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

February 15, 2018

7 Lecture 7: Nesterov's accelerated gradient descent

In this lecture, we present Nesterov's accelerated gradient descent for smooth convex functions. We will derive an algorithm that obtains a runtime of $\mathcal{O}\left(\frac{\beta}{t^2}\right)$ for β-smooth functions. For functions which are α-strongly convex, we will achieve a rate of $\mathcal{O}\left(\exp\left(-\sqrt{\frac{\beta}{\alpha}}t\right)\right)$. The algorithm proceeds iteratively as follows.

$$\begin{aligned} x_0 &= y_0 = z_0, \\ x_{k+1} &= \tau z_k + (1-\tau) y_k \\ y_k &= x_k - \frac{1}{\beta}\nabla f(x_k) \\ z_k &= z_{k-1} - \eta \|\nabla f(x_k)\| \end{aligned}$$

The definitions are constructed so that we can apply bounds related to smooth and convex functions. We will simplify our proof of the runtime of the algorithm by "restarting" the procedure in the lemma several times; namely, after the error is reduced from d to $\frac{d}{2}$, we run our algorithm again with a new $x_0^{'} = x_T$

Lemma 7.1. *Suppose $\|x_0 - x^*\| \leqslant R$, f attains its minimum at x^* and is β-smooth. Also assume $f(x_0) - f(x^*) \leqslant d$. 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 $\bar{x} = \frac{1}{T}\sum_{k=0}^{T-1} x_k$.

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

$$f(y_k) - f(x_k) \leqslant -\frac{1}{2\beta}\|\nabla f(x_k)\|^2. \tag{1}$$

By the "Fundamental Theorem of Optimization" (see Lecture 2), we have

$$\forall u: \ \eta\langle \nabla f(x_{k+1}), z_k - u\rangle = \frac{\eta^2\|\nabla f(x_{k+1})\|^2 + \|z_k - u\|^2 - \|z_{k+1} - u\|^2}{2} \tag{2}$$

$$\leqslant \frac{\eta^2}{2}\|\nabla f(x_{k+1})\|^2 + \|z_k - u\|^2 - \|z_{k+1} - u\|^2 \tag{3}$$

Substituting the first equation yields

$$\eta\langle \nabla f(x_{k+1}, z_k - u\rangle \leqslant \eta^2\beta(f(x_{k+1}) - f(y_{k+1})) + \|z_k - u\|^2 - \|z_{k+1} - u\|^2 \tag{4}$$

Besides,

$$\begin{aligned}
&\eta\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle - \eta\langle \nabla f(x_{k+1}), z_k - u\rangle \\
&= \eta\langle \nabla f(x_{k+1}), x_{k+1} - z_k\rangle \\
&= \frac{1-\tau}{\tau}\eta\langle \nabla f(x_{k+1}), y_k - x_{k+1}\rangle \\
&\leqslant \frac{1-\tau}{\tau}\eta(f(y_k) - f(x_{k+1})) \text{ (by convexity).}
\end{aligned} \tag{5}$$

Combining (4) and (5), and setting $\frac{1-\tau}{\tau} = \eta\beta$ yield

$$\forall u, \ \eta\langle \nabla f(x_{k+1}), x_{k+1} - u\rangle \leqslant \eta^2\beta(f(y_k) - f(y_{k+1})) + \|z_k - u\|^2 - \|z_{k+1} - u\|^2.$$

This implies

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

which can be rewritten as

$$f(\bar{x}) - f(x^*) \leqslant \frac{\eta\beta d}{T} + \frac{R^2}{\eta T}.$$

Setting $\eta = \frac{R}{\sqrt{\beta d}}$, this gives us

$$f(\bar{x}) - f(x) \leqslant \frac{2\sqrt{\beta d}}{T} R.$$

Therefore, for $T \geqslant 4\sqrt{\frac{\beta}{d}} R$, we have

$$f(\bar{x}) - f(x) \leqslant \frac{d}{2}.$$

■

This lemma was initially proven by Allen-Zhu and Orecchia [AZO17]. From the lemma, our analysis of the runtime of the algorithm naturally follows.

Theorem 7.2. *If a function f is β-smooth and convex (not strongly convex), the lemma gives that after $T(d) = 4R\sqrt{\frac{\beta}{d}}$ steps,*

$$f(\bar{x}) - f(x^*) \leqslant \frac{d}{2} \qquad \text{, where } \bar{x} = \frac{1}{T} \sum_{\tau=0}^{T-1} x_k.$$

This means after $T(d)$ steps, Nesterov's accelerated gradient descent method reduces the function value error by half. For the complete Nesterov's method, after every $T(d)$ steps, let $\bar{x}$ be the new initial state. We restart the iterative updating algorithm from the beginning.

Proof of Theorem 7.2. By applying the Lemma n times, we calculate the total iterations before $f(\bar{x}_{(n)}) - f(x^*) \leqslant \frac{d}{2^n} \leqslant \epsilon$,

$$\begin{aligned} t(\epsilon) &= \mathcal{O}\left(4R\sqrt{\frac{\beta}{d/2^{n-1}}} + 4R\sqrt{\frac{\beta}{d/2}} + 4R\sqrt{\frac{\beta}{d}}\right) \\ &= \mathcal{O}\left(4R\sqrt{\frac{\beta}{d/2^{n-1}}}\right) \\ &= \mathcal{O}\left(\sqrt{\frac{\beta}{\epsilon}}\right). \end{aligned}$$

This can also be written as:

$$\epsilon(t) = \mathcal{O}\left(\frac{\beta}{t^2}\right)$$

■

We can make a stronger statement for functions which are also strongly convex.

Theorem 7.3. *If a function f is β-smooth and α-strongly convex, the Lemma gives that after constant $T = 4\sqrt{\frac{2\beta}{\alpha}}$ steps,*

$$\|\bar{x} - x^*\|^2 \leqslant \frac{1}{2}\|x_0 - x^*\|^2 \qquad \text{, where } \bar{x} = \frac{1}{T}\sum_{\tau=0}^{T-1} x_k.$$

Proof of Theorem 7.3. Using the same trick in 1), we apply the Lemma n times before $\|\bar{x}_{(n)} - x^*\|^2 \leqslant \frac{1}{2^n}\|x_0 - x^*\|^2 \leqslant \epsilon$. Then the total iterations are:

$$\begin{aligned} t(\epsilon) &= \mathcal{O}(nT) \\ &= \mathcal{O}\left(4\sqrt{\frac{2\beta}{\alpha}}\log_2\left(\frac{\|x_0 - x^*\|^2}{\epsilon}\right)\right) \\ &= \mathcal{O}\left(-\sqrt{\frac{\beta}{\alpha}}\log_2\epsilon\right). \end{aligned}$$

■

In this case, Nesterov's accelerated gradient descent method yields linear convergence:

$$\epsilon(t) = 2^{-\Omega(t\sqrt{\frac{\alpha}{\beta}})}.$$

Table 1 compares the bounds on error $\epsilon(t)$ while applying Nesterov's method and ordinary gradient descent method to different functions. Nesterov's accelerated gradient descent method has a faster convergence rate than the ordinary gradient descent method.

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

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. *ICTS*, 2017.