Lectures on Optimization

Sinho Chewi

January 19, 2025

Contents

1 [1/14] Introduction and basics of convex functions

These lecture notes supplement S&DS 432/632 (Advanced Optimization Techniques), taught in Spring 2025. They are not meant to be comprehensive.

The notes are primarily based on the books [Bub15; Nes18], as well as my personal understanding of the subject formed through discussions with many people over the years. Please send me corrections and feedback via email. I thank Linghai Liu, Ruixiao Wang, and Ilias Zadik for correcting my mistakes.

Logistics. The problem sets, syllabus, and all other information can be found at the course website and the Canvas page. Grading is based on six problem sets and one take-home final exam, each of which counts for 1/7 of the total grade. All questions related to logistics should be directed to my email: sinho.chewi@yale.edu.

Audience. This course focuses on the theory of optimization. In particular, the course is **mathematical** in nature and taught in a theorem–proof format. The course assumes familiarity with basic proofs and logical reasoning, as well as linear algebra, multivariate calculus, and probability theory.

The reader should also be familiar with asymptotic notions (big-O notation). We use the shorthand notation $a \lesssim b$ (resp. $a \gtrsim b$) to mean that $a \leq Cb$ (resp. $a \geq b/C$) for an absolute constant $C > 0$ (i.e., a constant that does not depend on other parameters of the problem), and $a \asymp b$ to mean that both $a \lesssim b$ and $a \gtrsim b$ hold. We use $a = O(b)$ and $a \lesssim b$ interchangeably.

1.1 Overview of the course

The basic problem of optimization is to compute an approximate minimizer of a given function $f : \mathcal{X} \to \mathbb{R}$. In this course, $\mathcal{X}$ is always taken to be a subset of $\mathbb{R}^d$, although generalizations are possible (e.g., to manifolds).

Black-box optimization and the oracle model. What does it mean to "compute"? The answer depends on the representation of f and our model of computation. We start by studying *black-box optimization*. In this model, we presume that we can *evaluate f*, and possibly its derivatives, at any chosen point $x \in \mathcal{X}$.

The advantage of the black-box model is that it applies very *generally*: it is difficult to find situations in which we need to optimize a function but we cannot even evaluate

it! Consequently, algorithms developed in this model can be applied to the majority of problems encountered in practice[1]—witness the ubiquity of gradient descent.

The disadvantage is that by its very generality, it cannot take advantage of additional structural information about f which can bring computational savings. That is why, later in the course, we turn toward the study of *structured* optimization problems.

It is easy, at least at an intuitive level, to describe algorithms which are valid in the black-box model. Namely, they are algorithms which only "interact" with f through evaluations of f and its derivatives. The existence of an algorithm, together with a corresponding mathematical analysis of the number of iterations to reach an approximate minimizer contingent upon assumptions on f, provide an *upper bound* on the complexity of the optimization task. In this course, we are also interested in *lower bounds*, which delineate fundamental limitations encountered by *any* algorithm. In order to prove such a lower bound, we need to formalize the notion of "interaction" alluded to above, and this leads to the important concept of an *oracle*.

First, observe that it does not make sense to discuss the complexity of optimizing a *single* function f. For if $x_\star$ is the minimizer of f, we can consider the algorithm "output $x_\star$", which yields the correct answer in one iteration. But this algorithm is silly, since it utterly fails at optimizing any other function whose minimizer does not happen to be $x_\star$. Reflecting upon this situation, we do not consider an optimization algorithm to be sensible when it happens to succeed for one particular problem; rather, we expect it to succeed on many similar problems. Hence, we talk about a *class* of functions $\mathcal{F}$ of interest, and we require our algorithms to succeed on *every* $f \in \mathcal{F}$.

The algorithm is designed to succeed on $\mathcal{F}$ and thus, in an anthropomorphic sense, it "knows" $\mathcal{F}$. However, it does not know which particular $f \in \mathcal{F}$ it is trying to optimize. (If it possessed knowledge of f, then we run into the issue from before, namely it could simply output the minimizer.) The role of the oracle is to act as an intermediary between the algorithm and the function. Namely, we assume that the algorithm is allowed to ask certain questions ("queries") to the oracle for f, and this is the only means by which the algorithm can gather more information about f. The allowable queries and responses determine the nature of the oracle, e.g.:

- a **zeroth-order oracle** accepts a query point $x \in \mathbb{R}^d$ and outputs $f(x)$;

- a **first-order oracle** accepts a query point $x \in \mathbb{R}^d$ and outputs $(f(x), \nabla f(x))$.

Most of the course focuses on optimization with a first-order oracle, but other oracles are possible (e.g., *linear optimization oracles* and *proximal oracles*). The zeroth-order and

---

[1]There is a caveat: in this course, we solely consider continuous optimization problems. Combinatorial optimization is an entirely different beast.

first-order oracles are easy to justify, as they correspond to the black-box model described above. As the oracles become more exotic, it becomes necessary to show that they are reasonable, by describing important applications in which such access to f is feasible.

The *query complexity* of $\mathcal{F}$ for a particular choice of oracle, as a function of the prescribed tolerance ε, is then (informally) defined to be the minimum number N such that there exists an algorithm which, for any $f \in \mathcal{F}$, makes N queries to the oracle for f and outputs a point x with $f(x) - \min f \le \varepsilon$.

It is worth noting that query complexity is not the same as computational complexity. Indeed, query complexity only counts the number of interactions with the oracle, and the algorithm is allowed to perform unlimited computations between interactions. In principle, this could lead to a situation in which query complexity is wholly unrepresentative of the true computational cost of optimization—this would be the case if optimal algorithms in the oracle model were contrived and impractical. Thankfully, this is not the case. The oracle model is widely adopted as the standard model for optimization because it is the setting in which we can make precise claims about complexity, and because it generally aligns with optimization in practice.

This summarizes the conceptual framework for optimization theory—the "identity cards of the field" [Nes18], although a careful treatment of the framework only becomes necessary when discussing lower bounds (and hence we elaborate on the details then). As a branch of mathematics, the theory of optimization could be defined as the quest to characterize the query complexity of various classes $\mathcal{F}$, under various oracle models, and thereby identify optimal algorithms. This indeed remains a core element of the field, but as query complexity reaches maturity, research has shifted toward different types of questions, often inspired by practical developments.

The role of convexity.   In order to optimize efficiently, we need to place assumptions on f, ideally minimal ones. For example, we can assume that f is continuous. In this course, however, we are interested in *quantitative* rates of convergence for algorithms, and for this purpose, a *qualitative* assumption such as continuity is not enough. A quantitative form of continuity is to assume that f is *L-Lipschitz in the ℓ_∞ norm*:

$$|f(x) - f(y)| \le L \max_{i \in [d]} |x[i] - y[i]| \qquad \text{for all } x, y \in \mathcal{X}. \tag{1.1}$$

Also, for concreteness, let us take $\mathcal{X}$ to be the cube, $\mathcal{X} = [0, 1]^d$. In the language of the framework above, we consider the class

$$\mathcal{F} = \{f : [0, 1]^d \to \mathbb{R} \mid f \text{ satisfies (1.1)}\}. \tag{1.2}$$

One can then prove the following negative result.

Theorem 1.1. For any $0 < \varepsilon < L/2$ and any deterministic algorithm, the complexity of ε-approximately minimizing functions in the class defined in (1.2) to within ε using a zeroth-order oracle is at least $\lfloor \frac{L}{2\varepsilon} \rfloor^d$.

Thus, for $\varepsilon < L/4$, the complexity grows *exponentially* with the dimension. The proof is not difficult; see, e.g., [Nes18, Theorem 1.1.2]. It is also robust: variants of the result can be proven when the notion of Lipschitzness is w.r.t. the ℓ_2 norm; when the oracle is taken to be a first-order oracle; when the algorithm is allowed to be randomized; etc. The message is clear: in order for optimization to be tractable in the worst case, we must impose some structural assumptions.

The black-box oracles we have been considering are *local* in nature: given a query point $x \in \mathbb{R}^d$, the oracle reveals some information about the behavior of f in a local neighborhood of x. Assumptions such as Lipschitzness effectively govern how large this local neighborhood is. But ultimately, to render optimization tractable, we must ensure that local information yields global consequences. As justified in the next subsection, a key assumption that makes this possible is *convexity*.

Of course, not every problem is convex, and non-convex optimization often still succeeds. But for the purpose of understanding the core principles underlying optimization, there is no better starting place. It is important to remember that convex problems abound in every application domain; here, we give two classical examples from statistics.

Example 1.2 (logistic regression). The data consists of n pairs $(X_i, Y_i) \in \mathbb{R}^d \times \{0, 1\}$, where X_i is a vector of covariates and Y_i is a binary response. The statistical model assumes that the pairs are independently drawn, the covariates are deterministic, and Y_i has a Bernoulli distribution with parameter $\exp(\langle \theta, X_i \rangle)/\{1 + \exp(\langle \theta, X_i \rangle)\}$. The goal is to infer the parameter θ.

The maximum likelihood estimator (MLE) for this model is the solution to the convex optimization problem

$$\widehat{\theta}_{\mathrm{MLE}} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(\log(1 + \exp \langle \theta, X_i \rangle) - Y_i \langle \theta, X_i \rangle \right).$$

Example 1.3 (LASSO). The data consists of n pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$. The statistical model assumes that the pairs are independently drawn, and that $Y_i = \langle \theta, X_i \rangle + \xi_i$, where the ξ_i's are i.i.d. noise variables independent of the X_i's. When the parameter θ is assumed to be sparse, it is standard to use the LASSO estimator, which is the solution to the convex optimization problem

$$\widehat{\theta}_{\mathrm{LASSO}} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \langle \theta, X_i \rangle)^2 + \lambda \left\| \theta \right\|_1 \right\}.$$

Here, $\lambda > 0$ is the regularization parameter and $\|\cdot\|_1$ denotes the ℓ_1 norm, defined via $\|\theta\|_1 := \sum_{i=1}^{d} |\theta[i]|$.

In these examples, the estimator is defined as the solution to a convex problem which is not solvable in closed form, necessitating the use of numerical optimization. Actually, it is not that most problems in the "wild" are convex and hence there was a need to develop convex optimization. In fact, it often goes the other way around: convex optimization is such a powerful tool that problems are intentionally formulated to be convex. This is the case for the LASSO estimator, which can be motivated as a convex relaxation of the (statistically superior) ℓ_0-constrained least-squares estimator.

First-order methods. This course largely focuses on first-order methods, namely, gradient descent and its variants. This class of methods is natural from the perspective of the theory. Equally importantly, first-order methods are lightweight and therefore scalable to large problem sizes, making them the method of choice even for highly non-convex settings which fall squarely outside of the theory.

Beyond the black-box model. After developing results for the black-box model, we study structured problems which admit more efficient solutions. The LASSO estimator of Example 1.3 can be treated as a "composite" optimization problem (a sum of a smooth and a non-smooth function), and the estimators in both Example 1.2 and Example 1.3 (and empirical risk minimization more generally) are "finite sum" problems whose computation can be sped up via the use of stochastic gradients. Other examples include the use of alternative geometries (mirror descent) and the use of coordinate-wise structure (alternating maximization/coordinate descent).

We also study interior-point methods, which are a practically effective suite of algorithms which solve linear programs (LPs) and semidefinite programs (SDPs) with polynomial iteration complexities.

Further topics are considered as time permits.

1.2 Preliminaries on convexity and smoothness

We assume familiarity with the basic notion of convexity, and we briefly review it here.

Definition 1.4. A subset $\mathcal{C} \subseteq \mathbb{R}^d$ is **convex** if for all $x, y \in \mathcal{C}$ and all $t \in [0, 1]$, the point $(1 - t)\, x + t\, y$ also lies in $\mathcal{C}$.

Definition 1.5. Let $\mathcal{C}$ be convex and let $\alpha \geq 0$. A function $f : \mathcal{C} \to \mathbb{R}$ is α-**convex** if for all $x, y \in \mathcal{C}$ and all $t \in [0, 1]$,

$$f((1 - t)\, x + t\, y) \leq (1 - t)\, f(x) + t\, f(y) - \frac{\alpha}{2}\, t\, (1 - t)\, \|y - x\|^2 . \tag{1.3}$$

When $\alpha = 0$, this is just the usual definition of a convex function. When $\alpha > 0$, we say that the function is *strongly* convex.

The definition above has the advantage that it does not require f to be differentiable. However, for the purposes of checking and utilizing convexity, it is convenient to have the following equivalent reformulations, which should be committed to memory. For simplicity, we focus on the case $\mathcal{C} = \mathbb{R}^d$.

Proposition 1.6 (convexity equivalences)**.** Let $\mathcal{C} = \mathbb{R}^d$ and $\alpha \geq 0$.

1. If f is continuously differentiable, (1.3) is equivalent to each of the following:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|y - x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d . \tag{1.4}$$

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \alpha \|y - x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d . \tag{1.5}$$

2. If f is twice continuously differentiable, (1.3) is equivalent to

$$\langle v, \nabla^2 f(x)\, v \rangle \geq \alpha \|v\|^2 \qquad \text{for all } v, x \in \mathbb{R}^d . \tag{1.6}$$

Proof. Assume that f is continuously differentiable.

(1.3) $\Longrightarrow$ (1.4): Rearranging (1.3) yields, for $t > 0$,

$$f(y) \geq f(x) + \frac{f((1 - t)\, x + t\, y) - f(x)}{t} + \frac{\alpha\, (1 - t)}{2} \|y - x\|^2 .$$

Sending $t \searrow 0$ yields (1.4).

7

$(1.4) \Rightarrow (1.5)$: Swap x and y in (1.4) and add the resulting inequality back to (1.4).

$(1.5) \Rightarrow (1.3)$: By the fundamental theorem of calculus, for $v := y - x$,

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + sv), v \rangle \, ds \,,$$

$$f((1 - t) x + t y) = f(x) + \int_0^1 \langle \nabla f(x + stv), tv \rangle \, ds \,.$$

Hence, (1.5) yields

$$f((1 - t) x + t y) - (1 - t) f(x) - t f(y) = -t \int_0^1 \langle \nabla f(x + sv) - \nabla f(x + stv), v \rangle \, ds$$

$$\leq -t \int_0^1 \alpha s (1 - t) \|v\|^2 \, ds = -\frac{\alpha}{2} t (1 - t) \|v\|^2 \,.$$

Finally, assume that f is twice continuously differentiable. Letting $y = x + \varepsilon v$ in (1.5) and sending $\varepsilon \searrow 0$ establishes (1.6). Conversely, the fundamental theorem of calculus shows that

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

and hence (1.6) implies (1.5). □

The equivalent statements each have their own interpretation: for $\alpha = 0$, (1.3) states that f lies below each of its secant lines between the intersection points; (1.4) states that f globally lies above each of its tangent lines; (1.6) states that ∇f is a monotone vector field; and (1.6) is a statement about curvature.

As noted above, the key feature of convexity is that local information yields global conclusions. Before describing this, let us first recall some basic facts about optimization. For simplicity, we consider unconstrained optimization throughout.

> **Lemma 1.7** (existence of minimizer)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuous and its level sets be bounded. Then, there exists a global minimizer of f.

Proof. The proof uses some analysis. Let $x_0 \in \mathbb{R}^d$ and let $\mathcal{K} := \{f \leq f(x_0)\}$ denote the level set. By the continuity assumption, $\mathcal{K}$ is closed and bounded, thus compact. Let $\{x_n\}_{n \in \mathbb{N}}$ be a minimizing sequence, $f(x_n) \to \inf f$. By compactness, it admits a subsequence, still denoted $\{x_n\}_{n \in \mathbb{N}}$, which converges to some $x_\star \in \mathbb{R}^d$. By continuity, $f(x_\star) = \lim_{n \to \infty} f(x_n) = \inf f$. □

Lemma 1.8 (necessary conditions for optimality). Let $f : \mathbb{R}^d \to \mathbb{R}$ be minimized at $x_\star$.

1. If f is continuously differentiable, then $\nabla f(x_\star) = 0$.

2. If f is twice continuously differentiable, then $\nabla^2 f(x_\star) \geq 0$.

Proof. Let $v \in \mathbb{R}^d$ and $\varepsilon > 0$; then, $f(x_\star + \varepsilon v) - f(x_\star) \geq 0$. If f is continuously differentiable, this yields $\int_0^1 \langle \nabla f(x_\star + \varepsilon t v), v \rangle \, dt \geq 0$. By continuity of ∇f, sending $\varepsilon \searrow 0$ proves that $\langle \nabla f(x_\star), v \rangle \geq 0$ for all $v \in \mathbb{R}^d$, which entails $\nabla f(x_\star) = 0$.

If f is twice continuously differentiable, we can expand once more to obtain $0 \leq \int_0^1 \int_0^1 \langle \nabla^2 f(x_\star + \varepsilon s t v) \, v, v \rangle \, ds \, dt$. By continuity of $\nabla^2 f$, sending $\varepsilon \searrow 0$ then proves that $\langle \nabla^2 f(x_\star) \, v, v \rangle \geq 0$ for all $v \in \mathbb{R}^d$. $\square$

The conditions $\nabla f(x_\star) = 0$, $\nabla^2 f(x_\star) \geq 0$ are necessary for optimality, but not sufficient in general. The issue is that the proof of Lemma 1.8 is entirely local, so the same conclusion holds even if $x_\star$ is only assumed to be a *local* minimizer. On the other hand, under the assumption of convexity, the first-order necessary condition becomes sufficient.

Lemma 1.9 (sufficient condition for optimality). Let $f : \mathbb{R}^d \to \mathbb{R}$ be *convex* and continuously differentiable, and let $\nabla f(x_\star) = 0$. Then, $x_\star$ is a global minimizer of f.
In particular, every local minimizer of f is a global minimizer.

Proof. This easily follows from (1.4) with $x = x_\star$. $\square$

Next, we note that the minimizer is unique if f is strictly convex.

Lemma 1.10 (uniqueness of minimizer). Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is strictly convex, i.e., for all distinct $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$, $f((1 - t) x + t y) < (1 - t) f(x) + t f(y)$. Then, if f admits a minimizer $x_\star$, it is unique.

Proof. If we had two distinct minimizers $x_\star, \tilde{x}_\star$, so that $f(x_\star) = f(\tilde{x}_\star)$, then strict convexity would imply $f(\frac{1}{2} x_\star + \frac{1}{2} \tilde{x}_\star) < f(x_\star)$, which is a contradiction. $\square$

If f is strongly convex, then it is strictly convex. Also, from, e.g., (1.4), we see that f grows at least quadratically at ∞, which implies that it has bounded level sets. We can therefore conclude:

Corollary 1.11. Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex and continuously differentiable. Then, it admits a unique minimizer $x_\star$, which is characterized by $\nabla f(x_\star) = 0$.

9

Finally, when discussing algorithms, we also need a dual condition—an *upper* bound on the Hessian—which in this context is called *smoothness*.[2]

Definition 1.12. Let $\beta \geq 0$. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is β-**smooth** if it is continuously differentiable and

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d . \qquad (1.7)$$

The following proposition is established in the same way as Proposition 1.6, so we omit the proof.

Proposition 1.13 (smoothness equivalences)**.** Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $\beta \geq 0$. Then, f is β-smooth if and only if

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \leq \beta \|y - x\|^2 \qquad \text{for all } x, y \in \mathbb{R}^d .$$

If f is twice continuously differentiable, this is also equivalent to

$$\langle v, \nabla^2 f(x) v \rangle \leq \beta \|v\|^2 \qquad \text{for all } v, x \in \mathbb{R}^d .$$

If f is convex, β-smooth, and twice continuously differentiable, then $0 \preceq \nabla^2 f \preceq \beta I$, which implies that the gradient ∇f is β-Lipschitz:

$$\|\nabla f(y) - \nabla f(x)\| \leq \beta \|y - x\| \qquad \text{for all } x, y \in \mathbb{R}^d . \qquad (1.8)$$

This remains true even without assuming twice differentiability (Exercise 3.1).

Bibliographical notes

For further discussion on the oracle model, see [NY83, §1].

Exercises

Exercise 1.1. Let $f = \frac{\alpha}{2} \|\cdot\|^2$, where $\alpha \geq 0$. Show via direct computation that (1.3) holds with equality.

---

[2]This is not to be confused with the mathematical usage of "smoothness" as "infinitely differentiable".

2   [1/16] Gradient flow

Before we turn toward our main first-order algorithm of interest, namely gradient descent, we first study the situation in continuous time via the gradient flow. Throughout this section, we let $(x_t)_{t\geq 0}$ denote the gradient flow for f:

$$\dot{x}_t = -\nabla f(x_t)\,. \tag{GF}$$

This is an ordinary differential equation (ODE), and since the main purpose of this section is to develop intuition, we assume that f is twice continuously differentiable and do not worry about showing that (GF) is well-posed. We use the following notation throughout these notes:

$$x_\star \in \arg\min f\,, \qquad f_\star := \min f = f(x_\star)\,.$$

Generally, we always assume that f admits a minimizer.

The most basic property of GF is that it always decreases the function value.

Lemma 2.1 (descent property of GF). For any $f : \mathbb{R}^d \to \mathbb{R}$, the gradient flow $(x_t)_{t\geq 0}$ of f satisfies

$$\partial_t f(x_t) = -\|\nabla f(x_t)\|^2 \leq 0\,.$$

Proof. By the chain rule, $\partial_t f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\nabla f(x_t)\|^2$. $\square$

To obtain quantitative convergence results, we now use the assumption of convexity. Our first result shows that under strong convexity, the gradient flow *contracts*.

Theorem 2.2 (contraction of GF). Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex. Let $(y_t)_{t\geq 0}$ be another gradient flow for f, i.e., $\dot{y}_t = -\nabla f(y_t)$. Then, for all $t \geq 0$,

$$\|y_t - x_t\| \leq \exp(-\alpha t)\,\|y_0 - x_0\|\,.$$

Proof. We differentiate the squared distance between the two flows:

$$\partial_t(\|y_t - x_t\|^2) = 2\,\langle y_t - x_t, \dot{y}_t - \dot{x}_t \rangle = -2\,\langle y_t - x_t, \nabla f(y_t) - \nabla f(x_t) \rangle \leq -2\alpha\,\|y_t - x_t\|^2\,,$$

where the last inequality is (1.5). The proof is concluded by applying Grönwall's lemma (see Lemma 2.3) below. $\square$

11

The proof above arrives at what is called a *differential inequality*, that is, an inequality which holds between a quantity and its derivative(s). This is a common strategy for analyzing ODEs/PDEs, and it can be loosely viewed as the continuous-time analogue of induction. The following standard lemma is useful for handling such inequalities.

Lemma 2.3 (Grönwall). Suppose that $u : [0, T] \to \mathbb{R}$ is a continuously differentiable curve that satisfies the differential inequality

$$\dot{u}(t) \le Au(t) + B(t), \qquad t \in [0, T].$$

Then, it holds that

$$\dot{u}(t) \le u(0) \exp(At) + \int_0^t B(s) \exp(A\,(t - s)) \, ds, \qquad t \in [0, T].$$

Proof. The idea is to differentiate $t \mapsto \exp(-At)\,u(t)$:

$$\partial_t \left[\exp(-At)\,u(t) \right] = \exp(-At) \left\{ -Au(t) + \dot{u}(t) \right\} \le B(t) \exp(-At).$$

By the fundamental theorem of calculus,

$$\exp(-At)\,u(t) - u(0) \le \int_0^t B(s) \exp(-As) \, ds.$$

Rearranging yields the result. □

There are many variants of Grönwall's lemma that can be proven in similar ways, e.g., we can allow time-varying A as well.

Returning to Theorem 2.2, we can apply Lemma 2.3 with $A = -2\alpha$ and $B = 0$ to conclude that $\|y_t - x_t\|^2 \le \exp(-2\alpha t) \|y_0 - x_0\|^2$, which proves the theorem. Note in particular that we can take $y_t = x_\star$ for all $t \ge 0$, so it yields the following statement about convergence to the minimizer: $\|x_t - x_\star\| \le \exp(-\alpha t) \|x_0 - x_\star\|$.

The next result is about convergence in function value, and unlike Theorem 2.2, it yields convergence for the case $\alpha = 0$ as well.

Theorem 2.4 (convergence of GF in function value). Let $f : \mathbb{R}^d \to \mathbb{R}$ be α-convex, $\alpha \ge 0$. Then, for all $t \ge 0$,

$$f(x_t) - f_\star \le \frac{\alpha}{2 \left(\exp(\alpha t) - 1 \right)} \|x_0 - x_\star\|^2.$$

When $\alpha = 0$, the right-hand side should be interpreted as its limiting value as $\alpha \to 0$, namely, $\frac{1}{2t} \|x_0 - x_\star\|^2$.

Proof. We differentiate $t \mapsto \|x_t - x_\star\|^2$, but this time we apply (1.4):

$$\partial_t(\|x_t - x_\star\|^2) = -2 \langle \nabla f(x_t), x_t - x_\star \rangle \leq -\alpha \|x_t - x_\star\|^2 - 2 (f(x_t) - f_\star).$$

Applying Grönwall's lemma (Lemma 2.3) with $A = -\alpha$, $B(t) = -2 (f(x_t) - f_\star)$,

$$0 \leq \|x_t - x_\star\|^2 \leq \exp(-\alpha t) \|x_0 - x_\star\|^2 - 2 \int_0^t \exp(-\alpha (t - s)) (f(x_s) - f_\star) \, ds.$$

By the descent property (Lemma 2.1), $f(x_s) \geq f(x_t)$, so that

$$\int_0^t \exp(-\alpha (t - s)) (f(x_s) - f_\star) \, ds \geq (f(x_t) - f_\star) \int_0^t \exp(-\alpha (t - s)) \, ds$$

$$= (f(x_t) - f_\star) \frac{1 - \exp(-\alpha t)}{\alpha}.$$

Rearranging yields the result. □

When $\alpha > 0$, Theorem 2.4 shows that $f(x_t) - f_\star = O(\exp(-\alpha t))$. When $\alpha = 0$, the rate becomes $f(x_t) - f_\star = O(1/t)$. Actually, the rate in Theorem 2.4 is not sharp (see Exercise 2.1 and Exercise 2.2). However, the statement and proof are chosen because they form the basis of our approach in discrete time.

Next, we observe that convexity is not needed for convergence in function value. Due to the descent property (Lemma 2.1), it suffices to have a lower bound on the norm of the gradient to ensure that we make sufficient progress. For example, we can impose the following condition.

Definition 2.5. Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and $\alpha > 0$. We say that f satisfies a **Polyak–Łojasiewicz (PŁ) inequality** with constant α if

$$\|\nabla f(x)\|^2 \geq 2\alpha (f(x) - f(x_\star)) \qquad \text{for all } x \in \mathbb{R}^d. \tag{PŁ}$$

The next statement is an immediate corollary of Lemma 2.1, (PŁ), and Grönwall's lemma (Lemma 2.3).

Corollary 2.6 (convergence of GF under PŁ). Let $f : \mathbb{R}^d \to \mathbb{R}$ satisfy (PŁ) with constant $\alpha > 0$. Then, for all $t \geq 0$,

$$f(x_t) - f_\star \leq (f(x_0) - f_\star) \exp(-2\alpha t).$$

13

We present a few key properties of the PŁ inequality.

> **Proposition 2.7** (strong convexity $\Rightarrow$ PŁ $\Rightarrow$ quadratic growth). Let $f : \mathbb{R}^d \to \mathbb{R}$ and $\alpha > 0$. The following implications hold.
>
> 1. If f is α-convex, then f satisfies (PŁ) with constant α.
>
> 2. If f satisfies (PŁ) with constant α, then it satisfies the following **quadratic growth** property:
>
> $$f(x) - f_\star \geq \frac{\alpha}{2} \inf_{x_\star \in \mathcal{X}_\star} \|x - x_\star\|^2, \qquad \text{for all } x \in \mathbb{R}^d, \qquad \text{(QG)}$$
>
> where $\mathcal{X}_\star$ denotes the set of minimizers of f.

Proof.

1. Setting $y = x_\star$ in (1.4), we obtain

$$-(f(x) - f_\star) \geq \langle \nabla f(x), x_\star - x \rangle + \frac{\alpha}{2} \|x - x_\star\|^2$$

$$\geq -\|\nabla f(x)\| \, \|x_\star - x\| + \frac{\alpha}{2} \|x - x_\star\|^2 \geq -\frac{1}{2\alpha} \|\nabla f(x)\|^2,$$

where the last inequality uses $ab \leq \frac{\lambda}{2} a^2 + \frac{1}{2\lambda} b^2$ for all $\lambda > 0$.

2. Let $(x_t)_{t \geq 0}$ denote the gradient flow for f started at $x_0 = x$. For simplicity, we present a proof *assuming* that the gradient flow converges to a point $x_\star$, although this assumption can be avoided (cf. [KNS16]). By Corollary 2.6, we see that $x_\star \in \mathcal{X}_\star$.

   We start by observing that

$$\partial_t (\|x_t - x_0\|^2) = -2 \langle \nabla f(x_t), x_t - x_0 \rangle \leq 2 \|\nabla f(x_t)\| \, \|x_t - x_0\|$$

   and hence

$$\partial_t \|x_t - x_0\| \leq \|\nabla f(x_t)\|.$$

   We differentiate the following quantity: $\mathcal{L}_t := \sqrt{\frac{\alpha}{2}} \|x_t - x_0\| + \sqrt{f(x_t) - f_\star}$.

$$\dot{\mathcal{L}}_t \leq \sqrt{\frac{\alpha}{2}} \|\nabla f(x_t)\| - \frac{\|\nabla f(x_t)\|^2}{2\sqrt{f(x_t) - f_\star}} \leq 0,$$

   where we applied (PŁ). Since $\mathcal{L}_0 = \sqrt{f(x) - f_\star}$ and $\mathcal{L}_\infty = \sqrt{\frac{\alpha}{2}} \|x - x_\star\|$, we deduce the result from $\mathcal{L}_0 \geq \mathcal{L}_\infty$.

14

□

Hence, strong convexity implies (PŁ), but is (PŁ) truly weaker than convexity? Indeed, there are examples. In particular, the PŁ condition has been of interest in recent years because it holds for certain overparametrized models (Exercise 2.3).

We conclude this section by studying the implication of Lemma 2.1 alone. The fundamental theorem of calculus shows that

$$\frac{1}{t} \int_0^t \|\nabla f(x_s)\|^2 \, ds \leq \frac{f(x_0) - f(x_t)}{t} \leq \frac{f(x_0) - f_\star}{t}.$$

We therefore arrive at the following simple consequence.

Corollary 2.8 (convergence of GF in gradient norm). For any $f : \mathbb{R}^d \to \mathbb{R}$,

$$\min_{s \in [0,t]} \|\nabla f(x_s)\| \leq \sqrt{\frac{f(x_0) - f_\star}{t}}.$$

(In contrast, note that if we additionally assume convexity, then Exercise 2.1 shows that $\|\nabla f(x_t)\| = O(1/t)$.)

This implies there exists a sequence of times $\{t_n\}_{n \in \mathbb{N}} \nearrow \infty$ such that $\|\nabla f(x_{t_n})\| \to 0$. (Indeed, $\min_{s \in [n,2n]} \|\nabla f(x_s)\| = O(1/n^{1/2})$, so we can choose $t_n \in [n, 2n]$.) However, the gradient flow may not converge. Famously, it is a result of [Łoj63] that for *real analytic* f, if the gradient flow remains bounded, then it does converge, and hence necessarily to a stationary point. Of course, such a stationary point may not be a global minimizer.

The idea of subsequent sections is to replicate the preceding analysis in discrete time.

Bibliographical notes

My understanding of Theorem 2.4, Exercise 2.1, and Exercise 2.2 is based on extensive discussions with Jason M. Altschuler, Adil Salim, Andre Wibisono, and Ashia Wilson. The proof in Exercise 2.1 is taken from [OV01], and the extension in Exercise 2.2 to $\alpha > 0$ is recorded in [LMW24, §F]. Both of these references pertain to the Langevin diffusion, but underneath the hood they make use of principles from optimization; see [Che25] for an introduction to this perspective.

The PŁ inequality is attributed to [Łoj63; Pol63] and it was popularized in [KNS16]. The proof that (PŁ) implies the quadratic growth inequality goes back at least to the celebrated work of [OV00].

Exercises

Exercise 2.1. Let f be convex. Show that the following quantity is decreasing, $\dot{\mathscr{L}}_t \leq 0$:

$$\mathscr{L}_t := t^2 \, \|\nabla f(x_t)\|^2 + 2t \, (f(x_t) - f_\star) + \|x_t - x_\star\|^2 \, .$$

Deduce the following gradient bound:

$$\|\nabla f(x_t)\|^2 \leq \frac{1}{t^2} \, \|x_0 - x_\star\|^2 \, .$$

Moreover, use (1.4) to argue that $2t \, (f(x_t) - f_\star) \leq t^2 \, \|\nabla f(x_t)\|^2 + \|x_t - x_\star\|^2$, hence

$$f(x_t) - f_\star \leq \frac{1}{4t} \, \|x_0 - x_\star\|^2 \, . \tag{2.1}$$

Note that this improves upon Theorem 2.4 by a factor of 2. Furthermore, show that (2.1) is sharp, as follows: for any $R, t > 0$, let $f : x \mapsto \frac{R}{2t} \max\{0, x\}$, $x_0 = R$, and show that (2.1) holds with equality.

Exercise 2.2. Extend Exercise 2.1 to the case $\alpha > 0$. Toward this end, consider

$$\mathscr{L}_t := A_t \, \|\nabla f(x_t)\|^2 + 2B_t \, (f(x_t) - f_\star) + \|x_t - x_\star\|^2 \, .$$

Choose A_t, B_t carefully to ensure that $\dot{\mathscr{L}}_t \leq -\alpha \mathscr{L}_t$, and thereby deduce the following sharp bounds:

$$\|\nabla f(x_t)\|^2 \leq \frac{\alpha^2 \, \|x_0 - x_\star\|^2}{\exp(2\alpha t) \, (1 - \exp(-\alpha t))^2} \, , \qquad f(x_t) - f_\star \leq \frac{\alpha \, \|x_0 - x_\star\|^2}{2 \, (\exp(2\alpha t) - 1)} \, .$$

Exercise 2.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be α-convex with $\alpha > 0$, and let $g : \mathbb{R}^d \to \mathbb{R}^n$ with $d \geq n$. Assume that g is surjective and that for all $x \in \mathbb{R}^d$, if $\nabla g(x)$ denotes the Jacobian at x (interpreted as a $d \times n$ matrix), then $\nabla g(x)^\mathsf{T} \nabla g(x) \succeq \sigma I_n$. Show that the composition $f \circ g$ satisfies (PŁ) with constant $\alpha\sigma$. Note that for $d > n$, there are typically multiple minimizers of $f \circ g$.

3   [1/21] Gradient descent: smooth case

In this section, we study the **gradient descent** algorithm:

$$x_{n+1} := x_n - h \, \nabla f(x_n) \, . \tag{GD}$$

16

From the perspective of numerical analysis, this is the *Euler* or *forward* discretization of (GF). Our aim is to show that if f is smooth, and the step size is sufficiently small (as a function of the smoothness), then the conclusions for (GF) transfer to (GD). Throughout this section, we assume that f is twice continuously differentiable and β-smooth.

Some of the results in this section pertain to a single step of (GD), so we use the following notation:

$$x^+ := x - h \, \nabla f(x) \,.$$

The first step is to establish the descent property.

Lemma 3.1 (descent lemma). For any β-smooth $f : \mathbb{R}^d \to \mathbb{R}$, if $h \le 1/\beta$, then

$$f(x^+) - f(x) \le -\frac{h}{2} \, \|\nabla f(x)\|^2 \,.$$

Proof. By the smoothness inequality (1.7),

$$f(x^+) \le f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{\beta}{2} \, \|x^+ - x\|^2 = f(x) - h \, \|\nabla f(x)\|^2 + \frac{\beta h^2}{2} \, \|\nabla f(x)\|^2 \,.$$

If $h \le 1/\beta$, then $-h \, (1 - \beta h/2) \le -h/2$. $\qquad\square$

It is natural to state the subsequent results in terms of the following parameter.

Definition 3.2. Let f be α-convex and β-smooth. Then, the **condition number** of f is defined to be the ratio $\kappa := \beta/\alpha \ge 1$.

When f is quadratic, $f(x) = \frac{1}{2} \langle x, A \, x \rangle$ with A symmetric, then α, β correspond to the minimum and maximum eigenvalues of A respectively, and the ratio β/α is known in numerical linear algebra as the condition number of the matrix A. Thus, Definition 3.2 provides a natural generalization of this notion. With this definition in hand, we now arrive at our first convergence result for (GD).

Theorem 3.3 (contraction of GD). Let f be α-convex and β-smooth. For all $x, y \in \mathbb{R}^d$ and step size $h \le 1/\beta$,

$$\|y^+ - x^+\| \le (1 - \alpha h)^{1/2} \, \|y - x\| \,.$$

Proof. Expanding the square,

$$\|y^+ - x^+\|^2 = \|y - x\|^2 - 2h \langle y - x, \nabla f(y) - \nabla f(x) \rangle + h^2 \|\nabla f(y) - \nabla f(x)\|^2.$$

By (3.3) in Exercise 3.1 below, for $h \le 1/\beta$ and from (1.5) we have

$$\|y^+ - x^+\|^2 \le \|y - x\|^2 - h \langle \nabla y - x, \nabla f(y) - \nabla f(x) \rangle \le (1 - \alpha h) \|y - x\|^2. \qquad \square$$

In particular, if we take $y = x_\star$, $h = 1/\beta$, and iterate, it yields

$$\|x_N - x_\star\| \le \left(1 - \frac{1}{\kappa}\right)^{N/2} \|x_0 - x_\star\| \le \exp\left(-\frac{N}{2\kappa}\right) \|x_0 - x_\star\|.$$

Thus, to obtain $\|x_N - x_\star\| \le \varepsilon$, it suffices to take $N \ge 2\kappa \log(\|x_0 - x_\star\|/\varepsilon)$.

The essence of these proofs is that the first-order term (scaling as h) replicates the continuous-time calculation, and we must apply smoothness in an appropriate way to control the second-order term (scaling as h^2). In the above proof, note that if we naïvely use Lipschitzness of the gradient (1.8) to control the second-order term, it leads to the suboptimal choice of step size $h = 1/(\beta\kappa)$, and a contraction factor of $(1 - 1/\kappa^2)^{1/2}$. To obtain $\|x_N - x_\star\| \le \varepsilon$, we would then have the estimate $N \ge 2\kappa^2 \log(\|x_0 - x_\star\|/\varepsilon)$, which is substantially worse. In conclusion, a bit of finesse is necessary. (In fact, Theorem 3.3 can also be improved, and the sharp rate is derived in Exercise 3.2.)

Next, we turn toward the analogue of Theorem 2.4.

> **Theorem 3.4** (convergence of GD in function value). Let f be α-convex and β-smooth. For any step size $h \le 1/\beta$,
>
> $$\|x^+ - x_\star\|^2 \le (1 - \alpha h) \|x - x_\star\|^2 - 2h \left(f(x^+) - f_\star \right). \tag{3.1}$$
>
> Therefore,
>
> $$f(x_N) - f_\star \le \frac{\alpha}{2 \left\{ (1 - \alpha h)^{-N} - 1 \right\}} \|x_0 - x_\star\|^2. \tag{3.2}$$
>
> When $\alpha = 0$, the right-hand side should be interpreted as its limiting value as $\alpha \to 0$, namely, $\frac{1}{2Nh} \|x_0 - x_\star\|^2$.

Proof. Expanding the square and applying convexity via (1.4),

$$\begin{aligned}
\|x^+ - x_\star\|^2 &= \|x - x_\star\|^2 - 2h \langle \nabla f(x), x - x_\star \rangle + h^2 \|\nabla f(x)\|^2 \\
&\le (1 - \alpha h) \|x - x_\star\|^2 - 2h \left(f(x) - f_\star \right) + h^2 \|\nabla f(x)\|^2.
\end{aligned}$$

18

For $h \le 1/\beta$, the descent lemma (Lemma 3.1) now implies (3.1).

The proof of (3.2), based on iterating the recursive inequality (3.1), is justified after Lemma 3.5 below. □

Iterating (3.1) is a matter of unrolling the recursion, but in order to maintain the analogy with continuous time, we refer to the lemma below as "discrete Grönwall".

Lemma 3.5 (discrete Grönwall). Suppose that for some $A > 0$,

$$u_{n+1} \le Au_n + B_n \qquad \text{for } n = 0, 1, \ldots, N-1.$$

Then,

$$u_N \le A^N u_0 + \sum_{n=1}^{N} A^{N-n} B_{n-1}.$$

Proof. We multiply the given inequality by $A^{-(n+1)}$ to form a telescoping sum:

$$A^{-N} u_N - u_0 = \sum_{n=0}^{N-1} A^{-(n+1)} (u_{n+1} - Au_n) \le \sum_{n=0}^{N-1} A^{-(n+1)} B_n.$$

Rearrange to obtain the result. □

To complete the proof of Theorem 3.4, we apply Lemma 3.5 with $u_n = \|x_n - x_\star\|^2$, $A = 1 - \alpha h$, and $B_n = -2h\,(f(x_{n+1}) - f_\star)$, yielding

$$2h \sum_{n=1}^{N} (1 - \alpha h)^{N-n} (f(x_n) - f_\star) \le (1 - \alpha h)^N \|x_0 - x_\star\|^2.$$

For $h \le 1/\beta$, the descent lemma (Lemma 3.1) implies $f(x_n) - f_\star \ge f(x_N) - f_\star$, so

$$f(x_N) - f_\star \le \frac{\|x_0 - x_\star\|^2}{2h \sum_{n=1}^{N} (1 - \alpha h)^{-n}} = \frac{\alpha \|x_0 - x_\star\|^2}{2\left\{(1 - \alpha h)^{-N} - 1\right\}}.$$

In particular, let us set $h = 1/\beta$. For $\alpha > 0$ it yields

$$f(x_N) - f_\star \le \frac{\alpha \|x_0 - x_\star\|^2}{2\left\{(1 - 1/\kappa)^{-N} - 1\right\}}$$

19

and for $\alpha = 0$, it yields

$$f(x_N) - f_\star \leq \frac{\beta \|x_0 - x_\star\|^2}{2N}.$$

The proof of convergence under (PŁ) is strikingly easy.

Theorem 3.6 (convergence of GD under PŁ). Let f be β-smooth and satisfy (PŁ) with constant $\alpha > 0$. Then, for all $h \leq 1/\beta$,

$$f(x_N) - f_\star \leq (1 - \alpha h)^N (f(x_0) - f_\star).$$

Proof. By the descent lemma (Lemma 3.1) and (PŁ),

$$f(x^+) - f_\star = f(x) - f_\star + f(x^+) - f(x) \leq f(x) - f_\star - \frac{h}{2} \|\nabla f(x)\|^2$$

$$\leq (1 - \alpha h) (f(x) - f_\star). \qquad \square$$

Finally, we present the result for obtaining a stationary point.

Theorem 3.7. Let f be β-smooth and $h \leq 1/\beta$. Then,

$$\min_{n=0,1,\ldots,N-1} \|\nabla f(x_n)\| \leq \sqrt{\frac{2 (f(x_0) - f_\star)}{Nh}}.$$

Proof. Telescope the descent lemma (Lemma 3.1):

$$\frac{h}{2N} \sum_{n=0}^{N-1} \|\nabla f(x_n)\|^2 \leq \frac{1}{N} \sum_{n=0}^{N-1} (f(x_n) - f(x_{n+1})) \leq \frac{f(x_0) - f_\star}{N}. \qquad \square$$

We summarize the results for GD in Table 1.

Assumptions	Criterion	Iterations
α-convex, β-smooth	$\|x_N - x_\star\| \le \varepsilon$	$O(\kappa \log(R/\varepsilon))$
α-convex, β-smooth	$f(x_N) - f_\star \le \varepsilon$	$O(\kappa \log(\alpha R^2/\varepsilon))$
convex, β-smooth	$f(x_N) - f_\star \le \varepsilon$	$O(\beta R^2/\varepsilon)$
α-(PŁ), β-smooth	$f(x_N) - f_\star \le \varepsilon$	$O(\kappa \log(\Delta_0/\varepsilon))$
β-smooth	$\min\limits_{n=0,1,\dots,N-1} \|\nabla f(x_n)\| \le \varepsilon$	$O(\beta \Delta_0/\varepsilon^2)$

Table 1: Rates for GD with step size $1/\beta$. Here, $R := \|x_0 - x_\star\|$ and $\Delta_0 := f(x_0) - f_\star$.

Example 3.8 (logistic regression revisited)**.** For fun, let us revisit logistic regression (Example 1.2) from a statistical lens. For concreteness, we consider Gaussian design, $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{normal}(0, I)$, and assume that the data is generated from the model with a true parameter $\theta^\star$. Let $\widehat{\mathcal{L}}$ denote the MLE objective, let $\mathcal{L} := \mathbb{E}\,\widehat{\mathcal{L}}$ denote the population risk, and let $R := \|\theta^\star\| \ge 1$. The state-of-the-art result [CLM24] shows that if $n \gtrsim Rd$ for a sufficiently large implied constant, $\widehat{\theta}_{\mathrm{MLE}}$ exists with probability $\ge 1 - \exp(-d)$ and satisfies the optimal risk bound $\mathcal{L}(\widehat{\theta}_{\mathrm{MLE}}) - \mathcal{L}(\theta^\star) \lesssim d/n$.

In practice, we cannot compute $\widehat{\theta}_{\mathrm{MLE}}$ exactly, so we use optimization. From [CLM24], any estimator $\widehat{\theta}$ satisfying $\widehat{\mathcal{L}}(\widehat{\theta}) - \widehat{\mathcal{L}}(\widehat{\theta}_{\mathrm{MLE}}) \lesssim d/n$ satisfies the same statistical risk bound as $\widehat{\theta}_{\mathrm{MLE}}$, up to a universal constant. We take $\widehat{\theta} = \widehat{\theta}_{\mathrm{GD}}$ to be the output of GD after N steps, and check how large N must be in order for this to hold. As justified in Exercise 3.3, we can expect an iteration complexity of $N \asymp R^2 n/d$.

Bibliographical notes

My understanding of Theorem 3.4 is again based on extensive discussions with Jason M. Altschuler, Adil Salim, Andre Wibisono, and Ashia Wilson.

Exercises

Exercise 3.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and β-smooth. Apply Lemma 3.1 to the function $y \mapsto f(y) - \langle \nabla f(x), y \rangle$ and observe that this function is minimized at x in order to prove

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

From this, deduce that

$$\|\nabla f(y) - \nabla f(x)\|^2 \le \beta \langle \nabla f(y) - \nabla f(x), y - x \rangle. \tag{3.3}$$

Finally, use the Cauchy–Schwarz inequality to show that ∇f is β-Lipschitz, i.e., that (1.8) holds. Note that this proof that convexity and β-smoothness together imply (1.8) does not require f to be twice differentiable.

Exercise 3.2. Let f be α-convex and β-smooth. Let $T := \mathrm{id} - h\,\nabla f$ denote the one-step GD mapping. By the fundamental theorem of calculus,

$$\|y^+ - x^+\| = \|T(y) - T(x)\| = \left\| \int_0^1 \nabla T((1-t)\,x + t\,y)\,(y-x)\,\mathrm{d}t \right\|$$

$$\leq \left(\int_0^1 \|\nabla T((1-t)\,x + ty)\|_{\mathrm{op}}\,\mathrm{d}t \right) \|y - x\| .$$

For any $z \in \mathbb{R}^d$, bound the eigenvalues of $\nabla T(z)$ and show that the choice of step size h which minimizes the bound on $\|\nabla T(z)\|_{\mathrm{op}}$ is $h = 2/(\alpha + \beta)$. Deduce the sharp rate

$$\|y^+ - x^+\| \leq \frac{\kappa - 1}{\kappa + 1} \|y - x\| .$$

Note that for large κ, the contraction factor is approximately $\exp(-2/\kappa)$, so this improves upon the iteration complexity implied by Theorem 3.3 by a factor of nearly 4.

Exercise 3.3. What does Theorem 3.4 imply for logistic regression (Example 1.2)? In the setting of Example 3.8, use the fact that $\lambda_{\max}(\frac{1}{n} \sum_{i=1}^n X_i X_i^\mathsf{T}) \lesssim 1$ with high probability[3] to justify the claimed $R^2 n / d$ iteration complexity.

4   [1/23] Lower bounds for smooth optimization

The goal of this section is to establish lower complexity bounds for convex smooth optimization. Refer to §1.1 for a conceptual first discussion of the oracle model.

Before doing so, we present some reductions between the convex and strongly convex settings which save us some effort.

4.1   Reductions between the convex and strongly convex settings

For brevity, let us say that an algorithm *successfully optimizes* a function class $\mathscr{F}$ in $\phi(\mathscr{F}, R)$ iterations if, given any $f \in \mathscr{F}$ and $x_0 \in \mathbb{R}^d$ with $\|x_0 - x_\star\| \leq R$, it outputs a point x with $f(x) - f_\star \leq \varepsilon$ using no more than $\phi(\mathscr{F}, R)$ queries to a first-order oracle for f.

---

[3]This is a standard fact about the Wishart distribution; see, e.g., [Ver18, Theorem 4.4.5].

Lemma 4.1. Assume there is an algorithm which successfully optimizes the class of convex and β-smooth functions in $\phi(\beta R^2/\varepsilon)$ iterations.

Then, there is an explicit algorithm which successfully optimizes the class of α-convex and β-smooth functions in $O(\phi(8\kappa)\log(\alpha R^2/\varepsilon))$ iterations.

Proof. Let f be α-strongly convex and β-smooth, and apply the given algorithm to f to obtain a new point x_1 with tolerance ε_1. By (QG), we have

$$\frac{\alpha}{2}\|x_1 - x_\star\|^2 \le f(x_1) - f_\star \le \varepsilon_1 .$$

Set $\varepsilon_1 = \alpha R^2/8$, so that

$$\|x_1 - x_\star\| \le \frac{1}{2} R = \frac{1}{2}\|x_0 - x_\star\| . \tag{4.1}$$

For $\kappa := \beta/\alpha$, this requires $\phi(8\kappa)$ iterations. From (4.1), if we now repeat this procedure $O(\log(\alpha R^2/\varepsilon))$ times, we can reach a point $\tilde{x}$ satisfying $\tilde{R} := \|\tilde{x} - x_\star\| \le \sqrt{\varepsilon/\alpha}$. Finally, apply the given algorithm one more time starting from $\tilde{x}$ with target accuracy ε to obtain a point x with $f(x) - f_\star \le \varepsilon$. The complexity of this final step is $\phi(\beta\tilde{R}^2/\varepsilon) = \phi(\kappa)$. □

For example, if we combine the $\alpha = 0$ case of Theorem 3.4 with Lemma 4.1, taking $\phi(x) = O(x)$, we recover the $\alpha > 0$ case of Theorem 3.4, up to constants.

Lemma 4.2. Assume there is an algorithm which successfully optimizes the class of α-convex and β-smooth functions in $\phi(\kappa)\log(\alpha R^2/\varepsilon)$ iterations.

Then, there is an explicit algorithm which successfully optimizes the class of convex and β-smooth functions in $O(\phi(2\beta R^2/\varepsilon))$ iterations.

Proof. Let f be convex and β-smooth. We apply the given algorithm to the regularized function $f_\delta := f + \frac{\delta}{2}\|\cdot - x_0\|^2$, obtaining a point x such that $f_\delta(x) \le \min f_\delta + \varepsilon/2$. If $x_{\delta,\star}$ denotes the minimizer of f_δ, then

$$f(x) \le f_\delta(x) \le f_\delta(x_{\delta,\star}) + \frac{\varepsilon}{2} \le f_\delta(x_\star) + \frac{\varepsilon}{2} = f_\star + \frac{\delta}{2}\|x_0 - x_\star\|^2 + \frac{\varepsilon}{2} .$$

We now set $\delta = \varepsilon/R^2$, so that $f(x) - f_\star \le \varepsilon$.

It remains to estimate the complexity. We first note that $f_\delta(x_{\delta,\star}) \le f_\delta(x_\star)$ implies $\|x_0 - x_{\star,\delta}\| \le \|x_0 - x_\star\|$, so the initial distance to the minimizer of f_δ is also bounded by R. We can assume that $\varepsilon \le \beta R^2$ (or else the minimization problem is trivial). Then, the smoothness of f_δ is bounded by $\beta + \delta \le 2\beta$, and the condition number of f_δ is bounded by $2\beta R^2/\varepsilon$. Substitute these quantities into the complexity of the given algorithm. □

23

Thus, the $\alpha > 0$ case of Theorem 3.4 and Lemma 4.2 recover the $\alpha = 0$ case of Theorem 3.4 up to constants.

Taken together, Lemma 4.1 and Lemma 4.2 show that the 0-convex and strongly convex settings are essentially equivalent to each other, in that an optimal method for one class yields an optimal method for the other class. Thus, we now aim to address the following question: what is the smallest possible $\phi(\cdot)$?

4.2 Lower bounds

According to the discussion in §1.1, establishing a lower complexity bound requires showing that *any* algorithm which interacts with the first-order oracle using at most a prescribed number of queries cannot have performance better than the lower bound. Actually, although this is possible (see [NY83]), it is not especially easy. It was shown by Nesterov in an earlier edition of [Nes18] that by imposing natural restrictions on the class of algorithms under consideration, it is possible to establish the lower bounds in a more transparent way. Accordingly, his approach has become standard in the field, and it is the approach we adopt here as well. It does, however, have the drawback of not applying to general query algorithms; for example, it does not apply against randomized algorithms.

The class of algorithms we consider is the following one.

Definition 4.3. An algorithm is called a **gradient span** algorithm if it deterministically generates a sequence of points $\{x_n\}_{n \in \mathbb{N}}$ such that for all $n \in \mathbb{N}$,

$$x_{n+1} \in x_0 + \mathrm{span}\{\nabla f(x_0), \ldots, \nabla f(x_n)\} .$$

For example, GD is a gradient span algorithm. On the basis of this assumption, we now establish the following result; recall the asymptotic notation $\gtrsim$, which only hides a universal constant.

Theorem 4.4 (lower bound for convex, smooth minimization). For any $1 \le N \le \frac{d-1}{2}$, $\beta > 0$, and $x_0 \in \mathbb{R}^d$, there exists a convex and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any gradient span algorithm,

$$f(x_N) - f_\star \gtrsim \frac{\beta \|x_0 - x_\star\|^2}{N^2} .$$

In other words, in order to obtain $f(x_N) - f_\star \le \varepsilon$, the number of iterations must satisfy

$$N \gtrsim \sqrt{\frac{\beta \|x_0 - x_\star\|^2}{\varepsilon}} .$$

24

Before proving this result, we observe that by applying Lemma 4.2 with $\phi(x) \asymp \sqrt{x}$, it yields the following corollary.

> **Theorem 4.5** (lower bound for strongly convex, smooth minimization). For any $0 < \alpha < \beta$, any $\varepsilon > 0$, any d sufficiently large, and any $x_0 \in \mathbb{R}^d$, there exists an α-convex and β-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that for any gradient span algorithm, in order to obtain $f(x_N) - f_\star \leq \varepsilon$, the number of iterations must satisfy
> $$N \gtrsim \sqrt{\kappa} \log \frac{\alpha \|x_0 - x_\star\|^2}{\varepsilon}.$$

Proof of Theorem 4.4. By translating the problem, we may assume $x_0 = 0$. The construction is based on the following function:

$$f_n : \mathbb{R}^d \to \mathbb{R}, \qquad f_n(x) := \frac{\beta}{4} \left\{ \frac{1}{2} \left(x[1]^2 + \sum_{k=1}^{n-1} (x[k] - x[k+1])^2 + x[n]^2 \right) - x[1] \right\}.$$

For any $v \in \mathbb{R}^d$,

$$\langle v, \nabla^2 f_n(x) v \rangle = \frac{\beta}{4} \left(v[1]^2 + \sum_{k=1}^{n-1} (v[k] - v[k+1])^2 + v[n]^2 \right) \leq \beta \|v\|^2,$$

so each f_n is convex and β-smooth.

We prove by induction that when we apply a gradient span algorithm to f_d, the n-th iterate x_n belongs to the subspace

$$\mathcal{V}_n := \{x \in \mathbb{R}^d : x[k] = 0 \text{ for all } k = n+1, \ldots, d\}.$$

Clearly, $x_0 \in \mathcal{V}_0$. Inductively, suppose that $x_k \in \mathcal{V}_k$ for all $k \leq n$. Then,

$$\nabla f_d(x_k) = \frac{\beta}{4} \left(x_k[1] \, e_1 + \sum_{j=1}^{k} (x_k[j] - x_k[j+1]) \, (e_j - e_{j+1}) \right) - \frac{\beta}{4} e_1 \in \mathcal{V}_{k+1},$$

hence

$$x_{n+1} \in \operatorname{span}\{\nabla f_d(x_0), \ldots, \nabla f_d(x_n)\} \subseteq \mathcal{V}_{n+1}.$$

This completes the induction. Also, since $f_N = f_d$ on $\mathcal{V}_N$, it follows that

$$f_d(x_N) = f_N(x_N) \geq (f_N)_\star.$$

The next step is to estimate $(f_n)_\star := \min f_n$ for all n. By setting the gradient to zero, $\nabla f_n((x_n)_\star) = 0$, we obtain the following system of equations:

$$2 (x_n)_\star[1] - (x_n)_\star[2] = 1,$$
$$(x_n)_\star[k-1] - 2 (x_n)_\star[k] + (x_n)_\star[k+1] = 0, \qquad \text{for } k = 1, \ldots, n,$$
$$-(x_n)_\star[n-1] + 2 (x_n)_\star[n] = 0.$$

The solution is $(x_n)_\star[k] = 1 - \frac{k}{n+1}$ for all $k \in [n]$. Writing $f_n(x) = \frac{\beta}{4} \{\frac{1}{2} \langle x, A_n x \rangle - \langle e_1, x \rangle\}$, the system above reads $A_n x = e_1$, hence

$$(f_n)_\star = f_n((x_n)_\star) = -\frac{\beta}{8} \langle e_1, (x_n)_\star \rangle = -\frac{\beta}{8} \left(1 - \frac{1}{n+1}\right).$$

Moreover, $\|x_0 - (x_n)_\star\|^2 = \|(x_n)_\star\|^2 \le n$. Finally, it yields

$$f_d(x_N) - (f_d)_\star \ge (f_N)_\star - (f_d)_\star \ge \frac{\beta}{8} \left(\frac{1}{N+1} - \frac{1}{d+1}\right)$$
$$\ge \frac{\beta \|x_0 - (x_d)_\star\|^2}{8d} \left(\frac{1}{N+1} - \frac{1}{d+1}\right).$$

Choosing $d \asymp N$, e.g., $d = 2N + 1$, yields the stated lower bound. $\qquad \square$

Notably, the iteration complexity lower bounds Theorem 4.4 and Theorem 4.5 are smaller than the bounds attained by GD in Theorem 3.4 by a square root. As developed in the next sections, in fact the lower bounds are tight and GD is suboptimal.

We make two further remarks. First, it is perhaps surprising that the lower bound construction is a *quadratic* function; in some sense, quadratics are the hardest convex and smooth functions to optimize. Second, the lower bound requires the ambient dimension to be larger than the iteration count; this is crucial for the proof technique, which relies on the algorithm discovering one new dimension per iteration. This turns out to be fundamental because there are better methods in low dimension, for quadratics and even for general convex functions.

Exercises

Exercise 4.1. In the setting of Theorem 4.4 and using the same construction as in the proof, show that $\|x_N - x_\star\|^2 \gtrsim \|x_0 - x_\star\|^2$. In other words, in the 0-convex case, it is not possible to make progress in the sense of distance to the minimizer by more than a constant factor.

Exercise 4.2. We used the reductions from §4.1 to reduce the strongly convex lower bound to the 0-convex lower bound for the sake of brevity, but it is of course possible to develop the strongly convex lower bound directly. Consider the function

$$f : \mathbb{R}^\infty \to \mathbb{R}, \qquad f(x) := \frac{\beta - \alpha}{8} \left\{ x[1]^2 + \sum_{n=1}^{\infty} (x[n] - x[n+1])^2 - 2x[1] \right\} + \frac{\alpha}{2} \|x\|^2.$$

By adapting the proof of Theorem 4.4, show that any gradient span algorithm satisfies

$$f(x_N) - f_\star \geq \frac{\alpha}{2} \|x_N - x_\star\|^2 \geq \frac{\alpha}{2} \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{2N} \|x_0 - x_\star\|^2.$$

References

[Bub15]   S. Bubeck. "Convex optimization: algorithms and complexity". In: *Foundations and Trends® in Machine Learning* 8.3-4 (2015), pp. 231–357.

[Che25]   S. Chewi. *Log-concave sampling*. Available online at chewisinho.github.io. Forthcoming, 2025.

[CLM24]   H. Chardon, M. Lerasle, and J. Mourtada. "Finite-sample performance of the maximum likelihood estimator in logistic regression". In: *arXiv preprint 2411.02137* (2024).

[KNS16]   H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition". In: *European Conference on Machine Learning and Knowledge Discovery in Databases—Volume 9851*. ECML PKDD 2016. Riva del Garda, Italy: Springer-Verlag, 2016, pp. 795–811.

[LMW24]   J. Liang, S. Mitra, and A. Wibisono. "On independent samples along the Langevin diffusion and the unadjusted Langevin algorithm". In: *arXiv preprint 2402.17067* (2024).

[Łoj63]   S. Łojasiewicz. "Une propriété topologique des sous-ensembles analytiques réels". In: *Les Équations aux Dérivées Partielles (Paris, 1962)*. Éditions du Centre National de la Recherche Scientifique (CNRS), Paris, 1963, pp. 87–89.

[Nes18]   Y. Nesterov. *Lectures on convex optimization*. Vol. 137. Springer Optimization and Its Applications. Springer, 2018, pp. xxiii+589.

[NY83]   A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. Translated from the Russian and with a preface by E. R. Dawson. John Wiley & Sons, Inc., New York, 1983, pp. xv+388.

[OV00]   F. Otto and C. Villani. "Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality". In: *J. Funct. Anal.* 173.2 (2000), pp. 361–400.

[OV01]   F. Otto and C. Villani. "Comment on: "Hypercontractivity of Hamilton–Jacobi equations" [J. Math. Pures Appl. (9) **80** (2001), no. 7, 669–696] by S. G. Bobkov, I. Gentil and M. Ledoux". In: *J. Math. Pures Appl. (9)* 80.7 (2001), pp. 697–700.

[Pol63]   B. T. Polyak. "Gradient methods for minimizing functionals". In: *Ž. Vyčisl. Mat i Mat. Fiz.* 3 (1963), pp. 643–653.

[Ver18]   R. Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018, pp. xiv+284.