Differentiable functions

As we will see below, there are alternative (but equally natural) ways to think about a convex function f when it is **differentiable**. Before delving into this, we're first going to do a brief review of the key notions from multivariable calculus that lie at the heart of how we think about many convex optimization problems.

The gradient and the Hessian

First, recall that a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable if its derivative, defined as

$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta},$$

exists for all $x \in \operatorname{dom} f$. To extend this notion to functions of multiple variables, we must first extend our notion of a derivative. For a function $f : \mathbb{R}^N \to \mathbb{R}$ that is defined on N-dimensional vectors, recall that the **partial derivative** with respect to x_n is

$$\frac{\partial f(\boldsymbol{x})}{\partial x_n} = \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{e}_n) - f(\boldsymbol{x})}{\delta},$$

where $\boldsymbol{e}_n$ is the n^{th} "standard basis element", i.e., the vector of all zeros with a single 1 in the n^{th} entry.

The **gradient** of a function $f : \mathbb{R}^N \to \mathbb{R}$ is the vector of partial derivatives given by:

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_N} \end{bmatrix}.$$

Similar to the scalar case, we say that f is differentiable if the gradient exists for each $\boldsymbol{x} \in \operatorname{dom} f$.

We will use the term gradient in two subtly different ways. Sometimes we use $\nabla f(\boldsymbol{x})$ to describe a *vector-valued function* or a *vector field*, i.e., a function that takes an arbitrary $\boldsymbol{x} \in \mathbb{R}^N$ and produces another vector. However, we also use the term gradient, and the same notation $\nabla f(\boldsymbol{x})$, to refer to vector given by the gradient at a particular point $\boldsymbol{x}$. So sometimes when we say "gradient" we mean a vector-valued function, and sometimes we mean a single vector, and in both cases we use the notation $\nabla f(\boldsymbol{x})$. Which one will usually be obvious by the context.[1]

Note that in some cases we will use the notation $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ to indicate that we are taking the gradient with respect to $\boldsymbol{x}$. This can be helpful when f is a function of more variables than just $\boldsymbol{x}$, but most of the time this is not necessary so we will typically use the simpler $\nabla f(\boldsymbol{x})$.

Here we adopt the convention that the gradient is a *column vector*. This is the most common choice and is most convenient in this class, but some texts will instead treat the gradient as a row vector. The reason for this is to align with the standard convention for the *Jacobian*.[2] Thus, it is always worth double-checking what notation is being used when consulting outside resources.

[1] This is just like in the scalar case, where the notation $f(x)$ can sometimes refer to the function f and sometimes the function evaluated at x.

[2] The Jacobian of a vector-valued function $f : \mathbb{R}^N \to \mathbb{R}^M$ is the $M \times N$ matrix of partial derivatives with respect to each dimension in the range. In this course we will mostly be concerned with functions mapping to a single dimension, in which case the Jacobian would be the $1 \times N$ matrix $\nabla^{\mathrm{T}} f(\boldsymbol{x})$, i.e., the gradient but treated as a row vector. Directly defining the gradient as a row vector instead of a column vector is thus more convenient in some contexts.

Finally, we will also occasionally need to make use of second derivatives. For a function $f : \mathbb{R}^N \to \mathbb{R}$, this is captured by the **Hessian**, which is the matrix of all possible pairwise partial derivatives:[3]

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N \partial x_1} & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f(\boldsymbol{x})}{\partial x_N^2} \end{bmatrix}.$$

Interpretation of the gradient

The gradient is one of the most fundamental concepts of this course. We can interpret the gradient in many ways. One way to think of the gradient when evaluated at a particular point $\boldsymbol{x}$ is that it defines a linear mapping from $\mathbb{R}^N$ to $\mathbb{R}$. Specifically, given a $\boldsymbol{u} \in \mathbb{R}^N$, we can use $\nabla f(\boldsymbol{x})$ to define a mapping of $\boldsymbol{u}$ to $\mathbb{R}$ by simply taking the inner product between the two vectors:

$$\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle.$$

What does this mapping tell us? It computes the **directional derivative** of f in the direction of $\boldsymbol{u}$, i.e.,

$$\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})) \rangle = \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{u}) - f(\boldsymbol{x})}{\delta}. \tag{1}$$

This tells us how fast f is changing at $\boldsymbol{x}$ when we move in the direction of $\boldsymbol{u}$.

[3] Note that if we view the gradient $\nabla f(\boldsymbol{x})$ as a vector valued function mapping from $\mathbb{R}^N$ to $\mathbb{R}^N$, then the Hessian is the same as the Jacobian of the gradient.

This fundamental fact is a direct consequence of Taylor's theorem (see the Technical Details section below). Specifically, let $f : \mathbb{R}^N \to \mathbb{R}$ be any differentiable function. Then for any $\boldsymbol{u} \in \mathbb{R}^N$, we can write

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

where $h(\boldsymbol{u}) : \mathbb{R}^N \to \mathbb{R}$ is some function satisfying $h(\boldsymbol{u}) \to 0$ as $\boldsymbol{u} \to \boldsymbol{0}$.

If we substitute $\delta \boldsymbol{u}$ in place of $\boldsymbol{u}$ above and rearrange, we obtain the identity

$$\begin{aligned} \langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle &= \frac{f(\boldsymbol{x}+\delta\boldsymbol{u}) - f(\boldsymbol{x}) - h(\delta\boldsymbol{u})\|\delta\boldsymbol{u}\|_2}{\delta} \\ &= \frac{f(\boldsymbol{x}+\delta\boldsymbol{u}) - f(\boldsymbol{x})}{\delta} - h(\delta\boldsymbol{u})\|\boldsymbol{u}\|_2. \end{aligned}$$

Note that this holds for any $\delta > 0$. Since $h(\delta\boldsymbol{u}) \to 0$ as $\delta \to 0$, we can arrive at (1) by simply taking the limit as $\delta \to 0$.

A related way to think of $\nabla f(\boldsymbol{x})$ is as a vector that is pointing in the direction of *steepest ascent*, i.e., the direction in which f increases the fastest when starting at $\boldsymbol{x}$. To justify this, note that we just observed that we can interpret $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle$ as measuring how quickly f increases when we move in the direction of $\boldsymbol{u}$. How can we find the direction $\boldsymbol{u}$ that maximizes this quantity? You may recall that the Cauchy-Schwarz inequality tells us that

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

and that this holds with equality when $\boldsymbol{u}$ is co-linear with $\nabla f(\boldsymbol{x})$, i.e., when $\boldsymbol{u}$ points in the same direction as $\nabla f(\boldsymbol{x})$. Specifically, this implies that $\nabla f(\boldsymbol{x})$ is the direction of steepest *ascent*, and $-\nabla f(\boldsymbol{x})$ is the direction of steepest *descent*.

More broadly, this characterizes the entire sets of ascent/descent directions. Suppose that $f : \mathbb{R}^N \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$. If $\boldsymbol{u} \in \mathbb{R}^N$ is a vector obeying $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle < 0$, then we say that $\boldsymbol{u}$ is a **descent direction** from $\boldsymbol{x}$, and for small enough $t > 0$,

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

Similarly, if $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}) \rangle > 0$, then we say that $\boldsymbol{u}$ is an **ascent direction** from $\boldsymbol{x}$, and for small enough $t > 0$,

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

It should hopefully not be a huge stretch of the imagination to see that being able to compute the direction of steepest ascent (or steepest descent) will be useful in the context of finding a maximum/minimum of a function.

Equivalent characterizations of convexity

Now that we can talk intelligently about what it means for a function to be differentiable, we can look more carefully at functions that are both convex *and* differentiable. For such functions, there are equivalent (possibly simpler) ways to think about convexity.

First order conditions for convexity

If f is differentiable, then it is convex if and only if

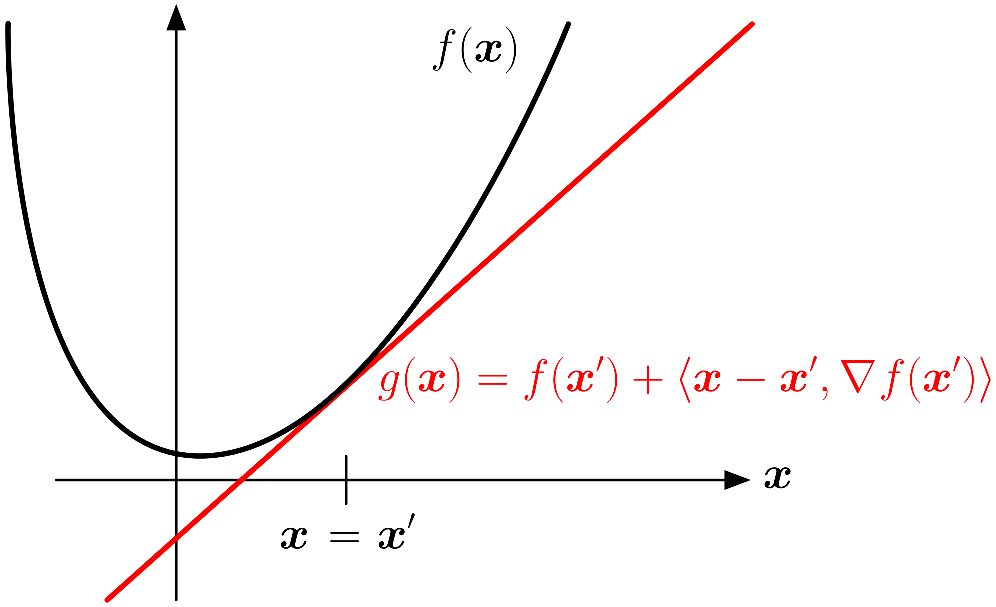
$$f(\boldsymbol{x}) \;\geq\; f(\boldsymbol{x}') + \langle \boldsymbol{x} - \boldsymbol{x}', \nabla f(\boldsymbol{x}') \rangle \tag{2}$$

for all $\boldsymbol{x}, \boldsymbol{x}' \in \operatorname{dom} f$.

This says that the linear approximation of f formed from the tangent line (or plane, or hyperplane, as we move to higher dimensions) will always remain *below* f. Specifically, in (2) we are comparing two functions: $f(\boldsymbol{x})$ and the tangent

$$g(\boldsymbol{x}) = f(\boldsymbol{x}') + \langle \boldsymbol{x} - \boldsymbol{x}', \nabla f(\boldsymbol{x}') \rangle.$$

What (2) is saying is that $g(\boldsymbol{x})$ is a *global underestimator* of $f(\boldsymbol{x})$.


$f(\boldsymbol{x})$
$g(\boldsymbol{x}) = f(\boldsymbol{x}') + \langle \boldsymbol{x} - \boldsymbol{x}', \nabla f(\boldsymbol{x}') \rangle$
$\boldsymbol{x}$
$\boldsymbol{x} = \boldsymbol{x}'$

This is an incredibly useful fact, and if we never had to worry about functions that were not differentiable, we might actually just take this as the definition of a convex function.

We now prove this result. It is easy to show that if f is convex and differentiable, then we must have (2). Specifically, since f is convex, we have that for any $\theta \in [0, 1]$,

$$f(\theta \boldsymbol{x} + (1-\theta)\boldsymbol{x}') \leq \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{x}').$$

Rearranging this, we have

$$\begin{aligned} f(\boldsymbol{x}) &\geq \frac{f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{x}') - (1-\theta)f(\boldsymbol{x}')}{\theta} \\ &= f(\boldsymbol{x}') + \frac{f(\boldsymbol{x}' + \theta(\boldsymbol{x} - \boldsymbol{x}')) - f(\boldsymbol{x}')}{\theta}. \end{aligned}$$

The inequality in (2) follows from this by taking the limit as $\theta \to 0$. To see this, recall (from our review of multivariable calculus) that the inner product between the gradient of f evaluated at $\boldsymbol{x}'$ and another vector $\boldsymbol{u}$ is the directional derivative of f in the direction of $\boldsymbol{u}$; setting $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{x}'$ this is exactly the same as

$$\langle \boldsymbol{x} - \boldsymbol{x}', \nabla f(\boldsymbol{x}') \rangle = \lim_{\theta \to 0} \frac{f(\boldsymbol{x}' + \theta(\boldsymbol{x} - \boldsymbol{x}')) - f(\boldsymbol{x}')}{\theta}.$$

We next need to show that if (2) holds, then f is convex. To do so, let $\boldsymbol{x} \neq \boldsymbol{y}$ be arbitrary vectors in $\operatorname{dom} f$ and fix $\theta \in [0,1]$. Set $\boldsymbol{z} = \theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}$. From (2) we have

$$f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \langle \boldsymbol{x} - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle$$

and

$$f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \langle \boldsymbol{y} - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle$$

If we multiply the first inequality by θ, the second by $1-\theta$, and then add the two, then since $\theta(\boldsymbol{x} - \boldsymbol{z}) + (1-\theta)(\boldsymbol{y} - \boldsymbol{z}) = \boldsymbol{0}$, we obtain

$$\theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y}) \geq f(\boldsymbol{z}) = f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}),$$

which is exactly the definition of a convex function.

Second-order conditions for convexity

Recall that we say that $f : \mathbb{R}^N \to \mathbb{R}$ is **twice differentiable** if the Hessian matrix $\nabla^2 f(\boldsymbol{x})$ exists for every $\boldsymbol{x} \in \operatorname{dom} f$.

If f is twice differentiable, then it is convex if and only if

$$\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$$

for all $\boldsymbol{x} \in \operatorname{dom} f$.

Note that for a one-dimensional function $f : \mathbb{R} \to \mathbb{R}$, the above condition just reduces to $f''(x) \geq 0$. You can prove the one-dimensional version relatively easy (although we will not do so here) using the first-order characterization of convexity described above and the definition of the second derivative. You can then prove the general case by considering the function $g(t) = f(\boldsymbol{x} + t\boldsymbol{v})$. To see how, note that if f is convex and twice differentiable, then so is g. Using the chain rule, we have

$$g''(t) = \boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{x} + t\boldsymbol{v})\boldsymbol{v}.$$

Since g is convex, the one-dimensional result above tells us that $g''(0) \geq 0$, and hence $\boldsymbol{v}^{\mathrm{T}} \nabla^2 f(\boldsymbol{x})\boldsymbol{v} \geq 0$. Since this has to hold for any $\boldsymbol{v}$, this means that $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$. The proof that $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$ implies convexity follows a similar strategy.

Examples

- **Quadratic functionals:** The function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}^{\mathrm{T}}\boldsymbol{x} + r,$$

where $\boldsymbol{P}$ is symmetric, has $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{P}$, so $f(\boldsymbol{x})$ is convex if and only if $\boldsymbol{P} \succeq \mathbf{0}$.

- **Least-squares:** The least squares objective function

  $$f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2,$$

  where $\boldsymbol{A}$ is an arbitrary $M \times N$ matrix, has $\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, so $f(\boldsymbol{x})$ is convex for any $\boldsymbol{A}$.

Strict convexity

It is relatively straightforward to show that for f differentiable, *strict* convexity is equivalent to (2) holding with a strict inequality. It is also easy to show that if $\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0}$, then f is strictly convex.

However, it is *not* the case that f being strictly convex implies $\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0}$ for all $\boldsymbol{x}$. As an example, consider the function $f(x) = x^4$. This function *is* strictly convex, but also has $f''(0) = 0$.

Why convexity?

Convex functions satisfy a number of properties that are desirable in the context of optimization. Here we will first discuss two fundamental facts.

Recall the unconstrained optimization problem:

$$\underset{\boldsymbol{x} \in \mathbb{R}^N}{\text{minimize}} \;\; f(\boldsymbol{x}). \tag{3}$$

Below we will first show that for any convex f, if $\boldsymbol{x}^\star$ is a local minimizer of (3), 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 (3) 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 is also true for non-differentiable (but still convex) f. We will explore this later in the course.

Local minima are also global minima

The most important property of convex functions from an optimization perspective is that any local minimum is also a global minimum, or more formally:

Let $f(\boldsymbol{x})$ be a convex function on $\mathbb{R}^N$, and suppose that $\boldsymbol{x}^\star$ 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 $\boldsymbol{x}^\star$ is a local minimum. We want to show that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}')$ for any $\boldsymbol{x}'$. We already have that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}')$ if $\|\boldsymbol{x}' - \boldsymbol{x}^\star\|_2 \leq \epsilon$, so all we need to do is show that this also holds for $\boldsymbol{x}'$ with $\|\boldsymbol{x}' - \boldsymbol{x}^\star\|_2 > \epsilon$. Note that from convexity, we have

$$f(\theta\boldsymbol{x}' + (1-\theta)\boldsymbol{x}^\star) \leq \theta f(\boldsymbol{x}') + (1-\theta) f(\boldsymbol{x}^\star)$$

for any $\theta \in [0, 1]$. This has to hold for any $\theta \in [0, 1]$, and in particular, it must hold for $\theta = \epsilon / \|\boldsymbol{x}' - \boldsymbol{x}^\star\|_2$ (which is less than 1 since

$\|\boldsymbol{x}' - \boldsymbol{x}^\star\|_2 > \epsilon$). For this choice of θ we have

$$\|\theta\boldsymbol{x}' + (1-\theta)\boldsymbol{x}^\star - \boldsymbol{x}^\star\|_2 = \theta\|\boldsymbol{x}' - \boldsymbol{x}^\star\|_2 = \epsilon,$$

thus $\theta\boldsymbol{x}' + (1-\theta)\boldsymbol{x}^\star$ lives in the neighborhood where $\boldsymbol{x}^\star$ is a local minimum, and hence

$$f(\boldsymbol{x}^\star) \le f(\theta\boldsymbol{x}' + (1-\theta)\boldsymbol{x}^\star).$$

Combining this with the inequality above we have

$$f(\boldsymbol{x}^\star) \le \theta f(\boldsymbol{x}') + (1-\theta) f(\boldsymbol{x}^\star).$$

Rearranging this gives us $\theta f(\boldsymbol{x}^\star) \le \theta f(\boldsymbol{x}')$, or simply $f(\boldsymbol{x}^\star) \le f(\boldsymbol{x}')$, which is exactly what we wanted to prove.

Note that for functions f that are *not* convex, any number of things are possible. It *might* be the case that there is only one local minimum and that it corresponds to the global minimum. We are typically not so lucky, though. There may be many local minima, some of which may be very far from actually minimizing f.

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 differentiable functions

We have just shown that if we want to find a global minimum of a convex function, it is sufficient to find any local minimum. This raises the question: How do we know when we have found a minimum of a function (local or global)? Here we provide an answer to this question in the special case where f is differentiable.

Let f be convex and differentiable 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 $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$.

To prove this, we first assume that $\boldsymbol{x}^\star$ is a local minimum of f and show that this implies that $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$. This follows almost immediately. If $\boldsymbol{x}^\star$ is a local minimum of f, then this means that *every* direction must be an ascent direction, i.e., $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x})\rangle \geq 0$ for all $\boldsymbol{d} \in \mathbb{R}^N$. However, the only way we can make $\langle \boldsymbol{d}, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0$ for all $\boldsymbol{d}$ is if $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$. Thus, for differentiable f

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

Note that this fact does not actually require f to be convex.

Now we will show that for convex f we also have that $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$ implies that f is a minimizer. Specifically, it is a direct consequence of our first order characterization of convexity in (2) that

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

for all choices of $\boldsymbol{u} \in \mathbb{R}^N$. This now makes it clear that for convex f

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

This fact will lie at the heart of the algorithms for unconstrained convex optimization that we will begin discussing next time – if we can find an $\boldsymbol{x}$ that makes the gradient vanish, then we have solved the problem.

Existence and uniqueness

We close this discussion on a couple of minor technical notes. First, it is important to realize that it is not always the case that a convex function will actually have a minimizer. That is, there may be sometimes be no $\boldsymbol{x}^\star$ such that $f(\boldsymbol{x}^\star) \le f(\boldsymbol{x})$ 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 convex (and differentiable). We will not worry much about this in this course, but it is worth realizing that one can encounter a convex optimization problem for which no solution exists.

Moreover, even when a minimizer does exist, that does not always guarantee that it is *unique*. That is, there might be multiple distinct $\boldsymbol{x}$ that achieve the minimum value of f. However, there are certainly lots of scenarios where there is only one unique minimizer. One prominent example is when f is *strictly* convex.

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

This is 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 $0 < \theta < 1$,

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

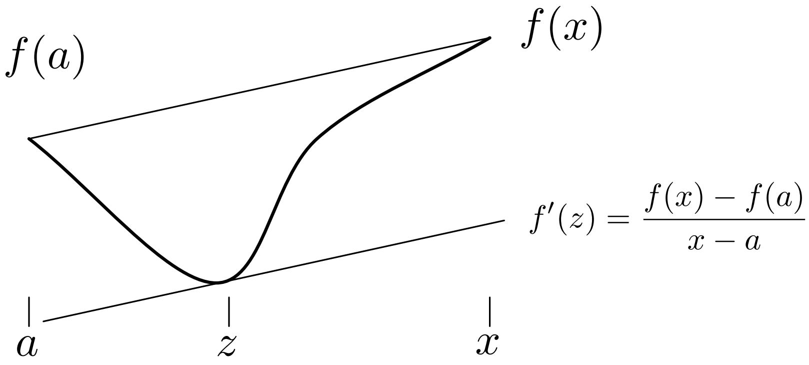
This would contradict the assertion that $\boldsymbol{x}^\star$ is a global minimizer, and hence no such $\widehat{\boldsymbol{x}}$ can exist.

Technical Details: Taylor's Theorem

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 $\boldsymbol{a}$,

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

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

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

We will use these two particular multidimensional results in this course, referring to them generically as “Taylor’s theorem”.