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

May 3, 2018

9 Lower bounds and trade-offs with robustness

In the first part of this lecture, we study whether the convergence rates derived in previous lectures are tight. For several classes of optimization problems (smooth, strongly convex, etc), we prove the answer is indeed yes. The highlight of this analysis is to show the $O(1/t^2)$ rate achieved by Nesterov's accelerated gradient method is optimal (in a weak technical sense) for smooth, convex functions.

In the second part of this lecture, we go beyond studying convergence rates and look towards other ways of comparing algorithms. We show the improved rates of accelerated gradient methods come at a cost in robustness to noise. In particular, if we restrict ourselves to only using approximate gradients, the standard gradient method suffers basically no slowdown, whereas the accelerated gradient method accumulates errors linearly in the number of iterations.

9.1 Lower bounds

Before launching into a discussion of lower bounds, it's helpful to first recap the upper bounds obtained thus far. For a convex function f, Table 1 summarizes the assumptions and rates proved in the first several lectures.

Each of the rates in Table 1 is obtained using some variant of the gradient method. These algorithms can be thought of as a procedure that maps a history of points and subgradients $(x_1, g_1, \dots, x_t, g_t)$ to a new point x_{t+1}. To prove lower bounds, we restrict

Table 1: Upper Bounds from Lectures 2-8

Function class	Algorithm	Rate
Convex, Lipschitz	Gradient descent	$RL/\sqrt{t}$
Strongly convex, Lipschitz	Gradient descent	$L^2/(\alpha t)$
Convex, smooth	Accelerated gradient descent	$\beta R^2/t^2$

the class of algorithms to similar procedures. Formally, define a black-box procedure as follows.

Definition 9.1 (Black-Box Procedure). A *black-box procedure* generates a sequence of points $\{x_t\}$ such that

$$x_{t+1} \in x_0 + \operatorname{span}\{g_1, \ldots, g_t\},$$

and $g_s \in \partial f(x_s)$.

Throughout, we will further assume $x_0 = 0$. As expected, gradient descent is a black-box procedure. Indeed, unrolling the iterates, x_{t+1} is given by

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla f(x_t) \\ &= x_{t-1} - \eta \nabla f(x_{t-2}) - \eta \nabla f(x_{t-1}) \\ &= x_0 - \sum_{i=0}^{t} \eta \nabla f(x_i). \end{aligned}$$

We now turn to proving lower bounds on the convergence rate for any black-box procedure. Our first theorem concerns the constrained, non-smooth case. The theorem is originally from [Nes83], but the presentation will follow [Nes04].

Theorem 9.2 (Constrainted, Non-Smooth f). *Let $t \leqslant n$, $L, R > 0$. There exists a convex L-Lipschitz function f such that any black-box procedure satisfies*

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(R)} f(x) \geqslant \frac{RL}{2(1+\sqrt{t})}. \tag{1}$$

Furthermore, there is an α-strongly convex, L-Lipschitz function f such that

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(\frac{L}{2\alpha})} f(x) \geqslant \frac{L^2}{8\alpha t}. \tag{2}$$

The proof strategy is to exhibit a convex function f so that, for any black-box procedure, $\operatorname{span}\{g_1, g_2, \ldots, g_i\} \subset \operatorname{span}\{e_1, \ldots, e_i\}$, where e_i is the i-th standard basis vector. After t steps for $t < n$, at least $n - t$ coordinates are exactly 0, and the theorem follows from lower bounding the error for each coordinate that is identically zero.

Proof. Consider the function

$$f(x) = \gamma \max_{1 \leqslant i \leqslant t} x[i] + \frac{\alpha}{2}\|x\|^2,$$

for some $\gamma, \alpha \in \mathbb{R}$. In the strongly convex case, γ is a free parameter, and in the Lipschitz case both α and γ are free parameters. By the subdifferential calculus,

$$\partial f(x) = \alpha x + \gamma \operatorname{conv}\{e_i \colon i \in \operatorname*{argmax}_{1 \leqslant j \leqslant t} x(j)\}.$$

The function f is evidently α-strongly convex. Furthermore, if $\|x\| \leqslant R$ and $g \in \partial f(x)$, then $\|g\| \leqslant \alpha R + \gamma$, so f is $(\alpha R + \gamma)$-Lipschitz on $B_2(R)$.

Suppose the gradient oracle returns $g_i = \alpha x + \gamma e_i$, where i is the first coordinate such that $x[i] = \max_{1 \leqslant j \leqslant t} x[j]$. An inductive argument then shows

$$x_s \in \operatorname{span}\{e_1, \ldots, e_{s-1}\}$$

Consequently, for $s \leqslant t$, $f(x_s) \geqslant 0$. However, consider $y \in \mathbb{R}^n$ such that

$$y[i] = \begin{cases} -\frac{\gamma}{\alpha t} & \text{if } 1 \leqslant i \leqslant t \\ 0 & \text{otherwise.} \end{cases}$$

Since $0 \in \partial f(y)$, y is an minimizer of f with objective value

$$f(y) = \frac{-\gamma^2}{\alpha t} + \frac{\alpha}{2}\frac{\gamma^2}{\alpha^2 t} = -\frac{\gamma^2}{2\alpha t},$$

and hence $f(x_s) - f(y) \geqslant \frac{\gamma^2}{2\alpha t}$. We conclude the proof by appropriately choosing α and γ. In the convex, Lipschitz case, set

$$\alpha = \frac{L}{R}\frac{1}{1+\sqrt{t}} \quad \text{and} \quad \gamma = L\frac{\sqrt{t}}{1+\sqrt{t}}.$$

Then, f is L-Lipschitz,

$$\|y\| = \sqrt{t\left(\frac{-\gamma}{\alpha t}\right)^2} = \frac{\gamma}{\alpha\sqrt{t}} = R$$

and hence

$$f(x_s) - \min_{x \in B_2(R)} f(x) = f(x_s) - f(y) \geqslant \frac{\gamma^2}{2\alpha t} = \frac{RL}{2(1+\sqrt{t})}.$$

In the strongly-convex case, set $\gamma = \frac{L}{2}$ and take $R = \frac{L}{2\alpha}$. Then, f is L-Lipschitz,

$$\|y\| = \frac{\gamma}{\alpha\sqrt{t}} = \frac{L}{2\alpha\sqrt{t}} = \frac{R}{\sqrt{t}} \leqslant R,$$

and therefore

$$f(x_s) - \min_{x \in B_2(L/2\alpha)} f(x) = f(x_s) - f(y) \geqslant \frac{LR}{4t} = \frac{L^2}{8\alpha t}.$$

■

Next, we study the smooth, convex case and show the $O(1/t^2)$ rate achieved by accelerated gradient descent is optimal.

Theorem 9.3 (Smooth-f)**.** *Let $t \leqslant \frac{n-1}{2}$, $\beta > 0$. There exists a β-smooth convex quadratic f such that any black-box method satisfies*

$$\min_{1 \leqslant s \leqslant t} f(x_s) - f(x^\star) \geqslant \frac{3\beta \|x_0 - x^\star\|_2^2}{32(t+1)^2}. \tag{3}$$

Similar to the previous theorem, the proof strategy is to exhibit a pathological convex function. In this case, we choose what Nesterov calls "the worst-function in the world" [Nes04].

Proof. Without loss of generality, let $n = 2t+1$. Let $L \in \mathbb{R}^{n \times n}$ be the tridiagonal matrix

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ 0 & 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 2 \end{bmatrix}.$$

The matrix L is almost the Laplacian of the cycle graph (in fact, it's the Laplacian of the chain graph).[1] Notice

$$x^\top L x = x[1]^2 + x[n]^2 + \sum_{i=1}^{n-1} (x[i] - x[i+1])^2,$$

and, from this expression, it's a simple to check $0 \preceq L \preceq 4I$. Define the following β-smooth convex function

$$f(x) = \frac{\beta}{8} x^\top L x - \frac{\beta}{4} \langle x, e_1 \rangle.$$

The optimal solution $x^\star$ satisfies $Lx^\star = e_1$, and solving this system of equations gives

$$x^\star[i] = 1 - \frac{i}{n+1},$$

[1] https://en.wikipedia.org/wiki/Laplacian_matrix

which has objective value

$$f(x^\star) = \frac{\beta}{8} x^{\star\top} L x^\star - \frac{\beta}{4}\langle x^\star, e_1\rangle$$
$$= -\frac{\beta}{8}\langle x^\star, e_1\rangle = -\frac{\beta}{8}(1 - \frac{1}{n+1}).$$

Similar to the proof of Theorem 9.2, we can argue

$$x_s \in \operatorname{span}\{e_1, \ldots, e_{s-1}\},$$

so if $x_0 = 0$, then $x_s[i] = 0$ for $i \geqslant s$ for any black-box procedure. Let $x_s^\star = \operatorname{argmin}_{x:\, i \geqslant s, x[i]=0} f(x)$. Notice $x_s^\star$ is the solution of a smaller $s \times s$ Laplacian system formed by the first s rows and columns of L, so

$$x_s^\star[i] = \begin{cases} 1 - \frac{i}{s+1} & \text{if } i < s \\ 0 & \text{otherwise,} \end{cases}$$

which has objective value $f(x_s^\star) = -\frac{\beta}{8}(1 - \frac{1}{s+1})$. Therefore, for any $s \leqslant t$,

$$\begin{aligned} f(x_s) - f(x^\star) &\geqslant f(x_t^\star) - f(x^\star) \\ &\geqslant \frac{\beta}{8}\left(\frac{1}{t+1} - \frac{1}{n+1}\right) \\ &= \frac{\beta}{8}\left(\frac{1}{t+1} - \frac{1}{2(t+1)}\right) \\ &= \frac{\beta}{8}\frac{1}{2(t+1)}. \end{aligned}$$

To conclude, we bound the initial distance to the optimum. Recalling $x_0 = 0$,

$$\begin{aligned} \|x_0 - x^\star\|^2 = \|x^\star\|^2 &= \sum_{i=1}^{n}(1 - \frac{i}{n+1})^2 \\ &= n - \frac{2}{n+1}\sum_{i=1}^{n} i + \frac{1}{(n+1)^2}\sum_{i=1}^{n} i^2 \\ &\leqslant n - \frac{2}{n+1}\sum_{i=1}^{n} i + \frac{1}{(n+1)^2}\int_1^{n+1} x^2\,dx \\ &\leqslant n - \frac{2}{n+1}\frac{n(n+1)}{2} + \frac{1}{(n+1)^2}\frac{(n+1)^3}{3} \\ &= \frac{(n+1)}{3} \\ &= \frac{2(t+1)}{3}. \end{aligned}$$

Combining the previous two displays, for any $s \leqslant t$,

$$f(x_s) - f(x^\star) \geqslant \frac{\beta}{8}\frac{1}{2(t+1)} \geqslant \frac{3\beta\|x_0 - x^\star\|^2}{32(t+1)^2}.$$

■

9.2 Robustness and acceleration trade-offs

The first part of the course focused almost exclusively on convergence rates for optimization algorithms. From this perspective, a better algorithm is one with a faster rate of convergence. A theory of optimization algorithms that stops with rates of convergence is incomplete. There are often other important algorithm design goals, e.g. robustness to noise or numerical errors, that are ignored by focusing on converges rates, and when these goals are of primary importance, excessive focus on rates can lead practitioners to choose the wrong algorithm. This section deals with one such case.

In the narrow, technical sense of the previous section, Nesterov's Accelerated Gradient Descent is an "optimal" algorithm, equipped with matching upper and lower bounds on it's rate of convergence. A slavish focus on convergence rates suggests one should then always use Nesterov's method. Before coronating Nesterov's method, however, it is instructive to consider how it performs in the presence of noise.

Figure 1 compares the performance of vanilla gradient descent and Nesterov's accelerated gradient descent on the function f used in the proof of Theorem 9.3. In the noiseless case, the accelerated method obtains the expected speed-up over gradient descent. However, if we add a small amount of spherical noise to the gradients, the speed-up not only disappears, but gradient descent begins to outperform the accelerated method, which begins to diverge after a large number of iterations.

The preceding example is not wickedly pathological in any sense. Instead, it is illustrative of a much broader phenomenon. Work by Devolder, Glineur and Nesterov [DGN14] shows there is a fundamental trade-off between acceleration and robustness, in a sense made precise below.

First, define the notion of an inexact gradient oracle. Recall for a β-smooth convex function f and any $x, y \in \Omega$,

$$0 \leqslant f(x) - (f(y) + \langle \nabla f(y), x - y\rangle) \leqslant \frac{\beta}{2}\|x - y\|^2. \tag{4}$$

For any $y \in \Omega$, an exact first-order oracle then returns a pair $(f(y), g(y)) = (f(y), \nabla f(y))$ that satisfies (4) exactly for every $x \in \Omega$. An inexact oracle, returns a pair so that (4) holds up to some slack δ.

Definition 9.4 (Inexact oracle)**.** Let $\delta > 0$. For any $y \in \Omega$, a δ-inexact oracle returns a pair $(f_\delta(y), g_\delta(y))$ such that for every $x \in \Omega$,

$$0 \leqslant f(x) - (f(y) + \langle \nabla f(y), x - y\rangle) \leqslant \frac{\beta}{2}\|x - y\|^2 + \delta\,.$$

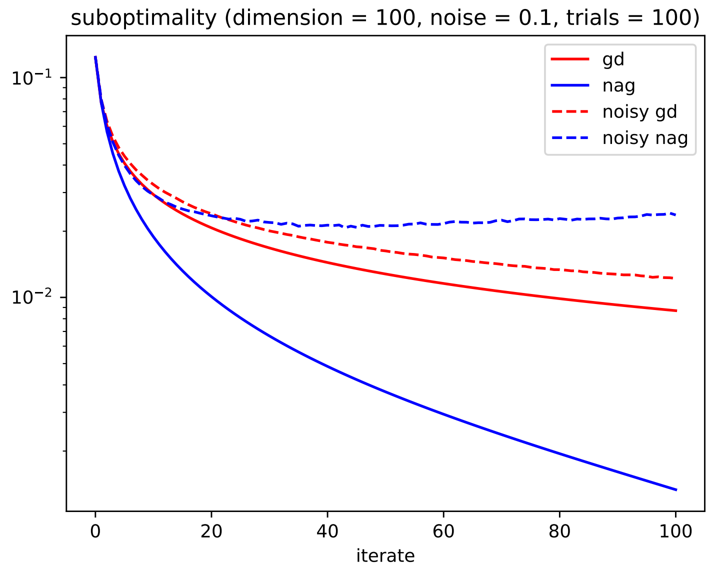

Figure 1: The optimality gap for iterations of gradient descent and Nesterov accelerated gradient descent applied to the worst function in the world with dimension $n = 100$. Notice with exact oracle gradients, acceleration helps significantly. However, when adding uniform spherical random noise with radius $\delta = 0.1$ to the gradient, stochastic gradient descent remains robust while stochastic accelerated gradient accumulates error. The stochastic results are averaged over 100 trials.

Consider running gradient descent with a δ-inexact oracle. Devolder et al. [DGN14] show, after t steps,

$$f(x_t) - f(x^\star) \leqslant \frac{\beta R^2}{2t} + \delta.$$

Comparing this rate with Table 1, the plain gradient method is not affected by the inexact oracle and doesn't accumulate errors. On the other hand, if the accelerated gradient method is run with a δ-inexact oracle, then after t steps,

$$f(x_t) - f(x^\star) \leqslant \frac{4\beta R^2}{(t+1)^2} + \frac{1}{3}(t+3)\delta.$$

In other words, the accelerated gradient method accumulates errors linearly with the number of steps! Moreover, this slack is not an artifact of the analysis. Any black-box method must accumulate errors if it is accelerated in the exact case, as the following theorem makes precise.

Theorem 9.5 (Theorem 6 in [DGN14]). *Consider a black-box method with convergence rate $O\left(\frac{\beta R^2}{t^p}\right)$ when using an exact oracle. With a δ-inexact oracle, suppose the algorithm achieves a* *rate*

$$f(x_t) - f(x^\star) \leqslant O\left(\frac{\beta R^2}{t^p}\right) + O\left(t^q \delta\right), \tag{5}$$

then $q \geqslant p - 1$.

In particular, for any accelerated method has $p > 1$, and consequently $q > 1$ so the method accumulates at least $O(t^{p-1}\delta)$ error with the number of iterations.

References

[DGN14] Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

[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.