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 2, 2018

9 Lower Bounds and Trade-offs Between Robustness and Acceleration

In the first part of this lecture, we study whether the convergence rates derived in previous lectures are tight. For several classes of optimzation problems we've considered (smooth, strongly convex, etc), we will 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 the convergence rates of different methods and look towards other ways of comparing algorithms. We will give evidence showing the improved rates of accelerated gradient methods come at a cost in robustness to noise. Indeed, 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

Table 1: Upper Bounds from Lectures 2-8

f	Algorithm	Rate
Convex, Lipschitz	Gradient Descent	$RL/\sqrt{\{t\}}$
Strongly Convex, Lipschitz	Gradient Descent	$L^2/(\alpha t)$
Convex, Smooth	Accelerated Gradient Descent	$\beta R^2/t^2$

subgradients $(x_1, g_1, \ldots, x_t, g_t)$ to a new point x_{t+1}. To prove lower bounds, we restrict 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

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

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

$$\begin{aligned} x_{t+1} &= x_t - \eta \nabla f(x_t) && (2)\\ &= x_{t-1} - \eta \nabla f(x_{t-2}) - \eta \nabla f(x_{t-1}) && (3)\\ &= x_0 - \sum_{i=0}^{t} \eta \nabla f(x_i). && (4) \end{aligned}$$

We now prove 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 [**?**], but the presentation in [**?**] is more readable.

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*

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(R)} f(x) \geqslant \frac{RL}{2(1+\sqrt{t})}. \tag{5}$$

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

$$\min_{1 \leqslant s \leqslant t} f(x_s) - \min_{x \in B_2(\frac{L}{2\alpha})} f(x) \geqslant \frac{L^2}{8\alpha t}. \tag{6}$$

The proof strategy is to exhibit a convex function f so that, for any black-box procedure, $\text{span}\{g_1, g_2, \ldots, g_i\} \subset \text{span}\{e_1, \ldots, e_i\}$, 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

$$f(x) = \gamma \max_{1 \leqslant i \leqslant t} x[i] + \frac{\alpha}{2}\|x\|^2, \tag{7}$$

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

$$\partial f(x) = \alpha x + \gamma \operatorname{conv}\{e_i \colon i \in \operatorname*{argmax}_{1 \leqslant j \leqslant t} x(j)\}. \tag{8}$$

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

$$x_s \in \operatorname{span}\{e_1, \ldots, e_{s-1}\} \tag{9}$$

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

$$y[i] = \begin{cases} -\frac{\gamma}{\alpha t} & \text{if } 1 \leqslant i \leqslant t \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

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

$$f(y) = \frac{-\gamma^2}{\alpha t} + \frac{\alpha}{2}\frac{\gamma^2}{\alpha^2 t} = -\frac{\gamma^2}{2\alpha t}, \tag{11}$$

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

$$\alpha = \frac{L}{R}\frac{1}{1+\sqrt{t}} \quad \text{and} \quad \gamma = L\frac{\sqrt{t}}{1+\sqrt{t}}. \tag{12}$$

Then, f is L-Lipschitz,

$$\|y\| = \sqrt{t\left(\frac{-\gamma}{\alpha t}\right)^2} = \frac{\gamma}{\alpha\sqrt{t}} = R \tag{13}$$

and hence

$$f(x_s) - \min_{x \in B_2(R)} f(x) = f(x_s) - f(y) \geqslant \frac{\gamma^2}{2\alpha t} = \frac{RL}{2(1+\sqrt{t})}. \tag{14}$$

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

$$\|y\| = \frac{\gamma}{\alpha\sqrt{t}} = \frac{L}{2\alpha\sqrt{t}} = \frac{R}{\sqrt{t}} \leqslant R, \tag{15}$$

and therefore

$$f(x_s) - \min_{x \in B_2(L/2\alpha)} f(x) = f(x_s) - f(y) \geqslant \frac{LR}{4t} = \frac{L^2}{8\alpha t}. \tag{16}$$

■

Next, we study the smooth, convex case. We 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*

$$\min_{1 \leqslant s \leqslant t} f(x_s) - f(x^\star) \geqslant \frac{3\beta \|x_0 - x^\star\|_2^2}{32(t+1)^2}. \tag{17}$$

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" [**?**].

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

$$L = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots -1 & 2 \end{bmatrix}. \tag{18}$$

The matrix L is almost the Laplacian of the cycle graph.[1] Notice

$$x^\top L x = x[1]^2 + x[n]^2 + \sum_{i=1}^{n-1} (x[i] - x[i+1])^2, \tag{19}$$

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

$$f(x) = \frac{\beta}{8} x^\top L x - \frac{\beta}{4} \langle x, e_1 \rangle. \tag{20}$$

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

$$x^\star[i] = 1 - \frac{i}{n+1}, \tag{21}$$

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

which has objective value

$$f(x^\star) = \frac{\beta}{8} x^{\star\top} L x^\star - \frac{\beta}{4}\langle x^\star, e_1\rangle \tag{22}$$

$$= -\frac{\beta}{8}\langle x^\star, e_1\rangle = -\frac{\beta}{8}(1 - \frac{1}{n+1}). \tag{23}$$

Similar to the proof of (9.2), we can argue

$$x_s \in \text{span}\{e_1, \ldots, e_{s-1}\}, \tag{24}$$

so if $x_0 = 0$, then $x_s[i] = 0$ for $i \geqslant s$ for any black-box procedure. Let $x_s^\star = \text{argmin}_{x:\, i \geqslant s, x[i]=0} f(x)$. 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

$$x_s^\star[i] = \begin{cases} 1 - \frac{i}{s+1} & \text{if } i < s \\ 0 & \text{otherwise,} \end{cases} \tag{25}$$

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

$$f(x_s) - f(x^\star) \geqslant f(x_t^\star) - f(x^\star) \tag{26}$$

$$\geqslant \frac{\beta}{8}\left(\frac{1}{t+1} - \frac{1}{n+1}\right) \tag{27}$$

$$= \frac{\beta}{8}\left(\frac{1}{t+1} - \frac{1}{2(t+1)}\right) \tag{28}$$

$$= \frac{\beta}{8}\frac{1}{2(t+1)}. \tag{29}$$

To conclude, we bound the norm of $x^\star$,

$$\|x_0 - x^\star\|^2 = \|x^\star\|^2 = \sum_{i=1}^{n}(1 - \frac{i}{n+1})^2 \tag{30}$$

$$= n - \frac{2}{n+1}\sum_{i=1}^{n} i + \frac{1}{(n+1)^2}\sum_{i=1}^{n} i^2 \tag{31}$$

$$\leqslant n - \frac{2}{n+1}\sum_{i=1}^{n} i + \frac{1}{(n+1)^2}\int_1^{n+1} x^2\, dx \tag{32}$$

$$\leqslant n - \frac{2}{n+1}\frac{n(n+1)}{2} + \frac{1}{(n+1)^2}\frac{(n+1)^3}{3} \tag{33}$$

$$= \frac{(n+1)}{3} \tag{34}$$

$$= \frac{2(t+1)}{3}. \tag{35}$$

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

$$f(x_s) - f(x^\star) \geqslant \frac{\beta}{8} \frac{1}{2(t+1)} \geqslant \frac{3\beta \|x_0 - x^\star\|^2}{32(t+1)^2}. \tag{36}$$

■

9.2 Robustness and Acceleration Trade-offs

References