apply steepest descent with constant step length to the ridge regression problem. Despite the availability of closed form solutions, iterative approaches are essential for large-scale applications, where forming $A^T A$ is not feasible. Indeed, for many real-world applications, practitioners may have access to A and A^T only through the action of these operators on vectors.

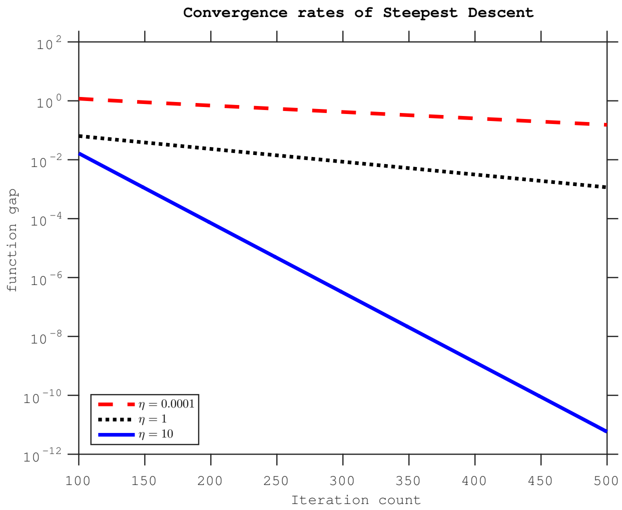

Figure 2.4: Convergence rate of steepest descent for Ridge Regression. In this example, the condition number of A is set to 10, and we show convergence of functional iterates $f(x_k) - f(x^*)$ for several values of η.

Since the eigenvalues of $A^TA + \eta I$ are simply eigenvalues of A^TA shifted by η, it is clear that the Lipschitz constant of the gradient of (2.6) is $\beta = \lambda_{\max}(A^TA) + \eta$. Each iteration of steepest descent is therefore given by

$$x_{k+1} = x_k - \frac{1}{\lambda_{\max}(A^TA) + \eta}\left(A^T(Ax_k - b) + \eta x_k\right). \tag{2.7}$$

The strong convexity constant α of the objective functions is

$$\alpha = \lambda_{\min}(A^TA) + \eta.$$

Therefore, Theorem 3.15 guarantees

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{\eta + \lambda_{\min}(A^TA)}{\eta + \lambda_{\max}(A^TA)}\right)^k \|x_0 - x^*\|^2.$$

The convergence rates of the steepest descent algorithm (for both iterates and function values) for ridge regression is shown in Figure 2.4. The linear rate is evident.

2.5.2 Newton's method

In this section we consider Newton's method, an algorithm much different from gradient descent. Consider the problem of minimizing a C^2-smooth function $f\colon \mathbf{E} \to \mathbf{R}$. Finding a critical point x of f can always be recast as