Notes on Convex Optimization and Approximation

Daniele Zago

Based on the Berkeley EE227C course notes

November 23, 2021

Contents

Part I

Gradient methods

Gradient descent is one of the most broadly applied techniques for minimizing functions, both convex and nonconvex alike. At its core, it is a form of *local search*, greedily optimizating a function in a small region over many successive iterations. If $f : \mathbb{R} \to \mathbb{R}$ is twice-continuously differentiable, then Taylor's theorem tells us that that

$$f(x + \delta) \approx f(x) + \delta f'(x) + \frac{1}{2}\delta^2 f''(x)\,.$$

This approximation directly reveals that if we move from x to $x + \delta$ where $\delta = -\eta \cdot f'(x)$ for sufficiently small $\eta > 0$, we generally expect to decrease the function value by about $\eta f'(x)^2$. The simple greedy way of decreasing the function value is known as *gradient descent*, and it generalizes to idea to functions of many variables with the help of multivariate versions of Taylor's theorem.

Gradient descent converges to points at which the first derivatives vanish. For the broad class of *convex* functions, such points turn out to be globally minimal. Moreover, gradient descent can be amended for convex functions which are not even differentiable. Later on, we will see that gradient descent can be shown to converge to locally (and, on occasion, globally!) minimal points as well.

In this part of the text, we begin in **??** by introducing the preliminaries about convex functions which make them so amenable to gradient descent. Crucially, we also introduce the notion of subgradient, which generalizes the gradient to possibly non-convex function and is used in settings such as LASSO and Elastic-net optimization.

In Section 2.1, we formally introduce (sub-)gradient descent, and prove explicit convergence rates when gradient descent is applied to convex functions. Section introduces a stronger assumption know as *strong convexity*, which allows (sub-)gradient descent to enjoy even faster rates.

Lecture 1: Convexity

2020-09-30

This lecture provides the most important facts about convex sets and convex functions that we'll heavily make use of. When f is sufficiently smooth, these facts are often simple consequences of Taylor's theorem.

1.1 Convex sets

Def. (Convex set)
A set $K \subseteq \mathbb{R}^n$ is ***convex*** if it the line segment between any two points in K is also contained in K. Formally, for all $x, y \in K$ and all scalars $\gamma \in [0, 1]$ we have $\gamma x + (1 - \gamma)y \in K$.

Theorem 1 (Separation Theorem)
Let $C, K \subseteq \mathbb{R}^n$ be convex sets with empty intersection $C \cap K = \emptyset$. Then there exists a point $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$ such that

1. *for all $x \in C$, we have $\langle a, x\rangle \geqslant b$.*
2. *for all $x \in K$, we have $\langle a, x\rangle \leqslant b$.*

If C and K are closed and at least one of them is bounded, then we can replace the inequalities by strict inequalities.

The case we're most concerned with is when both sets are compact (i.e., closed and bounded). We highlight its proof here.

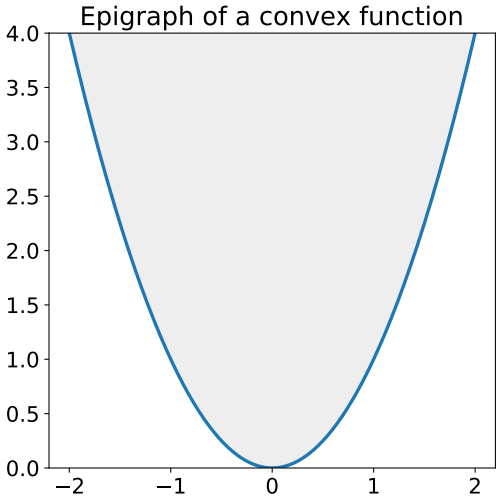
Proof of Theorem 1 for compact sets.
In this case, the Cartesian product $C \times K$ is also compact. Therefore, the distance function $\|x - y\|$ attains its minimum over $C \times K$. Take $p \in C, q \in K$ to be two points that achieve the minimum. Then, a separating hyperplane is given by the hyperplane perpendicular to $q - p$ that passes through the midpoint between p and q. That is, $a = q - p$ and $b = (\langle a, q\rangle - \langle a, p\rangle)/2$.

For the sake of contradiction, suppose there is a point r on this hyperplane contained in one of the two sets, say, C. Then the line segment from p to r is also contained in C by convexity. We can then find a point along the line segment that is closer to q than p is, thus contradicting our assumption of minimum distance.

■

1.1.1 Notable convex sets

- Linear spaces $\{x \in \mathbb{R}^n \mid Ax = 0\}$ and halfspaces $\{x \in \mathbb{R}^n \mid \langle a, x\rangle \geqslant 0\}$
- Affine transformations of convex sets. If $K \subseteq \mathbb{R}^n$ is convex, so is $\{Ax + b \mid x \in K\}$ for any $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. In particular, affine subspaces and affine halfspaces are convex.
- Intersections of convex sets. In fact, every convex set is equivalent to the intersection of all affine halfspaces that contain it (a consequence of the separating hyperplane theorem).



› The cone of positive semidefinite matrices, $S_+^n = \{A \in \mathbb{R}^{n\times n} \mid A \succeq 0\}$. Here we write $A \succeq 0$ to indicate that $x^\top Ax \geqslant 0$ for all $x \in \mathbb{R}^n$. The fact that S_+^n is convex can be verified directly from the definition, but it also follows from what we already knew. Indeed, denoting by $S_n = \{A \in \mathbb{R}^{n\times n} \mid A^\top = A\}$ the set of all $n \times n$ symmetric matrices, we can write S_+^n as an (infinite) intersection of halfspaces $S_+^n = \bigcap_{x\in\mathbb{R}^n\setminus\{0\}} \{A \in S_n \mid x^\top Ax \geqslant 0\}$.

› See Boyd and Vandenberghe (2004) for lots of other examples.

1.2 Convex functions

Def. (Convex function)
A function $f\colon \Omega \to \mathbb{R}$ is ***convex*** if for all $x, y \in \Omega$ and all scalars $\gamma \in [0, 1]$ we have

$$f(\gamma x + (1-\gamma)y) \leqslant \gamma f(x) + (1-\gamma)f(y).$$

Jensen (1905) showed that for continuous functions, convexity follows from the "midpoint" condition that for all $x, y \in \Omega$,

$$f\left(\frac{x+y}{2}\right) \leqslant \frac{f(x)+f(y)}{2}\,.$$

This result sometimes simplifies the proof that a function is convex in cases where we already know that it's continuous.

Def. (Epigraph)
The ***epigraph*** of a function $f\colon \Omega \to \mathbb{R}$ is defined as

$$\operatorname{epi}(f) = \{(x, t) \mid f(x) \leqslant t\}\,.$$

Fact A function is convex if and only if its epigraph is convex.

Convex functions enjoy the property that local minima are also global minima. Indeed, suppose that $x \in \Omega$ is a local minimum of $f\colon \Omega \to \mathbb{R}$ meaning that any point in a neighborhood around x has larger function value. Now, for every $y \in \Omega$, we can find a small enough γ such that

$$f(x) \overset{\text{min.}}{\leqslant} f((1-\gamma)x + \gamma y) \overset{\text{conv.}}{\leqslant} (1-\gamma)f(x) + \gamma f(y) \overset{\gamma=0}{\leqslant} f(y).$$

Therefore, $f(x) \leqslant f(y)$ and so x must be a global minimum.

1.2.1 First-order characterization

It is helpful to relate convexity to Taylor's theorem, which we recall now. We define the *gradient* of a differentiable function $f\colon \Omega \to \mathbb{R}$ at $x \in \Omega$ as the column vector of partial derivatives

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_i}\right)_{i=1,\dots,n}.$$

We note the following simple fact that relates linear forms of the gradient to a one-dimensional derivative evaluated at 0. It's a consequence of the multivariate chain rule:

Fact Assume $f\colon \Omega \to \mathbb{R}$ is differentiable and let $x, y \in \Omega$. Then,

$$\nabla f(x)^\top y = \left.\frac{\partial f(x+\gamma y)}{\partial \gamma}\right|_{\gamma=0}.$$

Taylor's theorem implies the following statement:

Prop. 1 (Linear approximation to f)
Assume $f\colon \Omega \to \mathbb{R}$ is continuously differentiable along the line segment between two points x and y. Then,

$$f(y) = f(x) + \nabla f(x)^\top (y-x) + \int_0^1 (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2}\,\mathrm{d}\gamma.$$

Proof.
Apply a second order Taylor's expansion to $g(\gamma) = f(x+\gamma(y-x))$ and apply the fact above to the first-order term.

■

Among differentiable functions, convexity is equivalent to the property that the first-order Taylor approximation provides a global lower bound on the function (Figure 1).

Prop. 2 (Characterization of convexity by gradient)
Assume $f\colon \Omega \to \mathbb{R}$ is differentiable. Then, f is convex if and only if for all $x, y \in \Omega$ we have

$$f(y) \geqslant f(x) + \nabla f(x)^\top (y-x)\,. \tag{1}$$

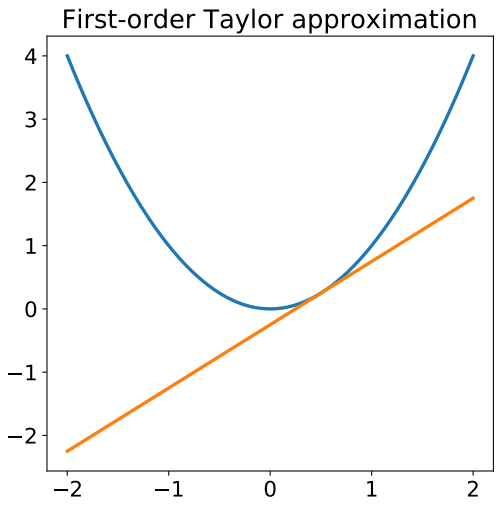


Figure 1: Taylor approximation of $f(x) = x^2$ at 0.5.

Proof.

Convex $\implies$ thm : Suppose f is convex, then by definition

$$(1-\gamma)f(x) + \gamma f(y) \geqslant f\big((1-\gamma)x + \gamma y\big),$$

and by rearranging the terms we have

$$\begin{aligned} f(y) &\geqslant \frac{f((1-\gamma)x + \gamma y) - (1-\gamma)f(x)}{\gamma} \\ &= f(x) + \frac{f(x + \gamma(y-x)) - f(x)}{\gamma} \\ &\xrightarrow{\gamma \to 0} f(x) + \nabla f(x)^\top (y - x) \qquad \text{(by Fact ??.)} \end{aligned}$$

Thm $\implies$ convex : On the other hand, fix two points $x, y \in \Omega$ and $\gamma \in [0,1]$. Putting $z = \gamma x + (1-\gamma)y$ we get from applying Equation 1 twice,

$$f(x) \geqslant f(z) + \nabla f(z)^\top (x - z) \quad \text{and} \quad f(y) \geqslant f(z) + \nabla f(z)^\top (y - z)$$

Adding these inequalities scaled by γ and $(1-\gamma)$, respectively, we get

$$\begin{aligned} \gamma f(x) + (1-\gamma) f(y) &\geqslant f(z) + \overbrace{\gamma \nabla f(z)^\top (x - z) + (1-\gamma)\nabla f(z)^\top (y - z)}^{=0} \\ \gamma f(x) + (1-\gamma) f(y) &\geqslant f(z), \end{aligned}$$

which establishes convexity.

■

A direct consequence of Proposition 2 is that if $\nabla f(x) = 0$ vanishes at a point x, then x must be a global minimizer of f.

Subgradients

Of course, not all convex functions are differentiable. The absolute value $f(x) = |x|$, for example, is convex but not differentiable at 0. Nonetheless, for every x, we can find a vector g such that

$$f(y) \geqslant f(x) + g^\top (y - x)\,.$$

Such a vector is called a *subgradient* of f at x. The existence of subgradients is often sufficient for optimization.

1.2.2 Second-order characterization

Def. (Hessian matrix)

The ***Hessian matrix*** of $f\colon \Omega \to \mathbb{R}$ at a point $x \in \Omega$ is the matrix of second order partial derivatives:

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{i,j\in[n]}\,.$$

Schwarz's theorem implies that the Hessian at a point x is symmetric provided that f has continuous second partial derivatives in an open set around x.

In analogy with **??**, we can relate quadratic forms in the Hessian matrix to one-dimensional derivatives using the chain rule.

Fact Assume that $f\colon \Omega \to \mathbb{R}$ is twice differentiable along the line segment from x to y. Then,

$$y^\top \nabla^2 f(x + \gamma y) y = \frac{\partial^2 f(x + \gamma y)}{\partial \gamma^2}\,.$$

Prop. 3 (Characterization of convexity by Hessian matrix)

If f is twice continuously differentiable on its domain Ω, then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \Omega$.

Proof.

$\boxed{\Longleftarrow}$: Suppose f is convex and our goal is to show that the Hessian is positive semidefinite. Let $y = x + \alpha u$ for some arbitrary vector u and scalar α. Proposition 2 shows

$$f(y) - f(x) - \nabla f(x)^\top (y - x) \geqslant 0$$

Hence, by Proposition 1,

$$\begin{aligned} 0 &\leqslant \int_0^1 (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2}\,\mathrm{d}\gamma \\ &= (1-\gamma)\frac{\partial^2 f(x+\gamma(y-x))}{\partial \gamma^2} \quad \text{for some } \gamma \in (0,1) \qquad \text{(by the mean value theorem)} \\ &= (1-\gamma)(y-x)^\top \nabla^2 f(x+\gamma(y-x))(y-x)\,. \qquad \text{(by Fact ??)} \end{aligned}$$

Plugging in our choice of y, this shows $0 \leqslant u^\top \nabla^2 f(x+\alpha\gamma u)u$. Letting α tend to zero establishes that $\nabla^2 f(x) \succeq 0$. (Note that γ generally depends on α but is always bounded by 1.)

$\boxed{\Longrightarrow}$: Now, suppose the Hessian is positive semidefinite everywhere in Ω and our goal is to show that the function f is convex. Using the same derivation as above, we can see that the second-order error term in Taylor's theorem must be non-negative. Hence, the first-order approximation is a global lower bound and so the function f is convex by Proposition 2.

■

1.3 Convex optimization

Much of this course will be about different ways of minimizing a convex function $f\colon \Omega \to \mathbb{R}$ over a convex domain Ω :

$$\min_{x\in\Omega} f(x)$$

Convex optimization is not necessarily easy! For starters, convex sets do not necessarily enjoy compact descriptions. When solving computational problems involving convex sets, we need to worry about how to represent the convex set we're dealing with. Rather than asking for an explicit description of the set, we can instead require a computational abstraction that highlights essential operations that we can carry out. The Separation Theorem motivates an important computational abstraction called *separation oracle*.

Def. (Separation oracle)

A *separation oracle* for a convex set K is a device, which given any point $x \notin K$ returns a hyperplane separating x from K.

Another computational abstraction is a *first-order oracle* that given a point $x \in \Omega$ returns the gradient $\nabla f(x)$. Similarly, a *second-order oracle* returns $\nabla^2 f(x)$. A function value oracle or *zeroth-order oracle* only returns $f(x)$. First-order methods are algorithms that make do with a first-order oracle. Analogously, we can define zeroth-order method, and second-order method.

1.3.1 What is efficient?

Classical complexity theory typically quantifies the resource consumption (primarily running time or memory) of an algorithm in terms of the bit complexity of the input. We say things like "we can multiply two n-bit numbers in time $O(n^2)$ using long multiplication method."

This computational approach can be cumbersome in convex optimization and most textbooks shy away from it. Instead, it's customary in optimization to quantify the cost of the algorithm in terms of more abstract resources, like, how often it accesses one of the oracles we mentioned. Counting oracle can give us a rough sense of how well we expect a method to work.

The definition of "efficient" is not completely cut and dry in optimization. Typically, our goal is to show that an algorithm finds a solution x with

$$f(x) \leqslant \min_{x \in \Omega} f(x) + \epsilon$$

for some additive error $\epsilon > 0$. Since the cost of the algorithm will depend on the target error, highly practical algorithms often have a polynomial dependence on ϵ, such as $O(1/\epsilon)$ or even $O(1/\epsilon^2)$. Other algorithms achieve $O(\log(1/\epsilon))$ steps in theory, but are prohibitive in their actual computational cost.

Technically, if we think of the parameter ϵ as being part of the input, it takes only $O(\log(1/\epsilon))$ bits to describe the error parameter. Therefore, an algorithm that depends more than logarithmically on $1/\epsilon$ may not be a polynomial-time algorithm in its input size.

In this course, we will make an attempt to highlight both the theoretical performance and practical appeal of an algorithm. Moreover, we will discuss other performance criteria such as *robustness to noise*. How well an algorithm performs is rarely decided by a single criterion, and usually depends on the application at hand.

Lecture 2: Gradient method

2020-09-30

In this lecture we encounter the fundamentally important *gradient method* and a few ways to analyze its convergence behavior. The goal here is to solve a problem of the form

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

To solve this problem, we will need to make some assumptions on both the *objective function* $f\colon \Omega \to \mathbb{R}$ and the constraint set Ω. In case $\Omega = \mathbb{R}^n$, we speak of an *unconstrained* optimization problem.
The proofs closely follow the corresponding chapter in Bubeck (2015).

2.1 Gradient descent

For a differentiable function f, the basic gradient method starting from an initial point x_1 is defined by the iterative update rule

$$x_{t+1} = x_t - \eta_t \nabla f(x_t)\,, \qquad t = 1, 2, \ldots$$

where the scalar η_t is the so-called *step size*, sometimes called *learning rate*, that may vary with t. There are numerous ways of choosing step sizes that have a significant effect on the performance of gradient descent. What we will see in this lecture are several choices of step sizes that ensure the convergence of gradient descent by virtue of a theorem. Note however that these "optimal" step sizes are not necessarily ideal for practical applications.

2.1.1 Projections

In cases where the constraint set Ω is not all of $\mathbb{R}^n$, the gradient update can take us outside the domain Ω. How can we ensure that $x_{t+1} \in \Omega$? One natural approach is to "project" each iterate back onto the domain Ω. As it turns out, this won't really make our analysis more difficult and so we include from the get-go.

Def. (Projection)
The ***projection*** of a point x onto a set Ω is defined as

$$\Pi_\Omega(x) = \arg\min_{y\in\Omega} \|x - y\|_2\,.$$

Example
A projection of x onto the Euclidean ball B_2 is just the normalization of x:

$$\Pi_{B_2}(x) = \frac{x}{\|x\|}$$

A crucial property of projections is that when $x \in \Omega$, we have for any y (possibly outside Ω):

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2$$

That is, the projection of y onto a convex set containing x is closer to x. In fact, a stronger claim is true that follows from the Pythagorean theorem.

Lemma 1 (Pythagoraean thorem)
If $x \in \Omega$, then for any y (possibly outside Ω), the projection $\Pi_\Omega(y)$ of y onto Ω is such that

$$\|\Pi_\Omega(y) - x\|^2 \leqslant \|y - x\|^2 - \|y - \Pi_\Omega(y)\|^2$$

So, now we can modify our original procedure as displayed in Algorithm 1.

Algorithm 1 Projected gradient descent

while $\|x_t - x_{t-1}\| < \varepsilon$ **do**
  $y_{t+1} = x_t - \eta_t \nabla f(x_t)$ ▷ gradient step
  $x_{t+1} = \Pi_\Omega(y_{t+1})$ ▷ projection
end while

And we are guaranteed that $x_{t+1} \in \Omega$. Note that computing the step may be computationally the hardest part of the problem. However, there are convex sets for which we know explicitly how to compute the projection (see Example). We will see several other non-trivial examples in later lectures.

2.2 Lipschitz functions

The first assumption that leads to a convergence analysis is that the gradients of the objective function aren't too big over the domain. This turns out to follow from a natural Lipschitz continuity assumption.

Def. (L-Lipschitz)
A function $f\colon \Omega \to \mathbb{R}$ is *L-Lipschitz* if for every $x, y \in \Omega$, we have

$$|f(x) - f(y)| \leqslant L\|x - y\|$$

Fact If the function f is L-Lipschitz, differentiable, and convex, then its gradient is bounded,

$$\|\nabla f(x)\| \leqslant L\,.$$

We can now prove our first convergence rate for gradient descent.

Theorem 2 (Rate of convergence of GD)
Assume that function f is convex, differentiable, and L-Lipschitz over the convex domain Ω. Let R be the upper bound on the distance $\|x_1 - x^*\|_2$ from the initial point x_1 to an optimal point $x^* \in \arg\min_{x\in\Omega} f(x)$. Let $x_1, \ldots, x_t$ be the sequence of iterates computed by t steps of projected gradient descent with constant step size $\eta = \frac{R}{L\sqrt{t}}$. Then,

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leqslant \frac{RL}{\sqrt{t}}.$$

This means that the difference between the functional value of the average point during the optimization process from the optimal value is bounded above by a constant proportional to $\frac{1}{\sqrt{t}}$.
Before proving the theorem, recall the "Fundamental Theorem of Optimization", which is that an inner product can be written as a sum of norms:

$$u^\top v = \frac{1}{2}(\|u\|^2 + \|v\|^2 - \|u - v\|^2) \tag{2}$$

This property follows from the more familiar identity $\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2u^\top v$.

Proof of Theorem 2.
The proof begins by first bounding the difference in function values $f(x_s) - f(x^*)$.

$$\begin{aligned}
f(x_s) - f(x^*) &\leqslant \nabla f(x_s)^\top (x_s - x^*) && \text{(by convexity)}\\
&= \frac{1}{\eta}(x_s - y_{s+1})^\top (x_s - x^*) && \text{(by the update rule)}\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2\right) && \text{(by Equation 2)}\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|\nabla f(x_s)\|^2 && \text{(by the update rule)}\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && \text{(Lipschitz condition)}\\
&\leqslant \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta L^2}{2} && \text{(Lemma 1)}
\end{aligned}$$

Now, sum these differences from $s = 1$ to $s = t$:

$$\begin{aligned}
\sum_{s=1}^{t} f(x_s) - f(x^*) &\leqslant \frac{1}{2\eta}\sum_{s=1}^{t}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{\eta L^2 t}{2}\\
&= \frac{1}{2\eta}\left(\|x_1 - x^*\|^2 - \|x_t - x^*\|^2\right) + \frac{\eta L^2 t}{2} && \text{(telescoping sum)}\\
&\leqslant \frac{1}{2\eta}\|x_1 - x^*\|^2 + \frac{\eta L^2 t}{2} && \text{(since } \|x_t - x^*\| \geqslant 0\text{)}\\
&\leqslant \frac{R^2}{2\eta} + \frac{\eta L^2 t}{2} && \text{(since } \|x_1 - x^*\| \leqslant R\text{)}
\end{aligned}$$

Finally,

$$
\begin{aligned}
f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) &\leqslant \frac{1}{t}\sum_{s=1}^{t} f(x_s) - f(x^*) && \text{(by convexity)} \\
&\leqslant \frac{R^2}{2\eta t} + \frac{\eta L^2}{2} && \text{(inequality above)} \\
&= \frac{RL}{\sqrt{t}} && \text{(for } \eta = R/L\sqrt{t}.\text{)}
\end{aligned}
$$

■

2.3 Smooth functions

The next property we'll encounter is called *smoothness*. The main point about smoothness is that it allows us to control the second-order term in the Taylor approximation. This often leads to stronger convergence guarantees at the expense of a relatively strong assumption.

Def. (Smoothness)

A continuously differentiable function f is ***β-smooth*** if the gradient map $\nabla f\colon \mathbb{R}^n \to \mathbb{R}^n$ is β-Lipschitz, i.e,

$$\|\nabla f(x) - \nabla f(y)\| \leqslant \beta\|x - y\|.$$

We will need a couple of technical lemmas before we can analyze gradient descent for smooth functions. It's safe to skip the proof of these technical lemmas on a first read.

Lemma 2 (Tightness of gradient characterization)

Let f be a β-smooth function on $\mathbb{R}^n$. Then, for every $x, y \in \mathbb{R}^n$,

$$\left|f(y) - f(x) - \nabla f(x)^\top (y - x)\right| \leqslant \frac{\beta}{2}\|y - x\|^2.$$

Proof.

Express $f(x) - f(y)$ as an integral, then apply Cauchy-Schwarz and β-smoothness as follows:

$$
\begin{aligned}
|f(y) - f(x) - \nabla f(x)^\top (y - x)| &= \left|\int_0^1 \nabla f(x + t(y - x))^\top (y - x) dt - \nabla f(x)^\top (y - x)\right| \\
&\leqslant \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| dt \\
&\leqslant \int_0^1 \beta t \|y - x\|^2 dt \\
&= \frac{\beta}{2}\|y - x\|^2
\end{aligned}
$$

■

The significance of this lemma is that we can choose $y = x - \frac{1}{\beta}\nabla f(x)$ and get that

$$f(y) - f(x) \leqslant -\frac{1}{2\beta}\|\nabla f(x)\|^2 .$$

This means that the gradient update decreases the function value by an amount proportional to the squared norm of the gradient.
We also need the following lemma.

Lemma 3
Let f be a β-smooth convex function, then for every $x, y \in \mathbb{R}^n$, we have

$$f(x) - f(y) \leqslant \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

Proof.
Let $z = y - \frac{1}{\beta}(\nabla f(y) - \nabla f(x))$. Then,

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\leqslant \nabla f(x)^\top (x - z) + \nabla f(y)^\top (z - y) + \frac{\beta}{2}\|z - y\|^2 \\ &= \nabla f(x)^\top (x - y) + (\nabla f(x) - \nabla f(y))^\top (y - z) + \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \\ &= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 \end{aligned}$$

Here, the inequality follows from convexity and smoothness.

■

We will show that gradient descent with the update rule

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

attains a faster rate of convergence under the smoothness condition.

Theorem 3 (Rate of convergence for β-smooth functions)
Let f be convex and β-smooth on $\mathbb{R}^n$ then gradient descent with $\eta = \frac{1}{\beta}$ satisfies

$$f(x_t) - f(x^*) \leqslant \frac{2\beta\|x_1 - x^*\|^2}{t - 1}$$

To prove this we will need the following two lemmas.

Proof.
By the update rule and Lemma 2 we have

$$f(x_{s+1}) - f(x_s) \leqslant -\frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

In particular, denoting $\delta_s = f(x_s) - f(x^*)$ this shows

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta}\|\nabla f(x_s)\|^2$$

One also has by convexity

$$\delta_s \leqslant \nabla f(x)s)^\top (x_s - x^*) \leqslant \|x_s - x^*\| \cdot \|\nabla f(x_s)\|$$

We will prove that $\|x_s - x^*\|$ is decreasing with s, which with the two above displays will imply

$$\delta_{s+1} \leqslant \delta_s - \frac{1}{2\beta\|x_1 - x^*\|^2}\delta_s^2$$

We solve the recurrence as follows. Let $w = \frac{1}{2\beta\|x_1 - x^*\|^2}$, then

$$w\delta_s^2 + \delta_{s+1} \leqslant \delta_s \iff w\frac{\delta_s}{\delta_{s+1}} + \frac{1}{\delta_s} \leqslant \frac{1}{\delta_{s+1}} \implies \frac{1}{\delta_{s+1}} - \frac{1}{\delta_s} \geqslant w \implies \frac{1}{\delta_t} \geqslant w(t-1)$$

To finish the proof it remains to show that $\|x_s - x^*\|$ is decreasing with s. Using Lemma 3, we get

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geqslant \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

We use this and the fact that $\nabla f(x^*) = 0$, to show

$$\begin{aligned}
\|x_{s+1} - x^*\|^2 &= \|x_s - \frac{1}{\beta}\nabla f(x_s) - x^*\|^2 \\
&= \|x_s - x^*\|^2 - \frac{2}{\beta}\nabla f(x_s)^\top (x_s - x^*) + \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\
&\leqslant \|x_s - x^*\|^2 - \frac{1}{\beta^2}\|\nabla f(x_s)\|^2 \\
&\leqslant \|x_s - x^*\|^2 .
\end{aligned}$$

■

Lecture 3: Strong convexity

2020-09-30

This lecture introduces the notion of strong convexity and combines it with smoothness to develop the concept of condition number. While smoothness gave as an upper bound on the second-order term in Taylor's approximation, strong convexity will give us a lower bound. Taken together, these two assumptions are quite powerful as they lead to a much faster convergence rate of the form $\exp(-\Omega(t))$. In words, gradient descent on smooth and strongly convex functions decreases the error multiplicatively by some factor strictly less than 1 in each iteration.

The technical part follows the corresponding chapter in Bubeck's text (Bubeck, 2015).

3.1 Reminders

Recall that we had (at least) two definitions apiece for convexity and smoothness: a general definition for all functions and a more compact definition for (twice-)differentiable functions.

A function f is convex if, for each input, there exists a globally valid *linear* lower bound on the function: $f(y) \geqslant f(x) + g^\top(x)(y - x)$. For differentiable functions, the role of g is played by the gradient.

A function f is β-smooth if, for each input, there exists a globally valid *quadratic* upper bound on the function, with (finite) quadratic parameter β: $f(y) \leqslant f(x) + g^\top(x)(y - x) + \frac{\beta}{2}\|x - y\|^2$. More poetically, a smooth, convex function is "trapped between a parabola and a line". Since β is covariant with affine transformations, e.g. changes of units of measurement, we will frequently refer to a β-smooth function as simply smooth.

For twice-differentiable functions, these properties admit simple conditions for smoothness in terms of the Hessian, or matrix of second partial derivatives. A $\mathcal{D}^2$ function f is convex if $\nabla^2 f(x) \succeq 0$ and it is β-smooth if $\nabla^2 f(x) \preceq \beta I$.

We furthermore defined the notion of L-Lipschitzness. A function f is L-Lipschitz if the amount that it "stretches" its inputs is bounded by L: $|f(x) - f(y)| \leqslant L\|x - y\|$. Note that for differentiable functions, β-smoothness is equivalent to β-Lipschitzness of the gradient.

3.2 Strong convexity

With these three concepts, we were able to prove two error decay rates for gradient descent (and its projective, stochastic, and subgradient flavors). However, these rates were substantially slower than what's observed in certain settings in practice.

Noting the asymmetry between our linear lower bound (from convexity) and our quadratic upper bound (from smoothness) we introduce a new, more restricted function class by upgrading our lower bound to second order.

Def. (Strong convexity)
A function $f\colon \Omega \to \mathbb{R}$ is *α-strongly convex* if, for all $x, y \in \Omega$, the following inequality holds for some $\alpha > 0$:

$$f(y) \geqslant f(x) + g(x)^\top (y - x) + \frac{\alpha}{2} \|x - y\|^2$$

As with smoothness, we will often shorten "α-strongly convex" to "strongly convex". A strongly convex, smooth function is one that can be "squeezed between two parabolas". If β-smoothness is a good thing, then α-convexity guarantees we don't have too much of a good thing.

A twice differentiable function is α-strongly convex if $\nabla^2 f(x) \succeq \alpha I$.

Once again, note that the parameter α changes under affine transformations. Conveniently enough, for α-strongly convex, β-smooth functions, we can define a basis-independent quantity called the *condition number*.

Def. (Condition Number)
An α-strongly convex, β-smooth function f has *condition number* $\frac{\beta}{\alpha}$.

For a positive-definite quadratic function f, this definition of the condition number corresponds with the perhaps more familiar definition of the condition number of the matrix defining the quadratic.

A look back and ahead. The following table summarizes the results from the previous lecture and the results to be obtained in this lecture. In both, the value ϵ is the difference between f at some value x' computed from the outputs of gradient descent and f calculated at an optimizer x^*.

	Convex	Strongly convex
Lipschitz	$\epsilon \leqslant O(1/\sqrt{t})$	$\epsilon \leqslant O(1/t)$
Smooth	$\epsilon \leqslant O(1/t)$	$\epsilon \leqslant e^{-\Omega(t)}$

Table 1: Bounds on error ϵ as a function of number of steps taken t for gradient descent applied to various classes of functions.

Since a rate that is exponential in terms of the magnitude of the error is linear in terms of the bit precision, this rate of convergence is termed *linear*. We now move to prove these rates.

3.3 Convergence rate strongly convex functions

For no good reason we begin with a convergence bound for strongly convex Lipschitz functions, in which we obtain a $O(1/t)$ rate of convergence.

Theorem 4

*Assume $f\colon \Omega \to \mathbb{R}$ is α-*strongly convex *and L-*Lipschitz*. Let x^* be an optimizer of f, and let x_s be the updated point at step s using projected gradient descent. Let the max number of iterations be t with an adaptive step size $\eta_s = \frac{2}{\alpha(s+1)}$, then*

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leqslant \frac{2L^2}{\alpha(t+1)}$$

The theorem implies the convergence rate of projected gradient descent for α-strongly convex functions is similar to that of β-smooth functions with a bound on error $\epsilon \leqslant O(1/t)$. In order to prove Theorem 4, we need the following proposition.

Prop. 4 (Jensen's inequality)

Assume $f\colon \Omega \to \mathbb{R}$ is a convex function and $x_1, x_2, ..., x_n \in \Omega$ are such that $\sum_{i=1}^n \gamma_i x_i / \sum_{i=1}^n \gamma_i \in \Omega$ with weights $\gamma_i > 0$, then

$$f\left(\frac{\sum_{i=1}^n \gamma_i x_i}{\sum_{i=1}^n \gamma_i}\right) \leqslant \frac{\sum_{i=1}^n \gamma_i f(x_i)}{\sum_{i=1}^n \gamma_i}$$

For a graphical "proof" follow this link.

Proof of Theorem 4.
Recall the two steps update rule of projected gradient descent

$$y_{s+1} = x_s - \eta_s \nabla f(x_s)$$

$$x_{s+1} = \Pi_\Omega(y_{s+1})$$

First, the proof begins by exploring an upper bound of difference between function values $f(x_s)$ and $f(x^*)$.

$$\begin{aligned}
f(x_s) - f(x^*) &\leqslant \nabla f(x_s)^\top (x_s - x^*) - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
&= \frac{1}{\eta_s}(x_s - y_{s+1})^\top (x_s - x^*) - \frac{\alpha}{2}\|x_s - x^*\|^2 && \text{(by update rule)} \\
&= \frac{1}{2\eta_s}(\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2) - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
& && \text{(by "Fundamental Theorem of Optimization")} \\
&= \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) + \frac{\eta_s}{2}\|\nabla f(x_s)\|^2 - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
& && \text{(by update rule)} \\
&\leqslant \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta_s}{2}\|\nabla f(x_s)\|^2 - \frac{\alpha}{2}\|x_s - x^*\|^2 && \text{(by Lemma 1)} \\
&\leqslant (\frac{1}{2\eta_s} - \frac{\alpha}{2})\|x_s - x^*\|^2 - \frac{1}{2\eta_s}\|x_{s+1} - x^*\|^2 + \frac{\eta_s L^2}{2} && \text{(by Lipschitzness)}
\end{aligned}$$

By multiplying s on both sides and substituting the step size η_s by $\frac{2}{\alpha(s+1)}$, we get

$$s(f(x_s) - f(x^*)) \leqslant \frac{L^2}{\alpha} + \frac{\alpha}{4}(s(s-1)\|x_s - x^*\|^2 - s(s+1)\|x_{s+1} - x^*\|^2)$$

Finally, we can find the upper bound of the function value shown in Theorem 4 obtained using t steps projected gradient descent

$$\begin{aligned}
f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) &\leqslant \sum_{s=1}^{t} \frac{2s}{t(t+1)} f(x_s) && \text{(by Proposition 4)} \\
&\leqslant \frac{2}{t(t+1)} \sum_{s=1}^{t} \left(s f(x^*) + \frac{L^2}{\alpha} + \frac{\alpha}{4}(s(s-1)\|x_s - x^*\|^2 - s(s+1)\|x_{s+1} - x^*\|^2)\right) \\
&= \frac{2}{t(t+1)} \sum_{s=1}^{t} s f(x^*) + \frac{2L^2}{\alpha(t+1)} - \frac{\alpha}{2}\|x_{t+1} - x^*\|^2 && \text{(by telescoping sum)} \\
&\leqslant f(x^*) + \frac{2L^2}{\alpha(t+1)}
\end{aligned}$$

This concludes that solving an optimization problem with a strongly convex objective function with projected gradient descent has a convergence rate is of the order $\frac{1}{t+1}$, which is faster compared to the case purely with Lipschitzness.

■

3.4 Convergence rate for smooth and strongly convex functions

Theorem 5

*Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ is α-*strongly convex *and β-*smooth*. Let x^* be an optimizer of f, and let x_t be the updated point at step t using gradient descent with a constant step size $\frac{1}{\beta}$, i.e. using the update rule $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$. Then,*

$$\|x_{t+1} - x^*\|^2 \leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2$$

In order to prove Theorem 5, we require use of the following lemma.

Lemma 4

Assume f as in Theorem 5. Then $\forall x, y \in \mathbb{R}^n$ and an update of the form $x^+ = x - \frac{1}{\beta}\nabla f(x)$,

$$f(x^+) - f(y) \leqslant \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2$$

Proof of Lemma 4.

$$\begin{aligned}
f(x^+) - f(x) + f(x) - f(y) &\leqslant \nabla f(x)^\top (x^+ - x) + \frac{\beta}{2}\|x^+ - x\|^2 && \text{(Smoothness)} \\
&\quad + \nabla f(x)^\top (x - y) - \frac{\alpha}{2}\|x - y\|^2 && \text{(Strong convexity)} \\
&= \nabla f(x)^\top (x^+ - y) + \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 && \text{(Definition of } x^+\text{)} \\
&= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 && \text{(Definition of } x^+\text{)}
\end{aligned}$$

■

Now with Lemma 4 we are able to prove Theorem 5.

Proof of Theorem 5.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \frac{1}{\beta}\nabla f(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 - \frac{2}{\beta}\nabla f(x_t)^\top (x_t - x^*) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\
&\leqslant \big(1 - \frac{\alpha}{\beta}\big)\|x_t - x^*\|^2 && \text{(Use of Lemma 4 with } y = x^*,\ x = x_t\text{)} \\
&\leqslant \big(1 - \frac{\alpha}{\beta}\big)^t \|x_1 - x^*\|^2 \\
&\leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2 && \blacksquare
\end{aligned}$$

We can also prove the same result for the constrained case using projected gradient descent.

Theorem 6

Assume $f\colon \Omega \to \mathbb{R}$ is α-strongly convex and β-smooth. Let x^* be an optimizer of f, and let x_t be the updated point at step t using projected gradient descent with a constant step size $\frac{1}{\beta}$, i.e. using the update rule $x_{t+1} = \Pi_\Omega(x_t - \frac{1}{\beta}\nabla f(x_t))$ where Π_Ω is the projection operator. Then,

$$\|x_{t+1} - x^*\|^2 \leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2$$

As in Theorem 5, we will require the use of the following Lemma in order to prove Theorem 6.

Lemma 5

Assume f as in Theorem 5. Then $\forall x, y \in \Omega$, define $x^+ \in \Omega$ as $x^+ = \Pi_\Omega(x - \frac{1}{\beta}\nabla f(x))$ and the function $g\colon \Omega \to \mathbb{R}$ as $g(x) = \beta(x - x^+)$. Then

$$f(x^+) - f(y) \leqslant g(x)^\top (x - y) - \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2$$

Proof of Lemma 5.
The following is given by the Projection Lemma, for all x, x^+, y defined as in Theorem 6.

$$\nabla f(x)^\top (x^+ - y) \leqslant g(x)^\top (x^+ - y)$$

Therefore, following the form of the proof of Lemma 4,

$$\begin{aligned} f(x^+) - f(x) + f(x) - f(y) &\leqslant \nabla f(x)^\top (x^+ - y) + \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &\leqslant \nabla g(x)^\top (x^+ - y) + \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &= \nabla g(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \end{aligned}$$

■

The proof of Theorem 6 is exactly as in Theorem 5 after substituting the appropriate projected gradient descent update in place of the standard gradient descent update, with Lemma 5 used in place of Lemma 4.

References

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. 1 edition. Cambridge University Press.

Bubeck, S. (2015). "Convex Optimization: Algorithms and Complexity". In: *arXiv:1405.4980 [cs, math, stat]*. arXiv: 1405.4980 [cs, math, stat].