Convex minimization

Kent Quanrud

April 28, 2021

This section is about the general, *continuous* optimization problem

$$\text{minimize } f(x) \text{ over } x \in \mathbb{R}^n,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a real-valued function. We want to design algorithms as general as possible - consequently, we will try to identify the minimum structure necessary to make the above problem tractable (step 1: convexity), and only query f in a black-box manner.

All of the algorithms presented here have been extended to more substantially more sophisticated settings (e.g., constrained optimization) and also improved (e.g., accelerated methods). Our primary goal to expose the reader to some of the principles and approaches of continuous optimization, while introducing a relatively small number of parameters and definitions. We refer the reader to [3] for an advanced and comprehensive text on continuous optimization.

Convex minimization is a basic and practical algorithmic problem. However our discussion be in a different language from previous discussions, due to the *continuous* nature of the analysis. It may take some getting use to. There will be some basic mathematical objects (e.g., vector spaces, linear maps, and the Hessian) that not all CS students have background in. Here we can only give a brief introduction. We recommend [1] for an introduction to linear algebra and [2] for an introduction to convex analysis. In the long run, you will want to develop a balanced skill set and appreciation of both the continuous and discrete perspectives.

1 Preliminaries

Let $\mathbb{R}^n$ be the set of real-valued, n-dimensional **vectors** of the form

$$x = (x_1, \ldots, x_n)$$

where $x_1, \ldots, x_n \in \mathbb{R}$. Two vectors $x, y \in \mathbb{R}^n$ sum to make a vector $x + y \in \mathbb{R}^n$, defined by

$$x + y = (x_1 + y_1, \ldots, x_n + y_n).$$

For $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$, we can *scale* x by a to obtain the vector

$$ax = (ax_1, \ldots, ax_n).$$

Note that $a(x+y) = ax + ay$ for $a \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$.

The **inner product** of two vectors $x, y \in \mathbb{R}^n$, denoted $\langle x, y \rangle$, is the quantity $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$. The **Euclidean norm** of a vector x, denoted $\|x\|$, is defined by

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}.$$

Two helpful inequalities to keep in mind are as follows. The **triangle inequality** states that for all $x, y \in \mathbb{R}^n$,

$$\|x - y\| \leq \|x\| + \|y\|.$$

The Cauchy-Schwartz inequality states that for all $x, y \in \mathbb{R}^n$,

$$\langle x, y \rangle \leq \|x\| \|y\|.$$

We note that most of our discussion generalizes to more general normed vector spaces, but we will stick to Euclidean space for simplicity.

2 Convex functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real-valued function. f is **convex** if all $x, y \in \mathbb{R}^n$, and all $\lambda \in [0, 1]$, we have

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

Some simple, one-dimensional examples of convex functions include:

- $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = x^2$.
- $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|$.

Some higher dimensional examples include $f : \mathbb{R}^n \to \mathbb{R}$ defined by $f(x) = \langle x, x \rangle$, or more generally, $f(x) = \langle Ax, Ax \rangle$, where $A \in \mathbb{R}^{m \times n}$ is a matrix.

Certain combinations of convex functions are also convex. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $\alpha > 0$, then αf is convex. If $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are both convex, then $f + g$ is convex. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, and $A : \mathbb{R}^m \to \mathbb{R}^n$ is a linear map, then the function $g(y) = f(Ay)$ is convex.

Derivatives. Our definition above did not require f to be continuously differentiable. In fact, the example $f(x) = |x|$ is not differentiable at $x = 0$. Henceforth, we assume that all our functions f are continuously differentiable. Algorithmically, we will also assume that we are able to compute the derivative of f.

Recall that the **derivative** of f is given by a vector

$$f'(x) \in \mathbb{R}^n, \text{ where } f'_i(x) = \frac{d}{dx_i} f(x) \text{ for each } i.$$

One way to interpret $f'(x)$ is as follows. Fix an input point $x \in \mathbb{R}^n$, as well as a direction $u \in \mathbb{R}^n$. Consider the one-variable function $g(t)$ defined by

$$g(t) = f(x + tu).$$

Then by the chain rule, we have

$$g'(t) = \langle f'(x + tu), u \rangle.$$

In particular, at $t = 0$,

$$g'(0) = \langle f'(x), u \rangle$$

is the (infinitesimal) change in f from moving in the direction u.

Another interpretation, building on the last one, is as follows. Let $x_0, x_1 \in \mathbb{R}^n$ be two input points. For $t \in \mathbb{R}$, let $x_t = tx_1 + (1-t)x_0$. Consider the function $g(t)$ defined by

$$g(t) \stackrel{\text{def}}{=} f(x_t) = f(tx_1 + (1-t)x_0).$$

In particular, from $t = 0$ to 1, $g(t)$ traces the value $f(x_t)$ along the line segment from x_0 to x_1. We have

$$f(x_1) - f(x_0) = g(1) - g(0) = \int_0^1 g'(t)\, dt = \int_0^1 \langle f'(x_t), x_1 - x_0 \rangle\, dt.$$

The point is that the rate of change $\langle f'(x_t), x_1 - x_0 \rangle$ along the line segment from x_0 to x_1 adds up to the total difference $f(x_1) - f(x_0)$.

Derivatives of convex functions. We now examine the case where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable *and* convex.

Lemma 2.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Then f is convex iff for all $x, y \in \mathbb{R}^m$,*

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle. \tag{1}$$

Proof. Suppose f is convex. We have

$$\begin{aligned} \langle f'(x), y - x \rangle &\stackrel{(a)}{=} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} = \frac{f((1-\lambda)x + \lambda y) - f(x)}{\lambda} \\ &\stackrel{(b)}{\leq} \frac{(1-\lambda)f(x) + \lambda f(y) - f(x)}{\lambda} \stackrel{(c)}{=} f(y) - f(x), \end{aligned}$$

which gives the desired inequality up to rearrangement. Here (a) is the chain rule. (b) is by definition of the derivative. *(c) is by convexity.*

Conversely, suppose f satisfies the inequality (1). Let $x_0, x_1 \in \mathbb{R}^n$. Let $x_t = tx_1 + (1-t)x_0$. We want to show that $f(x_t) \leq tf(x_1) + (1-t)f(x_0)$. We have

$$f(x_1) \geq f(x_t) + \langle f'(x_t), x_1 - x_t \rangle = f(x_t) + \langle f'(x_t), (1-t)(x_1 - x_0) \rangle,$$

and

$$f(x_0) \geq f(x_t) + \langle f'(x_t), x_0 - x_t \rangle = f(x_t) + \langle f'(x_t), t(x_0 - x_t) \rangle,$$

by assumption. By adding t times the first inequality and $(1-t)$ times the second inequality, we have

$$tf(x_1) + (1-t)f(x_0) \geq f(x_t),$$

as desired. ■

First-order conditions of optimality. We say that a point x is a **global minimum** of f if $f(x) \leq f(y)$ for all y. We point out that, because the input space is continuous, it is impossible to tell if x is the global minimum by querying values $f(x)$ alone. (What if we nudge x a little bit this way? Or that way?) The following theorem is thus extremely important: it gives us a *certificate* of optimality.

Theorem 2.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and convex. A point x is a global minimum of f iff $f'(x) = \mathbb{0}$.*

Proof. Suppose $f'(x) = \mathbb{0}$. Then for all y, we have

$$f(y) \overset{\text{(a)}}{\geq} f(x) + \langle f'(x), y - x \rangle \overset{\text{(b)}}{=} f(x).$$

Here (a) is convexity (i.e., Lemma 2.1). (b) is because $f'(x) = \mathbb{0}$.

Conversely, suppose x is the global minimum. For all u, we have

$$\langle f'(x), u \rangle = \lim_{t \to 0} \frac{f(x + tu) - f(x)}{t} \overset{\text{(c)}}{\geq} 0.$$

Here (c) is because $f(x + tu) \geq f(x)$. In particular, for $u = -f'(x)$, we have

$$-\|f'(x)\|^2 \geq 0,$$

hence $\|f'(x)\| \leq 0$. But this implies that $f'(x) = \mathbb{0}$. ■

3 Gradient descent for smooth functions

Lipschitz-continuity and smoothness. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. For a fixed parameter $L > 0$, f is ***L*-Lipschitz continuous** if we have

$$|f(x_0) - f(y)| \leq L\|x_0 - y\|$$

for all x_0, y. The *derivative of f* is ***L*-Lipschitz** continuous if we have

$$\|f'(x_0) - f'(x_1)\| \leq L\|x - y\|.$$

In the convex optimization literature, f is called ***L*-smooth** if its derivative is L-Lipschitz.

Lemma 3.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth, and let $x_1, x_0 \in \mathbb{R}^n$. Then*

$$f(x_1) \leq f(x_0) + \langle f'(x_0), x_1 - x_0 \rangle + \frac{L}{2}\|x_1 - x_0\|^2.$$

Proof. For $t \in [0, 1]$, let $x_t = tx_1 + (1-t)x_0$. We have

$$\begin{aligned} f(x_1) - f(x_0) - \langle f'(x_0), x_1 - x_0 \rangle &\overset{(a)}{=} \int_0^1 \frac{d}{dt} f(x_t)\, dt - \langle f'(x_0), x_1 - x_0 \rangle \\ &\overset{(b)}{=} \int_0^1 \langle f'(x_t) - f'(x_0), x_1 - x_0 \rangle, dt \\ &\overset{(c)}{\leq} \int_0^1 \|f'(x_t) - f'(x_0)\| \|x_1 - x_0\|\, dt \\ &\overset{(d)}{\leq} \int_0^1 \|x_t - x_0\| \|x_1 - x_0\|\, dt \\ &\overset{(e)}{=} \|x_1 - x_0\|^2 \int_0^1 t\, dt = \frac{1}{2}\|x_1 - x_0\|^2. \end{aligned}$$

Here (a) is the fundamental theorem of calculus. (b) is the chain rule. (c) is by the Cauchy-Schwartz inequality. *(d) is because f is L-smooth.* (e) observes that $x_t - x_0 = t(x_1 - x_0)$. ∎

3.1 Gradient step for smooth functions

Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function. Then, as observed above, the function

$$g(y) = f(x) + \langle f'(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

gives an upper bound for $f(y)$ for all y. Moreover, $g(y)$ is *not* a black box, and rather simple. Thus a natural *algorithmic idea* to improve from a point x is to identify the point y minimizing $g(y)$. Let us denote

$$x^+ = \arg\min_y f(x) + \langle f'(x), y - x \rangle + \frac{L}{2}\|x - y\|^2. \tag{2}$$

We can interpret x^+ as a *local search* heuristic (pending rigorous analysis). Given a current point x, we use some local information (namely $f'(x)$) to find a new point x^+ that we hope is an improvement on x.

Lemma 3.2. *Let f be L-smooth and $x \in \mathbb{R}^n$. Let x^+ be defined by* (2)*. Then*

$$x^+ = x - \frac{f'(x)}{L}.$$

Moreover,

$$f(x^+) \leq f(x) - \frac{1}{2L}\|f'(x)\|^2.$$

Proof. The function $g(y)$ defined above is convex in y, and the first-order conditions say it is minimized when

$$g'(y) = f'(x) + L(y - x) = 0 \iff y = x - \frac{1}{L} f'(x).$$

The RHS gives x^+.

Moreover, we have

$$\begin{aligned} f(x^+) &\overset{(a)}{\leq} f(x) + \langle f'(x), x^+ - x \rangle + \frac{L}{2} \|x^+ - x\|^2 \\ &\overset{(b)}{=} f(x) - \frac{1}{L} \|f'(x)\|^2 + \frac{1}{2L} \|f'(x)\|^2 \\ &= f(x) - \frac{1}{2L} \|f'(x)\|^2, \end{aligned}$$

as desired. Here (a) is because L is smooth. (b) substitutes $x^+ = x - f'(x)/L$. ■

Minimizing the gradient. A point x is a **critical point** of f is $f'(x) = \mathbb{0}$. Above we observed that critical points of a convex function are a global minimizer. The following theorem gives an algorithm for minimizing the gradient $\|f'(x)\|$. Note that the following does not require f to be convex.

Theorem 3.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function. Let $\epsilon > 0$ and x_0 be given. Let $x^\star$ be a global minimizer for f. Let $x_0, x_1, \dots \in \mathbb{R}^n$ be defined by $x_{t+1} = x_t^+$ (per equation* (2) *above). Then for all $T \in \mathbb{N}$,*

$$\min_{t=1,\dots,T} \|f'(x_t)\|^2 \leq \frac{2L(f(x_0) - f(x^\star))}{T}.$$

Proof. We have

$$\begin{aligned} f(x_0) - f(x^\star) \geq f(x_0) - f(x_T) &\overset{(a)}{=} \sum_{t=1}^{T} f(x_{t-1}) - f(x_t) \\ &\overset{(b)}{\geq} \sum_{t=1}^{T} \frac{\|f'(x_t)\|^2}{2L} \geq \frac{T}{2L} \min_{t=1,\dots,T} \|f'(x_t)\|^2, \end{aligned}$$

as desired. Here (a) is by telescoping series. (b) is by Lemma 3.2. ■

3.2 Gradient step for smooth and convex functions

In this section, we continue to let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth for a known parameter $L > 0$. For a point $x \in \mathbb{R}^n$, let x^+ be the gradient step as defined in equation (2) above. In addition to being smooth, we assume f is *convex*. Let $x^\star$ be a global minimizer of f.

We want to argue that x^+ makes progress towards optimality. We can measure this in two ways. The first is the difference in objective value between the current point and optimum, $f(x) - f(x^\star)$. The second is in the Euclidean distance in the input space, $\|x - x^\star\|$. For both we will show that x^+ improves x.

The first lemma shows that x^+ is closer to $x^\star$ than x.

Lemma 3.4. *For all x,*

$$\|x^+ - x^\star\|^2 \leq \|x - x^\star\|^2.$$

Proof. First we have

$$\frac{1}{2L}\|f'(x)\|^2 \leq f(x) - f(x^+) \overset{(a)}{\leq} f(x) - f(x^\star) \overset{(b)}{\leq} \langle f'(x), x - x^\star \rangle \tag{3}$$

Here (a) is by Lemma 3.2. (b) is because $x^\star$ is the global minimum. (c) is by convexity of f. Now, we have

$$\begin{aligned}\|x^+ - x^\star\|^2 &= \|x - f'(x)/L - x^\star\|^2 \\ &= \|x - x^\star\|^2 - \frac{2}{L}\langle f'(x), x - x^\star \rangle + \frac{1}{L^2}\|f'(x)\|^2 \\ &\overset{(c)}{\leq} \|x - x^\star\|^2.\end{aligned}$$

Here (d) is by the inequality obtained in (3). ■

The next lemma is about the decrease in objective value. Previously we showed that the decrease in objective is proportional to $\|f'(x)\|^2$. The following lemma goes one step further, showing that the decrease in $f(x)$ is proportional to the current gap from the optimum squared, $(f(x) - f(x^\star))^2$.

Lemma 3.5. *For all x,*

$$f(x) - f(x^+) \geq \frac{1}{2L}\left(\frac{f(x) - f(x^\star)}{\|x - x^\star\|}\right)^2.$$

Proof. We have

$$f(x) - f(x^\star) \overset{(a)}{\leq} \langle f'(x), x - x^\star \rangle \overset{(b)}{\leq} \|f'(x)\|\|x - x^\star\| \overset{(c)}{\leq} \|x - x^\star\|\sqrt{2L(f(x) - f(x^+))}.$$

Here (a) is by convexity. (b) is by Cauchy-Schwartz. (c) is by Lemma 3.2. Squaring both sides (where we note that both sides are nonnegative) and rearranging gives the desired inequality. ■

3.3 Gradient descent

We now extend our analysis to an overall algorithm for minimizing a convex function. In this section, Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex and L-smooth function. Let $x^\star$ be a global minimizer for f.

Consider the following algorithm that takes as input an initial point $x_0 \in \mathbb{R}^n$. We compute a sequence of points $x_0, x_1, x_2, \ldots$, where x_{t+1} takes a gradient step from x_t; i.e.,

$$\begin{aligned}x_{t+1} = x_t^+ &= \arg\min_y f(x_t) + \langle f'(x_t), y - x_t \rangle + \frac{L}{2}\|y - x\|^2 \\ &= x_t - \frac{1}{L}f'(x_t).\end{aligned}$$

Then $f(x_t)$ converges to $f(x_0)$ at the following rate.

Theorem 3.6. *For $T = O\left(L\|x_0 - x^\star\|^2\right)$, we have*

$$f(x_T) - f(x^\star) \le \epsilon.$$

Proof. We have

$$f(x_t) - f(x_{t+1}) \overset{(a)}{\ge} \frac{(f(x_t) - f(x^\star))^2}{2L\|x_t - x^\star\|^2} \overset{(b)}{\ge} \frac{(f(x_t) - f(x^\star))^2}{2L\|x_0 - x^\star\|^2}.$$

Here (a) is by Lemma 3.5 and (b) is by Lemma 3.4.

Now, fix $\delta > 0$. Suppose the error is bounded above by $f(x_t) - f(x^\star) \le 2\delta$, and consider the number of iterations until $f(x_t) - f(x^\star) \le \delta$. As long as $f(x_t) - f(x^\star) \ge \delta$, the decrease in error is at least

$$f(x_t) - f(x_{t+1}) \ge \frac{\delta^2}{2L\|x_0 - x^\star\|^2}.$$

Thus there can be at most $2L\|x_0 - x^\star\|^2/\delta$ iterations before has decreased from at most 2δ to below δ.

Our goal is to understand the number of iterations to reduce the error to $\epsilon > 0$. For each i, let $\epsilon_i = 2^i\epsilon$. Then for each i, we have at most $O\left(L\|x_0 - x^\star\|^2/\epsilon_i\right)$ iterations where the error, $f(x_t) - x^\star$, is between ϵ_i and ϵ_{i-1}. Thus the total number of iterations to reach error $\epsilon = \epsilon_0$ is at most

$$\sum_{i=0}^{\infty} O\left(\frac{L\|x_0 - x^\star\|^2}{\epsilon_i}\right) = O\left(\frac{L\|x_0 - x^\star\|^2}{\epsilon}\right)\sum_{i=0}^{\infty}\frac{1}{2^i} \le O\left(\frac{L\|x_0 - x^\star\|^2}{\epsilon}\right),$$

as desired. ∎

4 Exercises

Exercise 1. Prove that for all $x, y \in \mathbb{R}^n$,

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle.$$

Exercise 2. Prove that, for all $a, b \in \mathbb{R}$,

$$2ab \le a^2 + b^2.$$

Exercise 3. Prove the AM-GM inequality (for two variables): For two variables $a, b > 0$,

$$\sqrt{ab} \le \frac{a+b}{2}.$$

Exercise 4. Prove the Cauchy-Schwartz inequality.

Exercise 5. Prove the triangle inequality.

Exercise 6. Prove the following inequality.

$$\|x\| - \|y\| \leq \|x - y\|$$

for all x, y.

Exercise 7. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex and continuously differentiable function over $\mathbb{R}$. You may assume that querying a value $f(x)$, or a gradient $f'(x)$, each take $O(1)$ time.

Consider the problem of minimizing f. Suppose we are promised that there is a unique global minimizer $x^\star$ that lies in the open interval $(0, 1)$. Design and analyze an algorithm that, given an additional parameter $\epsilon > 0$, finds a point such that $x^\star - x \leq \epsilon$. (The running time should depend on ϵ (or rather $1/\epsilon$). As usual, the better the dependency on ϵ, the better.)

Exercise 8. Suppose we want to minimize a smooth convex function f, but did not actually know the value of the parameter L. Instead suppose you knew that L was a value between $1/\text{poly}(n)$ and $\text{poly}(n)$. Design and analyze a variation of the gradient descent algorithm in Section 3 that does not explicitly use L, but still obtains the same iteration count depending on (the unknown value) L. In particular, one needs to figure out how to choose an appropriate step size since the algorithm we discussed choose the step size as a function of L.

Exercise 9. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible map such that $\|f(x)\| \geq \mu\|x\|$ for all x. Prove that

$$\left\|f^{-1}(y)\right\| \leq \frac{1}{\mu}\|y\|$$

for all y.

References

[1] Sheldon Jay Axler. *Linear Algebra Done Right*. Third. Undergraduate Texts in Mathematics. Springer International Publishing, 2015. URL: https://purdue-primo-prod.hosted.exlibrisgroup.com/permalink/f/1c3q7im/TN_cdi_springer_books_10_1007_978_3_319_11080_6.

[2] Jonathan M. Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization, Theory and Examples*. 2000.

[3] Yurii Nesterov. *Lectures on Convex Optimization*. 2nd ed. Vol. 137. Springer Optimization and Its Applications. Springer International Publishing, 2018. URL: https://purdue-primo-prod.hosted.exlibrisgroup.com/permalink/f/1c3q7im/TN_cdi_springer_books_10_1007_978_3_319_91578_4.