CS 726: Basic Descent Methods

Jelena Diakonikolas

Fall 2023

In this lecture note, we discuss basic descent methods. Throughout this part, we assume that in our problem (P) the objective function f is L-smooth (usually w.r.t. the Euclidean norm, though not always) but not necessarily convex and the problem is unconstrained ($\mathcal{X} \equiv \mathbb{R}^d$). We further assume that f is bounded below by some $f^* \in \mathbb{R}$.

1 Descent Methods

Basic descent methods will start with some point $\mathbf{x}_0 \in \mathbb{R}^d$ iteratively construct points $\mathbf{x}_{k+1}$ for $k \geq 0$ according to a rule

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

where $\mathbf{z}_k$ is a descent direction (defined below) and α_k is chosen to be sufficiently small to guarantee that the function value decreases at each iteration $k \geq 1$.

Definition 1.1. We say that $\mathbf{z} \in \mathbb{R}^d$ is a descent direction for $f : \mathbb{R}^d \to \mathbb{R}$ at $\mathbf{x} \in \mathbb{R}^d$ if there exists $t > 0$ such that $f(\mathbf{x} + t'\mathbf{z}) \leq f(\mathbf{x})$ for all $0 < t' \leq t$.

It is not immediately clear how to construct a descent direction. The following proposition gives a sufficient condition for a direction to be a descent direction, which is a bit more helpful (but still insufficient to guarantee reasonable progress of an algorithm).

Proposition 1.2. *If a function $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable (in a neighborhood of $\mathbf{x} \in \mathbb{R}^d$), then any $\mathbf{z} \in \mathbb{R}^d$ such that $\langle \nabla f(\mathbf{x}), \mathbf{z} \rangle < 0$ is a descent direction for f at $\mathbf{x}$.*

Proof. Using Taylor Theorem, for any $t > 0$, there exists $\gamma \in (0, 1)$ such that

$$f(\mathbf{x} + t\mathbf{z}) = f(\mathbf{x}) + t \langle \nabla f(\mathbf{x} + t\gamma\mathbf{z}), \mathbf{z} \rangle .$$

As ∇f is continuous in the neighborhood of $\mathbf{x}$, there exists a sufficiently small $t > 0$ such that for all t', $0 < t' \leq t$ we have $\langle \nabla f(\mathbf{x} + t'\mathbf{z}), \mathbf{z} \rangle < 0$. Hence $\mathbf{z}$ is a descent direction for f at $\mathbf{x}$. □

2 Gradient Descent

Based on Proposition 1.2, it should be immediately clear that for any $\mathbf{x}$ such that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, $\mathbf{z} = -\nabla f(\mathbf{x})$ is a descent direction, as in that case $\langle \nabla f(\mathbf{x}), -\nabla f(\mathbf{x}) \rangle = -\|\nabla f(\mathbf{x})\|_2^2 < 0$. On the other hand, if $\nabla f(\mathbf{x}) = \mathbf{0}$, then $\mathbf{x}$ is a stationary point, which without any further assumptions about f is the best we can hope to find (as discussed in past lectures). Thus, it seems reasonable that an algorithm of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \tag{GD}$$

could work under a suitable choice of step sizes α_k. This is indeed the case (in the sense discussed below) and this algorithm is known as (standard) gradient descent (GD). It was first proposed by Cauchy in 1847, though its convergence properties were not studied until about a 100 years later.

To argue about convergence of (GD), we need some measure of progress toward a "solution." For nonconvex functions, the "solution" we can hope to find if we are only assuming that the function is smooth (but not necessarily

convex) would be a stationary point. Since for this part we are assuming that the problem is unconstrained, this goal translates to finding a point with a small (ideally zero) gradient.

Since in this lecture we are discussing *descent* methods, the least we should argue is that under a suitable choice of α_k, (GD) reduces the function value (unless already at a "solution"). To do so, we recall a lemma that we proved in previous lectures, which tells us that an L-smooth function f satisfies $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{2}$$

Standard gradient descent is typically analyzed assuming that f is L-smooth w.r.t. the Euclidean norm $\|\cdot\| = \|\cdot\|_2$, so we will make this assumption here too.[1] Setting $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{y} = \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$, we now have

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha_k \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{L{\alpha_k}^2}{2}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{3}$$

Hence, we can see that if $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ (otherwise we would be done), we can ensure that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ whenever $\frac{L{\alpha_k}^2}{2} - \alpha_k < 0$, or, equivalently, $\alpha_k \in (0, \frac{2}{L})$. We can say even more: for a function that is bounded below, we can estimate how many iterations it would take us to find a point whose gradient is smaller than some target error $\epsilon > 0$.

Lemma 2.1. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is L-smooth w.r.t. the Euclidean norm and an initial point $\mathbf{x}_0 \in \mathbb{R}^d$, consider (GD) with step size $\alpha_k = \alpha \in (0, \frac{1}{L}]$. Then, for all $k \geq 0$,*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{4}$$

Further, if $f(\mathbf{x}) \geq f^* > -\infty, \forall \mathbf{x} \in \mathbb{R}^d$, then $\forall k \geq 0$,

$$\min_{0 \leq i \leq k} \|\nabla f(\mathbf{x}_i)\|_2^2 \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\alpha(k+1)}. \tag{5}$$

As a consequence, for any $\epsilon > 0$, (GD) takes at most $k = \left\lfloor \frac{2(f(\mathbf{x}_0) - f^*)}{\alpha \epsilon^2} \right\rfloor$ iterations to construct a point $\mathbf{x}_i$, $i \in \{1, \ldots, k\}$ such that

$$\|\nabla f(\mathbf{x}_i)\|_2 \leq \epsilon.$$

Proof. The first part of the lemma follows by simply plugging $\alpha_k = \alpha \leq \frac{1}{L}$ into (3).

For the second part, telescoping (4), we get

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_0) \leq -\frac{\alpha}{2}\sum_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|_2^2. \tag{6}$$

Rearranging (6) and using that $f(\mathbf{x}_{k+1}) \geq f^*$, we have

$$\sum_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|_2^2 \leq \frac{2}{\alpha}(f(\mathbf{x}_0) - f^*). \tag{7}$$

Since the minimum of a sequence is always smaller than its average, we have

$$\min_{0 \leq i \leq k} \|\nabla f(\mathbf{x}_i)\|_2^2 \leq \frac{1}{k+1}\sum_{i=0}^{k} \|\nabla f(\mathbf{x}_i)\|_2^2.$$

Thus, (5) follows by dividing both sides of (7) by $k+1$.

Finally, the last claim is equivalent to stating that after at most k iterations, $\min_{0 \leq i \leq k} \|\nabla f(\mathbf{x}_i)\|_2^2 \leq \epsilon^2$. Due to (5), a sufficient condition for this to hold is that

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

Solving for k, we get that it suffices that $k \geq \frac{2(f(\mathbf{x}_0) - f^*)}{\alpha \epsilon^2} - 1$. Thus, it suffices to choose $k = \left\lfloor \frac{2(f(\mathbf{x}_0) - f^*)}{\alpha \epsilon^2} \right\rfloor$. □

[1] Recall once again that because we are in a finite dimensional space, if f is L-smooth with respect to any norm $\|\cdot\|$ then it is also smooth w.r.t. Euclidean norm, but its smoothness constant can be much bigger.

Observe that Lemma 2.1 only guarantees that the minimum gradient we see up to iteration k is decreasing, but it says nothing about the gradient at the last point we construct (apart from it being bounded). You may be wondering whether we could make a similar claim for the gradient at the last point constructed by (GD). It doesn't take long to realize that this is not possible without any additional assumptions, for any descent method. To see this, consider, for example, a univariate function that locally behaves as $-x^2$. Starting from $x_0 \neq 0$ and descending on this function necessarily increases the slope (the absolute value of the derivative), thus we cannot hope to guarantee that the last point we see will be the "best one" (in the sense of a small slope).

2.1 An Alternative Way of Arriving at Gradient Descent

There is an alternative way we could have arrived at (GD), by simply using the upper quadratic approximation (2) that holds for any smooth function. In particular, setting $\mathbf{x} = \mathbf{x}_k$, we know that the function is bounded as

$$f(\mathbf{y}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}_k\|_2^2. \tag{8}$$

As, when choosing $\mathbf{x}_{k+1}$ we know the values of $\mathbf{x}_k, \nabla f(\mathbf{x}_k)$, the right-hand side of (8) is only a function of $\mathbf{y}$ and it is a nice, convex quadratic function of $\mathbf{y}$ that is easy to minimize over $\mathbb{R}^d$ just by differentiating it and setting its gradient equal to zero. So another reasonable approach to deriving an algorithm update $\mathbf{x}_{k+1}$ would be to choose $\mathbf{x}_{k+1}$ as the minimizer of the upper quadratic approximation of the function given on the right-hand side of (8). This leads to the update $\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$ ((GD) update with $\alpha = \frac{1}{L}$) and descent

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

which is the same as what we have derived for (GD) with $\alpha = \frac{1}{L}$ in Lemma 2.1.

This alternative view is useful because we can conceivably derive different versions of "gradient descent" by minimizing any other upper bound on the function value that we may have. As an immediate consequence, we can think of deriving "gradient descent" for functions that are L-smooth with respect to a non-Euclidean norm by minimizing (2) with $\mathbf{x} = \mathbf{x}_k$ (with norm set to the norm w.r.t. which the function is L-smooth). We will see an example of this for ℓ_p norms and randomized coordinate descent methods in Homework #2. Section 3 further provides additional examples.

2.2 Analyzing Gradient Descent in Convex and Strongly Convex Cases

Lemma 2.1 provided us with a convergence bound for (GD) assuming only that the objective function f is L-smooth. You might wonder whether there is more to say for functions that are convex (or even strongly convex). This is indeed the case, and the reason we can say more is not only that every stationary point is a global minimum (as we saw in previous lectures), but that (strong) convexity gives us a way of bounding below the minimum function value. In other words, 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^*, not what the value $f(\mathbf{x})$ of the point $\mathbf{x}$ we have produced is. Thus, our goal is to show that $f(\mathbf{x}_k) - f^*$ (known as the *optimality gap*) goes down with k, and from now let's assume that the minimum is attained so that there exists $\mathbf{x}^* \in \mathbb{R}^d$ such that $f^* = f(\mathbf{x}^*)$.

Convexity allows us argue about reducing $f(\mathbf{x}_k) - f^*$ with k. How? Well, we saw in class that, because the function is differentiable and convex, we have, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle . \tag{9}$$

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.

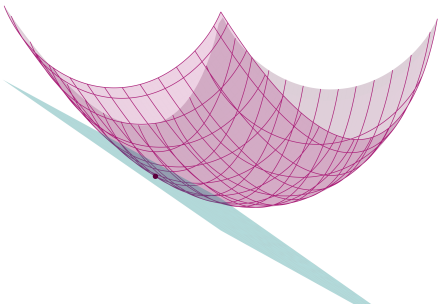

Figure 1: A convex function and a lower bound on $f(\mathbf{x}^*)$ based on (9).

More concretely, recall that we have shown that gradient descent with step size $\alpha \in (0, \frac{1}{L}]$ guarantees descent of the form $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2$. We can try directly bounding $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$ by bounding below $f(\mathbf{x}^*)$ using (9) with $\mathbf{y} = \mathbf{x}^*$ and $\mathbf{x} = \mathbf{x}_k$. This gives us

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

Combining with the descent progress, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq -\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{10}$$

Now observe that (completing the square), we have

$$-\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle - \frac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 - \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 = -\frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}^* - \alpha\nabla f(\mathbf{x}_k)\|_2^2.$$

Thus, combining with (10), we conclude that

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}^* - \alpha\nabla f(\mathbf{x}_k)\|_2^2 \\ &= \frac{1}{2\alpha}\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \frac{1}{2\alpha}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2, \end{aligned} \tag{11}$$

where we have used that, by definition, $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)$.

Observing that (11) telescopes, we now have

$$\sum_{i=0}^{k}(f(\mathbf{x}_{i+1}) - f(\mathbf{x}^*)) \leq \frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq \frac{1}{2\alpha}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2.$$

Recalling that (GD) is a descent method (and, thus, $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) \leq \cdots \leq f(\mathbf{x}_1)$), we can now conclude that (GD) has the following convergence guarantee

Lemma 2.2. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is convex and L-smooth w.r.t. the Euclidean norm and an initial point $\mathbf{x}_0 \in \mathbb{R}^d$, consider (GD) with step size $\alpha_k = \alpha \in (0, \frac{1}{L}]$. Then, for all $k \geq 0$,*

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\alpha(k+1)}. \tag{12}$$

As a consequence, for any $\epsilon > 0$, (GD) after at most $k = \left\lceil \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2\alpha\epsilon} \right\rceil$ iterations we have that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon.$$

Note also that we can conclude from (11) that $\forall k \geq 0$,

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2 \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2, \tag{13}$$

as $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \geq 0$. That is, (GD) never moves further away from the set of function minimizers. Sequences with the property from (13) are said to be "Fejér monotone" with respect to the set of minimizers $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. However, without further assumptions, we cannot guarantee in general that the sequence of (GD) iterates $\{\mathbf{x}_k\}_{k\geq 0}$ approaches $\mathbf{x}^*$ in a non-asymptotic sense.

When f is μ-strongly convex, we can provide an even stronger convergence guarantee for (GD) by using the better lower bound we get from strong convexity:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2. \tag{14}$$

Following the same argument as above but using this stronger lower bound, we can show that in the strongly convex case we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \Big(\frac{1}{2\alpha} - \frac{\mu}{2}\Big)\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \frac{1}{2\alpha}\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2, \tag{15}$$

Since $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \geq 0$, it follows that

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \leq (1 - \alpha\mu)\|\mathbf{x}_k - \mathbf{x}^*\|_2^2. \tag{16}$$

(As a sanity check, argue why it must be $1 - \alpha\mu \geq 0$.) Recursively applying (15), we can now reach the following conclusions.

Lemma 2.3. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is μ-strongly convex and L-smooth w.r.t. the Euclidean norm and an initial point $\mathbf{x}_0 \in \mathbb{R}^d$, consider (GD) with step size $\alpha_k = \alpha \in (0, \frac{1}{L}]$. Then, for all $k \geq 0$,*

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 \leq (1 - \alpha\mu)^k\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \tag{17}$$

As a consequence, for any $\epsilon > 0$, (GD) after at most $k = \max\big\{0, \big\lceil \frac{2}{\alpha\mu}\log(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon})\big\rceil\big\}$ iterations we have that

$$\|\mathbf{x}_k - \mathbf{x}^*\|_2 \leq \epsilon.$$

Proof. (17) follows by iteratively applying (16). To obtain the bound on the iteration count, due to (17), it suffices to have that

$$(1 - \alpha\mu)^k\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq e^{-\alpha\mu k}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \epsilon^2.$$

Taking the natural logarithm on both sides of the last inequality and simplifying, we get

$$k \geq \frac{2}{\alpha\mu}\log\Big(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon}\Big),$$

as required. □

For functions that are both smooth and strongly convex, all measures of optimality (gradient norm, optimality gap, distance to optimum) are related to each other. Thus, if one shows convergence w.r.t. one of these measures, it is immediate to translate it into other measures of convergence, as summarized in the following corollary for the specific case of (GD).

Corollary 2.4. *Given a function $f : \mathbb{R}^d \to \mathbb{R}$ that is μ-strongly convex and L-smooth w.r.t. the Euclidean norm and an initial point $\mathbf{x}_0 \in \mathbb{R}^d$, consider (GD) with step size $\alpha_k = \alpha \in (0, \frac{1}{L}]$. Then,*

- *After at most $k = \max\big\{0, \big\lceil \frac{2}{\alpha\mu}\log(\frac{\sqrt{L}\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon})\big\rceil\big\}$ iterations we have that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$;*
- *After at most $k = \max\big\{0, \big\lceil \frac{2}{\alpha\mu}\log(\frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\epsilon})\big\rceil\big\}$ iterations we have that $\|\nabla f(\mathbf{x}_k)\|_2 \leq \epsilon$.*

The proof is left as an exercise.

3 Other Descent Methods

Gradient descent is only one example of a descent method. There are, of course, other descent methods that make sense and are used in practice. We provide some examples below.

Example 3.1 (Preconditioned Methods)**.** Preconditioned methods are the methods whose descent direction at $\mathbf{x}$ is chosen as $\mathbf{z} = -\mathbf{A}\nabla f(\mathbf{x})$, where $\mathbf{A}$ is a symmetric positive definite matrix (also called a preconditioner). If f is L-smooth w.r.t. the Euclidean norm, then we have

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \alpha\left\langle \mathbf{A}\nabla f(\mathbf{x}), \nabla f(\mathbf{x})\right\rangle + \frac{\alpha^2 L}{2}\|\mathbf{A}\nabla f(\mathbf{x})\|_2^2.$$

Let λ_1 be the maximum eigenvalue of $\mathbf{A}$, λ_d be its minimum eigenvalue. Then, we have

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \alpha\lambda_d\|\nabla f(\mathbf{x})\|_2^2 + \frac{\alpha^2 L{\lambda_1}^2}{2}\|\nabla f(\mathbf{x})\|_2^2.$$

Thus, choosing α from the interval $(0, \frac{2\lambda_d}{{\lambda_1}^2 L})$ guarantees that we have a descent method. As a specific example, for $\lambda = \frac{\lambda_d}{{\lambda_1}^2 L}$, we have

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \frac{{\lambda_d}^2}{2{\lambda_1}^2 L}\|\nabla f(\mathbf{x})\|_2^2.$$

Newton's method, which takes $\mathbf{A} = \nabla^2 f(\mathbf{x})$ (in which case we are, of course, assuming, that f is twice differentiable and Hessian is positive definite), can be seen as a special preconditioned method. Proving its global convergence requires other regularity assumptions (such as strong convexity), in addition to smoothness, and for large target error $\epsilon > 0$, Newton's method is no faster than gradient descent. However, Newton's method, quasi-Newton methods (which use approximations of the Hessian), and other preconditioned methods can have very fast *local* convergence, which is seen once their iterates are sufficiently close to a solution. We will learn more about this in the second part of the semester when we start talking about second-order methods.

Example 3.2 (Gauss-Southwell Rule — Greedy Coordinate Descent)**.** In greedy coordinate descent, the descent direction at $\mathbf{x}$ is chosen as $\mathbf{z} = -\nabla_{i^*} f(\mathbf{x})\mathbf{e}_{i^*}$, where $i^* = \mathrm{argmax}_{1\leq i\leq d}|\nabla_i f(\mathbf{x})|$ and $\mathbf{e}_i$ denotes the i^{th} standard basis vector. When f is L-smooth, we have

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \alpha|\nabla_{i^*} f(\mathbf{x})|^2 + \frac{\alpha^2 L}{2}|\nabla_{i^*} f(\mathbf{x})|^2.$$

Thus, for $\alpha \in (0, \frac{2}{L})$, we obtain a descent method. Specifically, for $\alpha = \frac{1}{L}$,

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \frac{1}{2L}\max_{1\leq i\leq d}|\nabla_i f(\mathbf{x})|^2.$$

Since $\max_{1\leq i\leq d}|\nabla_i f(\mathbf{x})|^2 \geq \frac{1}{d}\|\nabla f(\mathbf{x})\|_2^2$, we also have

$$f(\mathbf{x}+\alpha\mathbf{z}) \leq f(\mathbf{x}) - \frac{1}{2Ld}\|\nabla f(\mathbf{x})\|_2^2.$$

Looking at greedy coordinate descent, it appears to be a useless method: in general, to compute i^*, we would need to compute the full gradient, while the resulting descent progress $-\frac{1}{2Ld}\|\nabla f(\mathbf{x})\|_2^2$ is strictly worse than what we would have with gradient descent. Nevertheless, this method is sometimes employed in practice and it can be quite effective. The reason is that for some problems, i^* and $\nabla_{i^*} f(\mathbf{x})$ can be computed without looking at the full gradient and the required computation to do so is much lower (ideally, by a factor of order-d) than computing the full gradient. Then, as $\max_{1\leq i\leq d}|\nabla_i f(\mathbf{x})|^2$ can be much larger than $\frac{1}{d}\|\nabla f(\mathbf{x})\|_2^2$ (how much in the best case?), the algorithm can make more progress than gradient descent with the same amount of computation.

There are also other descent methods that one can consider, and two such examples (gradient descent in ℓ_p norms and randomized coordinate descent) are provided in Homework #2.

A common thread for the algorithms we have seen so far is that they can all guarantee progress of the form

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \frac{\beta}{2}\|\nabla f(\mathbf{x})\|_*^2, \tag{18}$$

for some $\beta > 0$, where $\|\cdot\|_*$ is the norm dual to the norm with respect to which f is L-smooth. Based on the analysis from the beginning of this note, it is immediate that condition (1) is sufficient to guarantee that

$$\min_{0\leq i\leq k} \|\nabla f(\mathbf{x}_i)\|_*^2 \leq \frac{2(f(\mathbf{x}_0) - f^*)}{\beta(k+1)},$$

whenever f is bounded below by $f^* > -\infty$.

However, the analysis we saw for the convex and the strongly convex cases of gradient descent does not immediately transfer to other descent methods. There, we crucially used that the update was of the form $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\nabla f(\mathbf{x}_k)$. In the next two sections, we show that it is still possible to obtain similar (though slightly weaker) guarantees as for gradient descent if we are only assuming that our method satisfies (1).

3.1 Convex Case

We now turn to the analysis of basic descent methods only assuming that the function f is L-smooth and we have sufficient descent progress of the form (18).

When we analyzed gradient descent and tried to bound the optimality gap at iteration k, we only used convexity that gave us a lower bounding hyperplane at point $\mathbf{x}_k$, that is, we used:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \rangle . \tag{19}$$

But up to iteration k our algorithm generates points $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k$, and so we have, $\forall i \in \{0, \ldots, k\}$:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle . \tag{20}$$

This allows us to potentially construct a better lower bound on $f(\mathbf{x}^*)$ than what we get by just using (19). In particular, we can take an arbitrary convex combination of the lower-bounding hyperplanes from (20). 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:

$$f(\mathbf{x}^*) \geq \frac{1}{A_k}\sum_{i=0}^{k} a_i\big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle\big), \tag{21}$$

and we will call the right-hand side of (21) 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 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 (18), and, thus, $f(\mathbf{x}_1) \leq f(\mathbf{x}_0) - \frac{\beta}{2}\|\nabla f(\mathbf{x}_0)\|_*^2$, for some $\beta > 0$. By the definition of A_k, we have $A_0 = a_0$. From the definition of G_k :

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

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

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

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 $p = \sqrt{\beta}\|\nabla f(\mathbf{x}_0)\|_*$ and $q = \|\mathbf{x}^* - \mathbf{x}_0\|/\sqrt{\beta}$ to get:

$$A_0 G_0 \leq \frac{a_0}{2\beta}\|\mathbf{x}^* - \mathbf{x}_0\|^2. \tag{22}$$

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

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

where we have used $f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{\beta}{2}\|\nabla f(\mathbf{x}_k)\|_2^2$, 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 $A_0 G_0$: duality of norms and then $-\frac{p^2}{2} + pq \leq \frac{q^2}{2}$ with $p = \sqrt{\beta A_k}\|\nabla f(\mathbf{x}_k)\|_*$, $q = \frac{a_k}{\sqrt{\beta A_k}}\|\mathbf{x}^* - \mathbf{x}_k\|$, to get:

$$A_k G_k - A_{k-1} G_{k-1} \leq \frac{{a_k}^2}{2\beta A_k}\|\mathbf{x}^* - \mathbf{x}_k\|^2. \tag{23}$$

Applying (23) recursively until we reach $k = 0$ and then using (22), we have:

$$A_k G_k \leq \sum_{i=0}^{k} \frac{{a_i}^2}{2\beta A_i}\|\mathbf{x}^* - \mathbf{x}_i\|^2. \tag{24}$$

Define: $R = \max\{\|\mathbf{x}^* - \mathbf{x}\| : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$, so that $\|\mathbf{x}^* - \mathbf{x}_i\|^2 \leq R^2$ (in general, this will be bounded, and if we are only assuming progress as in (18), we cannot do better). From our definition of the gap, we have:

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

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:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{2R^2}{\beta(k+2)}. \tag{25}$$

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 (GD) and take $\|\cdot\| = \|\cdot\|_2$, then we can do a little bit better. Namely, we can replace R in (25) with $\|\mathbf{x}^* - \mathbf{x}_0\|_2$. This is because we can show that $\forall k \geq 1 : \|\mathbf{x}^* - \mathbf{x}_k\|_2 \leq \|\mathbf{x}^* - \mathbf{x}_0\|_2$, as we saw in the analysis of gradient descent from the beginning of the note.

Thus, for standard gradient descent from (GD), this analysis gives:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq \frac{2\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{\alpha(k+2)}. \tag{26}$$

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

3.2 Strongly Convex Case

When f is μ-strongly convex for some $\mu > 0$, we can create an even better lower bound on $f(\mathbf{x}^*)$. In particular, instead of (9), we can use that, by strong convexity, $\forall i$:

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

Notice that here I was using the Euclidean norm. This is not by accident. It goes beyond the scope of this class to explain why, but you should know that if you are assuming that a function is both smooth and strongly convex, then the only norm that "makes sense" is the Euclidean norm. Any other norm would lead to a worse condition number $\kappa := \frac{L}{\mu}$, by a factor dependent on the dimension.

Going back to our discussion, we can construct the following lower bound, where, as before, we hold off on choosing a_i's and only assume they are positive:

$$f(\mathbf{x}^*) \geq L_k := \frac{1}{A_k}\sum_{i=0}^{k} a_i\Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i\rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_i\|_2^2\Big). \tag{27}$$

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 $A_0 = a_0$:

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

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

$$A_0 G_0 \leq \frac{a_0(\frac{1}{\beta} - \mu)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2}. \tag{28}$$

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

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

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

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

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

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

Thus:

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

In particular, if $\frac{a_k}{A_k} \leq \mu\beta$, we have that the rhs of the last inequality is non-positive, and, thus, $A_k G_k \leq A_0 G_0$. Using (28):

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq G_k \leq \frac{a_0(\frac{1}{\beta} - \mu)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2A_k}. \tag{29}$$

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 $\frac{a_k}{A_k} \leq \beta\mu$. It is not hard to see that the fastest growth for A_k (as $A_k = A_{k-1} + a_k$) is obtained for $\frac{a_k}{A_k} = \beta\mu$. In this case, $\frac{A_{k-1}}{A_k} = \frac{A_k - a_k}{A_k} = 1 - \beta\mu$, and we can write:

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

Combining with (29):

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (1-\beta\mu)^k\frac{(\frac{1}{\beta} - \mu)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2} = (1-\beta\mu)^{k+1}\frac{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2\beta}. \tag{30}$$

Did we really need strong convexity, or can we use something weaker? It turns out that a weaker condition than strong convexity suffices to obtain the same convergence bound as in (30). One such condition known as the Polyak-Łojasiewicz (PŁ) condition was introduced in previous lectures:

$$(\forall \mathbf{x} \in \mathbb{R}^d): \quad \|\nabla f(\mathbf{x})\|_2^2 \geq 2\mu(f(\mathbf{x}) - f(\mathbf{x}^*)). \tag{31}$$

An example of a function that satisfies the PŁ 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 PŁ condition holds for this function with μ being equal to the smallest nonzero eigenvalue of $\mathbf{A}$ (you can find the proof in the Appendix of the Wright-Recht book).

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 (30), 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?

3.3 Rate Comparison

We have shown the following for the basic descent methods that ensure the progress from (18), where, for simplicity, we take $\beta = \frac{1}{L}$ and only summarize results for the Euclidean norm:

1. If f is smooth, we have that $\min_{0\leq i\leq k}\|\nabla f(\mathbf{x}_k)\|_2 \leq \sqrt{\frac{2L(f(\mathbf{x}_0)-f(\mathbf{x}^*))}{k+1}}$. Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ with $\|\nabla f(\mathbf{x}_k)\|_2 \leq \epsilon$ 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 (25)) $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2LR^2}{k+1}$ or $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{k+1}$ (from (26), if we use steepest descent). Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ that satisfies $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon$ 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 (30), $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \left(1 - \frac{m}{L}\right)^k \frac{L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{2}$. Thus, for any $\epsilon > 0$, basic descent methods find a point $\mathbf{x}$ that satisfies $f(\mathbf{x}) - f(\mathbf{x}^*) \leq \epsilon$ after at most $k = O(\frac{L}{m}\log(\frac{L\|\mathbf{x}^*-\mathbf{x}_0\|_2^2}{\epsilon}))$ iterations.

Exercises

1. Recall that a function is convex and L-smooth w.r.t. $\|\cdot\|$ if and only if

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d): \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2 \leq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle. \tag{32}$$

Take $\|\cdot\| = \|\cdot\|_2$. Use (32) to argue that gradient descent (GD) with step size $\alpha = \frac{1}{L}$ guarantees that

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{1}{2L}\left(\|\nabla f(\mathbf{x}_k)\|_2^2 + \|\nabla f(\mathbf{x}_{k+1})\|_2^2\right). \tag{33}$$

2. Use (32) to show that if f is convex and L-smooth, we also have

$$(\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d): \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2. \tag{34}$$

Now take $\|\cdot\| = \|\cdot\|_2$ and argue that gradient descent with step size $\alpha \leq \frac{1}{L}$ in this case guarantees that

$$\|\nabla f(\mathbf{x}_{k+1})\|_2 \leq \|\nabla f(\mathbf{x}_k)\|_2. \tag{35}$$

3. Use the results from the previous two exercises to argue that (GD) with step size $\alpha = \frac{1}{L}$ when applied to a convex and L-smooth function guarantees that the following potential function

$$\mathcal{C}_k = \frac{k}{L}\|\nabla f(\mathbf{x}_k)\|_2^2 + f(\mathbf{x}_k)$$

is non-increasing with k. Conclude that

$$\|\nabla f(\mathbf{x}_k)\|_2^2 \leq \frac{L(f(\mathbf{x}_0) - f^*)}{k},$$

where $f^* > -\infty$ bounds below f, for any $\mathbf{x} \in \mathbb{R}^d$.

4. In this exercise, you are asked to obtain a more direct proof of convergence of basic descent methods under a PŁ condition, using only the PŁ condition of the form

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|_*^2 \tag{36}$$

and the bound

$$\min_{0 \leq i \leq k} \|\nabla f(\mathbf{x}_i)\|_*^2 \leq \frac{2(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\beta(k+1)}. \tag{37}$$

Combine (36) and (37) to conclude that within the first $k_1 = O(\frac{1}{\beta\mu})$ iterations, there exists an iteration i_1 such that

$$\|\nabla f(\mathbf{x}_{i_1})\|_*^2 \leq \frac{\|\nabla f(\mathbf{x}_0)\|_*^2}{2}.$$

Now consider using (37) with $\mathbf{x}_{i_1}$ as your initial point (this bound holds no matter where you start, due to the descent progress from (18)). Then, within the next $k_2 = O(\frac{1}{\beta\mu})$ iterations, there exists an iteration i_2 such that

$$\|\nabla f(\mathbf{x}_{i_2})\|_*^2 \leq \frac{\|\nabla f(\mathbf{x}_{i_1})\|_*^2}{2}. \tag{38}$$

Apply this argument recursively to argue that, for any $\epsilon > 0$, after $K = O(\frac{1}{\beta\mu})\log(\frac{\|\nabla f(\mathbf{x}_0)\|_*}{\epsilon})$ iterations there must exist at least one point $\mathbf{x}_i$, $i \in \{K - \frac{C}{\beta\mu}, K\}$, where C is an absolute constant, such that $\|\nabla f(\mathbf{x}_i)\|_* \leq \epsilon$.

5. Now consider gradient descent with step size $\alpha_k = \frac{1}{L}$ applied to an L-smooth convex function f. Let $\mathbf{x}^*$ be any minimizer of f. The convergence bound derived in the lecture for this case is

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k}. \tag{39}$$

Suppose now in addition that f is $(2, \mu)$-sharp, so that the following holds for any $\mathbf{x} \in \mathbb{R}^d$ and any $\mathbf{x}^* \in \mathcal{X}^*$, where $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\mu}{2}\operatorname{dist}(\mathbf{x}, \mathcal{X}^*)^2. \tag{40}$$

Use the strategy from the previous exercise to argue based on (39) and (40) that gradient descent in this case guarantees that, for any $\epsilon > 0$, $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$ after $k = O\big(\frac{L}{\mu}\log\big(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*)}{\epsilon}\big)\big)$ iterations.