III. Unconstrained Optimization of Smooth Functions

Unconstrained optimization

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].

We will first 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:

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

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.

[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.

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 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 in $\mathcal{K}$ — there exists an $\boldsymbol{x}^\star \in \mathcal{K}$ 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 in $\mathcal{K}$.

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,

$$\mathcal{S}_\alpha(f) = \{\boldsymbol{x} \in \mathbb{R}^N \;\; : \;\; f(\boldsymbol{x}) \leq \alpha\}$$

are *compact* (again, this means closed and bounded), then there will be at least one global minimizer. This should be easy to see — just

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

choose α such that $\mathcal{S}_\alpha(f)$ is non-empty, then

$$\underset{\boldsymbol{x} \in \mathcal{S}_\alpha(f)}{\text{minimize}} \; f(\boldsymbol{x})$$

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 was an $\widehat{\boldsymbol{x}} \neq \boldsymbol{x}^\star$ such that $f(\widehat{\boldsymbol{x}}) < f(\boldsymbol{x}^\star)$. Then by the convexity of f,

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

But choosing a small enough value of θ puts $\boldsymbol{x}^\star + \theta(\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star)$ in the neighborhood where $\boldsymbol{x}^\star$ is supposed to be a local min. Specifically,

if we take $\theta < \epsilon / \|\widehat{\boldsymbol{x}} - \boldsymbol{x}^\star\|_2$, then the inequality above directly contradicts the assertion that $\boldsymbol{x}^\star$ is a local minimizer. Thus no such $\widehat{\boldsymbol{x}}$ 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 an $\widehat{\boldsymbol{x}} \neq \boldsymbol{x}^\star$ with $f(\widehat{\boldsymbol{x}}) = f(\boldsymbol{x}^\star)$. But then there would be many $\boldsymbol{x}$ which achieve smaller values, as for all $0 < \theta < 1$,

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

As this would contradict the assertion that $\boldsymbol{x}^\star$ is a global minimizer, no such $\widehat{\boldsymbol{x}}$ 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 this course we have already 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 $\boldsymbol{u} \to \boldsymbol{0}$. 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 $\nabla f(\boldsymbol{x}^\star) = \boldsymbol{0}$. So clearly

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

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

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

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

Algorithms for unconstrained minimization

We want to solve

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

where f is convex. In this section, in addition to assuming that f is differentiable (so its gradient exists at every point), we will also assume that it is **smooth** (we will consider a few different definitions of "smooth" — qualitatively, this just means that the gradient changes in a controlled manner as we move from point to point).[3]

Since f is convex, a necessary and sufficient condition for $\boldsymbol{x}^\star$ to be a minimizer is that the gradient vanishes:

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

As noted above, it is not a given that such a $\boldsymbol{x}^\star$ exists, but in this section we will assume that f does have (at least one) minimizer, and we will denote the optimal value as $p^\star = f(\boldsymbol{x}^\star)$.

Every general-purpose optimization algorithm we will look at in this course is **iterative** — they will all have the basic form:

$k = 0$, $\boldsymbol{x}^{(0)} =$ initial guess
while not converged
    $k = k + 1$
    calculate a direction to move $\boldsymbol{d}^{(k)}$
    calculate a step size $t_k \geq 0$
    $\boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + t_k \, \boldsymbol{d}^{(k)}$
    check convergence criteria

[3] While many problems are smooth, methods for nonsmooth $f(\boldsymbol{x})$ are also of great interest, and will be covered later in the course. Nonsmooth methods are not much more involved algorithmically, but they are slightly harder to analyze.

In the coming lectures, we will focus primarily on two methods for computing the direction $\boldsymbol{d}^{(k)}$.

1. **Gradient descent**: we take

$$\boldsymbol{d}^{(k)} = -\nabla f(\boldsymbol{x}^{(k-1)}).$$

   This is the direction of "steepest descent" (where "steepest" is defined relative to the Euclidean norm). Gradient descent iterations are **cheap**, but typically many iterations are required for convergence.

2. **Newton's method**: we build a quadratic model around $\boldsymbol{x}^{(k)}$ then compute the exact minimizer of this quadratic by solving a system of equations. This corresponds to taking

$$\boldsymbol{d}^{(k)} = -(\nabla^2 f(\boldsymbol{x}^{(k-1)}))^{-1} \nabla f(\boldsymbol{x}^{k-1}),$$

   that is, the inverse of the Hessian evaluated at $\boldsymbol{x}^{(k-1)}$ applied to the gradient evaluated at the same point. Newton iterations tend to be **expensive** (as they require a system solve), but they typically converge in far fewer iterations than gradient descent.

Whichever direction we choose, it should be a **descent direction**; $\boldsymbol{d}^{(k)}$ should satisfy

$$\langle \boldsymbol{d}^{(k)}, \nabla f(\boldsymbol{x}^{(k-1)}) \rangle \leq 0.$$

Since f is convex, it is always true that

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

and so to decrease the value of the functional while moving in direction $\boldsymbol{d}$, it is necessary that the inner product above be negative.

Line search

After the step direction is chosen, we need to compute how far to move. There are many methods for doing this, here are three:

Exact: Solve the 1D optimization program

$$\underset{s \geq 0}{\text{minimize}} \;\; f(\boldsymbol{x}^{(k-1)} + s\boldsymbol{d}^{(k)}).$$

This is typically not worth the trouble, but there are instances (i.e. unconstrained convex quadratic programs) when it can be solved analytically.

Backtracking: Start with a step size of $t = 1$, then decrease by a factor of β until we are below a certain line.
Fix $\alpha \in (0, 0.5)$ and $\beta \in (0, 1)$.
Given a starting point $\boldsymbol{x}$ and direction $\boldsymbol{d}$,
$t = 1$
repeat
    if $f(\boldsymbol{x} + t\boldsymbol{d}) < f(\boldsymbol{x}) + \alpha\, t \langle \boldsymbol{d}, \nabla f(\boldsymbol{x}) \rangle$
        converged
    else
        $t = \beta t$
    endif
until converged

Here is a figure (from [BV04, Chapter 9]):

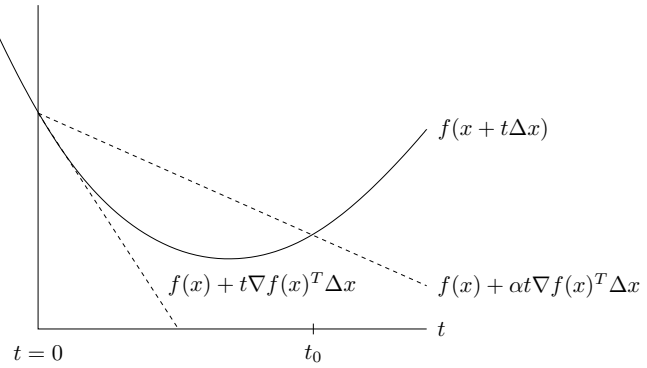


Figure 9.1 *Backtracking line search.* The curve shows f, restricted to the line over which we search. The lower dashed line shows the linear extrapolation of f, and the upper dashed line has a slope a factor of α smaller. The backtracking condition is that f lies below the upper dashed line, *i.e.*, $0 \leq t \leq t_0$.

The backtracking line search tends to be cheap, and works very well in practice. Typically, we take α small (to encourage a large step) and $\beta \in [0.3, 0.8]$.

Fixed: Finally, we can just use a constant step size $t_k = t$. This will actually work if the step size is small enough, but usually this results in way too many iterations.

Convergence of gradient descent

How quickly does gradient descent converge?

I'm glad you asked!

We will look at two different smoothness assumptions on f, and translate them into convergence rates. In the first case, we assume that f is twice differentiable (so that its Hessian exists everywhere), and that

$$m\mathbf{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq M\mathbf{I}.$$

That is, the eigenvalues of the Hessian (which is always in S^N for a convex function) are bounded between $m > 0$ and $M < \infty$. We call this assumption **strong convexity** (note that the lower bound by itself means that f is strictly convex).

Recall that the main consequence of convexity is that we have a way to compute a linear global lower bound at every point:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y}.$$

The main consequence of strong convexity is that we have in addition a quadratic lower and upper bound. By the Taylor theorem, we know that for any $\boldsymbol{x}, \boldsymbol{y}$,

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

for some point $\boldsymbol{z}$ on the line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. Thus we have the bounds

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2, \tag{2}$$

and

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{M}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{3}$$

An immediate consequence of the stronger lower bound in (2) is that we can tell at any point $\boldsymbol{x}$ how close to optimal we are. The smallest value the right hand side can take in that inequality is when $\widetilde{\boldsymbol{y}} = \boldsymbol{x} - m^{-1} \nabla f(\boldsymbol{x})$; plugging that value into the right hand side yields

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) - \frac{1}{2m} \|\nabla f(\boldsymbol{x})\|_2^2,$$

a bound which holds for every $f(\boldsymbol{y})$. In particular, it holds for the optimal value $p^\star$, and so

$$p^\star \geq f(\boldsymbol{x}) - \frac{1}{2m}\|\nabla f(\boldsymbol{x})\|_2^2. \tag{4}$$

Thus we get a bound on $f(\boldsymbol{x}) - p^\star$ just by calculating the norm of the gradient. It also re-iterates that if $\|\nabla f(\boldsymbol{x})\|_2$ is small, we are close to the solution.

We can also bound how close $\boldsymbol{x}$ is to the minimizer $\boldsymbol{x}^\star$. Using (2) with $\boldsymbol{y} = \boldsymbol{x}^\star$,

$$\begin{aligned} p^\star = f(\boldsymbol{x}^\star) &\geq f(\boldsymbol{x}) + \langle \boldsymbol{x}^\star - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle + \frac{m}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2^2 \\ &\geq f(\boldsymbol{x}) - \|\boldsymbol{x}^\star - \boldsymbol{x}\|_2 \|\nabla f(\boldsymbol{x})\|_2 + \frac{m}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2^2. \end{aligned}$$

Then since $p^\star \leq f(\boldsymbol{x})$,

$$-\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2 \|\nabla f(\boldsymbol{x})\|_2 + \frac{m}{2}\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2^2 \leq 0,$$

and so

$$\|\boldsymbol{x}^\star - \boldsymbol{x}\|_2 \leq \frac{2}{m}\|\nabla f(\boldsymbol{x})\|_2.$$

These facts are very useful for defining stopping criteria.

Suppose we have a strongly convex function that we minimize using gradient descent with an **exact line search**. We can show that at each iteration, the gap $f(\boldsymbol{x}^{(k)}) - p^\star$ gets cut down by a fixed factor. We consider a single iteration — we will use $\boldsymbol{x}$ to denote the current

point, and $\boldsymbol{x}^+ = \boldsymbol{x} - t_{\text{exact}} \nabla f(\boldsymbol{x})$ to denote the result of the gradient step. We choose t_{exact} by minimizing the following function:

$$\widetilde{f}(t) = f(\boldsymbol{x} - t\nabla f(\boldsymbol{x})).$$

By strong convexity, we know that

$$\widetilde{f}(t) \leq f(\boldsymbol{x}) - t\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\boldsymbol{x})\|_2^2.$$

By definition of t_{exact}, we know

$$f(\boldsymbol{x}^+) = \widetilde{f}(t_{\text{exact}}) \leq \widetilde{f}(1/M) \leq f(\boldsymbol{x}) - \frac{1}{2M}\|\nabla f(\boldsymbol{x})\|_2^2.$$

From (4), we also know that

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2m(f(\boldsymbol{x}) - p^\star),$$

and so

$$f(\boldsymbol{x}^+) - p^\star \;\leq\; f(\boldsymbol{x}) - p^\star - \frac{m}{M}(f(\boldsymbol{x}) - p^\star),$$

which means

$$\frac{f(\boldsymbol{x}^+) - p^\star}{f(\boldsymbol{x}) - p^\star} \leq \left(1 - \frac{m}{M}\right).$$

That is, the gap between the current functional evaluation and the optimal value has been cut down by a factor of $1 - m/M < 1$.

Applying this relationship recursively, we see that after k iterations of gradient descent, we have

$$\frac{f(\boldsymbol{x}^{(k)}) - p^\star}{f(\boldsymbol{x}^{(0)}) - p^\star} \leq \left(1 - \frac{m}{M}\right)^k.$$

Another way to say this is that we can achieve accuracy

$$f(\boldsymbol{x}^{(k)}) - p^\star \leq \epsilon,$$

by taking

$$k \geq \frac{\log(E_0/\epsilon)}{\log(1 - m/M)}, \quad E_0 = f(\boldsymbol{x}^{(0)}) - p^\star,$$

steps.

There are similar results for gradient descent on strongly convex functions using backtracking. They are similar in nature; they have the same linear convergence but with constants that depend on α and β along with m and M. See [BV04, p. 468].

Lipschitz gradient condition

We can also get (much weaker) convergence results when f is not strongly convex (or even necessarily twice differentiable), but rather has a **Lipschitz gradient**. This means that there exists an $L > 0$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{5}$$

The upshot here is that we still have a quadratic upper bound for f at every point, but not the lower bound.

To be precise, if f obeys (5), then

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2.$$

This follows immediately from the fundamental theorem of calculus[4]:

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

and so

$$\begin{aligned} f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle &= \int_0^1 \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f((1-t)\boldsymbol{x} + t\boldsymbol{y}) - \nabla f(\boldsymbol{x})\rangle \ \mathrm{d}t \\ &\leq \|\boldsymbol{y} - \boldsymbol{x}\|_2 \int_0^1 \|\nabla f((1-t)\boldsymbol{x} + t\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|_2 \ \mathrm{d}t \\ &\leq L\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \int_0^1 t \ \mathrm{d}t \\ &= \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \end{aligned}$$

Now, let's consider running gradient descent on such a function with a **fixed step size** $t \leq 1/L$. As before, we denote the current iterate as $\boldsymbol{x}$ and the next iterate at $\boldsymbol{x}^+ = \boldsymbol{x} - t\nabla f(\boldsymbol{x})$. We have

$$\begin{aligned} f(\boldsymbol{x}^+) &\leq f(\boldsymbol{x}) - \left(1 - \frac{Lt}{2}\right) t\|\nabla f(\boldsymbol{x})\|_2^2 \\ &\leq f(\boldsymbol{x}) - \frac{t}{2}\|\nabla f(\boldsymbol{x})\|_2^2, \end{aligned}$$

for the range of t we are considering. By convexity of f,

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x})\rangle,$$

where $\boldsymbol{x}^\star$ is a minimizer of f, and so

$$f(\boldsymbol{x}^+) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x})\rangle - \frac{t}{2}\|\nabla f(\boldsymbol{x})\|_2^2,$$

[4]The 1D function $\widetilde{f}(t) = f((1-t)\boldsymbol{x} + t\boldsymbol{y})$ has derivative $\widetilde{f}'(t) = \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f\left((1-t)\boldsymbol{x} + t\boldsymbol{y}\right)\rangle$. This is just a simple application of the chain rule.

and then substituting $\nabla f(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}^+)/t$ yields

$$\begin{aligned} f(\boldsymbol{x}^+) - f(\boldsymbol{x}^\star) &\leq \frac{1}{t}\langle \boldsymbol{x} - \boldsymbol{x}^\star, \boldsymbol{x} - \boldsymbol{x}^+ \rangle - \frac{1}{2t}\|\boldsymbol{x} - \boldsymbol{x}^+\|_2^2 \\ &= \frac{1}{2t}\left(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^+ - \boldsymbol{x}^\star\|_2^2\right). \end{aligned}$$

Summing this difference over k iterations yields:

$$\begin{aligned} \sum_{i=1}^{k} f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^\star) &\leq \frac{1}{2t}\left(\sum_{i=1}^{k} \|\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|_2^2\right) \\ &= \frac{1}{2t}\left(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2^2\right) \\ &\leq \frac{1}{2t}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2. \end{aligned}$$

Since $f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^\star)$ is monotonically decreasing in i, the k^{th} term will be smaller than the average:

$$\begin{aligned} f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) &\leq \frac{1}{k}\sum_{i=1}^{k} f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^\star) \\ &\leq \frac{1}{2tk}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2. \end{aligned}$$

Note that this convergence guarantee is much slower — it is $O(1/k)$ in place of $O(c^k)$ for some $c < 1$. This is the price we pay for allowing f to not be as smooth.

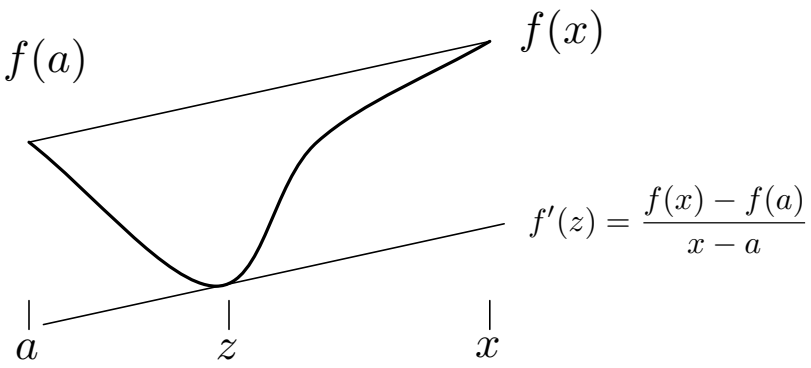
Technical Details: Taylor's Theorem

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 a $z \in [a, x]$ 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.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

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

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