Lectures on Optimization

Sinho Chewi

January 3, 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.

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).

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.

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

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 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*:

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

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

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

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

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

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 $(1-t)\,x + t\,y$ 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) $\Rightarrow$ (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).

(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 $v := y - x$,

$$f(y) = f(x) + \int_0^1 \langle \nabla f(x + sv), v \rangle \, \mathrm{d}s , \quad f((1-t)\,x + t\,y) = f(x) + \int_0^1 \langle \nabla f(x + stv), v \rangle \, \mathrm{d}s .$$

Hence, (1.5) yields

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

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

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = \int_0^1 \langle \nabla^2 f(x + t\,(y - x))\,(y - x),\ y - x \rangle \,\mathrm{d}t\,,$$

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 $f : \mathbb{R}^d \to \mathbb{R}$ satisfy the following conditions: (1) continuous; (2) bounded below; (3) has bounded level sets. Then, there exists a global minimizer of f.

Proof. The proof uses some analysis. Let $x_0 \in \mathbb{R}^d$ and let $\mathcal{K} := \{f \le f(x_0)\}$ 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 $\nabla^2 f(x_\star) \succeq 0$.

Proof. Let $v \in \mathbb{R}^d$ and $\varepsilon > 0$; then, $f(x_\star + \varepsilon v) - f(x) \ge 0$. If f is continuously differentiable, this yields $\int_0^1 \langle \nabla f(x_\star + \varepsilon t v), v\rangle \,\mathrm{d}t \ge 0$. By continuity of ∇f, sending $\varepsilon \searrow 0$ proves that $\langle \nabla f(x_\star), v\rangle \ge 0$ 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 $0 \le \int_0^1 \int_0^t \langle \nabla^2 f(x_\star + \varepsilon s t v)\, v, v\rangle \,\mathrm{d}s\,\mathrm{d}t$. By continuity of $\nabla^2 f$, sending $\varepsilon \searrow 0$ then proves that $\langle \nabla^2 f(x_\star)\, v, v\rangle \ge 0$ for all $v \in \mathbb{R}^d$. □

The conditions $\nabla f(x_\star) = 0$, $\nabla^2 f(x_\star) \succeq 0$ 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 $x = x_\star$. □

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. □

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 $\nabla f(x_\star) = 0$.

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

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

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

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

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:

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

This remains true even without assuming twice differentiability.

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 [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, $\partial_t f(x_t) = \langle \nabla f(x_t), \dot{x}_t \rangle = -\|\nabla f(x_t)\|^2$. □

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. □

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

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

Then, it holds that

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

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,

$$\exp(-At)\, u(t) - u(0) \leq \int_0^t B(s) \exp(-As)\, \mathrm{d}s\,.$$

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 $\|y_t - x_t\|^2 \leq \exp(-2\alpha t)\, \|y_0 - x_0\|^2$, 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 $f : \mathbb{R}^d \to \mathbb{R}$ 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 $\alpha \to 0$, 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 $A = -\alpha$, $B(t) = -2\, (f(x_t) - f_\star)$,

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

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

$$\int_0^t \exp(-\alpha\, (t - s))\, (f(x_s) - f_\star)\, \mathrm{d}s \geq (f(x_t) - f_\star) \int_0^t \exp(-\alpha\, (t - s))\, \mathrm{d}s$$

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

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 $t \geq 0$,

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

We present a few key properties of the PŁ inequality.

Proposition 2.7 (strong convexity ⇒ PŁ ⇒ 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:

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

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: $\mathscr{L}_t := \sqrt{\frac{\alpha}{2}} \|x_t - x_0\| + \sqrt{f(x_t) - f_\star}$.

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

   where we applied (PŁ). Since $\mathscr{L}_0 = \sqrt{f(x) - f_\star}$ and $\mathscr{L}_\infty = \sqrt{\frac{\alpha}{2}} \|x_0 - x_\star\|$, we deduce the result from $\mathscr{L}_0 \geq \mathscr{L}_\infty$.

□

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

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

We therefore arrive at the following simple consequence.

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

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

(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, $\dot{\mathscr{L}}_t \le 0$:

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

Deduce the following gradient bound:

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

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

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

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 $\dot{\mathscr{L}}_t \le -\alpha\mathscr{L}_t$, and thereby deduce the following sharp bounds:

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

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 $d \ge n$. 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 $f \circ g$.

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.

[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.