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:

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

Below we will first show that for any convex f, 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 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
>
> $$f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}) \quad \text{for all} \;\; \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq \epsilon.$$
>
> 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

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

Combining this with the inequality above we have

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

Rearranging this gives us $\theta f(\boldsymbol{x}^\star) \leq \theta f(\boldsymbol{x}')$, or simply $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}')$, 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. It might also be that there are many local minima, but that all of them achieve the same value of f and hence they are *all* global minima. We are typically not so lucky, though. In many conconvex problems there can be many local minima which are very far from actually minimizing f.

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., $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x})\rangle \geq 0$ for all $\boldsymbol{u} \in \mathbb{R}^N$. However, the only way we can make $\langle \boldsymbol{u}, \nabla f(\boldsymbol{x}^\star)\rangle \geq 0$ for all $\boldsymbol{u}$ 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, recall that if f is convex and differentiable then

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

for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^N$. By making the substitution $\boldsymbol{x} = \boldsymbol{x}^\star + \boldsymbol{u}$ and $\boldsymbol{x}' = \boldsymbol{x}^\star$ we can equivalently write this as

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

for all choices of $\boldsymbol{u} \in \mathbb{R}^N$. But if $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$ then this is simply

$$f(\boldsymbol{x}^\star + \boldsymbol{u}) \geq f(\boldsymbol{x}^\star)$$

for all $\boldsymbol{u}$. 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 lies at the heart of most algorithms for unconstrained convex optimization similar to gradient descent – if we can find an $\boldsymbol{x}$ that makes the gradient vanish, then we have solved the problem.

Existence of minimizers

Before turning to actual algorithms for unconstrained optimization, there are a couple of technical issues to consider. 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 $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x})$ 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 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.

Uniqueness of minimizers

It is also important to note that 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.

Algorithms for unconstrained minimization

One of the benefits of minimizing convex functions is that we can often use very simple algorithms to find solutions. Specifically, we want to solve

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

where f is convex. For now we will assume that f is also differentiable.[1] We have just seen that, in this case, a necessary and sufficient condition for $\boldsymbol{x}^\star$ to be a minimizer is that the gradient vanishes:

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

Thus, we can equivalently think of the problem of minimizing $f(\boldsymbol{x})$ as finding an $\boldsymbol{x}^\star$ that $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$. As noted before, it is not a given that such an $\boldsymbol{x}^\star$ exists, but for now we will assume that f does have (at least one) minimizer.

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

Iterative descent

Initialize: $k = 0$, $\boldsymbol{x}_0 =$ initial guess
while not converged **do**
  calculate a direction to move $\boldsymbol{d}_k$
  calculate a step size $\alpha_k \geq 0$
  $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$
  $k = k + 1$
end while

[1] We will also be interested in cases where f is not differentiable. We will revisit this later in the course.

The central challenge in designing a good algorithm mostly boils down to computing the direction $\boldsymbol{d}_k$. As a preview, here are some choices that we will discuss:

1. **Gradient descent**: We take

$$\boldsymbol{d}_k = -\nabla f\left(\boldsymbol{x}_k\right).$$

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

2. **Accelerated gradient descent**: We can sometimes reduce the number of iterations required by gradient descent by incorporating a *momentum* term. Specifically, we first compute

$$\boldsymbol{p}_k = \boldsymbol{x}_k - \boldsymbol{x}_{k-1}$$

   and then take

$$\boldsymbol{d}_k = -\nabla f\left(\boldsymbol{x}_k\right) + \frac{\beta_k}{\alpha_k}\boldsymbol{p}_k$$

   or

$$\boldsymbol{d}_k = -\nabla f\left(\boldsymbol{x}_k + \beta_k \boldsymbol{p}_k\right) + \frac{\beta_k}{\alpha_k}\boldsymbol{p}_k.$$

   The "heavy ball" method and conjugate gradient descent use the former update rule; Nesterov's method uses the latter. We will see later that by incorporating this momentum term, we can sometimes dramatically reduce the number of iterations required for convergence.

3. **Newton's method**: Gradient descent methods are based on building linear approximations to the function at each iteration. We can also 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

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

that is, the inverse of the Hessian evaluated at $\boldsymbol{x}_k$ 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.

4. **Quasi-Newton methods**: If the dimension N of $\boldsymbol{x}$ is large, Newton's method is not computationally feasible. In this case we can replace the Newton iteration with

$$\boldsymbol{d}_k = -\boldsymbol{Q}_k \nabla f\left(\boldsymbol{x}_k\right)$$

where $\boldsymbol{Q}_k$ is an approximation or estimate of $\left(\nabla^2 f(\boldsymbol{x}_k)\right)^{-1}$. Quasi-Newton methods may require more iterations than a pure Newton approach, but can still be very effective.

Whichever direction we choose, it should be a **descent direction**, i.e., $\boldsymbol{d}_k$ should satisfy

$$\left\langle \boldsymbol{d}_k, \nabla f\left(\boldsymbol{x}_k\right)\right\rangle \leq 0.$$

Since f is convex, it is always true that

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

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.