Unconstrained minimization of smooth functions

We will start our discussion about solving convex optimization programs by considering the unconstrained case. Our template problem is

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}), \tag{1}$$

where f is convex. While we state this problem as a search over all of $\mathbb{R}^N$, almost everything we say here can be applied to minimized a convex function over an *open* set[1].

In these notes, we discuss two fundamental results. First, for any convex f, we will give conditions under which a minimizer to (1) exists, and show that if $\boldsymbol{x}^\star$ is a local minimizer of (1), then it is also a global minimizer. Second, under the conditions that $f(\boldsymbol{x})$ is convex and differentiable, we will show that $\boldsymbol{x}^\star$ is a minimizer of (1) if and only if the derivative is equal to zero:

$$\boldsymbol{x}^\star \text{ is a global minimizer} \quad \Leftrightarrow \quad \nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}.$$

Something similar to this gradient condition is also true for non-differentiable (but still convex) f; we will explore this a little later in the course.

Existence of minimizers

We begin by recalling a fundamental result in analysis: the *Weierstrass extreme value theorem*. If $f(\boldsymbol{x})$ is a continuous functional on

[1] Intuition: Open sets don't have boundaries, closed sets do. The entire point of the study of constrained optimization, which we will get to next, comes down to treating the fact that the solution can (and probably is) on the boundary of your constraint set.

a compact[2] set $\mathcal{K} \subset \mathbb{R}^N$, then it attains its minimum value at least once. That is,

$$\underset{\boldsymbol{x}\in\mathcal{K}}{\text{minimize}}\, f(\boldsymbol{x})$$

has a minimizer on $\mathcal{K}$ — there exists a $\boldsymbol{x}^\star \in \mathcal{K}$ such that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{K}$. For a proof of this, see just about any introductory text on analysis. The same is also true for f achieving its maximum value on $\mathcal{K}$.

In the unconstrained setting, we are interested in

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{x}),$$

where f is convex. Simple examples illustrate that the minimum does not necessarily have to be achieved for any $\boldsymbol{x}$. That is, there is no $\boldsymbol{x}^\star$ such that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^N$. For example, $f(x) = e^{-x}$ does not have a minimizer on the real line, even though it is as convex and as smooth as can be.

There is, however, a class of functions for which we can guarantee a global minimizer in the unconstrained setting. If the sublevel sets of f,

$$\mathcal{S}(f,\beta) = \{\boldsymbol{x} \in \mathbb{R}^N \ : \ f(\boldsymbol{x}) \leq b\}$$

are *compact* (again, this means closed and bounded), then there will be at least one global minimizer. This should be easy to see — just choose β such that $\mathcal{S}(f,\beta)$ is non-empty, then

$$\underset{\boldsymbol{x}\in\mathcal{S}(f,\beta)}{\text{minimize}}\ f(\boldsymbol{x})$$

[2] In $\mathbb{R}^N$, saying a set is compact is the same as saying that it is closed and bounded.

has a minimizer (by the extreme value theorem), and this also clearly corresponds to a minimizer of f over $\mathbb{R}^N$. If f is continuous (which all convex functions with $\operatorname{dom} f = \mathbb{R}^N$ are), then having compact sublevel sets is the same as being *coercive*: for every sequence $\{\boldsymbol{x}_k\} \subset \mathbb{R}^N$ with $\|\boldsymbol{x}_k\|_2 \to \infty$, we have $f(\boldsymbol{x}_k) \to \infty$ as well. (I will let you prove that at home.)

Local minima are also global minima

Let's nail down a precise statement here right from the start:

Let $f(\boldsymbol{x})$ be convex function on $\mathbb{R}^N$, and suppose $\boldsymbol{x}^\star$ is a local minimizer of f in that there exists an $\epsilon > 0$ such that

$$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}) \quad \text{for all} \quad \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq \epsilon.$$

Then $\boldsymbol{x}^\star$ is also a global minimizer: $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^N$.

To prove this, suppose that there were a $\hat{\boldsymbol{x}} \neq \boldsymbol{x}^\star$ such that $f(\hat{\boldsymbol{x}}) \leq f(\boldsymbol{x}^\star)$. Then by the convexity of f,

$$\begin{aligned} f(\boldsymbol{x}^\star + \theta(\hat{\boldsymbol{x}} - \boldsymbol{x}^\star)) &\leq (1-\theta) f(\boldsymbol{x}^\star) + \theta f(\hat{\boldsymbol{x}}) \\ &\leq f(\boldsymbol{x}^\star) \quad \text{for all} \quad 0 \leq \theta \leq 1. \end{aligned}$$

But choosing a small enough value of θ puts $\boldsymbol{x}^\star + \theta(\hat{\boldsymbol{x}} - \boldsymbol{x}^\star)$ in the neighborhood where $\boldsymbol{x}^\star$ is supposed to be a local min. Specifically, if we take $\theta < \epsilon/\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2$, then the inequality above directly contradicts the assertion that $\boldsymbol{x}^\star$ is a local minimizer. Thus no such $\hat{\boldsymbol{x}}$ can exist.

Our final result in this section gives a sufficient (but definitely not necessary) condition for the minimizer to be unique.

Let f be strictly convex on $\mathbb{R}^N$. If f has a global minimizer, then it is unique.

This is again easy to argue by contradiction. Let $\boldsymbol{x}^\star$ be a global minimizer, and suppose that there existed a $\hat{\boldsymbol{x}} \neq \boldsymbol{x}^\star$ with $f(\hat{\boldsymbol{x}}) = f(\boldsymbol{x}^\star)$. But then there would be many $\boldsymbol{x}$ which achieve smaller values, as for all $0 < \theta < 1$,

$$\begin{aligned} f(\theta\boldsymbol{x}^\star + (1-\theta)\hat{\boldsymbol{x}}) &< \theta f(\boldsymbol{x}^\star) + (1-\theta) f(\hat{\boldsymbol{x}}) \\ &= f(\boldsymbol{x}^\star). \end{aligned}$$

As this would contradict the assertion that $\boldsymbol{x}^\star$ is a global minimizer, no such $\hat{\boldsymbol{x}}$ can exist.

We close this section by re-emphasizing that the entire discussion above would stay the same if we replaced $\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^N} f(\boldsymbol{x})$ with $\text{minimize}_{\boldsymbol{x}\in\mathcal{U}} f(\boldsymbol{x})$ for any open set $\mathcal{U} \subset \mathbb{R}^N$.

Optimality conditions for unconstrained optimization

How do we know when we have a minimizer of a convex function on our hands? What is our "certificate of optimality"? For the time being, we will assume that f is differentiable, that $\nabla f(\boldsymbol{x})$ exists at every point we are considering. There are a comparable set of results for non-smooth f that we will discuss in last segment of the course.

In our discussion on algorithms for unconstrained optimization over the past two weeks, we have often mentioned the following, but have never actually discussed exactly why it is true.

Let f be a convex differentiable function on $\mathbb{R}^N$. Then $\boldsymbol{x}^\star$ solves

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x})$$

if and only if $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$.

The proof of this relies on the critical fact that we can decrease f by moving in a direction which has an obtuse angle with the gradient.

Let f be a function on $\mathbb{R}^N$ that is differentiable at $\boldsymbol{x}$, and let $\boldsymbol{d} \in \mathbb{R}^N$ be a vector obeying $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle < 0$. Then for small enough $t > 0$,

$$f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x}).$$

We call such a $\boldsymbol{d}$ a **descent direction** from $\boldsymbol{x}$. Similarly, if $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle > 0$, then for small enough $t > 0$,

$$f(\boldsymbol{x} + t\boldsymbol{d}) > f(\boldsymbol{x}).$$

We call such a $\boldsymbol{d}$ an **ascent direction** from $\boldsymbol{x}$.

This fundamental fact is a direct consequence of the Taylor theorem (see the Technical Details section below): for any $\boldsymbol{u} \in \mathbb{R}^N$,

$$f(\boldsymbol{x} + \boldsymbol{u}) = f(\boldsymbol{x}) + \langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle + h(\boldsymbol{u})\|\boldsymbol{u}\|_2,$$

where $h(\boldsymbol{u}) : \mathbb{R}^N \to \mathbb{R}$ is some function satisfying $h(\boldsymbol{u}) \to 0$ as $\boldsymbol{u} \to \mathbf{0}$. Taking $\boldsymbol{u} = t\boldsymbol{d}$, we have

$$f(\boldsymbol{x} + t\boldsymbol{d}) = f(\boldsymbol{x}) + t\left(\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle + h(t\boldsymbol{d})\|\boldsymbol{d}\|_2\right).$$

For $t > 0$ small enough, we can make $|h(t\boldsymbol{d})| \cdot \|\boldsymbol{d}\|_2 < |\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle|$, and so the term inside the parentheses above is negative if $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle$ is negative, and it is positive if $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle$ is positive.

At a particular point $\boldsymbol{x}^\star$, the only way we can make $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0$ for all choices of $\boldsymbol{d}$ is if $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$. So clearly

$$\boldsymbol{x}^\star \text{ is a minimizer} \quad \Rightarrow \quad \nabla f(\boldsymbol{x}^\star) = \mathbf{0}.$$

On the other hand, if f is convex, then

$$f(\boldsymbol{x}^\star + t\boldsymbol{d}) \geq f(\boldsymbol{x}^\star) + t\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}^\star)\rangle,$$

for all $t \in \mathbb{R}$ and choices of $\boldsymbol{d} \in \mathbb{R}^N$. This now makes it clear that

$$\nabla f(\boldsymbol{x}^\star) = \mathbf{0} \quad \Rightarrow \quad \boldsymbol{x}^\star \text{ is a minimizer.}$$

Again, for everything we have said in this section, you can use any open domain $\boldsymbol{\mathcal{U}}$ in place of $\mathbb{R}^N$.

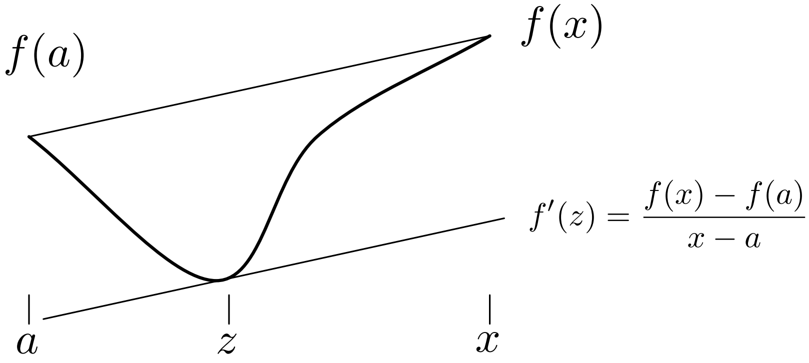
Technical Details: Taylor's Theorem

The following is an overview of classical results from real analysis. For details and proofs, see just about any text on this subject, including [Rud76, Apo74, Rud86].

You might recall the mean-value theorem from your first calculus class. If $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function on the interval $[a, x]$, then there is a point inside this interval where the derivative of f matches the line drawn between $f(a)$ and $f(x)$. More precisely, there exists $z \in [a, x]$ such that

$$f'(z) = \frac{f(x) - f(a)}{x - a}.$$

Here is a picture:



We can re-arrange the expression above to say that there is some z between a and x such that

$$f(x) = f(a) + f'(z)(x - a).$$

The mean-value theorem extends to derivatives of higher order; in this case it is known as *Taylor's theorem*. For example, suppose that f is twice differentiable on $[a, x]$, and that the first derivative f' is continuous. Then there exists a z between a and x such that

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(z)}{2}(x-a)^2.$$

In general, if f is $k+1$ times differentiable, and the first k derivatives are continuous, then there is a point z between a and x such that

$$f(x) = p_{k,a}(x) + \frac{f^{(k+1)}(z)}{k!}(x-a)^{k+1},$$

where $p_{k,a}(x)$ polynomial formed from the first k terms of the Taylor series expansion around a:

$$p_{k,a}(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k.$$

These results give us a way to quantify the accuracy of the Taylor approximation around a point. For example, if f is twice differentiable with f' continuous, then

$$f(x) = f(a) + f'(a)(x-a) + h_1(x)(x-a),$$

for a function $h_1(x)$ goes to zero as x goes to a:

$$\lim_{x \to a} h_1(x) = 0.$$

In fact, you do not even need two derivatives for this to be true. If f has a single derivative, then we can find such an h_1. When f has two derivatives, then we have an explicit form for h_1:

$$h_1(x) = \frac{f''(z_x)}{2}(x-a),$$

where z_x is the point returned by the (generalization of) the mean value theorem for a given x.

In general, if f has k derivatives, then there exists an $h_k(x)$ with $\lim_{x\to a} h_k(x) = 0$ such that

$$f(x) = p_{k,a}(x) + h_k(x)(x-a)^k.$$

All of the results above extend to functions of multiple variables. For example, if $f(\boldsymbol{x}) : \mathbb{R}^N \to \mathbb{R}$ is differentiable, then around any point $\boldsymbol{a}$,

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{x}-\boldsymbol{a}\rangle + h_1(\boldsymbol{x})\|\boldsymbol{x}-\boldsymbol{a}\|_2,$$

where $h_1(\boldsymbol{x}) \to 0$ as $\boldsymbol{x}$ approaches $\boldsymbol{a}$ from any direction. If $f(\boldsymbol{x})$ is twice differentiable and the first derivative is continuous, then there exists $\boldsymbol{z}$ on the line between $\boldsymbol{a}$ and $\boldsymbol{x}$ such that

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \langle \nabla f(\boldsymbol{a}), \boldsymbol{x}-\boldsymbol{a}\rangle + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{a})^{\mathrm{T}}\nabla^2 f(\boldsymbol{z})(\boldsymbol{x}-\boldsymbol{a}).$$

We will use these two particular multidimensional results repeatedly in our analysis throughout the course, referring to them generically as “Taylor’s theorem”.

References

[Apo74] T. M. Apostol. *Mathematical Analysis*. Pearson, 2nd edition, 1974.

[Rud76] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[Rud86] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 3rd edition, 1986.