Convex functions

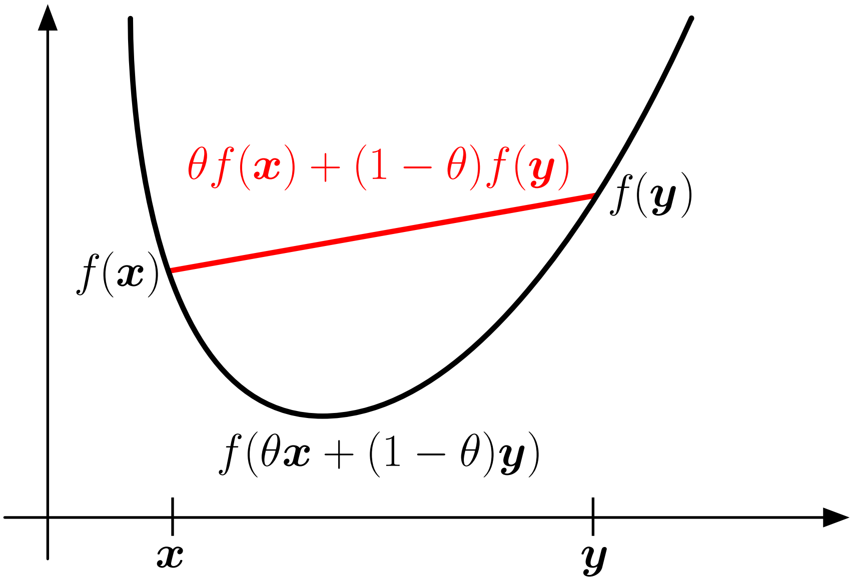
We have seen what it means for a set to be convex. In this set of notes, we start working towards what it means to be a convex *function*.

To define this concept rigorously, we must be specific about the subset of $\mathbb{R}^N$ where a function can be applied. Specifically, the **domain** $\operatorname{dom} f$ of a function $f : \mathbb{R}^N \to \mathbb{R}^M$ is the subset of $\mathbb{R}^N$ where f is well-defined. We then say that a function f is **convex** if $\operatorname{dom} f$ is a convex set, and

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \;\leq\; \theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$ and $0 \leq \theta \leq 1$.

This inequality is easier to interpret with a picture. The left-hand side of the inequality above is simply the function f evaluated along a line segment between $\boldsymbol{x}$ and $\boldsymbol{y}$. The right-hand side represents a straight line segment between $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$ as we move along this line segment, which for a convex function must lie above f.


$\theta f(\boldsymbol{x}) + (1-\theta) f(\boldsymbol{y})$
$f(\boldsymbol{y})$
$f(\boldsymbol{x})$
$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y})$
$\boldsymbol{x}$
$\boldsymbol{y}$

We say that f is **strictly convex** if $\operatorname{dom} f$ is convex and

$$f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \;<\; \theta f(\boldsymbol{x}) + (1-\theta)f(\boldsymbol{y})$$

for all $\boldsymbol{x} \neq \boldsymbol{y} \in \operatorname{dom} f$ and $0 < \theta < 1$.

Note also that we say that a function is f is **concave** if $-f$ is convex, and similarly for strictly concave functions. We are mostly interested in convex functions, but this is only because we are mostly restricting our attention to *minimization* problems. We justified this because any maximization problem can be converted to a minimization one by multiplying the objective function by -1. Everything that we say about minimizing convex functions also applies maximizing concave ones.

We make a special note here that *affine* functions of the form

$$f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{a} \rangle + b,$$

are both convex and concave (but neither strictly convex nor strictly concave). This is the only kind of function that has this property. (Why?)

Note that in the definition above, the domain matters. For example,

$$f(x) = x^3$$

is convex if $\operatorname{dom} f = \mathbb{R}_+ = [0, \infty]$ but not if $\operatorname{dom} f = \mathbb{R}$.

It will also sometimes be useful to consider the **extension** of f from $\operatorname{dom} f$ to all of $\mathbb{R}^N$, defined as

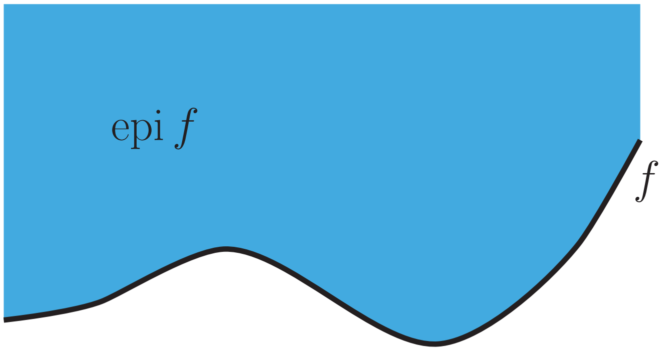
$$\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \operatorname{dom} f, \quad \tilde{f}(\boldsymbol{x}) = +\infty, \quad \boldsymbol{x} \notin \operatorname{dom} f.$$

If f is convex on $\operatorname{dom} f$, then its extension is also convex on $\mathbb{R}^N$.

The epigraph

A useful notion that we will encounter later in the course is that of the **epigraph** of a function. The epigraph of a function $f : \mathbb{R}^N \to \mathbb{R}$ is the subset of $\mathbb{R}^{N+1}$ created by filling in the space above f:

$$\operatorname{epi} f = \left\{ \begin{bmatrix} \boldsymbol{x} \\ t \end{bmatrix} \in \mathbb{R}^{N+1} \ : \ \boldsymbol{x} \in \operatorname{dom} f, \quad f(\boldsymbol{x}) \leq t \right\}.$$



It is not hard to show that f is convex if and only if epi f is a convex set. This connection should help to illustrate how even though the definitions of a convex set and convex function might initially appear quite different, they actually follow quite naturally from each other.

Examples of convex functions

Here are some standard examples for functions on $\mathbb{R}$:

- $f(x) = x^2$ is (strictly) convex.
- affine functions $f(x) = ax + b$ are both convex and concave for $a, b \in \mathbb{R}$.
- exponentials $f(x) = e^{ax}$ are convex for all $a \in \mathbb{R}$.

- powers x^α are:
  - convex on $\mathbb{R}_+$ for $\alpha \geq 1$,
  - concave for $0 \leq \alpha \leq 1$,
  - convex for $\alpha \leq 0$.
- $|x|^\alpha$ is convex on all of $\mathbb{R}$ for $\alpha \geq 1$.
- logarithms: $\log x$ is concave on $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$.
- the entropy function $-x \log x$ is concave on $\mathbb{R}_{++}$.

Here are some standard examples for functions on $\mathbb{R}^N$:

- affine functions $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{a} \rangle + b$ are both convex and concave on all of $\mathbb{R}^N$.
- any valid norm $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is convex on all of $\mathbb{R}^N$.
- if $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are both convex, then the sum $f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ is also convex.

A useful tool for showing that a function $f : \mathbb{R}^N \to \mathbb{R}$ is convex is the fact that f is convex if and only if the function $g_{\boldsymbol{v}} : \mathbb{R} \to \mathbb{R}$,

$$g_{\boldsymbol{v}}(t) = f(\boldsymbol{x} + t\boldsymbol{v}), \quad \operatorname{dom} g = \{t \ : \ \boldsymbol{x} + t\boldsymbol{v} \in \operatorname{dom} f\}$$

is convex for every $\boldsymbol{x} \in \operatorname{dom} f$, $\boldsymbol{v} \in \mathbb{R}^N$.

Example:

Let $f(\boldsymbol{X}) = -\log\det \boldsymbol{X}$ with $\operatorname{dom} f = \mathcal{S}^N_{++}$, where $\mathcal{S}^N_{++}$ denotes the set of symmetric and (strictly) positive definite matrices. For any $\boldsymbol{X} \in \mathcal{S}^N_{++}$, we know that

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}},$$

for some diagonal, positive $\boldsymbol{\Lambda}$, so we can define

$$\boldsymbol{X}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\mathrm{T}}, \quad \text{and} \quad \boldsymbol{X}^{-1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^{\mathrm{T}}.$$

Now consider any symmetric matrix $\boldsymbol{V}$ and t such that $\boldsymbol{X} + t\boldsymbol{V} \in \mathcal{S}^N_{++}$:

$$\begin{aligned} g_{\mathbf{V}}(t) &= -\log\det(\boldsymbol{X} + t\boldsymbol{V}) \\ &= -\log\det(\boldsymbol{X}^{1/2}(\mathbf{I} + t\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2})\boldsymbol{X}^{1/2}) \\ &= -\log\det\boldsymbol{X} - \log\det(\mathbf{I} + t\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2}) \\ &= -\log\det\boldsymbol{X} - \sum_{n=1}^{N}\log(1 + \sigma_i t), \end{aligned}$$

where the σ_i are the eigenvalues of $\boldsymbol{X}^{-1/2}\boldsymbol{V}\boldsymbol{X}^{-1/2}$. The function $-\log(1 + \sigma_i t)$ is convex, so the above is a sum of convex functions, which is convex.

Operations that preserve convexity

There are a number of useful operations that we can perform on a convex function while preserving convexity. Some examples include:

- **Positive weighted sum:** A **positive** weighted sum of convex functions is also convex, i.e., if $f_1, \ldots, f_m$ are convex and $w_1, \ldots, w_m \geq 0$, then $w_1 f_1 + \ldots + w_m f_m$ is also convex.
- **Composition with an affine function:** If $f : \mathbb{R}^N \to \mathbb{R}$ is convex, then $g : \mathbb{R}^D \to \mathbb{R}$ defined by

  $$g(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}),$$

  where $\boldsymbol{A} \in \mathbb{R}^{N \times D}$ and $b \in \mathbb{R}^N$, is convex.

- **Composition with scalar functions:** Consider the function $f(\boldsymbol{x}) = h(g(\boldsymbol{x}))$, where $g : \mathbb{R}^N \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$.
  - f is convex if g is convex and h is convex and non-decreasing. Example: $e^{g(\boldsymbol{x})}$ is convex if g is convex.
  - f is convex if g is concave and h is convex and non-increasing.
    Example: $\frac{1}{g(\boldsymbol{x})}$ is convex if g is concave and positive.
- **Max of convex functions:** If f_1 and f_2 are convex, then $f(\boldsymbol{x}) = \max\left(f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\right)$ is convex.

First-order conditions for convexity

We say that f is **differentiable** if $\operatorname{dom} f$ is an open set (all of $\mathbb{R}^N$, for example), and the gradient

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

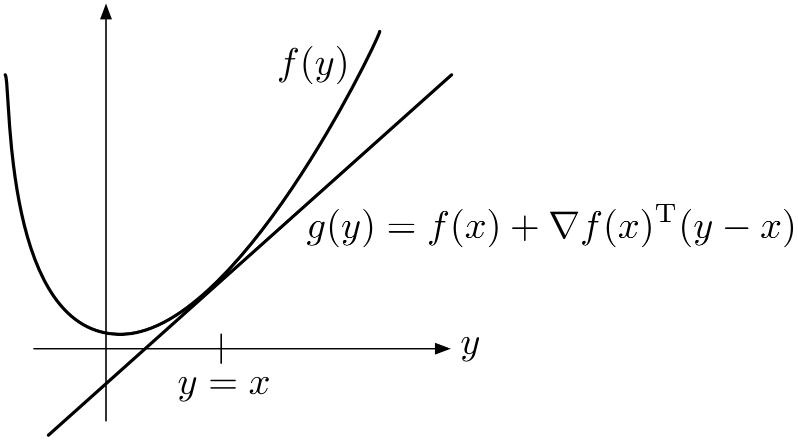
exists for each $\boldsymbol{x} \in \operatorname{dom} f$. The gradient of a function is a core concept in optimization and as such we review a little bit of what it means at the end of these notes.

The following characterization of convexity 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.

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

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}) \tag{1}$$

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



This means that the linear approximation

$$g(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{x}),$$

is a **global underestimator** of $f(\boldsymbol{y})$.

It is easy to show that f convex, differentiable $\Rightarrow$ (1). Since f is convex,

$$f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})) \;\leq\; (1-t)f(\boldsymbol{x}) + tf(\boldsymbol{y}), \quad 0 \leq t \leq 1,$$

and so

$$f(\boldsymbol{y}) \;\geq\; f(\boldsymbol{x}) + \frac{f(\boldsymbol{x}+t(\boldsymbol{y}-\boldsymbol{x})) - f(\boldsymbol{x})}{t}, \quad \forall 0 < t \leq 1.$$

Taking the limit as $t \to 0$ on the right yields (1).

It is also true that (1) $\Rightarrow f$ convex. To see this, choose arbitrary $\boldsymbol{x}, \boldsymbol{y}$ and set $\boldsymbol{z}_\theta = (1-\theta)\boldsymbol{x} + \theta\boldsymbol{y}$; then (1) tells us

$$f(\boldsymbol{w}) \ \geq \ f(\boldsymbol{z}_\theta) + \nabla f(\boldsymbol{z}_\theta)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{z}_\theta).$$

Applying this at $\boldsymbol{w} = \boldsymbol{y}$ and multiplying by θ, then applying it at $\boldsymbol{w} = \boldsymbol{x}$ and multiplying by $(1-\theta)$ yields

$$\begin{gathered}\theta f(\boldsymbol{y}) \geq \theta f(\boldsymbol{z}_\theta) + \theta \nabla f(\boldsymbol{z}_\theta)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{z}_\theta), \\ (1-\theta) f(\boldsymbol{x}) \geq (1-\theta) f(\boldsymbol{z}_\theta) + (1-\theta)\nabla f(\boldsymbol{z}_\theta)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{z}_\theta).\end{gathered}$$

Adding these inequalities together establishes the result.

Second-order conditions for convexity

We say that $f : \mathbb{R}^N \to \mathbb{R}$ is twice differentiable if dom f is an open set, and the $N \times N$ Hessian matrix

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

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

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

$$\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0} \quad (\text{i.e. } \nabla^2 f(\boldsymbol{x}) \in \mathcal{S}_+^N).$$

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 $\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$. The proof that $\nabla^2 f(\boldsymbol{x}) \succeq \boldsymbol{0}$ implies convexity follows a similar strategy.

In addition, if

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad (\text{i.e. } \nabla^2 f(\boldsymbol{x}) \in S^N_{++}). \quad \text{for all } \boldsymbol{x} \in \operatorname{dom} f,$$

then f is strictly convex. The converse is not quite true; it is possible that f is strictly convex even if $\nabla f(\boldsymbol{x})$ has eigenvalues that are zero at isolated points. For example, $f(x) = |x|^3$ is strictly convex but $f''(0) = 0$.

Standard examples (from [BV04])

Quadratic functionals:

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

where $\boldsymbol{P}$ is symmetric, has

$$\nabla f(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{x} + \boldsymbol{q}, \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{P},$$

so $f(\boldsymbol{x})$ is convex iff $\boldsymbol{P} \succeq \boldsymbol{0}$.

Least-squares:

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

where $\boldsymbol{A}$ is an arbitrary $M \times N$ matrix, has

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \quad \nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A},$$

and is convex for any $\boldsymbol{A}$.

Quadratic-over-linear:
In $\mathbb{R}^2$, if

$$f(\boldsymbol{x}) = x_1^2/x_2,$$

then

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 2x_1/x_2 \\ -x_1^2/x_2^2 \end{bmatrix}, \quad \nabla^2 f(\boldsymbol{x}) = \frac{2}{x_2^3}\begin{bmatrix} x_2^2 & -x_1x_2 \\ -x_1x_2 & x_1 \end{bmatrix}$$
$$= \frac{2}{x_2^3}\begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}\begin{bmatrix} x_2 & -x_1 \end{bmatrix},$$

and so f is convex on $\mathbb{R} \times [0, \infty]$ ($x_1 \in \mathbb{R}, x_2 \geq 0$).

Strong convexity and smoothness

We say that a function f is **strongly convex** if there is a $\mu > 0$ such that

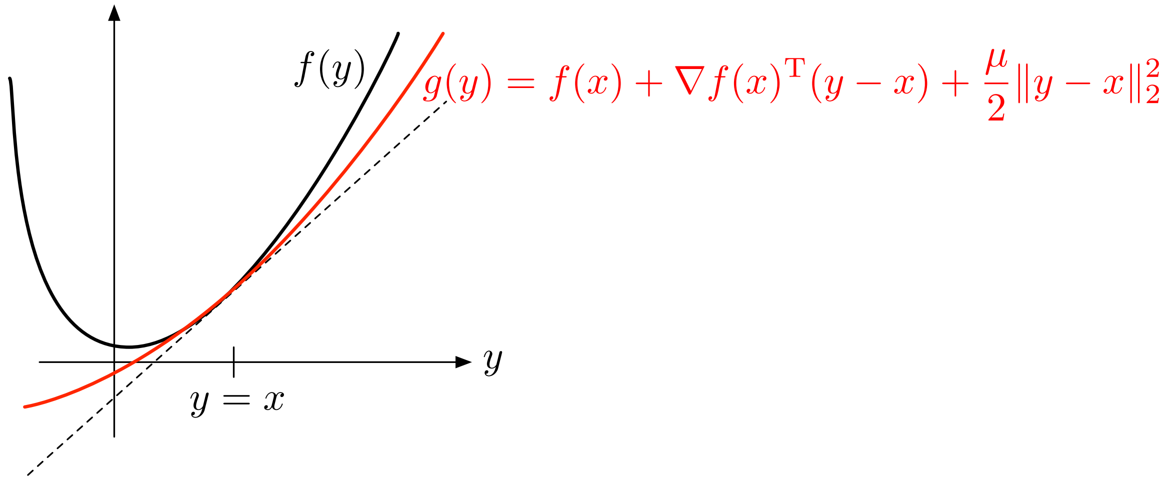
$$f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 \quad \text{is convex.} \tag{2}$$

We call μ the *strong convexity parameter* and will sometimes say that f is μ-strongly convex. In a sense, what we are saying is that f is so convex that we can subtract off a quadratic function and still preserve convexity.

If f is differentiable, there is another interpretation of strong convexity. We have seen that an equivalent definition of regular convexity is that the linear approximation formed using the gradient at a point $\boldsymbol{x}$ is a global underestimator of the function — see (1) and the picture below. If f obeys (2), then we can form a *quadratic* global underestimator as

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{3}$$

Here is a picture



$f(y)$
$g(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{\mu}{2}\|y - x\|_2^2$
y
$y = x$

We will show that (2) implies (3) in a future homework.

If f is twice differentiable, there is yet another interpretation of strong convexity. If f obeys (2) then we know that the Hessian of $f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2$ does not have any negative eigenvalues, i.e.

$$\nabla^2 \left(f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|_2^2 \right) \succeq \mathbf{0}.$$

Thus (since $\nabla^2(\|\boldsymbol{x}\|_2^2) = 2\mathbf{I}$),

$$\begin{gathered}\nabla^2 f(\boldsymbol{x}) - \mu\mathbf{I} \succeq \mathbf{0}, \\ \Downarrow \\ \nabla^2 f(\boldsymbol{x}) \succeq \mu\mathbf{I}.\end{gathered}$$

This is just a fancy way of saying that the smallest eigenvalue of the Hessian $\nabla^2 f(\boldsymbol{x})$ is uniformly bounded below by μ for all $\boldsymbol{x}$.

In addition to convexity, there is one more type of structure that we consider for functions $f : \mathbb{R}^N \to \mathbb{R}$. We say that differentiable f has a **Lipschitz gradient** if there is a L such that

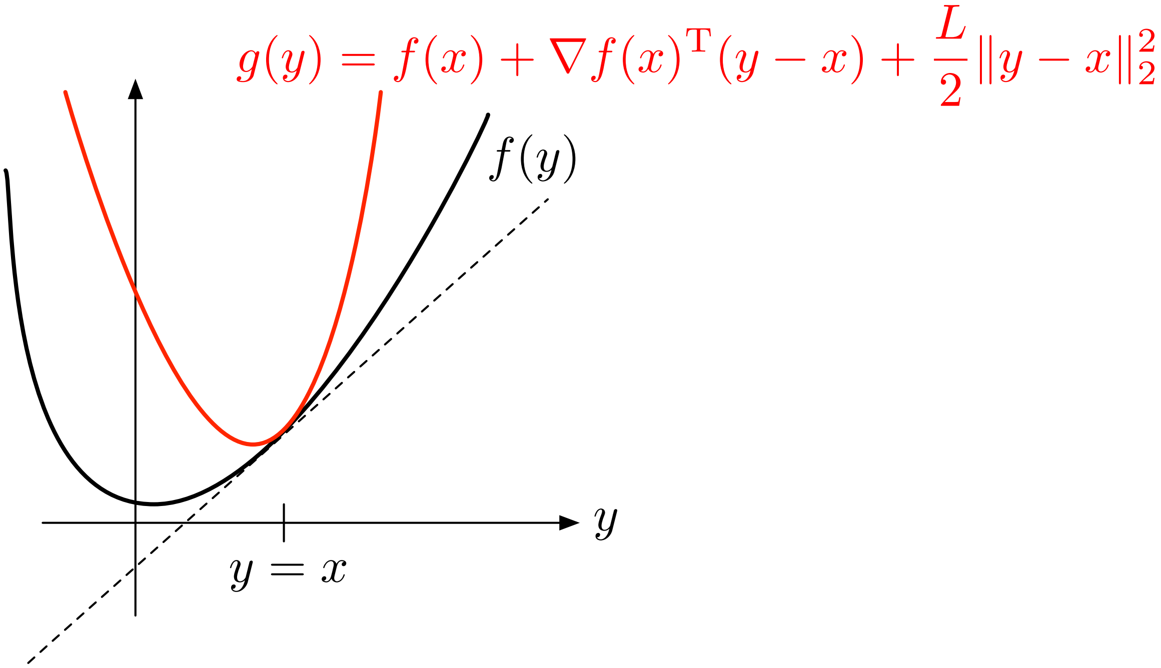
$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \ \leq \ L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \text{for all } \ \boldsymbol{x}, \boldsymbol{y}. \tag{4}$$

This means that the gradient $\nabla f(\boldsymbol{x})$ does not change radically as we change $\boldsymbol{x}$. Functions f that obey (4) are also referred to as ***L*-smooth**. This definition applies whether or not the function f is convex.

Whether or not f is convex, if it is L-smooth, it there is a natural quadtratic *over*estimator. Around any point $\boldsymbol{x}$, we have the upper bound

$$f(\boldsymbol{y}) \ \leq \ f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{5}$$

Here is a picture


$g(y) = f(x) + \nabla f(x)^{\mathrm{T}}(y - x) + \frac{L}{2}\|y - x\|_2^2$
$f(y)$
y
$y = x$

We will show that (4) implies (5) in a future homework.

If f is twice differentiable, then there is another way to interpret L-smoothness. If f obeys (4), then we have a uniform upper bound on the *largest* eigenvalue of the Hessian at every point:

$$\nabla^2 f(\boldsymbol{x}) \preceq L\mathbf{I}, \quad \text{for all} \;\; \boldsymbol{x}. \tag{6}$$

This makes intuitive sense, as (4) tells us that the first derivative cannot change too quickly, so there must be some kind of bound on the second derivative. We will establish that (4) implies (6) (again, regardless of whether f is convex) in a future homework.

Review: The gradient

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. When referring to this vector-valued function, we sometimes use the words **gradient map**, but sometimes we will overload the term "gradient"; we will use the notation $\nabla f(\boldsymbol{x})$ to refer to the vector given by the gradient map evaluated 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.

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.,

$$\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})) \rangle = \lim_{\delta \to 0} \frac{f(\boldsymbol{x} + \delta \boldsymbol{u}) - f(\boldsymbol{x})}{\delta}. \tag{7}$$

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

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 $\boldsymbol{u} \to \mathbf{0}$.

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 (7) 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 $\boldsymbol{x}$, meaning we can find a $t > 0$ small enough so that

$$f(\boldsymbol{x} + t\boldsymbol{u}) < f(\boldsymbol{x}) \tag{8}$$

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

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

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.

To show that $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle < 0$ implies (8), we again use the Taylor theorem to get

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

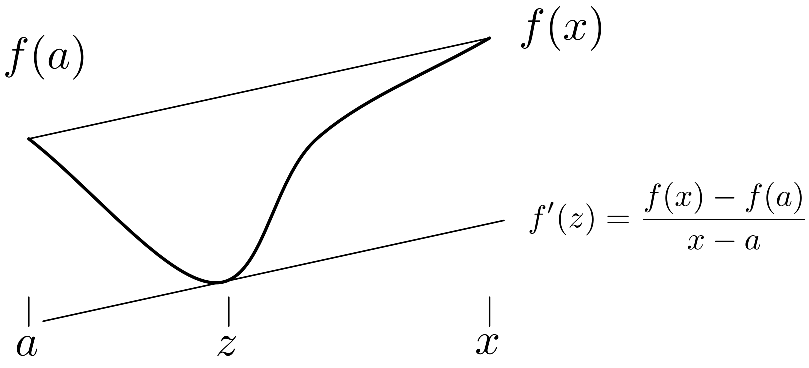
where now we have $h(t\boldsymbol{u}) \to 0$ as $t \to 0$. For $t > 0$ small enough, we can make $|h(t\boldsymbol{u})| \cdot \|\boldsymbol{u}\|_2 < |\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle|$, and so the term inside the parentheses above is negative if $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle$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

Technical Details: Taylor's Theorem

You might recall the mean-value theorem from your first calculus class. If $f : \mathbb{R} \rightarrow \mathbb{R}$ 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”.

References

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