Nesterov's "optimal" method

In the case where f is strictly convex, you can come up with examples that show that the convergence rate of the heavy ball method can't be improved in general. For non-strictly convex f, the story is more complicated.

Recall from a few lectures ago that we also had the convergence result for gradient descent in the case where we only know that the gradient is Lipschitz:

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \ \leq \ L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

We saw that for gradient descent with a step size $t = 1/L$,

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) \lesssim \frac{L}{k} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2.$$

Thus, to reduce the error by a factor of ϵ requires

$$k \gtrsim 1/\epsilon$$

iterations.

In 1983, Yuri Nesterov proposed a slight variation on the heavy ball method that can improve on this theory, and often works better in practice [Nes83].[1] Specifically, recall the heavy ball method, which can be represented via the iteration:

$$\begin{aligned} \boldsymbol{p}^{(k)} &= \beta_k \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \right) \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)}) + \boldsymbol{p}^{(k)}, \end{aligned}$$

[1] Note that this method remained to a large extent unknown in the wider community until his 2004 publication (in English) of [Nes04].

where we start with $\boldsymbol{p}^{(0)} = \mathbf{0}$. Nesterov's method makes a subtle, but significant, change to this iteration:

$$\begin{aligned}\boldsymbol{p}^{(k)} &= \beta_k \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \right) \\ \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)} + \boldsymbol{p}^{(k)}) + \boldsymbol{p}^{(k)}.\end{aligned}$$

Notice that this is the same as heavy ball *except* that there is also a momentum term *inside* the gradient expression. With this iteration, we will show that

$$f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) \lesssim \frac{L}{k^2} \|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2,$$

meaning that we can reduce the error by a factor of ϵ in

$$k \gtrsim 1/\sqrt{\epsilon},$$

iterations. When $\epsilon \sim 10^{-4}$, this is much, much better than $1/\epsilon$.

Nesterov's method is called "optimal" because it is impossible to beat the $1/k^2$ rate using only function and gradient evaluations. There are careful demonstrations of this in the literature.

Convergence analysis of Nesterov's method

It will be useful to first recall the rough structure of our convergence analysis of gradient descent under the Lipschitz gradient condition, as our approach below will contain many of the same elements.

First, recall that from our Lipschitz assumption, we have

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

From (1), we have that for any $\boldsymbol{x}$,

$$f\left(\boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})\right) \leq f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2. \tag{2}$$

Thus, if we define $\delta_k = f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star)$, then for step size $t = 1/L$,

$$\delta_{k+1} - \delta_k = f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}^{(k)})\|_2^2. \tag{3}$$

Moreover, as a direct consequence of convexity, we have

$$\delta_k = f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star) \leq \langle \boldsymbol{x}^{(k)} - \boldsymbol{x}^\star, \nabla f(\boldsymbol{x}^{(k)})\rangle. \tag{4}$$

By making the substitution $\nabla f(\boldsymbol{x}^{(k)}) = L(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k+1)})$ and then simply adding these inequalities, we obtain

$$\delta_{k+1} \leq \frac{L}{2}\left(2\langle \boldsymbol{x}^{(k)} - \boldsymbol{x}^\star, \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k+1)}\rangle - \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k+1)}\|_2^2\right).$$

This looks a little ugly, until we notice that for a pair of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, we have that $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \|\boldsymbol{a}\|_2^2 = \|\boldsymbol{a} + \boldsymbol{b}\|_2^2 - \|\boldsymbol{b}\|_2^2$. Applying this to our bound above with $\boldsymbol{a} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k+1)}$ and $\boldsymbol{b} = \boldsymbol{x}^\star - \boldsymbol{x}^{(k)}$, we obtain

$$\delta_{k+1} \leq \frac{L}{2}\left(\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^\star\|_2^2\right).$$

The beauty of this inequality is that it yields the *telescopic sum*

$$\begin{aligned}
\sum_{i=1}^k \delta_i &\leq \frac{L}{2}\left(\sum_{i=1}^k \|\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|_2^2\right) \\
&= \frac{L}{2}\left(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|_2^2\right) \\
&\leq \frac{L}{2}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2.
\end{aligned}$$

The proof for gradient descent concludes by noting that

$$\delta_k \leq \frac{1}{k}\sum_{i=1}^{k}\delta_i \leq \frac{L}{2k}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2.$$

We will follow an extremely similar argument to analyze Nesterov's method. We will take $\alpha_k = \frac{1}{L}$ and let $\boldsymbol{g}_t = -\frac{1}{L}\nabla f(\boldsymbol{x}^{(k)} + \boldsymbol{p}^{(k)})$. Our initial building blocks are again inequalities (1) and (2), from which we can derive the inequalities

$$\delta_{k+1} - \delta_k \leq -\frac{L}{2}\left(\|\boldsymbol{g}^{(k)}\|_2^2 + 2\langle \boldsymbol{g}^{(k)}, \boldsymbol{p}^{(k)}\rangle\right), \tag{5}$$

and

$$\delta_{k+1} \leq -\frac{L}{2}\left(\|\boldsymbol{g}^{(k)}\|_2^2 + 2\langle \boldsymbol{g}^{(k)}, \boldsymbol{x}^{(k)} + \boldsymbol{p}^{(k)} - \boldsymbol{x}^\star\rangle\right). \tag{6}$$

We would like to somehow replicate our analysis of gradient descent and somehow combine these two inequalities to obtain a telescopic sum. This time, the approach will be slightly different, and we will obtain an inequality that will yield a telescoping sum on both sides.

Towards this end, we will consider the inequality formed by the convex combination of $\lambda_k - 1$ times (5) plus 1 times (6):

$$\lambda_k\delta_{k+1} + (1-\lambda_k)\delta_k \leq -\frac{L}{2}\left(\lambda_k\|\boldsymbol{g}^{(k)}\|_2^2 + 2\langle \boldsymbol{g}^{(k)}, \boldsymbol{x}^{(k)} + \lambda_k\boldsymbol{p}^{(k)} - \boldsymbol{x}^\star\rangle\right).$$

Again using the identity $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle + \|\boldsymbol{a}\|_2^2 = \|\boldsymbol{a}+\boldsymbol{b}\|_2^2 - \|\boldsymbol{b}\|_2^2$, we can replace the right-hand side of the inequality above with

$$-\frac{L}{2\lambda_k}\left(\|\boldsymbol{x}^{(k)} + \lambda_k\boldsymbol{p}^{(k)} - \boldsymbol{x}^\star + \lambda_k\boldsymbol{g}^{(k)}\|_2^2 - \|\boldsymbol{x}^{(k)} + \lambda_k\boldsymbol{p}^{(k)} - \boldsymbol{x}^\star\|_2^2\right).$$

This is ugly, but remember that we have not yet chosen λ_k (or β_k). Recall that our goal is to obtain a telescopic sum; this will occur if

$$\boldsymbol{x}^{(k)} + \lambda_k \boldsymbol{p}^{(k)} - \boldsymbol{x}^\star + \lambda_k \boldsymbol{g}^{(k)} = \boldsymbol{x}^{(k+1)} + \lambda_{k+1} \boldsymbol{p}^{(k+1)} - \boldsymbol{x}^\star. \tag{7}$$

Observe that since $\boldsymbol{p}^{(k+1)} = \beta_{k+1}\left(\boldsymbol{g}^{(k)} + \boldsymbol{p}^{(k)}\right)$, we can write

$$\boldsymbol{x}^{(k+1)} + \lambda_{k+1} \boldsymbol{p}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{g}^{(k)} + \boldsymbol{p}^{(k)} + \lambda_{k+1}\beta_{k+1}\left(\boldsymbol{g}^{(k)} + \boldsymbol{p}^{(k)}\right).$$

Thus, to satisfy (7) we simply need to have

$$\lambda_k = 1 + \lambda_{k+1}\beta_{k+1} \Rightarrow \beta_k = \frac{\lambda_k - 1}{\lambda_{k+1}}. \tag{8}$$

Using β_k satisfying (8) and setting $\boldsymbol{u}_k = \boldsymbol{x}^{(k)} + \lambda_k \boldsymbol{p}^{(k)} - \boldsymbol{x}^\star$, we thus have

$$\lambda_k^2 \delta_{k+1} - (\lambda_k^2 - \lambda_k)\delta_k \leq \frac{L}{2}\left(\|\boldsymbol{u}^{(k)}\|_2^2 - \|\boldsymbol{u}^{(k+1)}\|_2^2\right). \tag{9}$$

All that remains now is to set λ_k so that we also obtain a telescopic sum on the left-hand side of our inequality, i.e. so that $\lambda_k^2 - \lambda_k = \lambda_{k=1}^2$. Using the quadratic formula to solve for λ, we derive the rule:

$$\lambda_0 = 0 \qquad \lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}.$$

For this choice of λ_k, summing (9) from $i = 0$ to $i = k - 1$ yields

$$\delta_k \leq \frac{L}{2\lambda_{k-1}^2}\|\boldsymbol{u}^{(0)}\|_2^2 = \frac{L}{2\lambda_{k-1}^2}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|_2^2.$$

It is straightforward to show (by induction) that $\lambda_{k-1} \geq k/2$, and thus we have exactly the $1/k^2$ convergence rate we wanted.

Note that in practice, α_k can be chosen using a standard line search, and a good choice of β (both in theory and in practice) turns out to be

$$\beta_k = \frac{k-1}{k+2}.$$

Note that this is very close to what we would obtain by simply setting $\lambda_k = k$ in our analysis above. Most significantly, note that in setting β_k we do *not* need to know anything about the function we are minimizing (such as strong convexity parameters). This represents an important advantage compared to the heavy ball method described before.

References

[Nes83] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Proc. USSR Acad. Sci.*, 269:543–547, 1983.

[Nes04] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Springer Science+Business Media, 2004.