Course Notes for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

April 27, 2018

7 Nesterov's accelerated gradient descent

Previously, we saw how we can accelerate gradient descent for minimizing quadratics $f(x) = x^\top Ax + b^\top x$, where A is a positive definite matrix. In particular, we achieved a quadratic improvement in the dependence on the condition number of the matrix A than what standard gradient descent achieved. The resulting update rule had the form

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) + \mu(x_t - x_{t-1})\,,$$

where we interpreted the last term as a form of "momentum". In this simple form, the update rule is sometimes called Polyak's *heavy ball method*.

To get the same accelerated convergence rate for general smooth convex functions that we saw for quadratics, we will have to work a bit harder and look into Nesterov's celebrated *accelerated gradient method* [Nes83, Nes04]

Specifically, we will see that Nesterov's method achieves a convergence rate of $\mathcal{O}\left(\frac{\beta}{t^2}\right)$ for β-smooth functions. For smooth functions which are also α-strongly convex, we will achieve a rate of $\exp\left(-\Omega\left(\sqrt{\frac{\beta}{\alpha}}t\right)\right)$.

The update rule is a bit more complicated than the plain momentum rule and

proceeds as follows:

$$\begin{aligned} x_0 &= y_0 = z_0, \\ x_{t+1} &= \tau z_t + (1-\tau) y_t && (t \geqslant 0) \\ y_t &= x_t - \frac{1}{\beta}\nabla f(x_t) && (t \geqslant 1) \\ z_t &= z_{t-1} - \eta \nabla f(x_t) && (t \geqslant 1) \end{aligned}$$

Here, the parameter β is the smoothness constant of the function we're minimizing. The step size η and the parameter τ will be chosen below so as to give us a convergence guarantee.

7.1 Convergence analysis

We first show that for a simple setting of the step sizes, the algorithm reduces its initial error from some value d to $\frac{d}{2}$. We will then repeatedly restart the algorithm to continue reducing the error. This is a slight departure from Nesterov's method which does not need restarting, albeit requiring a much more delicate step size schedule that complicates the analysis.

Lemma 7.1. *Suppose $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex, β-smooth function that attains its minimum at a point $x^* \in \mathbb{R}^n$. Assume that the initial point satisfies $\|x_0 - x^*\| \leqslant R$ and $f(x_0) - f(x^*) \leqslant d$. Put $\eta = \frac{R}{\sqrt{d\beta}}$, and τ s.t. $\frac{1-\tau}{\tau} = \eta\beta$.*

Then after $T = 4R\sqrt{\frac{\beta}{d}}$ steps, we have

$$f(\bar{x}) - f(x^*) \leqslant \frac{d}{2},$$

where $\bar{x} = \frac{1}{T}\sum_{k=0}^{T-1} x_t$.

Proof. In Lecture 2, we showed the following properties for smooth and convex functions:

$$f(y_t) - f(x_t) \leqslant -\frac{1}{2\beta}\|\nabla f(x_t)\|^2 \tag{1}$$

By the "Fundamental Theorem of Optimization" (see Lecture 2), we have for all $u \in \mathbb{R}^n$:

$$\eta\langle \nabla f(x_{t+1}), z_t - u\rangle = \frac{\eta^2}{2}\|\nabla f(x_{t+1})\|^2 + \frac{1}{2}\|z_t - u\|^2 - \frac{1}{2}\|z_{t+1} - u\|^2 . \tag{2}$$

Substituting the first equation yields

$$\eta\langle \nabla f(x_{t+1}, z_t - u\rangle \leqslant \eta^2\beta(f(x_{t+1}) - f(y_{t+1})) + \frac{1}{2}\|z_t - u\|^2 - \frac{1}{2}\|z_{t+1} - u\|^2 \tag{3}$$

Working towards a term that we can turn into a telescoping sum, we compute the

following difference

$$\begin{aligned}
&\eta\langle\nabla f(x_{t+1}), x_{t+1} - u\rangle - \eta\langle\nabla f(x_{t+1}), z_t - u\rangle \\
&= \eta\langle\nabla f(x_{t+1}), x_{t+1} - z_t\rangle \\
&= \frac{1-\tau}{\tau}\eta\langle\nabla f(x_{t+1}), y_t - x_{t+1}\rangle \\
&\leqslant \frac{1-\tau}{\tau}\eta(f(y_t) - f(x_{t+1})) \quad \text{(by convexity).}
\end{aligned} \tag{4}$$

Combining (3) and (4), and setting $\frac{1-\tau}{\tau} = \eta\beta$ yield for all $u \in \mathbb{R}^n$:

$$\eta\langle\nabla f(x_{t+1}), x_{t+1} - u\rangle \leqslant \eta^2\beta(f(y_t) - f(y_{t+1})) + \frac{1}{2}\|z_t - u\|^2 - \frac{1}{2}\|z_{t+1} - u\|^2.$$

Proceeding as in our basic gradient descent analysis, we apply this inequality for $u = x^*$, sum it up from $k = 0$ to T and exploit the telescoping effect:

$$\begin{aligned}
\eta T(f(\bar{x}) - f(x^*)) &\leqslant \sum_{k=0}^{T} \eta\langle\nabla f(x_{t+1}), x_{t+1} - x^*\rangle \\
&\leqslant \eta^2\beta d + R^2,
\end{aligned}$$

which can be rewritten as

$$\begin{aligned}
f(\bar{x}) - f(x^*) &\leqslant \frac{\eta\beta d}{T} + \frac{R^2}{\eta T} \\
&\leqslant \frac{2\sqrt{\beta d}}{T}R && \text{(since } \eta = R/\sqrt{\beta d}\text{)} \\
&\leqslant \frac{d}{2} && \text{(since } T \geqslant 4R\sqrt{\beta/D}.\text{)}
\end{aligned}$$

■

This lemma appears in work by Allen-Zhu and Orecchia [AZO17], who interpret Nesterov's method as a coupling of two ways of analyzing gradient descent. One is the the inequality in (1) that is commonly used in the analysis of gradient descent for smooth functions. The other is Equation 2 commonly used in the convergence analysis for non-smooth functions. Both were shown in our Lecture 2.

Theorem 7.2. *Under the assumptions of Lemma 7.1, by restarting the algorithm repeatedly, we can find a point x such that*

$$f(x) - f(x^*) \leqslant \epsilon$$

with at most $O(R\sqrt{\beta/\epsilon})$ gradient updates.

Proof. By Lemma 7.1, we can go from error d to $d/2$ with $CR\sqrt{\beta/d}$ gradient updates for some constant C. Initializing each run with the output of the previous run, we can there for successively reduce the error from an initial value d to $d/2$ to $d/4$ and so on until we reach error ϵ after $O(\log(d/\epsilon))$ runs of the algorithm. The total number of gradient steps we make is

$$CR\sqrt{\beta/d} + CR\sqrt{2\beta/d} + \cdots + CR\sqrt{\beta/\epsilon} = O\left(R\sqrt{\beta/\epsilon}\right) .$$

Note that the last run of the algorithm dominates the total number of steps up to a constant factor. ■

7.2 Strongly convex case

We can prove a variant of Lemma 7.1 that applies when the function is also α-strongly convex, ultimately leading to a linear convergence rate. The idea is just a general trick to convert a convergence rate for a smooth function to a convergence rate in domain using the definition of strong convexity.

Lemma 7.3. *Under the assumption of Lemma 7.1 and the additional assumption that the function f is α-strongly convex, we can find a point x with $T = O\left(\sqrt{\frac{\beta}{\alpha}}\right)$ gradient updates such that*

$$\|\bar{x} - x^*\|^2 \leqslant \frac{1}{2}\|x_0 - x^*\|^2 .$$

Proof. Noting that $\|x_0 - x^*\|^2 \leqslant R^2$, we can apply Theorem 7.2 with error parameter $\epsilon = \frac{\alpha}{4}\|x_0 - x^*\|^2$ to find a point x such that

$$f(x) - f(x^*) \leqslant \frac{\alpha}{4}\|x_0 - x^*\|^2 ,$$

while only making $O\left(\sqrt{\beta/\alpha}\right)$ many steps. From the definition of strong convexity it follows that

$$\frac{\alpha}{2}\|x - x^*\|^2 \leqslant f(x) - f(x^*) .$$

Combining the two inequalities gives the statement we needed to show. ■

We see from the lemma that for strongly convex function we actually reduce the distance to the optimum in domain by a constant factor at each step. We can therefore repeatedly apply the lemma to get a linear convergence rate.

Table 1 compares the bounds on error $\epsilon(t)$ as a function of the total number of steps when applying Nesterov's method and ordinary gradient descent method to different functions.

	Nesterov's Method	Ordinary GD Method
β-smooth, convex	$O\left(\beta/t^2\right)$	$O\left(\beta/t\right)$
β-smooth, α-strongly convex	$\exp\left(-\Omega(t\sqrt{\alpha/\beta})\right)$	$\exp\left(-\Omega(t\alpha/\beta)\right)$

Table 1: Bounds on error ϵ as a function of number of iterations t for different methods.

References

[AZO17] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proc. 8th ITCS*, 2017.

[Nes83] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, 269:543–547, 1983.

[Nes04] Yurii Nesterov. *Introductory Lectures on Convex Programming. Volume I: A basic course*. Kluwer Academic Publishers, 2004.