Nesterov's Lower Complexity Bounds

Michael L. Overton
following the derivation in Nesterov's book

Spring 2022

These notes follow Sec. 2.1.4 of Nesterov's book, giving only the main lower complexity bound results but with more details and a somewhat different notation. We assume as before that f is strongly convex and C^2 with parameters $M > m > 0$ (called L and μ by Nesterov) such that

$$mI \preceq \nabla^2 f(x) \preceq MI, \qquad \text{for all } x. \tag{1}$$

Before embarking on the lower complexity bounds, let's recall from the previous lecture that setting $t = 1/M$, we obtained[1]

$$f(x^{(k)}) - p^* \le \left(1 - \frac{1}{\kappa}\right)^k \left(f(x^{(0)}) - p^*\right),$$

where $\kappa = M/m$ and p^* is the minimal value of f. In his book, Nesterov derives yet another complexity result for the gradient method, this time using $t = 2/(m + M)$, and showing, in his Thm 2.1.15, that

$$\|x^{(k)} - x^*\| \le \left(\frac{1 - 1/\kappa}{1 + 1/\kappa}\right)^k \|x^{(0)} - x^*\|,$$

where x^* is the minimizer, and, using his Thm 2.1.8, that

$$f(x^{(k)}) - p^* \le \kappa \left(\frac{1 - 1/\kappa}{1 + 1/\kappa}\right)^{2k} \left(f(x^{(0)}) - p^*\right). \tag{2}$$

[1] Since we don't usually know M in practice, we also showed a slightly weaker result obtained using a backtracking line search.

Is this better or worse than the result we derived? It is *worse* for small k if κ is large, but no matter how large κ is, it will eventually be *better* for sufficiently large k.[2]

Now, let us move on to Nesterov's very interesting lower complexity bounds.

Assumption A. We use a method generating iterates $x^{(k)}$ for which a "first-order oracle" or "black box" computes $f(x^{(k)})$ and $\nabla f(x^{(k)})$, and assume that

$$x^{(k)} - x^{(0)} \in \text{Span}\left(\nabla f(x^{(0)}), \ldots, \nabla f(x^{(k-1)})\right).$$

For simplicity we assume

$$\text{dom} f = \mathbb{R}^\infty = \ell_2 = \left\{ x = [x_1, x_2, ...]^T : \|x\|^2 = \sum_{i=1}^{\infty} x_i^2 < \infty \right\}.$$

Let e_k denote the kth coordinate vector $[0, \ldots, 0, 1, 0, \ldots]^T$ in this space.

Now we define a "difficult" quadratic function F by

$$F(x) = \frac{M-m}{8}\left(x_1^2 + \sum_{i=1}^{\infty}(x_i - x_{i+1})^2 - 2x_1\right) + \frac{m}{2}\|x\|^2.$$

We have

$$\begin{aligned}\frac{\partial F}{\partial x_1} &= \frac{M-m}{8}\Big(2x_1 + 2(x_1 - x_2) - 2\Big) + mx_1 \\ &= \frac{M-m}{8}\Big(4x_1 - 2x_2 - 2\Big) + mx_1\end{aligned}$$

and, for $j = 2, 3, \ldots$,

$$\begin{aligned}\frac{\partial F}{\partial x_j} &= \frac{M-m}{8}\Big(2(x_j - x_{j+1}) - 2(x_{j-1} - x_j)\Big) + mx_j \\ &= \frac{M-m}{8}\Big(4x_j - 2x_{j+1} - 2x_{j-1})\Big) + mx_j.\end{aligned}$$

It follows that

$$\nabla^2 F(x) = \frac{M-m}{4} T + mI \equiv \frac{M-m}{4}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} + mI$$

[2] Both BV and Nesterov derive gradient complexity results for the non-strongly-convex case too, but these are much weaker, and we will not discuss them.)

with $mI \preceq \nabla^2 F(x) \preceq MI$ as required, because the tridiagonal matrix T and also $4I - T$ are symmetric and diagonally dominant – see pp. 6–7. Also

$$\nabla F(x) = \left(\frac{M-m}{4} T + mI \right) x - \frac{M-m}{4} e_1.$$

The solution x^* is defined by $\nabla F(x^*) = 0$, i.e., writing $x = x^*$,

$$\frac{M-m}{4}(2x_1 - x_2) + mx_1 = \frac{M-m}{4},$$

so

$$x_2 - 2\frac{M+m}{M-m} x_1 + 1 = 0,$$

and, for $j = 2, 3, \ldots$,

$$\frac{M-m}{4}(-x_{j-1} + 2x_j - x_{j+1}) + mx_j = 0$$

so

$$x_{j+1} - 2\frac{M+m}{M-m} x_j + x_{j-1} = 0.$$

This *second-order difference equation* can be solved by substituting $x_j = q^j$ and solving for $q \in \mathbb{R}$. We get

$$q^{j+1} - 2\frac{M+m}{M-m} q^j + q^{j-1} = 0, \qquad j = 2, 3, \ldots$$

implying

$$q^2 - 2\frac{M+m}{M-m} q + 1 = 0, \qquad j = 2, 3, \ldots,$$

which also holds for $j = 1$ (see above). We claim that the roots of this quadratic are

$$\frac{M + m \pm 2\sqrt{M}\sqrt{m}}{M-m}.$$

We can check this by noting that the sum of these roots is then $2\frac{M+m}{M-m}$, as desired, and their product is

$$\frac{(M+m)^2 - 4Mm}{(M-m)^2} = 1,$$

as desired. We are interested in the *smaller* root:

$$\begin{aligned} q &= \frac{M+m-2\sqrt{M}\sqrt{m}}{M-m} \\ &= \frac{\left(\sqrt{M}-\sqrt{m}\right)^2}{\left(\sqrt{M}-\sqrt{m}\right)\left(\sqrt{M}+\sqrt{m}\right)} \\ &= \frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}} \\ &= \frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}, \end{aligned}$$

where $\kappa = M/m$, the condition number of the convex function F. *Note the square root!*

Now we can prove

Theorem (Nesterov's Thm 2.1.13). For any $x^{(0)} \in \mathbb{R}^\infty$ and any $M > m > 0$, there exists a quadratic function F satisfying (1) such that, for methods satisfying Assumption A,

$$\|x^{(k)} - x^*\|^2 \geq \left(\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}\right)^{2k} \|x^{(0)} - x^*\|^2$$

where x^* minimizes F and $\kappa = M/m$, and furthermore

$$F(x^{(k)}) - p^* \geq \frac{m}{2}\left(\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}\right)^{2k} \|x^{(0)} - x^*\|^2.$$

Proof. WLOG we can take $x^{(0)} = 0$. Then we use F as already defined, getting

$$\|x^{(0)} - x^*\|^2 = \sum_{i=1}^{\infty}(x_i^*)^2 = \sum_{i=1}^{\infty} q^{2i} = \frac{q^2}{1-q^2}.$$

Now we claim that, for $j = 1, 2, \ldots$, we have

$$\nabla F(x^{(j-1)}) \in \operatorname{Span}(e_1, \ldots, e_j) \text{ and } x^j \in \operatorname{Span}(e_1, \ldots, e_j).$$

Let us prove this statement by induction. It holds for $j = 1$, because $\nabla F(0)$ is a multiple of e_1 and the iterate $x^{(1)}$ must be a multiple of $\nabla F(0)$ by

Assumption A. Now suppose the statement is true for iterate j; we want to prove it for iterate $j+1$. We know that $\nabla F(x^{(j)})$ is a linear combination of $Tx^{(j)}$, $x^{(j)}$ and e_1 so, since $x^{(j)}$ is a linear combination of $e_1, \ldots, e_j$ and T is tridiagonal, we see that $\nabla F(x^{(j)})$ is a linear combination of $e_1, \ldots, e_{j+1}$. Furthermore, $x^{(j+1)}$ is also in the same linear span by Assumption A.

It follows that

$$\|x^{(j)} - x^*\|^2 \geq \sum_{i=j+1}^{\infty} (x^*)_i^2 = \sum_{i=j+1}^{\infty} q^{2i} = \frac{q^{2(j+1)}}{1-q^2} = q^{2j}\|x^{(0)} - x^*\|^2$$

with

$$q = \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}. \tag{3}$$

This proves the first lower bound given by the Theorem. The second one follows from (1) in the Gradient Method notes, using $x = x^*$, $y = x^{(0)}$, which was derived from Taylor's theorem (equivalently, BV (9.8)).

To put this lower bound result in terms of function values only, let's use (3) from the Gradient Method notes, again using $x = x^*$, $y = x^{(0)}$, giving

$$\|x^{(0)} - x^*\|^2 \geq \frac{2}{M}\left(F(x^{(0)}) - p^*\right),$$

so the lower bound on F becomes

$$F(x^k) - p^* \geq \frac{1}{\kappa}\left(\frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}\right)^{2k}\left(F(x^{(0)}) - p^*\right)$$

Compare this to (2), Nesterov's result for the gradient method with $t = 2/(m+M)$. The key difference is that here, we see the square root of κ, not κ.

So, is there a method that attains the lower bound given in the Theorem? The remarkable answer is Yes, and Nesterov spends most of the rest of his Chapter 2 deriving such methods, but the development is far too complicated to cover in class. In the end, the simplest such method is:[3]

[3] Nesterov's book, p. 81

Nesterov's Optimal Gradient (Accelerated Gradient) Method
Choose $y^{(0)} = x^{(0)} \in \mathbb{R}^n$
For $k = 0, 1, 2, \ldots$

```
    Set $x^{(k+1)} = y^{(k)} - \frac{1}{M}\nabla f(y^{(k)})$
    Set $y^{(k+1)} = x^{(k+1)} + q(x^{(k+1)} - x^{(k)})$
```

where q is given as before by (3).

Important Historical Note.
The appearance of the factor $\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}$ in the convergence rate bound may remind you of the well-known convergence rate for the Conjugate Gradient (CG) method for solving symmetric positive definite linear systems, equivalently minimizing strongly convex quadratic functions. The solution of second-order difference equations also comes up in the classical analysis of related iterative methods such as the Chebyshev method. Assuming one knows the parameters M and m, Nesterov's optimal gradient method is much more powerful than CG because it applies to general strongly convex functions, although the lower complexity bound is derived using a quadratic example. On the other hand, an essential advantage of CG (which does not apply to other classical iterative methods for solving $Ax = b$) is that *it does not need to know any parameters*! There is a very nice paper by Karimi and Vavasis discussing the relationship between these topics; this could be the basis of an interesting final project. Note that well-known extensions of CG to non-quadratic functions, particularly "nonlinear CG" methods such as Fletcher-Reeves or Polak-Riebiere, do not have the same powerful complexity result, and nor do quasi-Newton methods such as BFGS, although BFGS especially has other very nice properties.

More about the tridiagonal matrix. The tridiagonal matrix T is a scaling of the well known discrete Laplacian in one variable. For any fixed dimension N, its eigenvalues are known to be $2-2\cos(j\pi/(N+1))$, which lie in the interval $[0, 4]$ and fill up the interval as $N \to \infty$. To argue about positive definiteness without using eigenvalues, observe that both T and $4I - T$ are symmetric and diagonally dominant, that is, their entries a_{ij} satisfy, for $i = 1, \ldots, N$,

$$a_{ii} \geq \sum_{j \neq i} |a_{ij}|.$$

It is well-known that symmetric diagonally dominant matrices are positive semidefinite, although the proof of this general fact is nontrivial. For this particular tridiagonal matrix, however, it's easy to prove that T (and likewise $4I - T$) is positive semidefinite.[4]

First, a small example. Let $N = 4$. Then

$$\begin{aligned}
x^T T_4 x &= \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 2x_1 - x_2 \\ -x_1 + 2x_2 - x_3 \\ -x_2 + 2x_3 - x_4 \\ -x_3 + 2x_4 \end{bmatrix} \\
&= (2x_1^2 - x_1 x_2) + (-x_1 x_2 + 2x_2^2 - x_2 x_3) \\
&\quad + (-x_2 x_3 + 2x_3^2 - x_3 x_4) + (-x_3 x_4 + 2x_4^2).
\end{aligned}$$

From this it follows straightforwardly that

$$\begin{aligned}
x^T T_4 x &= (2x_1^2 + 2x_2^2 + 2x_3^2 + 2x_4^2) - 2(x_1 x_2 + x_2 x_3 + x_3 x_4) \\
&= x_1^2 + (x_1^2 - 2x_1 x_2 + x_2^2) + (x_2^2 - 2x_2 x_3 + x_3^2) + (x_3^2 - 2x_3 x_4 + x_4^2) + x_4^2 \\
&= x_1^2 + (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_4)^2 + x_4^2 \geq 0.
\end{aligned}$$

The generalization to any dimension N is now immediate, and we have for x of length N:

$$x^T T_N x = x_1^2 + x_N^2 + \sum_{i=2}^{N} (x_i - x_{i-1})^2.$$

If we now consider the infinite version of T then not much needs to be changed except make this an infinite sum, and from the nonnegativity of the summands, the semidefiniteness result follows.

[4] Thanks to Chen Greif for providing this argument.