4 Calculus and Optimization

Much of machine learning is about minimizing a **cost function** (also called an **objective function** in the optimization community), which is a scalar function of several variables that typically measures how poorly our model fits the data we have.

4.1 Extrema

Optimization is about finding **extrema**, which depending on the application could be minima or maxima. When defining extrema, it is necessary to consider the set of inputs over which we're optimizing. This set $\mathcal{X} \subseteq \mathbb{R}^d$ is called the **feasible set**. If $\mathcal{X}$ is the entire domain of the function being optimized (as it often will be for our purposes), we say that the problem is **unconstrained**. Otherwise the problem is **constrained** and may be much harder to solve, depending on the nature of the feasible set.

Suppose $f : \mathbb{R}^d \to \mathbb{R}$. A point $\mathbf{x}$ is said to be a **local minimum** (resp. **local maximum**) of f in $\mathcal{X}$ if $f(\mathbf{x}) \leq f(\mathbf{y})$ (resp. $f(\mathbf{x}) \geq f(\mathbf{y})$) for all $\mathbf{y}$ in some neighborhood $N \subseteq \mathcal{X}$ about $\mathbf{x}$.[8] Furthermore, if $f(\mathbf{x}) \leq f(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{X}$, then $\mathbf{x}$ is a **global minimum** of f in $\mathcal{X}$ (similarly for global maximum). If the phrase "in $\mathcal{X}$" is unclear from context, assume we are optimizing over the whole domain of the function.

The qualifier **strict** (as in e.g. a strict local minimum) means that the inequality sign in the definition is actually a $>$ or $<$, with equality not allowed. This indicates that the extremum is unique within some neighborhood.

Observe that maximizing a function f is equivalent to minimizing $-f$, so optimization problems are typically phrased in terms of minimization without loss of generality. This convention (which we follow here) eliminates the need to discuss minimization and maximization separately.

4.2 Gradients

The single most important concept from calculus in the context of machine learning is the **gradient**. Gradients generalize derivatives to scalar functions of several variables. The gradient of $f : \mathbb{R}^d \to \mathbb{R}$, denoted ∇f, is given by

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \qquad \text{i.e.} \qquad [\nabla f]_i = \frac{\partial f}{\partial x_i}$$

Gradients have the following very important property: $\nabla f(\mathbf{x})$ points in the direction of **steepest ascent** from $\mathbf{x}$. Similarly, $-\nabla f(\mathbf{x})$ points in the direction of **steepest descent** from $\mathbf{x}$. We will use this fact frequently when iteratively minimizing a function via **gradient descent**.

4.3 The Jacobian

The **Jacobian** of $f : \mathbb{R}^n \to \mathbb{R}^m$ is a matrix of first-order partial derivatives:

$$\mathbf{J}_f = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \qquad \text{i.e.} \qquad [\mathbf{J}_f]_{ij} = \frac{\partial f_i}{\partial x_j}$$

[8] A **neighborhood** about $\mathbf{x}$ is an open set which contains $\mathbf{x}$.

Note the special case $m = 1$, where $\nabla f = \mathbf{J}_f^\top$.

4.4 The Hessian

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

Recall that if the partial derivatives are continuous, the order of differentiation can be interchanged (Clairaut's theorem), so the Hessian matrix will be symmetric. This will typically be the case for differentiable functions that we work with.

The Hessian is used in some optimization algorithms such as Newton's method. It is expensive to calculate but can drastically reduce the number of iterations needed to converge to a local minimum by providing information about the curvature of f.

4.5 Matrix calculus

Since a lot of optimization reduces to finding points where the gradient vanishes, it is useful to have differentiation rules for matrix and vector expressions. We give some common rules here. Probably the two most important for our purposes are

$$\nabla_{\mathbf{x}}(\mathbf{a}^\top \mathbf{x}) = \mathbf{a}$$
$$\nabla_{\mathbf{x}}(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}$$

Note that this second rule is defined only if $\mathbf{A}$ is square. Furthermore, if $\mathbf{A}$ is symmetric, we can simplify the result to $2\mathbf{A}\mathbf{x}$.

4.5.1 The chain rule

Most functions that we wish to optimize are not completely arbitrary functions, but rather are composed of simpler functions which we know how to handle. The chain rule gives us a way to calculate derivatives for a composite function in terms of the derivatives of the simpler functions that make it up.

The chain rule from single-variable calculus should be familiar:

$$(f \circ g)'(x) = f'(g(x))g'(x)$$

where $\circ$ denotes function composition. There is a natural generalization of this rule to multivariate functions.

Proposition 12. *Suppose $f : \mathbb{R}^m \to \mathbb{R}^k$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Then $f \circ g : \mathbb{R}^n \to \mathbb{R}^k$ and*

$$\mathbf{J}_{f \circ g}(\mathbf{x}) = \mathbf{J}_f(g(\mathbf{x}))\mathbf{J}_g(\mathbf{x})$$

In the special case $k = 1$ we have the following corollary since $\nabla f = \mathbf{J}_f^\top$.

Corollary 1. *Suppose $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$. Then $f \circ g : \mathbb{R}^n \to \mathbb{R}$ and*

$$\nabla(f \circ g)(\mathbf{x}) = \mathbf{J}_g(\mathbf{x})^\top \nabla f(g(\mathbf{x}))$$

4.6 Taylor's theorem

Taylor's theorem has natural generalizations to functions of more than one variable. We give the version presented in [1].

Theorem 6. *(Taylor's theorem) Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and let $\mathbf{h} \in \mathbb{R}^d$. Then there exists $t \in (0, 1)$ such that*

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x} + t\mathbf{h})^\top \mathbf{h}$$

Furthermore, if f is twice continuously differentiable, then

$$\nabla f(\mathbf{x} + \mathbf{h}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h} \, \mathrm{d}t$$

and there exists $t \in (0, 1)$ such that

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

This theorem is used in proofs about conditions for local minima of unconstrained optimization problems. Some of the most important results are given in the next section.

4.7 Conditions for local minima

Proposition 13. *If $\mathbf{x}^*$ is a local minimum of f and f is continuously differentiable in a neighborhood of $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.*

Proof. Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Let $\mathbf{h} = -\nabla f(\mathbf{x}^*)$, noting that by the continuity of ∇f we have

$$\lim_{t \to 0} -\nabla f(\mathbf{x}^* + t\mathbf{h}) = -\nabla f(\mathbf{x}^*) = \mathbf{h}$$

Hence

$$\lim_{t \to 0} \mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) = \mathbf{h}^\top \nabla f(\mathbf{x}^*) = -\|\mathbf{h}\|_2^2 < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) < 0$ for all $t \in [0, T]$. Now we apply Taylor's theorem: for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + t\mathbf{h}^\top \nabla f(\mathbf{x}^* + t'\mathbf{h}) < f(\mathbf{x}^*)$$

whence it follows that $\mathbf{x}^*$ is not a local minimum, a contradiction. Hence $\nabla f(\mathbf{x}^*) = \mathbf{0}$. □

The proof shows us why the vanishing gradient is necessary for an extremum: if $\nabla f(\mathbf{x})$ is nonzero, there always exists a sufficiently small step $\alpha > 0$ such that $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}))) < f(\mathbf{x})$. For this reason, $-\nabla f(\mathbf{x})$ is called a **descent direction**.

Points where the gradient vanishes are called **stationary points**. Note that not all stationary points are extrema. Consider $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 - y^2$. We have $\nabla f(\mathbf{0}) = \mathbf{0}$, but the point $\mathbf{0}$ is the minimum along the line $y = 0$ and the maximum along the line $x = 0$. Thus it is neither a local minimum nor a local maximum of f. Points such as these, where the gradient vanishes but there is no local extremum, are called **saddle points**.

We have seen that first-order information (i.e. the gradient) is insufficient to characterize local minima. But we can say more with second-order information (i.e. the Hessian). First we prove a necessary second-order condition for local minima.

Proposition 14. *If $\mathbf{x}^*$ is a local minimum of f and f is twice continuously differentiable in a neighborhood of $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite.*

Proof. Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla^2 f(\mathbf{x}^*)$ is not positive semi-definite. Let $\mathbf{h}$ be such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$, noting that by the continuity of $\nabla^2 f$ we have

$$\lim_{t\to 0} \nabla^2 f(\mathbf{x}^* + t\mathbf{h}) = \nabla^2 f(\mathbf{x}^*)$$

Hence

$$\lim_{t\to 0} \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} = \mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} < 0$ for all $t \in [0, T]$. Now we apply Taylor's theorem: for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + \underbrace{t\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}t^2 \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t'\mathbf{h})\mathbf{h} < f(\mathbf{x}^*)$$

where the middle term vanishes because $\nabla f(\mathbf{x}^*) = \mathbf{0}$ by the previous result. It follows that $\mathbf{x}^*$ is not a local minimum, a contradiction. Hence $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite. □

Now we give sufficient conditions for local minima.

Proposition 15. *Suppose f is twice continuously differentiable with $\nabla^2 f$ positive semi-definite in a neighborhood of $\mathbf{x}^*$, and that $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Then $\mathbf{x}^*$ is a local minimum of f. Furthermore if $\nabla^2 f(\mathbf{x}^*)$ is positive definite, then $\mathbf{x}^*$ is a strict local minimum.*

Proof. Let B be an open ball of radius $r > 0$ centered at $\mathbf{x}^*$ which is contained in the neighborhood. Applying Taylor's theorem, we have that for any $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, there exists $t \in (0, 1)$ such that

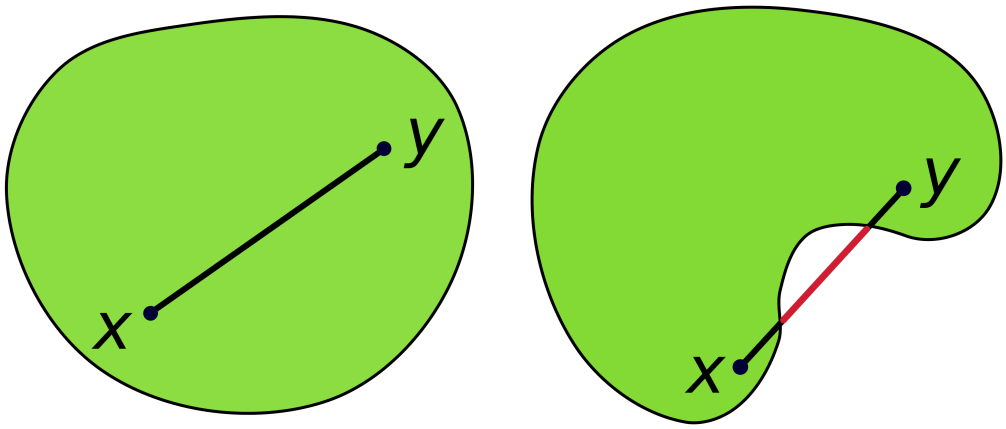
$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \underbrace{\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq f(\mathbf{x}^*)$$

The last inequality holds because $\nabla^2 f(\mathbf{x}^* + t\mathbf{h})$ is positive semi-definite (since $\|t\mathbf{h}\|_2 = t\|\mathbf{h}\|_2 < \|\mathbf{h}\|_2 < r$), so $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq 0$. Since $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \mathbf{h})$ for all $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, we conclude that $\mathbf{x}^*$ is a local minimum.

Now further suppose that $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite. Since the Hessian is continuous we can choose another ball B' with radius $r' > 0$ centered at $\mathbf{x}^*$ such that $\nabla^2 f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in B'$. Then following the same argument as above (except with a strict inequality now since the Hessian is positive definite) we have $f(\mathbf{x}^* + \mathbf{h}) > f(\mathbf{x}^*)$ for all $\mathbf{h}$ with $0 < \|\mathbf{h}\|_2 < r'$. Hence $\mathbf{x}^*$ is a strict local minimum. □

Note that, perhaps counterintuitively, the conditions $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ positive semi-definite are not enough to guarantee a local minimum at $\mathbf{x}^*$! Consider the function $f(x) = x^3$. We have $f'(0) = 0$ and $f''(0) = 0$ (so the Hessian, which in this case is the 1×1 matrix $\begin{bmatrix}0\end{bmatrix}$, is positive semi-definite). But f has a saddle point at $x = 0$. The function $f(x) = -x^4$ is an even worse offender – it has the same gradient and Hessian at $x = 0$, but $x = 0$ is a strict local maximum for this function!

For these reasons we require that the Hessian remains positive semi-definite as long as we are close to $\mathbf{x}^*$. Unfortunately, this condition is not practical to check computationally, but in some cases we can verify it analytically (usually by showing that $\nabla^2 f(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathbb{R}^d$). Also, if $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite, the continuity assumption on f implies this condition, so we don't have to worry.



(a) A convex set (b) A non-convex set

Figure 1: What convex sets look like

4.8 Convexity

Convexity is a term that pertains to both sets and functions. For functions, there are different degrees of convexity, and how convex a function is tells us a lot about its minima: do they exist, are they unique, how quickly can we find them using optimization algorithms, etc. In this section, we present basic results regarding convexity, strict convexity, and strong convexity.

4.8.1 Convex sets

A set $\mathcal{X} \subseteq \mathbb{R}^d$ is **convex** if

$$t\mathbf{x} + (1-t)\mathbf{y} \in \mathcal{X}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $t \in [0, 1]$.

Geometrically, this means that all the points on the line segment between any two points in $\mathcal{X}$ are also in $\mathcal{X}$. See Figure 1 for a visual.

Why do we care whether or not a set is convex? We will see later that the nature of minima can depend greatly on whether or not the feasible set is convex. Undesirable pathological results can occur when we allow the feasible set to be arbitrary, so for proofs we will need to assume that it is convex. Fortunately, we often want to minimize over all of $\mathbb{R}^d$, which is easily seen to be a convex set.

4.8.2 Basics of convex functions

In the remainder of this section, assume $f : \mathbb{R}^d \to \mathbb{R}$ unless otherwise noted. We'll start with the definitions and then give some results.

A function f is **convex** if

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and all $t \in [0, 1]$.

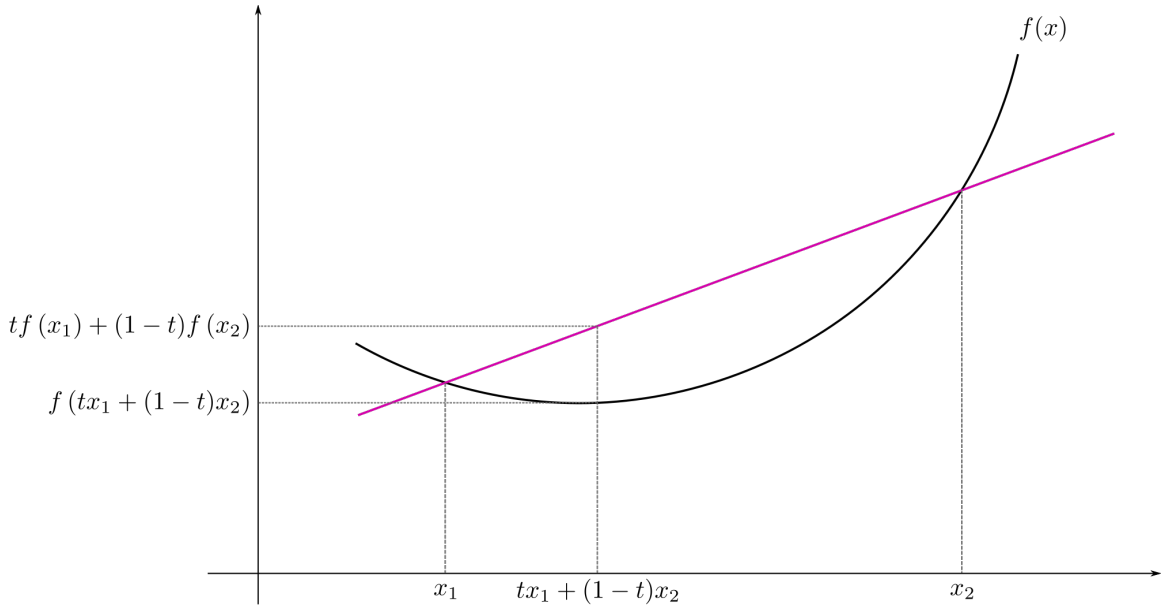


Figure 2: What convex functions look like

If the inequality holds strictly (i.e. $<$ rather than $\leq$) for all $t \in (0,1)$ and $\mathbf{x} \neq \mathbf{y}$, then we say that f is **strictly convex**.

A function f is **strongly convex with parameter** m (or m**-strongly convex**) if the function

$$\mathbf{x} \mapsto f(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

is convex.

These conditions are given in increasing order of strength; strong convexity implies strict convexity which implies convexity.

Geometrically, convexity means that the line segment between two points on the graph of f lies on or above the graph itself. See Figure 2 for a visual.

Strict convexity means that the line segment lies strictly above the graph of f, except at the segment endpoints. (So actually the function in the figure appears to be strictly convex.)

4.8.3 Consequences of convexity

Why do we care if a function is (strictly/strongly) convex?

Basically, our various notions of convexity have implications about the nature of minima. It should not be surprising that the stronger conditions tell us more about the minima.

Proposition 16. *Let $\mathcal{X}$ be a convex set. If f is convex, then any local minimum of f in $\mathcal{X}$ is also a global minimum.*

Proof. Suppose f is convex, and let $\mathbf{x}^*$ be a local minimum of f in $\mathcal{X}$. Then for some neighborhood $N \subseteq \mathcal{X}$ about $\mathbf{x}^*$, we have $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in N$. Suppose towards a contradiction that there exists $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $f(\tilde{\mathbf{x}}) < f(\mathbf{x}^*)$.

Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, noting that $\mathbf{x}(t) \in \mathcal{X}$ by the convexity of $\mathcal{X}$. Then by the convexity of f,

$$f(\mathbf{x}(t)) \leq tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) < tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

for all $t \in (0,1)$.

We can pick t to be sufficiently close to 1 that $\mathbf{x}(t) \in N$; then $f(\mathbf{x}(t)) \geq f(\mathbf{x}^*)$ by the definition of N, but $f(\mathbf{x}(t)) < f(\mathbf{x}^*)$ by the above inequality, a contradiction.

It follows that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, so $\mathbf{x}^*$ is a global minimum of f in $\mathcal{X}$. □

Proposition 17. *Let $\mathcal{X}$ be a convex set. If f is strictly convex, then there exists at most one local minimum of f in $\mathcal{X}$. Consequently, if it exists it is the unique global minimum of f in $\mathcal{X}$.*

Proof. The second sentence follows from the first, so all we must show is that if a local minimum exists in $\mathcal{X}$ then it is unique.

Suppose $\mathbf{x}^*$ is a local minimum of f in $\mathcal{X}$, and suppose towards a contradiction that there exists a local minimum $\tilde{\mathbf{x}} \in \mathcal{X}$ such that $\tilde{\mathbf{x}} \neq \mathbf{x}^*$.

Since f is strictly convex, it is convex, so $\mathbf{x}^*$ and $\tilde{\mathbf{x}}$ are both global minima of f in $\mathcal{X}$ by the previous result. Hence $f(\mathbf{x}^*) = f(\tilde{\mathbf{x}})$. Consider the line segment $\mathbf{x}(t) = t\mathbf{x}^* + (1-t)\tilde{\mathbf{x}},\ t \in [0,1]$, which again must lie entirely in $\mathcal{X}$. By the strict convexity of f,

$$f(\mathbf{x}(t)) < tf(\mathbf{x}^*) + (1-t)f(\tilde{\mathbf{x}}) = tf(\mathbf{x}^*) + (1-t)f(\mathbf{x}^*) = f(\mathbf{x}^*)$$

for all $t \in (0,1)$. But this contradicts the fact that $\mathbf{x}^*$ is a global minimum. Therefore if $\tilde{\mathbf{x}}$ is a local minimum of f in $\mathcal{X}$, then $\tilde{\mathbf{x}} = \mathbf{x}^*$, so $\mathbf{x}^*$ is the unique minimum in $\mathcal{X}$. □

It is worthwhile to examine how the feasible set affects the optimization problem. We will see why the assumption that $\mathcal{X}$ is convex is needed in the results above.

Consider the function $f(x) = x^2$, which is a strictly convex function. The unique global minimum of this function in $\mathbb{R}$ is $x = 0$. But let's see what happens when we change the feasible set $\mathcal{X}$.

(i) $\mathcal{X} = \{1\}$: This set is actually convex, so we still have a unique global minimum. But it is not the same as the unconstrained minimum!

(ii) $\mathcal{X} = \mathbb{R} \setminus \{0\}$: This set is non-convex, and we can see that f has no minima in $\mathcal{X}$. For any point $x \in \mathcal{X}$, one can find another point $y \in \mathcal{X}$ such that $f(y) < f(x)$.

(iii) $\mathcal{X} = (-\infty, -1] \cup [0, \infty)$: This set is non-convex, and we can see that there is a local minimum ($x = -1$) which is distinct from the global minimum ($x = 0$).

(iv) $\mathcal{X} = (-\infty, -1] \cup [1, \infty)$: This set is non-convex, and we can see that there are two global minima ($x = \pm 1$).

4.8.4 Showing that a function is convex

Hopefully the previous section has convinced the reader that convexity is an important property. Next we turn to the issue of showing that a function is (strictly/strongly) convex. It is of course possible (in principle) to directly show that the condition in the definition holds, but this is usually not the easiest way.

Proposition 18. *Norms are convex.*

Proof. Let $\|\cdot\|$ be a norm on a vector space V. Then for all $\mathbf{x}, \mathbf{y} \in V$ and $t \in [0,1]$,

$$\|t\mathbf{x} + (1-t)\mathbf{y}\| \leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y}\| = |t|\|\mathbf{x}\| + |1-t|\|\mathbf{y}\| = t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\|$$

where we have used respectively the triangle inequality, the homogeneity of norms, and the fact that t and $1-t$ are nonnegative. Hence $\|\cdot\|$ is convex. □

Proposition 19. *Suppose f is differentiable. Then f is convex if and only if*

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

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

Proof. ($\implies$) Suppose f is convex, i.e.

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) = f(\mathbf{y}) + t(f(\mathbf{x}) - f(\mathbf{y}))$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and all $t \in [0,1]$. Rearranging gives

$$\frac{f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{t} \leq f(\mathbf{x}) - f(\mathbf{y})$$

As $t \to 0$, the left-hand side becomes $\langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$, so the result follows.

($\impliedby$) Suppose

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$. Fix $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, $t \in [0,1]$, and define $\mathbf{z} = t\mathbf{x} + (1-t)\mathbf{y}$. Then

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \\ f(\mathbf{y}) &\geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \end{aligned}$$

so

$$\begin{aligned} tf(\mathbf{x}) + (1-t)f(\mathbf{y}) &\geq t\left(f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle\right) + (1-t)\left(f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle\right) \\ &= f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), t(\mathbf{x} - \mathbf{z}) + (1-t)(\mathbf{y} - \mathbf{z}) \rangle \\ &= f(t\mathbf{x} + (1-t)\mathbf{y}) + \langle \nabla f(\mathbf{z}), \underbrace{t\mathbf{x} + (1-t)\mathbf{y} - \mathbf{z}}_{\mathbf{0}} \rangle \\ &= f(t\mathbf{x} + (1-t)\mathbf{y}) \end{aligned}$$

implying that f is convex. □

Proposition 20. *Suppose f is twice differentiable. Then*

(i) *f is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \operatorname{dom} f$.*

(ii) *If $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \operatorname{dom} f$, then f is strictly convex.*

(iii) *f is m-strongly convex if and only if $\nabla^2 f(\mathbf{x}) \succeq mI$ for all $\mathbf{x} \in \operatorname{dom} f$.*

Proof. Omitted. □

Proposition 21. *If f is convex and $\alpha \geq 0$, then αf is convex.*

Proof. Suppose f is convex and $\alpha \geq 0$. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(\alpha f) = \operatorname{dom} f$,

$$\begin{aligned}(\alpha f)(t\mathbf{x} + (1-t)\mathbf{y}) &= \alpha f(t\mathbf{x} + (1-t)\mathbf{y}) \\ &\leq \alpha \left(tf(\mathbf{x}) + (1-t)f(\mathbf{y})\right) \\ &= t(\alpha f(\mathbf{x})) + (1-t)(\alpha f(\mathbf{y})) \\ &= t(\alpha f)(\mathbf{x}) + (1-t)(\alpha f)(\mathbf{y})\end{aligned}$$

so αf is convex. □

Proposition 22. *If f and g are convex, then $f + g$ is convex. Furthermore, if g is strictly convex, then $f + g$ is strictly convex, and if g is m-strongly convex, then $f + g$ is m-strongly convex.*

Proof. Suppose f and g are convex. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom}(f + g) = \operatorname{dom} f \cap \operatorname{dom} g$,

$$\begin{aligned}(f+g)(t\mathbf{x} + (1-t)\mathbf{y}) &= f(t\mathbf{x} + (1-t)\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) \\ &\leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + g(t\mathbf{x} + (1-t)\mathbf{y}) && \text{convexity of } f \\ &\leq tf(\mathbf{x}) + (1-t)f(\mathbf{y}) + tg(\mathbf{x}) + (1-t)g(\mathbf{y}) && \text{convexity of } g \\ &= t(f(\mathbf{x}) + g(\mathbf{x})) + (1-t)(f(\mathbf{y}) + g(\mathbf{y})) \\ &= t(f+g)(\mathbf{x}) + (1-t)(f+g)(\mathbf{y})\end{aligned}$$

so $f + g$ is convex.

If g is strictly convex, the second inequality above holds strictly for $\mathbf{x} \neq \mathbf{y}$ and $t \in (0, 1)$, so $f + g$ is strictly convex.

If g is m-strongly convex, then the function $h(\mathbf{x}) \equiv g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$ is convex, so $f + h$ is convex. But

$$(f+h)(\mathbf{x}) \equiv f(\mathbf{x}) + h(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2 \equiv (f+g)(\mathbf{x}) - \frac{m}{2}\|\mathbf{x}\|_2^2$$

so $f + g$ is m-strongly convex. □

Proposition 23. *If $f_1, \ldots, f_n$ are convex and $\alpha_1, \ldots, \alpha_n \geq 0$, then*

$$\sum_{i=1}^{n} \alpha_i f_i$$

is convex.

Proof. Follows from the previous two propositions by induction. □

Proposition 24. *If f is convex, then $g(\mathbf{x}) \equiv f(\mathbf{A}\mathbf{x} + \mathbf{b})$ is convex for any appropriately-sized $\mathbf{A}$ and $\mathbf{b}$.*

Proof. Suppose f is convex and g is defined like so. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} g$,

$$\begin{aligned}g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(\mathbf{A}(t\mathbf{x} + (1-t)\mathbf{y}) + \mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + \mathbf{b}) \\ &= f(t\mathbf{A}\mathbf{x} + (1-t)\mathbf{A}\mathbf{y} + t\mathbf{b} + (1-t)\mathbf{b}) \\ &= f(t(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)(\mathbf{A}\mathbf{y} + \mathbf{b})) \\ &\leq tf(\mathbf{A}\mathbf{x} + \mathbf{b}) + (1-t)f(\mathbf{A}\mathbf{y} + \mathbf{b}) && \text{convexity of } f \\ &= tg(\mathbf{x}) + (1-t)g(\mathbf{y})\end{aligned}$$

Thus g is convex. □

Proposition 25. *If f and g are convex, then $h(\mathbf{x}) \equiv \max\{f(\mathbf{x}), g(\mathbf{x})\}$ is convex.*

Proof. Suppose f and g are convex and h is defined like so. Then for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} h$,

$$\begin{aligned} h(t\mathbf{x} + (1-t)\mathbf{y}) &= \max\{f(t\mathbf{x} + (1-t)\mathbf{y}), g(t\mathbf{x} + (1-t)\mathbf{y})\} \\ &\leq \max\{tf(\mathbf{x}) + (1-t)f(\mathbf{y}), tg(\mathbf{x}) + (1-t)g(\mathbf{y})\} \\ &\leq \max\{tf(\mathbf{x}), tg(\mathbf{x})\} + \max\{(1-t)f(\mathbf{y}), (1-t)g(\mathbf{y})\} \\ &= t\max\{f(\mathbf{x}), g(\mathbf{x})\} + (1-t)\max\{f(\mathbf{y}), g(\mathbf{y})\} \\ &= th(\mathbf{x}) + (1-t)h(\mathbf{y}) \end{aligned}$$

Note that in the first inequality we have used convexity of f and g plus the fact that $a \leq c, b \leq d$ implies $\max\{a, b\} \leq \max\{c, d\}$. In the second inequality we have used the fact that $\max\{a+b, c+d\} \leq \max\{a, c\} + \max\{b, d\}$.

Thus h is convex. □

4.8.5 Examples

A good way to gain intuition about the distinction between convex, strictly convex, and strongly convex functions is to consider examples where the stronger property fails to hold.

Functions that are convex but not strictly convex:

(i) $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \alpha$ for any $\mathbf{w} \in \mathbb{R}^d, \alpha \in \mathbb{R}$. Such a function is called an **affine function**, and it is both convex and concave. (In fact, a function is affine if and only if it is both convex and concave.) Note that linear functions and constant functions are special cases of affine functions.

(ii) $f(\mathbf{x}) = \|\mathbf{x}\|_1$

Functions that are strictly but not strongly convex:

(i) $f(x) = x^4$. This example is interesting because it is strictly convex but you cannot show this fact via a second-order argument (since $f''(0) = 0$).

(ii) $f(x) = \exp(x)$. This example is interesting because it's bounded below but has no local minimum.

(iii) $f(x) = -\log x$. This example is interesting because it's strictly convex but not bounded below.

Functions that are strongly convex:

(i) $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$