Convergence of gradient descent

Here we will prove convergence guarantees for **gradient descent**, i.e., the version of our iterative algorithm where we set

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

resulting in the update rule

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f\left(\boldsymbol{x}_k\right).$$

It is hard to say much about the convergence properties of this approach for *arbitrary* convex functions. However, if f satisfies certain "regularity conditions", then we can get very nice guarantees, even for a fixed step size. Here we will look at two different regularity assumptions on f, and translate them into convergence rates. Throughout, we will assume that f is differentiable everywhere.[1]

Smoothness

First, we will see what we can show if we assume that f is **smooth** in a certain sense. Qualitatively, we would just like to assume that the gradient changes in a controlled manner as we move from point to point. Quantitatively, we will assume that f has a **Lipschitz gradient**. This means that there exists an $M > 0$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq M\|\boldsymbol{x} - \boldsymbol{y}\|_2, \tag{1}$$

[1] Methods for nondifferentiable $f(\boldsymbol{x})$ are also of great interest, and will be covered later in the course. These methods are not much more involved algorithmically (although, you obviously will have to replace the gradient with something else), but they are slightly harder to analyze.

for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$. We will say that such a function is M**-smooth** or **strongly smooth**.

In the homework you will show that f obeying (1) is actually equivalent to saying that

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

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

This provides some intuition for what kind of structure the Lipschitz gradient condition imposes on f. Recall that for any convex function, we have that

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

so if f is convex, then at any point $\boldsymbol{x}$ we can bound f from *below* by a linear approximation. If f has a Lipschitz gradient, (2) but we can also bound it from *above* using a quadratic approximation.

In the homework you will also argue that (2) is equivalent to the assumption that $\frac{M}{2}\|\boldsymbol{x}\|_2^2 - f(\boldsymbol{x})$ is convex. In the case that f is twice differentiable, it is not hard to use this fact to show that (2) is equivalent to

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

i.e., that the largest eigenvalue of the Hessian is bounded by M for all $\boldsymbol{x}$. Note, however, that the Lipschitz gradient condition and the analysis below does not require f to be twice differentiable.

Convergence of gradient descent: M-smoothness

Now, let's consider running gradient descent on such a function with a **fixed step size**[2] $\alpha_k = 1/M$. Recall that the central gradient descent iteration is just

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \frac{1}{M}\nabla f(\boldsymbol{x}_k).$$

From our assumption that f is M-smooth, we know that f satisfies (2), and thus plugging in $\boldsymbol{y} = \boldsymbol{x}_{k+1}$, we obtain

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) &\le f(\boldsymbol{x}_k) + \left\langle -\frac{1}{M}\nabla f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k)\right\rangle + \frac{M}{2}\left\|\frac{1}{M}\nabla f(\boldsymbol{x}_k)\right\|_2^2 \\ &= f(\boldsymbol{x}_k) - \frac{1}{M}\|\nabla f(\boldsymbol{x}_k)\|_2^2 + \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\ &= f(\boldsymbol{x}_k) - \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2. \end{aligned} \tag{3}$$

Note that (3) shows that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$ as long as we are not already at the solution, so we are at least guaranteed to make some progress at each iteration. In fact, it says a bit more, giving us a guarantee regarding *how much* progress we are making, namely that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \ge \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2,$$

so that if the gradient is large we are guaranteed to make a large amount of progress.

[2] This requires that you know M, which may not be possible in practice. In fact, if $\alpha < 1/M$ you will still get convergence, it will simply be slower. Moreover, it is not too hard to extend this approach to get a similar guarantee when using a backtracking line search.

In the technical addendum at the end of these notes, we show that by combining this result with the definition of convexity and doing some clever manipulations, we can get a guarantee of the form

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{M}{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Thus, for M-smooth functions, we can guarantee that the error is $O(1/k)$ after k iterations. Another way to put this is to say that we can guarantee accuracy

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \epsilon$$

as long as

$$k \geq \frac{M}{2\epsilon}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.$$

Note that if ϵ is very small, this says we can expect to need a very large number of iterations.

Strong convexity

We will now consider a stronger assumption on f and show that we can get greatly improved guarantees. Recall that before we assumed that f was M-smooth, meaning that

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

for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$. In the analysis below, we consider adding an additional assumption. Specifically, we will assume that f is also **strongly convex** (with strong convexity parameter $m > 0$), meaning that

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x})\rangle + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \qquad (4)$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$. This tells us that not only is f bounded below by a *linear* approximation (since it is convex), but also by a (nontrivial) convex *quadratic* approximation. Note also that strong convexity implies strict convexity, but strict convexity does not necessarily imply strong convexity.

In the homework you will argue that strong convexity as defined in (4) is equivalent to the assumption that $f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|_2^2$ is convex. In the case that f is twice differentiable, this implies that

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

That is, the eigenvalues of the Hessian are bounded below by $m > 0$ for all $\boldsymbol{x}$. When combined with the assumption of M-smoothness, this bounds the conditioning of the Hessian matrix so that its eigenvalues are bounded between $m > 0$ and $M < \infty$. However, again note that strong convexity does not require f to be twice differentiable.

Convergence of gradient descent: Strong convexity

Here we will show that if a function is strongly convex, in addition to being M-smooth, then we can obtain a significantly improved convergence guarantee compared to what we had in the case of M-smoothness alone. We begin our analysis in the same way as before, which began by showing in (3) that M-smoothness implies that

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Next we use strong convexity to obtain a lower bound on $\|\nabla f(\boldsymbol{x})\|_2^2$.

Specifically, recall from the definition of strong convexity in (4) that for any $\boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f$

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{y} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \frac{m}{2} \|\boldsymbol{y} - \boldsymbol{x}\|_2^2. \tag{5}$$

We can obtain a simpler lower bound for $f(\boldsymbol{y})$ by determining the smallest value that the right-hand side of (4) could ever take over all possible choices of $\boldsymbol{y}$ To do this, we simply minimize this lower bound by taking the gradient with respect to $\boldsymbol{y}$ and setting it equal to zero:

$$\nabla f(\boldsymbol{x}) + m(\boldsymbol{y} - \boldsymbol{x}) = 0,$$

From this we obtain that the lower bound in (5) will be minimized by

$$\boldsymbol{y} - \boldsymbol{x} = -\frac{1}{m} \nabla f(\boldsymbol{x}).$$

Plugging this into (5) yields

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) - \frac{1}{m} \|\nabla f(\boldsymbol{x})\|_2^2 + \frac{1}{2m} \|\nabla f(\boldsymbol{x})\|_2^2 \\ &= f(\boldsymbol{x}) - \frac{1}{2m} \|\nabla f(\boldsymbol{x})\|_2^2. \end{aligned}$$

In particular, this applies when $\boldsymbol{y} = \boldsymbol{x}^\star$, which after some rearranging yields

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq 2m \left(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\right). \tag{PL}$$

This is a famous and useful result, often referred to as the **Polyak-Łojasiewicz inequality**.

Combining the PL inequality with (3) we obtain

$$\begin{aligned} f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) &\leq f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) - \frac{m}{M} \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\right) \\ &= \left(1 - \frac{m}{M}\right) \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\right). \end{aligned}$$

That is, the gap between the current value of the objective function and the optimal value is cut down by a factor of $1 - m/M < 1$ at each iteration.

This is an example of *linear convergence*, and you will show on the homework that this implies that

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

as long as

$$k \geq \frac{\log(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)/\epsilon)}{\log((M-m)/M)}.$$

This is **much** faster convergence than what we obtained before – it is $O(\log \epsilon^{-1})$ versus $O(\epsilon^{-1})$. As an example, if we wanted to set $\epsilon = 10^{-6}$, $\log \epsilon^{-1} \approx 14$ (versus $\epsilon^{-1} = 10^6$). Of course, to get this we had to make a much stronger assumption (strong convexity), which may not always be applicable depending on the objective function you are optimizing.

Finally, we also note that the PL inequality above also provides some guidance in terms of setting a stopping criterion. Specifically, if we declare convergence when $\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \epsilon$ then the PL inequality allows us to conclude that

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{1}{2m}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \leq \frac{\epsilon^2}{2m}.$$

This provides a principled way of declaring convergence.

Technical Details: Convergence analysis for M-smooth functions

Here we complete the convergence analysis for gradient descent on M-smooth functions that is summarized above. Specifically, recall that above in (3) we showed that if f is M-smooth then

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Moreover, by the convexity of f,

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

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

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}^\star) + \langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x})\rangle - \frac{1}{2M}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Substituting $\nabla f(\boldsymbol{x}_k) = M(\boldsymbol{x}_k - \boldsymbol{x}_{k+1})$ then yields

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq M\langle \boldsymbol{x}_k - \boldsymbol{x}^\star, \boldsymbol{x}_k - \boldsymbol{x}_{k+1}\rangle - \frac{M}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|_2^2. \quad (6)$$

We can re-write this in a slightly more convenient way using the fact that

$$\|\boldsymbol{a} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - 2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \|\boldsymbol{b}\|_2^2$$

and thus

$$2\langle \boldsymbol{a}, \boldsymbol{b}\rangle - \|\boldsymbol{b}\|_2^2 = \|\boldsymbol{a}\|_2^2 - \|\boldsymbol{a} - \boldsymbol{b}\|_2^2.$$

Setting $\boldsymbol{a} = \boldsymbol{x}_k - \boldsymbol{x}^\star$ and $\boldsymbol{b} = \boldsymbol{x}_k - \boldsymbol{x}_{k+1}$ and applying this to (6), we obtain the bound

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star) \leq \frac{M}{2}\left(\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^\star\|_2^2\right).$$

This result bounds how far away $f(\boldsymbol{x}_{k+1})$ is from the optimal $f(\boldsymbol{x}^\star)$ in terms (primarily) of the error in the previous iteration: $\|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2$. We can use this result to bound $f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^\star)$ in terms of the initial error $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$ by a clever argument.

Specifically, this bound holds not only for iteration k, but for all iterations $i = 1, \dots, k$, so we can write down k inequalities and then sum them up to obtain

$$\sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) \leq \frac{M}{2}\left(\sum_{i=1}^{k} \|\boldsymbol{x}_{i-1} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{x}^\star\|_2^2\right).$$

The right-hand side of this inequality is what is called a *telescopic sum*: each successive term in the sum cancels out part of the previous term. Once you write this out, all the terms cancel except for two (one component from the $i = 1$ term and one from the $i = k$ term) giving us:

$$\begin{aligned}\sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star) &\leq \frac{M}{2}\left(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^\star\|_2^2\right) \\ &\leq \frac{M}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2.\end{aligned}$$

Since, as noted above, $f(\boldsymbol{x}_i)$ is monotonically decreasing in i, we also have that

$$k\left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star)\right) \leq \sum_{i=1}^{k} f(\boldsymbol{x}_i) - f(\boldsymbol{x}^\star),$$

and thus

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^\star) \leq \frac{M}{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2,$$

which is exactly what we wanted to show.