CS 726: Nesterov Acceleration

Jelena Diakonikolas

Fall 2022

In this lecture note, we discuss Nesterov accelerated method, which achieves optimal convergence rates for the classes of smooth convex and smooth strongly convex functions. Throughout this part, we assume that in our problem (P) the objective function f is convex and L-smooth w.r.t. the Euclidean norm and the problem is unconstrained ($\mathcal{X} \equiv \mathbb{R}^d$). We further assume that f attains its minimum at some $\mathbf{x}^* \in \mathbb{R}^d$.

As before, we will base our analysis on an optimality gap estimate $G_k \geq f(\mathbf{x}_k^{\text{out}}) - f(\mathbf{x}^*)$, where $\mathbf{x}_k^{\text{out}}$ is the point that our method outputs at iteration k. This estimate will be constructed as $G_k = U_k - L_k$, where $U_k \geq f(\mathbf{x}_k^{\text{out}})$ and $L_k \leq f(\mathbf{x}^*)$, as discussed in previous lectures.

1 Smooth Convex Setting

We saw in previous lectures that basic descent methods when applied to smooth unconstrained convex problems reduce the optimality gap $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ with rate $1/k$. An immediate question is whether we can do better than this. It turns out that the answer is "yes" and the improvement is not so small: all other parameters being equal, the convergence rate can be improved to $1/k^2$ using celebrated Nesterov accelerated method, which we cover in this note.

When first learning about optimization, it is not so clear why we can even do better than basic descent. After all, in basic descent methods, we are allowed to perform a line search and reduce the function value as much as possible along the descent direction. Still, even with a line search, the worst-case convergence is no better than what we get with gradient descent: $f(\mathbf{x}_k) - f(\mathbf{x}^*) = O\big(\frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k}\big)$.

1.1 Improving the Lower Bound

To see where we could possibly improve, observe that basic descent methods such as gradient descent do not use past information to learn anything about the function. Instead, they are greedy in the sense that they only explore the descent direction at the current point. Thus, one avenue for improvement would be to somehow utilize all of the information about the function that is collected up to iteration k. This is what Nesterov accelerated method does, in addition to using a descent step.

In particular, recall that up to iteration k, any first-order method has at its disposal all previously queried points $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$ and the gradient and the function value at all those points. As we have previously discussed, convexity of f allows us to construct a lower bound on $f(\mathbf{x}^*)$: for an arbitrary sequence of non-negative numbers $\{a_i\}_{i \geq 0}$, where $a_0 > 0$, and $A_k = \sum_{i=0}^k$, we have

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

In the past lectures, we used (1) as a means of lower bounding $f(\mathbf{x}^*)$ to analyze the convergence of basic descent methods. But the algorithm itself had nothing to do with this lower bound: it was only the analysis that utilized it. In particular, we made no use of the lower bound to decide on which points to query or output next.

Looking at (1), it is not clear what would be a useful point to query next, especially given that we do not know $\mathbf{x}^*$. Our desiderata may be to maximize the lower bound, but given that we do not know $\mathbf{x}^*$, it is not clear how to do that. Another thing that may come to mind is looking at the "worst case" lower bound by getting rid of $\mathbf{x}^*$ altogether and

replacing the left-hand side of (1) by its minimum value taken over $\mathbf{x} \in \mathcal{X}$:

$$\min_{\mathbf{x}\in\mathcal{X}} \frac{1}{A_k}\sum_{i=0}^{k} a_i\big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i\rangle\big).$$

However, this also turns out to be impractical in unconstrained settings, as, unless $\sum_{i=0}^{k} a_i \nabla f(\mathbf{x}_i) = \mathbf{0}$, the minimum is $-\infty$, which means that the lower bound becomes uninformative.

However, not all is lost when thinking about a minimization approach; we just need to make sure that we do not end up with the trivial lower bound $f(\mathbf{x}^*) \geq -\infty$. The key idea (which is by no means trivial) is to add a quadratic[1] function to the left-hand side of (1) before replacing $\mathbf{x}^*$ by the minimizer of the resulting function. In particular, we can consider the following:

$$\begin{aligned} f(\mathbf{x}^*) &\geq \frac{1}{A_k}\sum_{i=0}^{k} a_i\big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i\rangle\big) + \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \\ &\geq \frac{1}{A_k}\min_{\mathbf{x}\in\mathbb{R}^d}\Big\{\sum_{i=0}^{k} a_i\big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i\rangle\big) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2\Big\} - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 =: L_k. \end{aligned} \tag{2}$$

The constructed lower bound (2) has the following advantages: the minimum in its definition is easily computable, the lower bound is finite, and the change in the scaled lower bound $A_kL_k - A_{k-1}L_{k-1}$ is controlled, in the sense that it doesn't change too much over iterations. To simplify the notation a bit, let us denote by $m_k(\mathbf{x})$ the function under the minimum in (2); that is

$$m_k(\mathbf{x}) := \sum_{i=0}^{k} a_i\big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i\rangle\big) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2. \tag{3}$$

Further, let

$$\mathbf{v}_k = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} m_k(\mathbf{x}) = \mathbf{x}_0 - \sum_{i=0}^{k} a_i \nabla f(\mathbf{x}_i) \tag{4}$$

so that now we can write

$$L_k = \frac{1}{A_k} m_k(\mathbf{v}_k) - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2.$$

Since our goal will be to ensure that $A_kG_k - A_{k-1}G_{k-1}$ is small (ideally zero), it makes sense to compute $A_kL_k - A_{k-1}L_{k-1}$. This is simply:

$$A_kL_k - A_{k-1}L_{k-1} = m_k(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1}).$$

Looking at the definition of $m_k(\mathbf{v}_k)$, we have that

$$m_k(\mathbf{v}_k) = m_{k-1}(\mathbf{v}_k) + a_k f(\mathbf{x}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle.$$

Further, as a quadratic function, m_{k-1} satisfies, $\forall \mathbf{x}, \mathbf{y}$,

$$m_{k-1}(\mathbf{y}) = m_{k-1}(\mathbf{x}) + \langle \nabla m_{k-1}(\mathbf{x}), \mathbf{x} - \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2.$$

Hence, we have:

$$\begin{aligned} A_kL_k - A_{k-1}L_{k-1} &= a_k f(\mathbf{x}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + m_{k-1}(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1}) \\ &= a_k f(\mathbf{x}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + \langle \nabla m_{k-1}(\mathbf{v}_{k-1}), \mathbf{v}_k - \mathbf{v}_{k-1}\rangle + \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \end{aligned}$$

But m_{k-1} is minimized by $\mathbf{v}_{k-1}$ (by the definition of $\mathbf{v}_{k-1}$) and so it must be $\nabla m_{k-1}(\mathbf{v}_{k-1}) = \mathbf{0}$. Thus, we can conclude that

$$A_kL_k - A_{k-1}L_{k-1} = a_k f(\mathbf{x}_k) + a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{5}$$

[1] It is possible to also add a different convex and "sufficiently curved" function, but quadratics are good enough for now.

1.2 Upper Bound

We will take $\mathbf{y}_k$ to be the output point at iteration k (slightly easier for typing that $\mathbf{x}_k^{\text{out}}$; also the standard notation for this type of methods). Our "upper bound" at iteration k is going to be defined by simply

$$U_k = f(\mathbf{y}_k). \tag{6}$$

We have not yet defined our query points $\mathbf{x}_k$, but we will define here $\mathbf{y}_k$ as simply taking a gradient descent step from $\mathbf{x}_k$:

$$\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k). \tag{7}$$

This will simplify some of our calculations.

Observe that

$$A_k U_k - A_{k-1}U_{k-1} = A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}). \tag{8}$$

1.3 The Change in the Scaled Gap

Since we have formally defined the upper and the lower bound, we can now focus on bounding the change in $A_k G_k$. In particular, combining (5) and (8), we have

$$A_k G_k - A_{k-1}G_{k-1} = A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}) - a_k f(\mathbf{x}_k) - a_k \left\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \right\rangle - \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{9}$$

Observe that, by the definition of $\mathbf{y}_k$, we have (by the sufficient decent property of gradient descent) that

$$f(\mathbf{y}_k) \leq f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Thus, plugging this inequality into (9), we have

$$A_k G_k - A_{k-1}G_{k-1} \leq -\frac{A_k}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})) - a_k \left\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \right\rangle - \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{10}$$

By convexity of f, $f(\mathbf{x}_k) - f(\mathbf{y}_{k-1}) \leq \left\langle \nabla f(\mathbf{x}_k), \mathbf{y}_{k-1} - \mathbf{x}_k \right\rangle$, and so we can further simplify (10) to

$$A_k G_k - A_{k-1}G_{k-1} \leq -\frac{A_k}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + \left\langle \nabla f(\mathbf{x}_k), A_k\mathbf{x}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{v}_k \right\rangle - \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{11}$$

Now, if we could set $A_k\mathbf{x}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{v}_k = \mathbf{0}$, we would be done. In this case $A_k G_k - A_{k-1}G_{k-1}$ would be negative, irrespective of how fast A_k grows, so the method would be arbitrarily fast. This, of course, is not possible. The reason is that $\mathbf{v}_k$ depends on $\mathbf{x}_k$ and thus we cannot explicitly define $\mathbf{x}_k$ to make $A_k\mathbf{x}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{v}_k = \mathbf{0}$ hold. Instead, we choose

$$\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}. \tag{12}$$

This, in turn, gives us:

$$A_k G_k - A_{k-1}G_{k-1} \leq -\frac{A_k}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2 + a_k \left\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{v}_{k-1} \right\rangle - \frac{1}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{13}$$

To complete bounding $A_k G_k - A_{k-1}G_{k-1}$, it remains to use the definition of $\mathbf{v}_k$ from (4), by which we have $\mathbf{v}_k = \mathbf{v}_{k-1} - a_k\nabla f(\mathbf{x}_k)$. Plugging this into (13), we now finally get

$$A_k G_k - A_{k-1}G_{k-1} \leq \left(\frac{{a_k}^2}{2} - \frac{A_k}{2L}\right)\|\nabla f(\mathbf{x}_k)\|_2^2. \tag{14}$$

Thus, to ensure that $A_k G_k - A_{k-1}G_{k-1} \leq 0$, it suffices to have

$$\frac{{a_k}^2}{A_k} \leq \frac{1}{L}. \tag{15}$$

Nesterov accelerated method can now be summarized as performing the following updates for $k \geq 1$,

$$\begin{aligned} \mathbf{x}_k &= \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1} \\ \mathbf{v}_k &= \mathbf{v}_{k-1} - a_k \nabla f(\mathbf{x}_k), \\ \mathbf{y}_k &= \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k), \end{aligned} \tag{16}$$

where we require that $a_k > 0$ and (15) holds.

To fully specify the method, it remains to discuss the initialization. We do so by bounding the initial gap.

1.4 Bounding the Initial Gap

By definitions of lower and upper bounds in (2) and (6), we have

$$A_0 G_0 = A_0\Big(f(\mathbf{y}_0) - f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{v}_0 - \mathbf{x}_0\rangle - \frac{1}{2}\|\mathbf{v}_0 - \mathbf{x}_0\|_2^2\Big) + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2,$$

where $\mathbf{v}_0 = \operatorname{argmin}_{\mathbf{x}} m_0(\mathbf{x}) = \mathbf{x}_0 - a_0 \nabla f(\mathbf{x}_0)$. Taking $\mathbf{x}_0$ to be arbitrary and setting $\mathbf{y}_0 = \mathbf{x}_0 - \frac{1}{L}\nabla f(\mathbf{x}_0)$ immediately ensures that

$$A_0 G_0 \leq \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2, \tag{17}$$

provided that $a_0 \leq \frac{1}{L}$, by the same arguments as in bounding $A_k G_k - A_{k-1} G_{k-1}$. Thus, this is the choice of initialization that we make.

1.5 Putting Everything Together

We can now fully describe Nesterov's algorithm, summarized in Algorithm 1.

Algorithm 1 Nesterov-Smooth($\mathbf{x}_0, f, L, K$)

Initialization: $a_0 = A_0 = \frac{1}{L}$, $\mathbf{v}_0 = \mathbf{x}_0 - a_0 \nabla f(\mathbf{x}_0)$, $\mathbf{y}_0 = \mathbf{x}_0 - \frac{1}{L}\nabla f(\mathbf{x}_0)$
for $k = 1$ to K **do**
  $a_k = \frac{k+2}{L}$, $A_k = a_k + A_{k-1}$
  $\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}$
  $\mathbf{v}_k = \mathbf{v}_{k-1} - a_k \nabla f(\mathbf{x}_k)$
  $\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$
end for
return $\mathbf{y}_K$

Combining all the results from previous sections, we have that if $a_0 = A_0 \leq \frac{1}{L}$ and $\frac{a_k^2}{A_k} \leq \frac{1}{L}$ for $k \geq 1$, we have

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq G_k \leq \frac{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2A_k}.$$

It remains to choose the sequence a_k (and consequently A_k). It is not hard to verify that $a_k \propto \frac{k}{L}$ satisfies the requirements set above. In particular, choosing $a_k = \frac{k+2}{2L}$, we have $A_k = \frac{(k+1)(k+4)}{4L}$ and both $a_0 \leq \frac{1}{L}$ and $\frac{{a_k}^2}{A_k}$ for $k \geq 1$ hold. Thus, under this choice, the algorithm is now fully specified and we get

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq G_k \leq \frac{2L\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{(k+1)(k+4)}. \tag{18}$$

2 Smooth and Strongly Convex Optimization

You may be wondering if the algorithm we saw for the smooth convex case already gives faster, geometric (in optimization literature known as linear) convergence as gradient descent did. It turns out that this is not the case. If we just apply Nesterov algorithm (16) for smooth convex optimization to a smooth strongly convex problem, the convergence is known not to be faster than order-$1/k^3$. This is much slower than the convergence of the order $\exp(-\frac{\mu}{L}k)$ that we had for gradient descent, where L is the smoothness constant and μ is the modulus of strong convexity.

The phenomenon of acceleration, however, is not restricted to classes of smooth convex problems, and obtaining convergence that is faster than for standard gradient descent is possible also for smooth strongly convex problems. In the following, we will see two ways of obtaining faster convergence for such problems, by building on the results from the previous section.

2.1 Acceleration via Restarting

The original way of accelerating convergence when strong convexity holds is using restarting. We have already hinted on this idea in the exercises at the end of the last lecture note, though we did not need explicitly restart there. The basic idea comes from observing that strong convexity guarantees that

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \geq \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^*\|_2^2. \tag{19}$$

In fact, even a weaker property (sharpness) would give us the same inequality, which turns out to be sufficient for acceleration (assuming we have convexity and smoothness).

Combining (18) with (19), we have that after k iterations of Nesterov algorithm for smooth convex minimization,

$$\frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^*\|_2^2 \leq \frac{2L\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{(k+1)(k+4)}. \tag{20}$$

We can conclude from (20) that for $k \geq \lfloor\sqrt{\frac{8L}{\mu}}\rfloor$, we have that $\|\mathbf{y}_k - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2$. In other words, after the number of iterations of Nesterov algorithm for smooth convex optimization (Algorithm 1) that scales with $\sqrt{\frac{L}{\mu}}$, the distance to (any fixed, and thus the closest) optimal solution drops by a factor of two. Thus, the idea is to repeatedly call Algorithm 1 for $K = \lfloor\sqrt{\frac{8L}{\mu}}\rfloor$ iterations, and each time initialize it with the output of the previous call. Since each call to the algorithm halves the distance to the set of optimal solutions, we have that order-$\log(\|\mathbf{x}^* - \mathbf{x}_0\|_2/\epsilon)$ calls suffice to ensure the distance to optimum is of the order ϵ. This strategy is summarized in Algorithm 2 and formally analyzed in Lemma 2.1.

Algorithm 2 Nesterov-Restart($\mathbf{x}_0, f, L, \mu, R$)

1: **Initialization:** $\bar{\mathbf{x}}_0 = \mathbf{x}_0$
2: **for** $r = 1$ to R **do**
3: $\quad \bar{\mathbf{x}}_r = $ Nesterov-Smooth($\bar{\mathbf{x}}_{r-1}, f, L, \lfloor\sqrt{\frac{8L}{\mu}}\rfloor$)
4: **end for**
5: **return** $\bar{\mathbf{x}}_R$

Lemma 2.1. *Let f be a smooth, convex, and $(2,\mu)$-sharp function. Let $\mathcal{X}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Given an arbitrary $\mathbf{x}_0 \in \mathbb{R}^d$, consider running Algorithm 2 for $R \geq 1$ iterations. Then, given $\epsilon > 0$, if $R \geq 2\log_2(\frac{\operatorname{dist}(\mathbf{x}_0,\mathcal{X}^*)}{\epsilon})$, we have that*

$$\operatorname{dist}(\bar{\mathbf{x}}_R, \mathcal{X}^*) \leq \epsilon.$$

The total number of vector operations (such as addition, scalar multiplication, and gradient evaluation) is

$$O\left(\sqrt{\frac{L}{\mu}}\log_2\left(\frac{\operatorname{dist}(\mathbf{x}_0,\mathcal{X}^*)}{\epsilon}\right)\right).$$

Proof. Since f is $(2,\mu)$-sharp, we have that for any $\mathbf{x} \in \mathbb{R}^d$ and any $\mathbf{x}^* \in \mathcal{X}^*$,

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\mu}{2} \operatorname{dist}(\mathbf{x}, \mathcal{X}^*)^2. \tag{21}$$

On the other hand, the guarantee of Nesterov algorithm from (18) holds for an arbitrary $\mathbf{x}^* \in \mathcal{X}^*$. Thus, we can conclude that for any $\mathbf{x}^* \in \mathcal{X}^*$ and any $r \geq 1$,

$$f(\bar{\mathbf{x}}_r) - f(\mathbf{x}^*) \leq \frac{2L \operatorname{dist}(\bar{\mathbf{x}}_{r-1}, \mathcal{X}^*)^2}{(k+1)(k+4)}, \tag{22}$$

where $k = \lfloor \sqrt{\frac{8L}{\mu}} \rfloor$. Under this choice of k, we have that $\frac{2L \operatorname{dist}(\bar{\mathbf{x}}_{r-1}, \mathcal{X}^*)^2}{(k+1)(k+4)} \leq \frac{\operatorname{dist}(\bar{\mathbf{x}}_{r-1}, \mathcal{X}^*)^2}{2}$. Hence, combining with (21) and (22), we get that for all $r \geq 1$,

$$\operatorname{dist}(\bar{\mathbf{x}}_r, \mathcal{X}^*)^2 \leq \frac{1}{2} \operatorname{dist}(\bar{\mathbf{x}}_{r-1}, \mathcal{X}^*)^2. \tag{23}$$

Applying (23) recursively between 0 and R, we can conclude that

$$\operatorname{dist}(\bar{\mathbf{x}}_R, \mathcal{X}^*)^2 \leq \frac{1}{2^R} \operatorname{dist}(\bar{\mathbf{x}}_0, \mathcal{X}^*)^2 = \frac{1}{2^R} \operatorname{dist}(\mathbf{x}_0, \mathcal{X}^*)^2. \tag{24}$$

The right-hand side of (24) is at most ϵ^2 if $R \geq \log_2\left(\frac{\operatorname{dist}(\mathbf{x}_0, \mathcal{X}^*)^2}{\epsilon^2}\right) = 2\log_2\left(\frac{\operatorname{dist}(\mathbf{x}_0, \mathcal{X}^*)}{\epsilon}\right)$. Hence, from (24), if $R \geq 2\log_2\left(\frac{\operatorname{dist}(\mathbf{x}_0, \mathcal{X}^*)}{\epsilon}\right)$, we have that $\operatorname{dist}(\bar{\mathbf{x}}_R, \mathcal{X}^*) \leq \epsilon$, as claimed. As each iteration of Algorithm 2 (which consists of a call to Algorithm 1 with $O(\sqrt{\frac{L}{\mu}})$ iterations) takes $O(\sqrt{\frac{L}{\mu}})$ vector operations, we get the claimed bound on the total number of vector operations that Algorithm 2 makes. □

2.2 A Direct Algorithm

You may be wondering at this point whether restarting was required for achieving acceleration in the smooth and strongly convex (or sharp) case. The answer is no, and we can obtain a direct algorithm by appropriately modifying our analysis from Section 1 to account for strong convexity (or sharpness). For concreteness, we will carry out the analysis assuming that f is μ-strongly convex, though it is possible to extend this analysis to the case where f is just convex and sharp, as the only inequality that comes from strong convexity that we will be using is that $\forall \mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2. \tag{25}$$

The method we will obtain will look very similar to the method we obtained for the smooth convex case (Algorithm 1), with the primary difference being in how we define $\mathbf{v}_k$ (as the lower bound L_k will take advantage of strong convexity) and the sequences a_k, A_k (due to strong convexity, these sequences will be allowed grow much faster).

Lower Bound Construction. Given a sequence of gradient query points $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$ and positive numbers $\{a_i\}_{i \geq 0}$, $A_k = \sum_{i=0}^k a_i$, strong convexity (25) guarantees that

$$f(\mathbf{x}^*) \geq \frac{1}{A_k} \sum_{i=0}^k a_i \Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle + \frac{\mu}{2}\|\mathbf{x}_i - \mathbf{x}^*\|_2^2 \Big). \tag{26}$$

Similarly as in the smooth convex case, we add and subtract $\frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2$ from the right-hand side before taking a minimum. This is not necessary and we could just immediately replace $\mathbf{x}^*$ with the minimizer of the right-hand side of (26); the only reason for introducing $\frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2$ is that makes the computation of the initial lower bound (and the initial gap estimate G_0) slightly nicer. In particular, we have that

$$\begin{aligned} f(\mathbf{x}^*) &\geq \frac{1}{A_k} \sum_{i=0}^k a_i \Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x}^* - \mathbf{x}_i \rangle + \frac{\mu}{2}\|\mathbf{x}_i - \mathbf{x}^*\|_2^2 \Big) + \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2 \\ &\geq \frac{1}{A_k} \min_{\mathbf{x} \in \mathbb{R}^d} \Big\{ \sum_{i=0}^k a_i \Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i \rangle + \frac{\mu}{2}\|\mathbf{x}_i - \mathbf{x}\|_2^2 \Big) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \Big\} - \frac{1}{2A_k}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2. \end{aligned} \tag{27}$$

Similarly as before, to simplify the notation, we define:

$$m_k(\mathbf{x}) := \sum_{i=0}^{k} a_i \Big(f(\mathbf{x}_i) + \langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i \rangle + \frac{\mu}{2} \|\mathbf{x}_i - \mathbf{x}\|_2^2 \Big) + \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \tag{28}$$

and

$$\begin{aligned} \mathbf{v}_k &:= \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} m_k(\mathbf{x}) \\ &= \frac{\mathbf{x}_0 + \mu \sum_{i=0}^{k} a_i \mathbf{x}_i - \sum_{i=0}^{k} a_i \nabla f(\mathbf{x}_i)}{1 + \mu A_k} \\ &= \frac{(1 + \mu A_{k-1}) \mathbf{v}_{k-1} + a_k (\mu \mathbf{x}_k - \nabla f(\mathbf{x}_k))}{1 + \mu A_k}, \end{aligned} \tag{29}$$

so that we can define:

$$L_k := \frac{1}{A_k} m_k(\mathbf{v}_k) - \frac{1}{2A_k} \|\mathbf{x}^* - \mathbf{x}_0\|_2^2. \tag{30}$$

Observe that, due to (27), $L_k \leq f(\mathbf{x}^*)$, thus L_k is a valid lower bound.

Since we will later be bounding the change in $A_k G_k$, it is useful to bound the change in $A_k L_k$, as follows. By the definition of L_k from (30), we have

$$A_k L_k - A_{k-1} L_{k-1} = m_k(\mathbf{v}_k) - m_{k-1}(\mathbf{v}_{k-1}). \tag{31}$$

By the definition of m_k, we have

$$m_k(\mathbf{v}_k) = m_{k-1}(\mathbf{v}_k) + a_k \Big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{v}_k - \mathbf{x}_k\|_2^2 \Big). \tag{32}$$

Further, as m_{k-1} is $(1 + \mu A_{k-1})$-strongly convex and minimized by $\mathbf{v}_{k-1}$, we have

$$m_{k-1}(\mathbf{v}_k) \geq m_{k-1}(\mathbf{v}_{k-1}) + \frac{1 + \mu A_{k-1}}{2} \|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{33}$$

Combining (31)–(33), we get

$$A_k L_k - A_{k-1} L_{k-1} = a_k \Big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle + \frac{\mu}{2} \|\mathbf{v}_k - \mathbf{x}_k\|_2^2 \Big) + \frac{1 + \mu A_{k-1}}{2} \|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2. \tag{34}$$

Applying Jensen's inequality to combine the quadratic terms in (34), we get

$$\begin{aligned} A_k L_k - A_{k-1} L_{k-1} &\geq a_k \big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle \big) + \frac{1 + \mu A_k}{2} \Big\| \mathbf{v}_k - \frac{1 + \mu A_{k-1}}{1 + \mu A_k} \mathbf{v}_{k-1} - \frac{\mu a_k}{1 + \mu A_k} \mathbf{x}_k \Big\|_2^2 \\ &= a_k \big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle \big) + \frac{{a_k}^2}{2(1 + \mu A_k)} \|\nabla f(\mathbf{x}_k)\|_2^2, \end{aligned} \tag{35}$$

where the last equality is by (29).

Upper Bound. Same as for the smooth convex case, we define $\mathbf{y}_k := \mathbf{x}_k - \frac{1}{L} \mathbf{x}_k$ to be the output point at iteration k and $U_k := f(\mathbf{y}_k)$. By the sufficient decrease property of the gradient descent step, this gives us

$$U_k = f(\mathbf{y}_k) \leq f(\mathbf{x}_k) - \frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|_2^2. \tag{36}$$

Hence, the change in the scaled upper bound simplifies to

$$\begin{aligned} A_k U_k &= A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}) \\ &\leq A_k f(\mathbf{x}_k) - A_{k-1} f(\mathbf{y}_{k-1}) - \frac{A_k}{2L} \|\nabla f(\mathbf{x}_k)\|_2^2. \end{aligned} \tag{37}$$

Change in the Scaled Gap. Combining (35) and (37), we can conclude that

$$A_kG_k - A_{k-1}G_{k-1} \leq A_{k-1}(f(\mathbf{x}_k) - f(\mathbf{y}_{k-1})) - a_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k \rangle - \Big(\frac{{a_k}^2}{2(1+\mu A_k)} + \frac{A_k}{2L}\Big)\|\nabla f(\mathbf{x}_k)\|_2^2.$$

As f is convex, we have that $f(\mathbf{x}_k) - f(\mathbf{y}_{k-1}) \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_{k-1} \rangle$, and hence we get that

$$A_kG_k - A_{k-1}G_{k-1} \leq \langle \nabla f(\mathbf{x}_k), A_k\mathbf{x}_k - A_{k-1}\mathbf{y}_{k-1} - a_k\mathbf{v}_k \rangle - \Big(\frac{{a_k}^2}{2(1+\mu A_k)} + \frac{A_k}{2L}\Big)\|\nabla f(\mathbf{x}_k)\|_2^2.$$

Plugging in the definition of $\mathbf{v}_k$ from (29) into the last inequality, it follows that

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

Hence, choosing

$$\mathbf{x}_k = \frac{A_{k-1}}{A_k - \frac{\mu {a_k}^2}{1+\mu A_k}}\mathbf{y}_{k-1} + \frac{a_k(1+\mu A_{k-1})}{(1+\mu A_k)(A_k - \frac{\mu {a_k}^2}{1+\mu A_k})}\mathbf{v}_{k-1} \tag{38}$$

and ensuring that

$$\frac{{a_k}^2}{A_k(1+\mu A_k)} \leq \frac{1}{L} \tag{39}$$

guarantees that

$$A_kG_k - A_{k-1}G_{k-1} \leq 0.$$

Bounding the Initial Gap. To complete the argument (and the definition of the algorithm), it remains to bound the initial gap. By the definitions of U_k and L_k from (25) and (36), we have that

$$A_0G_0 \leq A_0\Big(-\frac{1}{2L}\|\nabla f(\mathbf{x}_0)\|_2^2 - \langle \nabla f(\mathbf{x}_0), \mathbf{v}_0 - \mathbf{x}_0 \rangle - \frac{\mu + 1/A_0}{2}\|\mathbf{v}_0 - \mathbf{x}_0\|_2^2\Big) + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2. \tag{40}$$

Since, by definition, $\mathbf{v}_0 = \frac{\mathbf{x}_0 + \mu a_0\mathbf{x}_0 - a_0\nabla f(\mathbf{x}_0)}{1+\mu A_0}$, (40) simplifies to

$$A_0G_0 \leq A_0\Big(-\frac{1}{2L} + \frac{a_0}{1+\mu A_0} - \frac{{a_0}^2}{2A_0(1+\mu A_0)}\Big)\|\nabla f(\mathbf{x}_0)\|_2^2 + \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2.$$

As $a_0 = A_0$ (by definition), it follows that it suffices to have

$$\frac{a_0}{1+\mu A_0} \leq \frac{1}{L}. \tag{41}$$

to guarantee that $A_0G_0 \leq \frac{1}{2}\|\mathbf{x}^* - \mathbf{x}_0\|_2^2$. Solving for a_0, (41) leads to

$$a_0 \leq \frac{1}{L-\mu}. \tag{42}$$

Algorithm 3 Nesterov-Smooth-Strongly-Convex($\mathbf{x}_0, f, L, \mu, K$)

Initialization: $a_0 = A_0 = \frac{1}{L-\mu}$, $\mathbf{v}_0 = \frac{\mathbf{x}_0+\mu a_0 \mathbf{x}_0 - a_0 \nabla f(\mathbf{x}_0)}{1+\mu A_0}$, $\mathbf{y}_0 = \mathbf{x}_0 - \frac{1}{L}\nabla f(\mathbf{x}_0)$
for $k = 1$ to K **do**
  Let a_k be the positive solution to $\frac{{a_k}^2}{(A_{k-1}+a_k)(1+\mu(A_{k-1}+a_k))} = \frac{1}{L}$, $A_k = a_k + A_{k-1}$
  $\mathbf{x}_k = \frac{A_{k-1}}{A_k(1-\mu/L)}\mathbf{y}_{k-1} + \left(1 - \frac{A_{k-1}}{A_k(1-\mu/L)}\right)\mathbf{v}_{k-1}$
  $\mathbf{v}_k = \frac{(1+\mu A_{k-1})\mathbf{v}_{k-1} + a_k(\mu \mathbf{x}_k - \nabla f(\mathbf{x}_k))}{1+\mu A_k}$
  $\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$
end for
return $\mathbf{y}_K$

Putting Everything Together. It remains to summarize the algorithm and its convergence guarantee using what we have shown so far. The largest value of A_k that we can achieve is obtained by making the inequalities (42) and (39) tight. This somewhat simplifies the expression for $\mathbf{x}_k$ and the full algorithm is provided in Algorithm 3.

We have already argued that for this algorithm, we have, for all $k \geq 1$,

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq G_k \leq \frac{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2A_k}, \tag{43}$$

so it just remains to bound the growth of A_k.

To avoid dealing with possibly infinite values, we will assume that $L > m$ and set $a_0 = \frac{1}{L-\mu}$. For $k \geq 1$, the algorithm ensures that $\frac{{a_k}^2}{A_k(1+\mu A_k)} = \frac{1}{L}$. Because $1+\mu A_k > \mu A_k$, we have that $\frac{{a_k}^2}{{A_k}^2} > \frac{\mu}{L}$, or, equivalently, $\frac{a_k}{A_k} > \sqrt{\frac{\mu}{L}}$. Now using that $A_k = A_{k-1} + a_k$, this gives us $a_k > \frac{\sqrt{\mu/L}}{1-\sqrt{\mu/L}}A_{k-1}$

$$A_k = A_{k-1} + a_k > A_{k-1}\left(1 + \frac{\sqrt{\mu/L}}{1-\sqrt{\mu/L}}\right) = \frac{A_{k-1}}{1-\sqrt{\mu/L}}.$$

Applying the last inequality recursively, we get

$$A_k > \frac{A_0}{(1-\sqrt{\mu/L})^k} = \frac{1}{(L-\mu)(1-\sqrt{\mu/L})^k}.$$

Hence, we can conclude that

$$f(\mathbf{y}_k) - f(\mathbf{x}^*) < \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \frac{(L-\mu)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2}. \tag{44}$$

From (44), we can now further conclude that for $k \geq \sqrt{\frac{L}{\mu}} \log\left(\frac{(L-\mu)\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}{2\epsilon}\right)$, we have $f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq \epsilon$. How does this bound compare to what we have obtained for the restarted algorithm in the previous subsection?

Exercises

1. There are many alternative ways of analyzing Nesterov algorithms, and in this exercise you will see one such example. Consider Nesterov algorithm for smooth convex optimization specified in Algorithm 1 applied to an L-smooth convex function f and let $\mathbf{x}^*$ be an arbitrary minimizer of f. Recall from Exercise 9 in the first lecture note that a function f is L-smooth and convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \leq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle .$$

Prove that the following potential function

$$\mathcal{C}_k = \sum_{i=0}^{k} \frac{A_i}{2}\|\nabla f(\mathbf{x}_i)\|_2^2 + LA_k(f(\mathbf{x}_k) - f(\mathbf{x}^*)) + \frac{L}{2}\|\mathbf{v}_k - \mathbf{x}^*\|_2^2$$

is non-increasing. Based on the non-increasing property of $\mathcal{C}_k$, conclude that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = O\Big(\frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k^2}\Big)$$

and, further,

$$\min_{0\le i\le k} \|\nabla f(\mathbf{x}_i)\|_2^2 = O\Big(\frac{L^2\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{k^3}\Big). \tag{45}$$

Compare these bounds to what you would get from standard gradient descent.

2. Suppose that you have a function f that is L-smooth and μ-strongly convex. Consider the following restarting approach. You modify Nesterov method for smooth convex optimization (Algorithm 1) so that it halts for a specified gradient norm, as summarized in Algorithm 4 below. You then run the (adaptive) restarting method described by

Algorithm 4 Nesterov-Smooth-Mod($\mathbf{x}_0, f, L, \epsilon$)

Initialization: $a_0 = A_0 = \frac{1}{L}$, $\mathbf{v}_0 = \mathbf{x}_0 - a_0\nabla f(\mathbf{x}_0)$, $\mathbf{y}_0 = \mathbf{x}_0 - \frac{1}{L}\nabla f(\mathbf{x}_0)$, $k = 0$
while $\|\nabla f(\mathbf{x}_k)\|_2 > \epsilon$ **do**
  $k = k + 1$
  $a_k = \frac{k+2}{L}$, $A_k = a_k + A_{k-1}$
  $\mathbf{x}_k = \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_{k-1}$
  $\mathbf{v}_k = \mathbf{v}_{k-1} - a_k\nabla f(\mathbf{x}_k)$
  $\mathbf{y}_k = \mathbf{x}_k - \frac{1}{L}\nabla f(\mathbf{x}_k)$
end while
return $\mathbf{x}_k$

Algorithm 5, for a specified $\epsilon > 0$. Prove that the restarted method specified in Algorithm 5 performs at most

Algorithm 5 Nesterov-Restart-Mod($\mathbf{x}_0, f, L, \mu, \epsilon$)

Initialization: $\bar{\mathbf{x}}_0 = \mathbf{x}_0$, $r = 0$
while $\|\nabla f(\bar{\mathbf{x}}_r)\|_2 > \epsilon$ **do**
  $r = r + 1$
  $\bar{\mathbf{x}}_r$ = Nesterov-Smooth-Mod($\bar{\mathbf{x}}_{r-1}, f, L, \frac{1}{2}\|\nabla f(\bar{\mathbf{x}}_{r-1})\|_2$)
end while
return $\bar{\mathbf{x}}_r$

$O\Big(\big(\frac{L}{\mu}\big)^{2/3}\log\big(\frac{\|\nabla f(\mathbf{x}_0)\|_2}{\epsilon}\big)\Big)$ vector operations and outputs a point $\bar{\mathbf{x}}_r$ with $\|\nabla f(\bar{\mathbf{x}}_r)\|_2 \le \epsilon$.

3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex L-smooth function and let $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. Given a target error $\epsilon > 0$, consider applying the direct Nesterov method for smooth strongly convex optimization (Algorithm 3) to function $\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$, where $\mu = \frac{\epsilon}{\|\mathbf{x}^* - \mathbf{x}_0\|_2^2}$. Prove that this approach allows you to obtain $\mathbf{y}_k \in \mathbb{R}^d$ such that $f(\mathbf{y}_k) - f(\mathbf{x}^*) \le \epsilon$ with $O\Big(\sqrt{\frac{L}{\epsilon}}\|\mathbf{x}_0 - \mathbf{x}^*\|_2\log\big(\frac{L\|\mathbf{x}^* - \mathbf{x}_0\|_2}{\epsilon}\big)\Big)$ vector operations. How does this compare to the guarantee of Algorithm 1?

Hint: Let $\bar{\mathbf{x}}^* = \operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^d}\bar{f}(\mathbf{x})$. Prove first that $\|\bar{\mathbf{x}}^* - \mathbf{x}_0\|_2 \le \|\mathbf{x}^* - \mathbf{x}_0\|_2$, starting from $f(\mathbf{x}^*) - f(\bar{\mathbf{x}}^*) \le 0$ and using the definition of $\bar{f}$.

4. Modify Nesterov method for L-smooth convex optimization (Algorithm 1) so that it becomes a descent method that makes at least as much progress per iteration as gradient descent. Argue that if, in addition, the input function turns out to be μ-strongly convex, your modified method, for any $\epsilon > 0$, outputs a point $\mathbf{y}_k$ such that $f(\mathbf{y}_k) - f(\mathbf{x}^*) \le \epsilon$ after at most

$$O\Bigg(\min\Bigg\{\sqrt{\frac{L}{\epsilon}}\|\mathbf{x}_0 - \mathbf{x}^*\|_2,\ \frac{L}{\mu}\log\Big(\frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2}{\mu\epsilon}\Big)\Bigg\}\Bigg)$$

iterations, where $\mathbf{x}^* \in \mathbb{R}^d$ minimizes f.