

Lectures on Optimization

Sinho Chewi

January 27, 2025

Contents

1	[1/14] Introduction and basics of convex functions	2
1.1	Overview of the course	2
1.2	Preliminaries on convexity and smoothness	7
2	[1/16] Gradient flow	11
3	[1/21] Gradient descent: smooth case	16
4	[1/23] Lower bounds for smooth optimization	22
4.1	Reductions between the convex and strongly convex settings	22
4.2	Lower bounds	24
5	[1/28–1/30] Acceleration	27
5.1	Quadratic case: the conjugate gradient method	27
5.2	General case: continuous time	32
5.3	General case: discrete time	34

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 $f : \mathcal{X} \rightarrow \mathbb{R}$. 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 practice¹—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

¹There 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 $f(x) - \min f \leq \varepsilon$.

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*:

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

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

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

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

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

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

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

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 $(1 - t)x + ty$ also lies in \mathcal{C} .

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

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

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:

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

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

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

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

Proof. Assume that f is continuously differentiable.

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

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

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

(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 $v := y - x$,

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

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

Hence, (1.5) yields

$$\begin{aligned} f((1-t)x + ty) - (1-t)f(x) - tf(y) &= -t \int_0^1 \langle \nabla f(x + sv) - \nabla f(x + stv), v \rangle ds \\ &\leq -t \int_0^1 \alpha s(1-t) \|v\|^2 ds = -\frac{\alpha}{2} t(1-t) \|v\|^2. \end{aligned}$$

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). \square

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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, $f(x_n) \rightarrow \inf f$. 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, $f(x_\star) = \lim_{n \rightarrow \infty} f(x_n) = \inf f$. \square

Lemma 1.8 (necessary conditions for optimality). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is strictly convex, i.e., for all distinct $x, y \in \mathbb{R}^d$ and $t \in (0, 1)$, $f((1-t)x + ty) < (1-t)f(x) + tf(y)$. 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be strongly convex and continuously differentiable. Then, it admits a unique minimizer x_\star , which is characterized by $\nabla f(x_\star) = 0$.

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is β -**smooth** if it is continuously differentiable and

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

The following proposition is established in the same way as [Proposition 1.6](#), so we omit the proof.

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

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

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

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

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

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

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.

²This 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 :

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

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:

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$, 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

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 $u : [0, T] \rightarrow \mathbb{R}$ is a continuously differentiable curve that satisfies the differential inequality

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

Then, it holds that

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

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

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

By the fundamental theorem of calculus,

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

Rearranging yields the result. \square

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 $\|y_t - x_t\|^2 \leq \exp(-2\alpha t) \|y_0 - x_0\|^2$, which proves the theorem. Note in particular that we can take $y_t = x_\star$ for all $t \geq 0$, so it yields the following statement about convergence to the minimizer: $\|x_t - x_\star\| \leq \exp(-\alpha t) \|x_0 - x_\star\|$.

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be α -convex, $\alpha \geq 0$. Then, for all $t \geq 0$,

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

When $\alpha = 0$, the right-hand side should be interpreted as its limiting value as $\alpha \rightarrow 0$, 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

$$\begin{aligned} \int_0^t \exp(-\alpha(t-s)) (f(x_s) - f_\star) ds &\geq (f(x_t) - f_\star) \int_0^t \exp(-\alpha(t-s)) ds \\ &= (f(x_t) - f_\star) \frac{1 - \exp(-\alpha t)}{\alpha}. \end{aligned}$$

Rearranging yields the result. \square

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

$$\|\nabla f(x)\|^2 \geq 2\alpha (f(x) - f_\star) \quad \text{for all } x \in \mathbb{R}^d. \quad (\text{PL})$$

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

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

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

We present a few key properties of the PL inequality.

Proposition 2.7 (strong convexity \Rightarrow PL \Rightarrow quadratic growth). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\alpha > 0$. The following implications hold.

1. If f is α -convex, then f satisfies (PL) with constant α .
2. If f satisfies (PL) with constant α , then it satisfies the following **quadratic growth** property:

$$f(x) - f_\star \geq \frac{\alpha}{2} \inf_{x_\star \in \mathcal{X}_\star} \|x - x_\star\|^2, \quad \text{for all } x \in \mathbb{R}^d, \quad (\text{QG})$$

where \mathcal{X}_\star denotes the set of minimizers of f .

Proof.

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

$$\begin{aligned} -(f(x) - f_\star) &\geq \langle \nabla f(x), x_\star - x \rangle + \frac{\alpha}{2} \|x - x_\star\|^2 \\ &\geq -\|\nabla f(x)\| \|x_\star - x\| + \frac{\alpha}{2} \|x - x_\star\|^2 \geq -\frac{1}{2\alpha} \|\nabla f(x)\|^2, \end{aligned}$$

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 (PL). 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 $\mathcal{L}_0 \geq \mathcal{L}_\infty$.

□

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

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

(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 $\|\nabla f(x_{t_n})\| \rightarrow 0$. (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 [Loj63] 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 [Loj63; 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, $\dot{\mathcal{L}}_t \leq 0$:

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

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

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

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

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

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

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

Exercise 2.3. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be α -convex with $\alpha > 0$, and let $g : \mathbb{R}^d \rightarrow \mathbb{R}^n$ 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 $\nabla g(x)^\top \nabla g(x) \geq \sigma I_n$. Show that the composition $f \circ g$ satisfies (PL) 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:

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$, if $h \leq 1/\beta$, then

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

Proof. By the smoothness inequality (1.7),

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

If $h \leq 1/\beta$, then $-h(1 - \beta h/2) \leq -h/2$. □

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 $\kappa := \beta/\alpha \geq 1$.

When f is quadratic, $f(x) = \frac{1}{2} \langle x, Ax \rangle$ 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 $h \leq 1/\beta$,

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

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 (3.4) in [Exercise 3.1](#) below, for $h \leq 1/\beta$ and from (1.5) we have

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

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

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

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

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 $\|x_N - x_\star\| \leq \varepsilon$, we would then have the estimate $N \geq 2\kappa^2 \log(\|x_0 - x_\star\|/\varepsilon)$, 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 $h \leq 1/\beta$,

$$\|x^+ - x_\star\|^2 \leq (1 - \alpha h) \|x - x_\star\|^2 - 2h (f(x^+) - f_\star). \quad (3.1)$$

Therefore,

$$f(x_N) - f_\star \leq \frac{\alpha}{2 \{(1 - \alpha h)^{-N} - 1\}} \|x_0 - x_\star\|^2. \quad (3.2)$$

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

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

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

For $h \leq 1/\beta$, 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. \square

We remark for later use that the proof of (3.1) goes through even if we replace x_\star with any other point $z \in \mathbb{R}^d$, i.e.,

$$\|x^+ - z\|^2 \leq (1 - \alpha h) \|x - z\|^2 - 2h (f(x^+) - f(z)), \quad \text{for all } z \in \mathbb{R}^d. \quad (3.3)$$

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 $A > 0$,

$$u_{n+1} \leq Au_n + B_n \quad \text{for } n = 0, 1, \dots, N-1.$$

Then,

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

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

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

Rearrange to obtain the result. \square

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

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

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

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

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

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

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),

$$\begin{aligned} f(x^+) - f_\star &= f(x) - f_\star + f(x^+) - f(x) \leq f(x) - f_\star - \frac{h}{2} \|\nabla f(x)\|^2 \\ &\leq (1 - \alpha h) (f(x) - f_\star). \end{aligned} \quad \square$$

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

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

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

Proof. Telescope the descent lemma (Lemma 3.1):

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

We summarize the results for GD in Table 1.

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

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, $X_i \stackrel{\text{i.i.d.}}{\sim} \text{normal}(0, I)$, and assume that the data is generated from the model with a true parameter θ^\star . Let $\widehat{\mathcal{L}}$ denote the MLE objective, let $\mathcal{L} := \mathbb{E} \widehat{\mathcal{L}}$ denote the population risk, and let $R := \|\theta^\star\| \geq 1$. The state-of-the-art result [[CLM24](#)] shows that if $n \gtrsim Rd$ for a sufficiently large implied constant, $\widehat{\theta}_{\text{MLE}}$ exists with probability $\geq 1 - \exp(-d)$ and satisfies the optimal risk bound $\mathcal{L}(\widehat{\theta}_{\text{MLE}}) - \mathcal{L}(\theta^\star) \lesssim d/n$.

In practice, we cannot compute $\widehat{\theta}_{\text{MLE}}$ exactly, so we use optimization. From [[CLM24](#)], any estimator $\widehat{\theta}$ satisfying $\widehat{\mathcal{L}}(\widehat{\theta}) - \widehat{\mathcal{L}}(\widehat{\theta}_{\text{MLE}}) \lesssim d/n$ satisfies the same statistical risk bound as $\widehat{\theta}_{\text{MLE}}$, up to a universal constant. We take $\widehat{\theta} = \widehat{\theta}_{\text{GD}}$ 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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

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

From this, deduce that

$$\|\nabla f(y) - \nabla f(x)\|^2 \leq \beta \langle \nabla f(y) - \nabla f(x), y - x \rangle. \quad (3.4)$$

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 $T := \text{id} - h \nabla f$ denote the one-step GD mapping. By the fundamental theorem of calculus,

$$\begin{aligned} \|y^+ - x^+\| &= \|T(y) - T(x)\| = \left\| \int_0^1 \nabla T((1-t)x + ty) (y-x) dt \right\| \\ &\leq \left(\int_0^1 \|\nabla T((1-t)x + ty)\|_{\text{op}} dt \right) \|y-x\|. \end{aligned}$$

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 $\|\nabla T(z)\|_{\text{op}}$ 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 $\lambda_{\max}(\frac{1}{n} \sum_{i=1}^n X_i X_i^\top) \lesssim 1$ with high probability³ 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 \mathcal{F} in $\phi(\mathcal{F}, R, \varepsilon)$ iterations if, given any $f \in \mathcal{F}$ and $x_0 \in \mathbb{R}^d$ with $\|x_0 - x_\star\| \leq R$, it outputs x with $f(x) - f_\star \leq \varepsilon$ using no more than $\phi(\mathcal{F}, R, \varepsilon)$ queries to a first-order oracle for f .

³This 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

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

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

$$\|x_1 - x_\star\| \leq \frac{1}{2} R = \frac{1}{2} \|x_0 - x_\star\|. \quad (4.1)$$

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 $\tilde{R} := \|\tilde{x} - x_\star\| \leq \sqrt{\varepsilon/\alpha}$. Finally, apply the given algorithm one more time starting from \tilde{x} with target accuracy ε to obtain a point x with $f(x) - f_\star \leq \varepsilon$. The complexity of this final step is $\phi(\beta \tilde{R}^2/\varepsilon) = \phi(\kappa)$. \square

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 $f_\delta(x) \leq \min f_\delta + \varepsilon/2$. If $x_{\delta,\star}$ denotes the minimizer of f_δ , then

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

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

It remains to estimate the complexity. We first note that $f_\delta(x_{\delta,\star}) \leq f_\delta(x_\star)$ implies $\|x_0 - x_{\delta,\star}\| \leq \|x_0 - x_\star\|$, so the initial distance to the minimizer of f_δ is also bounded by R . We can assume that $\varepsilon \leq \beta R^2$ (or else the minimization problem is trivial). Then, the smoothness of f_δ is bounded by $\beta + \delta \leq 2\beta$, 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. \square

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 $n \in \mathbb{N}$,

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

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 $1 \leq N \leq \frac{d-1}{2}$, $\beta > 0$, and $x_0 \in \mathbb{R}^d$, there exists a convex and β -smooth function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $f(x_N) - f_\star \leq \varepsilon$, the number of iterations must satisfy

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

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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that for any gradient span algorithm, in order to obtain $f(x_N) - f_\star \leq \varepsilon$, the number of iterations must satisfy

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

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

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

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

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

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

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

hence

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

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:

$$\begin{aligned} 2x_{n,\star}[1] - x_{n,\star}[2] &= 1, \\ x_{n,\star}[k-1] - 2x_{n,\star}[k] + x_{n,\star}[k+1] &= 0, \quad \text{for } k = 1, \dots, n, \\ -x_{n,\star}[n-1] + 2x_{n,\star}[n] &= 0. \end{aligned}$$

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 $A_n x_{n,\star} = e_1$, 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, $\|x_0 - x_{n,\star}\|^2 = \|x_{n,\star}\|^2 \leq n$. Finally, it yields

$$\begin{aligned} f_d(x_N) - (f_d)_\star &\geq (f_N)_\star - (f_d)_\star \geq \frac{\beta}{8} \left(\frac{1}{N+1} - \frac{1}{d+1} \right) \\ &\geq \frac{\beta \|x_0 - x_{d,\star}\|^2}{8d} \left(\frac{1}{N+1} - \frac{1}{d+1} \right). \end{aligned}$$

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

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

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

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

5 [1/28–1/30] Acceleration

We now show that the lower bounds of Theorem 4.4 and Theorem 4.5 can be attained via algorithms which improve upon GD. This is known as the *acceleration* phenomenon in optimization. We begin with the quadratic case.

5.1 Quadratic case: the conjugate gradient method

In this section, the objective function is quadratic:

$$f : \mathbb{R}^d \rightarrow \mathbb{R}, \quad f(x) = \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle,$$

where A is a symmetric matrix, $A \succ 0$. Note also that minimizing f corresponds to solving the system of equations $Ax_\star = b$. We now introduce the *conjugate gradient* method.

The method is succinctly described as follows:

$$x_{n+1} := \arg \min \{ f(x) \mid x \in x_0 + \text{span}\{\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_n)\} \}. \quad (\text{CG})$$

This scheme is very natural in light of the definition of a gradient span algorithm (Definition 4.3) that we encountered for the lower bounds. However, it is not yet clear that (CG) can be implemented cheaply. Using the fact that f is quadratic, our aim is to show that (CG) can be rewritten as a simple iteration that uses one gradient query per step.

As is usually the case in linear algebra, instead of working with the set of vectors $\{\nabla f(x_0), \nabla f(x_1), \dots, \nabla f(x_n)\}$, it is more convenient to work with an *orthonormal* set $\{p_0, p_1, \dots, p_n\}$. Here, orthonormality is with respect to the inner product $\langle \cdot, \cdot \rangle_A$, i.e., we will require $\langle p_i, A p_j \rangle = \delta_{i,j}$ for all $i, j \in [n]$. We start with $p_0 := \nabla f(x_0)$, and we write $\mathcal{K}_n := \text{span}\{p_0, p_1, \dots, p_n\}$. We must address the following two questions:

- Given \mathcal{K}_n and x_n , how can we compute $x_{n+1} = \arg \min_{x_0 + \mathcal{K}_n} f$?
- Given \mathcal{K}_n and $\nabla f(x_{n+1})$, how can we compute p_{n+1} and thus \mathcal{K}_{n+1} ?

For the first question, we may assume inductively that $x_n = \arg \min_{x_0 + \mathcal{K}_{n-1}} f$, which means that $\langle \nabla f(x_n), p_k \rangle = 0$ for all $k < n$. The next point is taken to be $x_{n+1} = x_n + h_n p_n$, chosen so that $\langle \nabla f(x_{n+1}), p_k \rangle = 0$ for all $k \leq n$. Since ∇f is linear,

$$\langle \nabla f(x_{n+1}), p_k \rangle = \langle \nabla f(x_n) + h_n A p_n, p_k \rangle.$$

For $k < n$, this equals zero by the inductive hypothesis on x_n , and the orthonormality of $\{p_0, p_1, \dots, p_n\}$. We choose h_n to ensure that this equals zero for $k = n$ too:

$$h_n = -\frac{\langle \nabla f(x_n), p_n \rangle}{\|p_n\|_A^2}.$$

For the second question, we want to compute the Gram–Schmidt orthogonalization of $\nabla f(x_{n+1})$ w.r.t. $\{p_0, p_1, \dots, p_n\}$ in the $\langle \cdot, \cdot \rangle_A$ inner product. We claim that $\nabla f(x_{n+1})$ is already A -orthogonal to p_k for $k < n$, so that

$$p_{n+1} = \nabla f(x_{n+1}) - \langle \nabla f(x_{n+1}), p_n \rangle_A \frac{p_n}{\|p_n\|_A^2}. \quad (5.1)$$

To justify this, we show that for $k < n$, $\boxed{A p_k \in \mathcal{K}_{k+1}}$, hence

$$\langle \nabla f(x_{n+1}), p_k \rangle_A = \langle \nabla f(x_{n+1}), A p_k \rangle = 0$$

using the fact shown above that $\nabla f(x_{n+1})$ is orthogonal (in the usual inner product) to \mathcal{K}_n . Finally, the boxed equation is shown through the following lemma.

Lemma 5.1. For all $n \in \mathbb{N}$,

$$\mathcal{K}_n = \text{span}\{p_0, A p_0, \dots, A^n p_0\}.$$

Proof. We proceed via induction, where the case $n = 0$ is obvious. Assuming it holds at iteration n , let us show that $\widetilde{\mathcal{K}}_{n+1} \in \widetilde{\mathcal{K}}_{n+1} := \text{span}\{p_0, A p_0, \dots, A^{n+1} p_0\}$. By (5.1), it suffices to show that $\nabla f(x_{n+1}) \in \mathcal{K}_{n+1}$. However, as discussed above, $\nabla f(x_{n+1}) = \nabla f(x_n) + h_n A p_n = p_0 + h_0 A p_0 + \dots + h_n A p_n \in \widetilde{\mathcal{K}}_{n+1}$.

Conversely, we must show that $A^{n+1} p_0 \in \mathcal{K}_{n+1}$. Since $A^n p_0 \in \mathcal{K}_n$, we can write $A^n p_0 = \sum_{k=0}^n c_k p_k$, thus $A^{n+1} p_0 = \sum_{k=0}^n c_k A p_k$. By the inductive hypothesis, each $A p_k$ for $k < n$ belongs to \mathcal{K}_n , so it suffices to have $A p_n \in \mathcal{K}_{n+1}$. However, we can observe that $A p_n = h_n^{-1} (\nabla f(x_{n+1}) - \nabla f(x_n)) \in \mathcal{K}_{n+1}$ by (5.1). \square

Definition 5.2. The subspaces $\{\mathcal{K}_n\}_{n \in \mathbb{N}}$ are called **Krylov subspaces**.

Finally, let us write the iterations in a form which is convenient for implementation. Note first that $\langle \nabla f(x_n), \nabla f(x_{n+1}) \rangle = 0$ (indeed, $\nabla f(x_{n+1})$ is orthogonal to all of \mathcal{K}_n). So,

$$\frac{\langle \nabla f(x_{n+1}), p_n \rangle_A}{\|p_n\|_A^2} = \frac{\langle \nabla f(x_{n+1}), \nabla f(x_{n+1}) - \nabla f(x_n) \rangle}{h_n \|p_n\|_A^2} = -\frac{\|\nabla f(x_{n+1})\|^2}{\langle \nabla f(x_n), p_n \rangle}$$

and $\|\nabla f(x_n)\|^2 = \langle \nabla f(x_n), \nabla f(x_n) \rangle = \langle \nabla f(x_n), p_n \rangle$ using (5.1) and the fact that $\nabla f(x_n)$ is orthogonal to \mathcal{K}_{n-1} . This yields the following iteration, where we write $r_n := Ax_n - b = \nabla f(x_n)$ for the residual.

$$x_{n+1} = x_n - \frac{\|r_n\|^2}{\langle p_n, A p_n \rangle} p_n, \quad r_{n+1} = r_n - \frac{\|r_n\|^2}{\langle p_n, A p_n \rangle} A p_n, \quad p_{n+1} = r_{n+1} + \frac{\|r_{n+1}\|^2}{\|r_n\|^2} p_n.$$

This algorithm requires one matrix-vector multiplication per iteration, namely, the computation of $A p_n$.

Note that if $p_{n+1} = 0$, then $\nabla f(x_{n+1}) \in \mathcal{K}_n$, yet $\nabla f(x_{n+1}) \perp \mathcal{K}_n$ and thus $\nabla f(x_{n+1}) = 0$, $x_{n+1} = x_\star$. Since $p_{d+1} = 0$ (an orthonormal set in \mathbb{R}^d cannot have more than d elements), we arrive at the following conclusion.

Theorem 5.3 (termination of CG). The **CG** algorithm returns the exact minimizer in at most d iterations.

Let us now show that **CG** can find an approximate minimizer at the accelerated rate.

Theorem 5.4 (accelerated convergence for CG). Let $0 < A \leq \beta I$. Then, **CG** satisfies

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

Proof. By the descent lemma (Lemma 3.1) and the defining property of **CG**,

$$f(x_{n+1}) \leq f\left(x_n - \frac{1}{\beta} \nabla f(x_n)\right) \leq f(x_n) - \frac{1}{2\beta} \|\nabla f(x_n)\|^2,$$

so that

$$f(x_N) - f_\star \geq \frac{1}{2\beta} \sum_{n=0}^{N-1} \|\nabla f(x_n)\|^2.$$

On the other hand, since $\nabla f(x_n) \perp x_{k+1} - x_k$ for $k < n$,

$$f_\star - f(x_n) \geq \langle \nabla f(x_n), x_\star - x_n \rangle = \langle \nabla f(x_n), x_\star - x_0 \rangle.$$

If we sum these inequalities and use orthogonality of the gradients,

$$\begin{aligned} N(f(x_N) - f_\star) &\leq \sum_{n=0}^{N-1} (f(x_n) - f_\star) \leq \left\langle \sum_{n=0}^{N-1} \nabla f(x_n), x_0 - x_\star \right\rangle \leq \left\| \sum_{n=0}^{N-1} \nabla f(x_n) \right\| \|x_0 - x_\star\| \\ &= \left(\sum_{n=0}^{N-1} \|\nabla f(x_n)\|^2 \right)^{1/2} \|x_0 - x_\star\| \leq \sqrt{2\beta(f(x_N) - f_\star)} \|x_0 - x_\star\|. \end{aligned}$$

Rearrange to complete the proof. \square

By applying the restart strategy as in [Lemma 4.1](#), one can also show an iteration complexity scaling with $\sqrt{\kappa}$ in the strongly convex case. However, we instead give a different proof in order to explain the classical link with polynomial approximation.

Due to [Lemma 5.1](#), $x_N - x_0 \in \mathcal{K}_{N-1}$ can be written in the form $x_N - x_0 = \sum_{n=0}^{N-1} c_n A^n p_0$, so $x_N - x_\star = x_0 - x_\star + \sum_{n=0}^{N-1} c_n A^{n+1} (x_0 - x_\star) = P_N(A) (x_0 - x_\star)$ where P_N is a polynomial of degree at most N satisfying $P_N(0) = 1$. Conversely, if Q_N is any other degree- N polynomial with $Q_N(0) = 1$, then $\tilde{x}_N := x_0 + A^{-1} (Q_N(A) - I) p_0 \in x_0 + \mathcal{K}_{N-1}$ satisfies $\tilde{x}_N - x_\star = x_0 - x_\star + A^{-1} (Q_N(A) - I) p_0 = Q_N(A) (x_0 - x_\star)$.

This equivalence, together with the fact that the output x_N of [CG](#) minimizes f over $x_0 + \mathcal{K}_{N-1}$, shows that

$$f(x_N) - f_\star \leq \frac{1}{2} \min \{ \|Q_N(A) (x_0 - x_\star)\|_A^2 : Q_N \in \mathbb{R}_{\leq N}[X], Q_N(0) = 1 \},$$

where $\mathbb{R}_{\leq N}[X]$ denotes the set of polynomials with real-valued coefficients and with degree at most N . Furthermore, since A and $Q_N(A)$ commute,

$$\|Q_N(A) (x_0 - x_\star)\|_A^2 \leq \|Q_N(A)\|_{\text{op}}^2 \|x_0 - x_\star\|_A^2 \leq \left(\max_{[\lambda_{\min}(A), \lambda_{\max}(A)]} |Q_N(\lambda)|^2 \right) \|x_0 - x_\star\|_A^2.$$

We have arrived at the following result.

Lemma 5.5 (CG and polynomial approximation). Assume that $0 < \alpha I \leq A \leq \beta I$. Then, the output x_N of [CG](#) satisfies

$$f(x_N) - f_\star \leq \min \left\{ \max_{\lambda \in [\alpha, \beta]} |Q_N(\lambda)|^2 : Q_N \in \mathbb{R}_{\leq N}[X], Q_N(0) = 1 \right\} (f(x_0) - f_\star).$$

Informally, this result states that **CG** performs as well as the best possible degree- N polynomial in A . To bound the rate of convergence of **CG**, it therefore remains to exhibit a judicious polynomial Q_N . This is accomplished by the family of Chebyshev polynomials, on which many volumes have been written.

Definition 5.6. The degree- n **Chebyshev polynomial** T_n is defined so that $\cos(n\theta) = T_n(\cos \theta)$ for all $\theta \in \mathbb{R}$.

It is not obvious at first glance that T_n is indeed a degree- n polynomial, but this can be established via trigonometric identities. The use of the Chebyshev polynomials to establish a rate of convergence for **CG** is explored in [Exercise 5.1](#).

Here, we point out another interesting fact that arises from this connection. Recall from the proof of [Lemma 5.5](#) that if we can compute $\tilde{x}_N := x_0 + A^{-1} (Q_N(A) - I) p_0$, then it incurs error at most $f(\tilde{x}_N) - f_\star \leq (\max_{\lambda \in [\alpha, \beta]} |Q_N(\lambda)|^2) (f(x_0) - f_\star)$. In particular, rather than using **CG**, we can try to compute the polynomial $x \mapsto (Q_N(x) - 1)/x$ directly, where Q_N is the polynomial in [Exercise 5.1](#) which witnesses the fast convergence of **CG**. Although we omit the details, it is worth noting that the family of Chebyshev polynomials satisfies a so-called three-term recurrence:

$$T_{n+1}(x) = 2x T_n(x) - T_{n-1}(x), \quad x \in \mathbb{R}.$$

In fact, orthogonal families of polynomials usually do.⁴ From an algorithmic standpoint, it leads to an optimization algorithm of the form

$$x_{n+1} = c_0 A x_n + c_1 x_{n-1} + c_2 b,$$

where $c_0, c_1, c_2 \in \mathbb{R}$ are fixed coefficients. Note that unlike **GD**, x_{n+1} depends on the previous *two* iterates. This is often referred to as *momentum*, and also forms the basis for acceleration for general convex functions.

Remark 5.7 (practicality of **CG).** Solving the linear system $Ax = b$ via Gaussian elimination requires $O(d^3)$ operations and is numerically unstable, whereas for well-conditioned matrices A , **CG** returns an approximate solution in $\tilde{O}(\sqrt{\kappa})$ iterations, each of which requires a matrix-vector multiplication. A matrix-vector multiplication requires $O(d^2)$ time in the worst case, but can be faster if A is sparse. In practice, **CG** is widely used, especially when combined with other strategies such as preconditioning.

⁴This arises in connection with second-order differential operators.

5.2 General case: continuous time

Although it does not follow the historical development of events, we begin our treatment of acceleration for general convex smooth functions in continuous time. As identified in [SBC16], the continuous-time ODE is

$$\begin{aligned}\dot{x}_t &= p_t, \\ \dot{p}_t &= -\nabla f(x_t) - \gamma_t p_t.\end{aligned}\tag{AGF}$$

We refer to (AGF) as the *accelerated gradient flow*, and the variable p_t admits the physical interpretation of momentum (for a particle with unit mass). The dynamics consists of two parts: the equations

$$\begin{aligned}\dot{x}_t &= p_t, \\ \dot{p}_t &= -\nabla f(x_t)\end{aligned}$$

are known as Hamilton's equations, and they are the standard first-order reformulation of Newton's law of motion $\ddot{x}_t = -\nabla f(x_t)$ with potential energy f . Hamilton's equations conserve the energy (or Hamiltonian) $H(x, p) := f(x) + \frac{1}{2} \|p\|^2$, and this conservation property is perhaps undesirable for an optimization algorithm which seeks to minimize f . Thus, the second part of the dynamics, $\dot{p}_t = -\gamma_t p_t$ adds a dissipative *friction* force, where $\gamma_t \geq 0$ is a possibly time-varying coefficient of friction.

In the case where f is merely assumed to be convex, it turns out that the right choice of friction coefficient is $\gamma_t = 3/t$. This is mysterious at first sight and was obtained by taking the continuous-time limit of Nesterov's discrete algorithm in the next subsection. We begin with a convergence analysis in this setting. (Similar caveats as for §2 apply here; we assume that f is smooth, that it admits a minimizer x_\star , and that (AGF) is well-posed.)

Theorem 5.8 (convergence of AGF under convexity). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and let $(x_t)_{t \geq 0}$ evolve along AGF with $\gamma_t = 3/t$ and $p_0 = 0$. Then, for all $t \geq 0$,

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

Proof. Consider the auxiliary point $z_t := x_t + \frac{t}{2} p_t$, and the Lyapunov function

$$\mathcal{L}_t := \frac{t^2}{2} (f(x_t) - f_\star) + \|z_t - x_\star\|^2.$$

The computation below shows that $\dot{\mathcal{L}}_t \leq 0$, which implies the result. The choice of Lyapunov function is mysterious, so we partially demystify it after the proof.

Straightforward differentiation and convexity yield

$$\begin{aligned}\dot{\mathcal{L}}_t &= t (f(x_t) - f_\star) + \frac{t^2}{2} \langle \nabla f(x_t), p_t \rangle - t \langle \nabla f(x_t), z_t - x_\star \rangle \\ &= t (f(x_t) - f_\star) - t \langle \nabla f(x_t), x_t - x_\star \rangle \leq 0.\end{aligned}\quad \square$$

Although the Lyapunov function above appears fortuitous, it can be derived in a reasonably systematic manner; see [Exercise 5.2](#). The strongly convex case is similar, and is left as [Exercise 5.3](#).

Theorem 5.9 (convergence of AGF under strong convexity). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be α -convex and let $(x_t)_{t \geq 0}$ evolve along AGF with $\gamma_t = 2\sqrt{\alpha}$ and $p_0 = 0$. For all $t \geq 0$,

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

Recall that under convexity and α -convexity, the objective gap $f(x_t) - f_\star$ for GF converges at the rates $O(1/t)$ and $O(\exp(-2\alpha t))$ respectively. On the other hand, for AGF, the convergence happens at the rates $O(1/t^2)$ and $O(\exp(-\sqrt{\alpha} t))$ respectively. This is strongly suggestive of the square root factor speed-up, that is, *acceleration*. However, we caution that it is dangerous to deduce conclusions from continuous-time analysis alone. For example, we can run any ODE faster, which can make the continuous-time convergence rate arbitrarily fast; however, this does not translate into a better discrete-time algorithm, since speeding up time makes the ODE more unstable and therefore requires a smaller step size for discretization.

So how, then, can we discretize AGF? Part of the subtlety of acceleration is that not all discretizations work. For example, we could consider

$$\begin{aligned}x_{n+1} &\approx x_n + h p_{n+1}, \\ p_{n+1} &\approx p_n - h \nabla f(x_n) - \gamma_n h p_n\end{aligned}$$

which is equivalent to the update

$$x_{n+1} = x_n - h^2 \nabla f(x_n) + h (1 - \gamma_n h) (x_n - x_{n-1}).$$

Or, if we do not presume to know the coefficients for the discrete-time scheme in advance, we could write the update as

$$x_{n+1} = x_n - \eta_n \nabla f(x_n) + \theta_n (x_n - x_{n-1}).$$

In other words, we take a gradient step and then apply momentum. This is known as Polyak's heavy ball method, and although it can be tuned to converge at the rate

of [CG](#) for quadratic objectives, this same tuning leads to divergence for general convex functions [\[LRP16\]](#). On the other hand, the optimal method in the next subsection can be written in the form

$$x_{n+1} = x_n + \theta_n (x_n - x_{n-1}) - \eta_n \nabla f(x_n + \theta_n (x_n - x_{n-1})) .$$

In other words, we add momentum and then take a gradient step.

5.3 General case: discrete time

The acceleration phenomenon is undoubtedly one of the most elusive and fascinating aspects of optimization, so it is no surprise that it has been explored through many different angles over the course of countless research papers. At this junction, we must choose how to present the method and in what level of detail.

Having explored acceleration carefully in the quadratic case and in continuous time, here we follow the expedient route by giving perhaps the most direct and shortest proof, at the cost of generality and intuition.⁵

We analyze the following method with $x_{-1} = x_0$:

$$x_{n+1} := x_n + \theta_n (x_n - x_{n-1}) - \frac{1}{\beta} \nabla f(x_n + \theta_n (x_n - x_{n-1})) . \quad (\text{AGD})$$

Theorem 5.10 (convergence of AGD). Let f be convex and β -smooth. Define the sequence: $\lambda_0 := 0$ and $\lambda_{n+1} := \frac{1}{2} (1 + \sqrt{1 + 4\lambda_n^2})$ for $n \in \mathbb{N}$. Set $\theta_n := (\lambda_n - 1)/\lambda_{n+1}$. Then, [AGD](#) satisfies

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

Proof. Let $y_n := x_n + \theta_n (x_n - x_{n-1})$, so that $x_{n+1} = y_n - \frac{1}{\beta} \nabla f(y_n)$. Recall from [\(3.3\)](#) that for any $z \in \mathbb{R}^d$, it holds that

$$\|x_{n+1} - z\|^2 \leq \|y_n - z\|^2 - \frac{2}{\beta} (f(x_{n+1}) - f(z)) .$$

Rearranging, it yields

$$f(x_{n+1}) - f(z) \leq \frac{\beta}{2} (\|y_n - z\|^2 - \|x_{n+1} - z\|^2) = -\frac{\beta}{2} \|x_{n+1} - y_n\|^2 - \beta \langle x_{n+1} - y_n, y_n - z \rangle .$$

⁵Perhaps I will change my mind in a future edition of these notes.

We apply this inequality with two points, $z = x_n$ and $z = x_\star$. By multiplying the first inequality by $\lambda_{n+1} - 1 \geq 0$ and adding it to the second inequality, it implies

$$\begin{aligned} & (\lambda_{n+1} - 1) (f(x_{n+1}) - f(x_n)) + f(x_{n+1}) - f_\star \\ & \leq -\frac{\beta\lambda_{n+1}}{2} \|x_{n+1} - y_n\|^2 - \beta \langle x_{n+1} - y_n, \lambda_{n+1} y_n - (\lambda_{n+1} - 1) x_n - x_\star \rangle \\ & = \frac{\beta}{2\lambda_{n+1}} (\|\lambda_{n+1} y_n - (\lambda_{n+1} - 1) x_n - x_\star\|^2 - \|\lambda_{n+1} x_{n+1} - (\lambda_{n+1} - 1) x_n - x_\star\|^2), \end{aligned}$$

where the last line uses the identity $\|a\|^2 + 2\langle a, b \rangle = \|a + b\|^2 - \|b\|^2$. Our goal is to produce a telescoping sum, which is the case if we ensure that

$$\lambda_{n+1} x_{n+1} - (\lambda_{n+1} - 1) x_n = \lambda_{n+2} y_{n+1} - (\lambda_{n+2} - 1) x_{n+1}.$$

By substituting in $y_{n+1} = x_{n+1} + \theta_{n+1} (x_{n+1} - x_n)$, some algebra shows that it suffices to take $\theta_{n+1} = (\lambda_{n+1} - 1)/\lambda_{n+2}$.

After multiplying the above inequality by λ_{n+1} and summing, we find that

$$\frac{\beta}{2} \|\lambda_1 y_0 - (\lambda_1 - 1) x_0 - x_\star\|^2 \geq \sum_{n=0}^{N-1} \{\lambda_{n+1}^2 (f(x_{n+1}) - f_\star) - \lambda_{n+1} (\lambda_{n+1} - 1) (f(x_n) - f_\star)\}.$$

We also want the right-hand side to telescope, so we set $\lambda_{n+1} (\lambda_{n+1} - 1) = \lambda_n^2$, which yields the recursion $\lambda_{n+1} = \frac{1}{2} (1 + \sqrt{1 + 4\lambda_n^2})$. With $\lambda_0 = 0$, it yields

$$f(x_N) - f_\star \leq \frac{\beta \|y_0 - x_\star\|^2}{2\lambda_N^2} = \frac{\beta \|x_0 - x_\star\|^2}{2\lambda_N^2}.$$

Finally, it is straightforward to show by induction that $\lambda_N \geq N/2$. □

By applying the reduction in [Lemma 4.1](#), it also yields an accelerated algorithm for the strongly convex case, i.e., an algorithm that achieves $f(x_N) - f_\star \leq \varepsilon$ in $O(\sqrt{\kappa} \log \frac{\alpha R^2}{\varepsilon})$ iterations, where $R := \|x_0 - x_\star\|$.

Example 5.11. If we apply the accelerated method to logistic regression (see [Example 3.8](#)), it improves the iteration complexity from $O(R^2 n/d)$ to $O(R\sqrt{n/d})$.

Bibliographical notes

The simple proof of [Theorem 5.4](#) is inspired by [\[NY83\]](#). The discussion on Chebyshev polynomials follows [\[Vis12\]](#).

The literature on acceleration is too large to be surveyed here, but we mention a recent result in a somewhat different direction: what is the best rate of [GD](#) just by changing the step sizes? Thus, we consider the iteration $x_{n+1} = x_n - h_n \nabla f(x_n)$, with the only freedom being to choose the sequence $\{h_n\}_{n \in \mathbb{N}}$. It turns out a constant step size schedule is not optimal, and as established in [\[AP24a; AP24b\]](#), the so-called silver step size schedule achieves the rates of [Lemma 4.1](#) and [Lemma 4.2](#) with $\phi(x) = x^{\log_\rho 2} \approx x^{0.786}$ with $\rho := 1 + \sqrt{2}$. This is a rate intermediate between the unaccelerated rate of [GD](#) and the accelerated rate of [AGD](#).

Exercises

Exercise 5.1. Define the polynomial $Q_n(x) = T_n(\frac{\alpha+\beta-2x}{\beta-\alpha})/T_n(\frac{\alpha+\beta}{\beta-\alpha})$. Show that $Q_n(0) = 1$, and by using [Definition 5.6](#) to establish the identity

$$T_n(x) = \frac{1}{2} \left((x - \sqrt{x^2 - 1})^n + (x + \sqrt{x^2 - 1})^n \right) \quad \text{for } x \in [-1, 1],$$

prove that

$$\max_{x \in [\alpha, \beta]} |Q_n(x)| \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^n.$$

Note that by combining this with [Lemma 5.5](#), it yields an exponential rate of convergence for [CG](#) matching the lower bound of [Exercise 4.2](#).

Exercise 5.2. To better understand the proof of [Theorem 5.8](#), consider a Lyapunov function of the form

$$\mathcal{L}_t = \|x_t - x_\star\|^2 + a_t \langle x_t - x_\star, p_t \rangle + b_t \|p_t\|^2 + c_t (f(x_t) - f_\star).$$

Note that this is the most general Lyapunov function consisting of a combination of a quadratic function in $x_t - x_\star$ and p_t , as well as the objective gap; here, it is crucial that we include the mixed term $a_t \langle x_t - x_\star, p_t \rangle$. Our goal is to choose the coefficients a_t, b_t, c_t so that $\dot{\mathcal{L}}_t \leq 0$.

Compute the derivative in time of \mathcal{L}_t along [AGF](#) with $\gamma_t = 3/t$, and apply convexity to the term $\langle \nabla f(x_t), x_t - x_\star \rangle$. In the resulting expression, since the terms $\langle x_t - x_\star, p_t \rangle$ and $\langle \nabla f(x_t), p_t \rangle$ do not have definite signs, ensure that the coefficients in front of these terms

vanish through a suitable choice of a_t, b_t, c_t . Show that this leads to $a_t = t + \bar{a}t^3$ for some $\bar{a} \geq 0$. Next, from the remaining terms, obtain the condition $\dot{b}_t \leq \max\{\frac{a_t}{2}, \frac{6b_t}{t} - a_t\}$, which implies $3\dot{b}_t \leq 6b_t/t$, hence we consider $b_t = b_0 + \bar{b}t^2$ for some $b_0, \bar{b} \geq 0$. Furthermore, argue that we must take $\bar{a} = 0$ and $\bar{b} = \frac{1}{4}$. To ensure that \mathcal{L}_0 only depends on $\|x_0 - x_\star\|$, we set $b_0 = c_0 = 0$. Finally, check that with these choices, we have $b_t \geq a_t^2/4$, which is necessary to ensure that $\mathcal{L}_t \geq c_t (f(x_t) - f_\star)$.

Show that the Lyapunov function derived in this way coincides with the one used in [Theorem 5.8](#).

Exercise 5.3. Prove [Theorem 5.9](#).

Hint: Let $z_t := x_t + \frac{2}{\gamma} p_t$ and consider

$$\mathcal{L}_t := f(x_t) - f_\star + \frac{\alpha}{2} \|z_t - x_\star\|^2.$$

References

- [AP24a] J. M. Altschuler and P. A. Parrilo. “Acceleration by stepsize hedging: multi-step descent and the silver stepsize schedule”. In: *J. ACM* (Dec. 2024).
- [AP24b] J. M. Altschuler and P. A. Parrilo. “Acceleration by stepsize hedging: silver stepsize schedule for smooth convex optimization”. In: *Mathematical Programming* (2024).
- [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.
- [LRP16] L. Lessard, B. Recht, and A. Packard. “Analysis and design of optimization algorithms via integral quadratic constraints”. In: *SIAM J. Optim.* 26.1 (2016), pp. 57–95.
- [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. Vychisl. Mat i Mat. Fiz.* 3 (1963), pp. 643–653.
- [SBC16] W. Su, S. Boyd, and E. J. Candès. “A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights”. In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 153, 43.
- [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.
- [Vis12] N. K. Vishnoi. “ $Lx = b$ Laplacian solvers and their algorithmic applications”. In: *Found. Trends Theor. Comput. Sci.* 8.1-2 (2012), front matter, 1–141.