CS 726: Basic Descent Methods in the Convex Case

Jelena Diakonikolas

So far we have analyzed the standard gradient (or steepest) descent method. But gradient descent is not the only possible descent method, and we have seen in class a few other examples of methods that take the following form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \tag{1}$$

and for which we can guarantee the following "descent lemma":

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

Recall that the basic gradient descent method is a special case of basic descent methods, described by:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla f(\mathbf{x}_k), \tag{3}$$

where $\alpha \in (0, 1/L]$.

We will assume throughout that the function is bounded below and that it has a minimizer $\mathbf{x}^* \in \mathbb{R}^d$.

Note that our analysis of gradient descent in the nonconvex setting from class had only utilized the descent property from Eq. (2). Hence the same guarantee we derived in class applies to all basic descent methods that satisfy this property, and in this lecture we will focus on other assumptions about f.

How Does Convexity Help?

Convexity is helpful because it allows us to estimate how "good" or "bad" we are doing compared to the minimum function value. This is the information that is actually useful to us in many cases.

Consider the following scenario: You are visiting a country that you know nearly nothing about; in particular, you do not know what are the standard prices there. You have decided to buy a souvenir. You go to a seller, and the seller tells you that the souvenir you like costs #5, in the local currency # you do not know anything about. Does this mean much to you? How about if I told you that the market price for the souvenir you wanted is #1? How would you feel if the market price was #4.99?

What I am getting at is that we are often interested in knowing how well we are approximating $f(\mathbf{x}^*)$, not what the value $f(\mathbf{x})$ of the point $\mathbf{x}$ we have produced is. Thus, our goal is to show that $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ goes down with k. Convexity allows us to do that. How? Well, we saw in class that, because the function is differentiable and convex, we have, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle . \tag{4}$$

In particular, taking $\mathbf{y} = \mathbf{x}^*$, we have an estimate of $f(\mathbf{x}^*)$ that is based on $\mathbf{x}$ and the function value and gradient at $\mathbf{x}$. This estimate may not seem particularly useful, as we do not know $\mathbf{x}^*$ (which appears on the right-hand side of the inequality), but we will soon see how we can get past that. Pictorially, our estimate looks like the blue plane in Fig. 1.

Now, if our algorithm has generated points $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k$, we have, $\forall i \in \{0, \ldots, k\}$:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle . \tag{5}$$

In particular, we can take an arbitrary convex combination of the lower-bounding hyperplanes from Eq. (5). To do so, let $a_0, a_1, ..., a_k$ be a sequence of positive real numbers, and let $A_k = \sum_{i=0}^k a_i$ (so that $\frac{1}{A_k}\sum_{i=0}^k a_i = 1$). We have:

$$f(\mathbf{x}^*) \geq \frac{1}{A_k}\sum_{i=0}^k a_i \big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle \big), \tag{6}$$

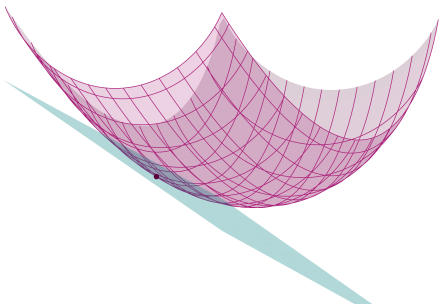

Figure 1: A convex function and a lower bound on $f(\mathbf{x}^*)$ based on Eq. (4).

and we will call the right-hand side of Eq. (6) L_k, so that we have $f(\mathbf{x}^*) \geq L_k$.

Our strategy is as follows. First, we keep track of the optimality gap[1] estimate $G_k = f(\mathbf{x}_{k+1}) - L_k$. As we have chosen L_k so that $L_k \leq f(\mathbf{x}^*)$, we have $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq G_k$. Our goal is to show that $A_k G_k$ does not increase, modulo some small error, and for A_k that grows as fast as possible. In particular, if $A_k G_k \leq A_{k-1} G_{k-1} + E_k$, then, unrolling this recursive relationship down to zero, we have $A_k G_k \leq A_0 G_0 + \sum_{i=1}^k E_i$. Rearranging, and using that, by design, $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq G_k$, we then have:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq G_k \leq \frac{A_0 G_0 + \sum_{i=1}^k E_i}{A_k}.$$

Thus, if $\sum_{i=1}^k E_i$ grows slowly compared to A_k (ideally, it would be zero) and $A_0 G_0$ is bounded, our solutions must be approaching the minimum function value at some nontrivial rate.

Let us now make this approach formal. Recall that we assume that every step of our method makes progress as in Eq. (2), and, thus, $f(\mathbf{x}_1) \leq f(\mathbf{x}_0) - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_0)\|_2^2$, for some $\alpha > 0$. By the definition of A_k, we have $A_0 = a_0$. From the definition of G_k :

$$\begin{aligned} A_0 G_0 &= a_0\big(f(\mathbf{x}_1) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0 \rangle\big) \\ &\leq a_0\Big(-\frac{\alpha}{2}\|\nabla f(\mathbf{x}_0)\|_2^2 - \langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0 \rangle\Big). \end{aligned}$$

Applying the Cauchy-Schwarz inequality to $\langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0 \rangle$, we further have:

$$A_0 G_0 \leq a_0\Big(-\frac{\alpha}{2}\|\nabla f(\mathbf{x}_0)\|_2^2 + \|\nabla f(\mathbf{x}_0)\|_2 \cdot \|\mathbf{x}^* - \mathbf{x}_0\|_2\Big).$$

Here is an important (and simple!) inequality: $\forall p, q \in \mathbb{R} : -\frac{p^2}{2} + pq \leq \frac{q^2}{2}$. (Can you prove it?) Apply this inequality with $p = \sqrt{\alpha}\|\nabla f(\mathbf{x}_0)\|_2$ and $q = \|\mathbf{x}^* - \mathbf{x}_0\|_2/\sqrt{\alpha}$ to get:

$$A_0 G_0 \leq \frac{a_0}{2\alpha}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2. \tag{7}$$

This looks good! Now, let us bound $A_k G_k - A_{k-1} G_{k-1}$ for $k \geq 1$. By definition:

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} &= A_k f(\mathbf{x}_{k+1}) - A_{k-1} f(\mathbf{x}_k) - a_k\big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle\big) \\ &= A_k(f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)) - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle \\ &\leq -A_k\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle, \end{aligned}$$

where we have used $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2$, which we assumed for our algorithm. To complete bounding $A_k G_k - A_{k-1} G_{k-1}$, we use the same approach as we did for $A_0 G_0$: Cauchy-Schwarz inequality and then $-\frac{p^2}{2} + pq \leq$

[1] Optimality gap is how far we are from $f(\mathbf{x}^*)$: for a point $\mathbf{x}$, it equals $f(\mathbf{x}) - f(\mathbf{x}^*)$.

$\frac{q^2}{2}$ with $p = \sqrt{\alpha A_k}\|\nabla f(\mathbf{x}_k)\|$, $q = \frac{a_k}{\sqrt{\alpha A_k}}\|\mathbf{x}^* - \mathbf{x}_k\|_k$, to get:

$$A_k G_k - A_{k-1} G_{k-1} \leq \frac{{a_k}^2}{2\alpha A_k}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2. \tag{8}$$

Applying Eq. (8) recursively until we reach $k = 0$ and then using Eq. (7), we have:

$$A_k G_k \leq \sum_{i=0}^{k} \frac{{a_i}^2}{2\alpha A_i}\|\mathbf{x}^* - \mathbf{x}_i\|_2^2. \tag{9}$$

Define: $R = \max\{\|\mathbf{x}^* - \mathbf{x}\| : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, so that $\|\mathbf{x}^* - \mathbf{x}_i\|_2^2 \leq R^2$ (in general, this will be bounded, and if we are only assuming progress as in Eq. (2), we cannot do better). From our definition of the gap, we have:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{R^2}{2\alpha A_k} \sum_{i=0}^{k} \frac{{a_i}^2}{A_i}. \tag{10}$$

To complete our analysis, it remains to choose the sequence a_k. There are different choices that work and give a similar result, but one that works well is $a_i = \frac{i+1}{2}$. Using the standard arithmetic series result, $A_i = \frac{(i+1)(i+2)}{4}$. Thus $\frac{{a_i}^2}{A_i} \leq 1$, and we have:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{2R^2}{\alpha(k+2)}. \tag{11}$$

Can we do better? Not without changing either the algorithm or the assumptions about f.

If we assume a little bit more about our algorithm; in particular, that we run gradient descent from Eq. (3), then we can do a little bit better. Namely, we can replace R in Eq. (11) with $\|\mathbf{x}^* - \mathbf{x}_0\|$. This is because we can show that $\forall k \geq 1 : \|\mathbf{x}^* - \mathbf{x}_k\| \leq \|\mathbf{x}^* - \mathbf{x}_0\|$, as we saw in class.

Thus, for standard steepest (or gradient) descent from Eq. (3), this analysis gives:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{2\|\mathbf{x}^* - \mathbf{x}_0\|^2}{\alpha(k+2)}. \tag{12}$$

We will see next how strong convexity allows us to get an even better bound on $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$.

Strongly Convex Case

When f is m-strongly convex for some $m > 0$, we can create an even better lower bound on $f(\mathbf{x}^*)$. In particular, instead of Eq. (4), we can use that, by strong convexity, $\forall i$:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle + \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_i\|_2^2.$$

This allows us to construct the following lower bound, where, as before, we hold off on choosing a_i's and only assume they are positive:

$$f(\mathbf{x}^*) \geq L_k := \frac{1}{A_k} \sum_{i=0}^{k} a_i \Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle + \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_i\|_2^2 \Big). \tag{13}$$

We now use the same strategy as for the convex case, where, as before $G_k = f(\mathbf{x}_{k+1}) - L_k \geq f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$. Let us first bound the initial gap. By definition of G_k and because $A_0 = a_0$:

$$\begin{aligned} A_0 G_0 &= a_0 \left(f(\mathbf{x}_1) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0 \rangle - \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \right) \\ &\leq a_0 \left(-\frac{\alpha}{2}\|\nabla f(\mathbf{x}_0)\|_2^2 - \langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0 \rangle - \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \right). \end{aligned}$$

Similarly as for the convex case, we can bound $-\frac{\alpha}{2}\|\nabla f(\mathbf{x}_0)\|_2^2 - \langle \nabla f(\mathbf{x}_0), \mathbf{x}^* - \mathbf{x}_0\rangle$ by $\frac{1}{2\alpha}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2$ to get:

$$A_0 G_0 \leq \frac{a_0(\frac{1}{\alpha} - m)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2}. \tag{14}$$

As a sanity check, note that for gradient descent $\frac{1}{\alpha} \geq L$ and it is always true that $L \geq m$ (why?).

Now let us bound the change in $A_k G_k$. We have:

$$\begin{aligned} A_k G_k - A_{k-1} G_{k-1} &= A_k f(\mathbf{x}_{k+1}) - A_{k-1} f(\mathbf{x}_k) - a_k \Big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle + \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2 \Big) \\ &\leq -\frac{A_k \alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle - a_k \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2, \end{aligned}$$

where, same as before, we have used that $A_k = A_{k-1} + a_k$ and $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2$.

We have already shown (while working on the convex case) that

$$-\frac{A_k \alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle \leq \frac{{a_k}^2}{2\alpha A_k}\|\mathbf{x}^* - \mathbf{x}_k\|_2^2.$$

Thus:

$$A_k G_k - A_{k-1} G_{k-1} \leq \Big(\frac{{a_k}^2}{\alpha A_k} - a_k m \Big) \frac{\|\mathbf{x}^* - \mathbf{x}_k\|_2^2}{2}.$$

In particular, if $\frac{a_k}{A_k} \leq m\alpha$, we have that the rhs of the last inequality is non-positive, and, thus, $A_k G_k \leq A_0 G_0$. Using Eq. (14):

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq G_k \leq \frac{a_0(\frac{1}{\alpha} - m)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2A_k}. \tag{15}$$

To make this bound be as good as possible, we want to make A_k grow as fast as possible. But, to obtain the bound, we have already used that $\frac{a_k}{A_k} \leq \alpha m$. It is not hard to see that the fastest growth for A_k (as $A_k = A_{k-1} + a_k$) is obtained for $\frac{a_k}{A_k} = \alpha m$. In this case, $\frac{A_{k-1}}{A_k} = \frac{A_k - a_k}{A_k} = 1 - \alpha m$, and we can write:

$$\frac{a_0}{A_k} = \frac{A_0}{A_k} = \frac{A_0}{A_1} \cdot \frac{A_1}{A_2} \cdots \frac{A_{k-1}}{A_k} = (1 - \alpha m)^k.$$

Combining with Eq. (15):

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (1 - \alpha m)^k \frac{(\frac{1}{\alpha} - m)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2} = (1 - \alpha m)^{k+1} \frac{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2\alpha}. \tag{16}$$

Did we really need strong convexity, or can we use something weaker? It turns out that there is a condition that is weaker than strong convexity that we could have used to obtain the same convergence bound as in Eq. (16). This condition is known as the Polyak-Lojasiewicz (PL) condition and is defined by:

$$(\forall \mathbf{x} \in \mathbb{R}^d): \quad \|\nabla f(\mathbf{x})\|_2^2 \geq 2m(f(\mathbf{x}) - f(\mathbf{x}^*)). \tag{17}$$

(Show that this condition is more general than strong convexity, i.e., that any m-strongly convex function satisfies Eq. (17)). An example of a function that satisfies the PL condition but is not strongly convex is $f(\mathbf{x}) = \frac{1}{2}\langle \mathbf{A}\mathbf{x}, \mathbf{x}\rangle$, where $\mathbf{A}$ is a symmetric PSD matrix that is singular. The PL condition holds for this function with m being equal to the smallest nonzero eigenvalue of $\mathbf{A}$ (you can find the proof in the Appendix of Recht-Wright).

As an exercise, you should adapt the proof for the strongly convex case to the case where only the PL condition holds. To obtain the same bound as in Eq. (16), you could use the following inequalities to bound the initial gap:

$$(\forall \mathbf{x} \in \mathbb{R}^d): \quad \frac{1}{2L}\|\nabla f(\mathbf{x})\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x}^* - \mathbf{x}\|_2^2.$$

How would you prove these inequalities?

Rate Comparison

We have shown the following for the basic descent methods that ensure the progress from Eq. (2), where, for simplicity, we take $\alpha = \frac{1}{L}$:

1. If f is smooth, we have that $\min_{0 \le i \le k} \|\nabla f(\mathbf{x}_k)\|_2 \le \sqrt{\frac{2L(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{k+1}}$. Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ with $\|\nabla f(\mathbf{x}_k)\|_2 \le \epsilon$ in no more than $k = \left\lceil \frac{2L(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\epsilon^2} - 1 \right\rceil = O(\frac{L(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\epsilon^2})$ iterations (assuming w.l.o.g. $\frac{2L(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{\epsilon^2} > 1$).

2. If f is, in addition, convex, then we have (from Eq. (11)) $f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{2LR^2}{k+1}$ or $f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{2L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{k+1}$ (from Eq. (12), if we use steepest descent). Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ that satisfies $f(\mathbf{x}) - f(\mathbf{x}^*) \le \epsilon$ in at most $k = \left\lceil \frac{2LR^2}{\epsilon} \right\rceil - 1 = O(\frac{LR^2}{\epsilon})$ or $k = \left\lceil \frac{2L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{\epsilon} \right\rceil - 1 = O(\frac{L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{\epsilon})$ iterations (where, again, w.l.o.g., $\frac{2LR^2}{\epsilon} > 1$).

3. Finally, if, in addition, f is also m-strongly convex (or satisfies the PL condition with parameter $m > 0$), we have, from Eq. (16), $f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \left(1 - \frac{m}{L}\right)^k \frac{L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{2}$. Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ that satisfies $f(\mathbf{x}) - f(\mathbf{x}^*) \le \epsilon$ after at most $k = O(\frac{L}{m} \log(\frac{L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{\epsilon}))$ iterations.