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 18, 2018

3 Lecture 3: Strong convexity

This lecture introduces the notion of α-strong convexity and combines it with β-smoothness to develop the concept of condition number. Adding an assumption of, respectively, strong convexity or conditioning improves the rates of error decay for gradient descent proved in the previous lecture from $O(1/\sqrt{t})$ to $O(1/t)$ and $O(1/t)$ to $O(\mathrm{e}^{-t})$.

The technical part follows the corresponding chapter in Bubeck's text [Bub15].

3.1 Reminders

Recall that we had (at least) two definitions apiece for convexity and smoothness: a general definition for all functions and a more compact definition for (twice-)differentiable functions.

A function f is convex if, for each input, there exists a globally valid *linear* lower bound on the function: $f(y) \geqslant f(x) + g^\top(x)(y - x)$. For differentiable functions, the role of g is played by the gradient.

A function f is β-smooth if, for each input, there exists a globally valid *quadratic* upper bound on the function, with (finite) quadratic parameter β: $f(y) \leqslant f(x) + g^\top(x)(y - x) + \frac{\beta}{2}\|x - y\|^2$. More poetically, a smooth, convex function is "trapped between a parabola and a line". Since β is covariant with affine transformations, e.g. changes of units of measurement, we will frequently refer to a β-smooth function as simply smooth.

For twice-differentiable functions, these properties admit simple conditions for smoothness in terms of the Hessian, or matrix of second partial derivatives. A $\mathcal{D}^2$ function f is convex if $\nabla^2 f(x) \succeq 0$ and it is β-smooth if $\nabla^2 f(x) \preceq \beta I$.

We furthermore defined the notion of L-Lipschitzness. A function f is L-Lipschitz if the amount that it "stretches" its inputs is bounded by L: $|f(x) - f(y)| \leqslant L \|x - y\|$. Note that for differentiable functions, β-smoothness is equivalent to β-Lipschitzness of the gradient.

3.2 Strong convexity

With these three concepts, we were able to prove two error decay rates for gradient descent (and its projective, stochastic, and subgradient flavors). However, these rates were substantially slower than what's observed in certain settings in practice.

Noting the asymmetry between our linear lower bound (from convexity) and our quadratic upper bound (from smoothness) we introduce a new, more restricted function class by upgrading our lower bound to second order.

Definition 3.1 (α-Strong Convexity)**.** A function $f : \Omega \to \mathbb{R}$ is *α-strongly convex* if, for all $x, y \in \Omega$, the following inequality holds for some $\alpha > 0$:

$$f(y) \geqslant f(x) + g(x)^\top (y - x) + \frac{\alpha}{2} \|x - y\|^2$$

As with smoothness, we will often shorten "α-strongly convex" to "strongly convex". A strongly convex, smooth function is one that can be "squeezed between two parabolas". If β-smoothness is a good thing, then α-convexity guarantees we don't have too much of a good thing.

Once again, twice-differentiable functions afford a quick condition: a D^2 function is $alpha$-strongly convex if $\nabla^2 f(x) \succeq \alpha I$.

Once again, note that α changes under affine transformations. Conveniently enough, for α-strongly convex, β-smooth functions, we can define a basis-independent quantity that combines these properties:

Definition 3.2 (Condition Number)**.** An α-strongly convex, β-smooth function f has *condition number* $\frac{\beta}{\alpha}$.

For a positive-definite quadratic function f, this definition of the condition number corresponds with the perhaps more familiar definition of the condition number of a matrix.

3.3 A look back and ahead

The following table summarizes the results from the previous lecture and the results to be obtained in this lecture. In both, the value ϵ is the difference between f at some value x' computed from the outputs of gradient descent and f calculated at an optimizer x^*.

	Convex	Strongly Convex
Lipschitz	$\epsilon \leqslant O(1/\sqrt{t})$	$\epsilon \leqslant O(1/t)$
Smooth	$\epsilon \leqslant O(1/t)$	$\epsilon \leqslant O(\mathrm{e}^{-t})$

Table 1: Bounds on error ϵ as a function of number of steps taken t for gradient descent applied to various classes of functions.

Since a rate that is exponential in terms of the magnitude of the error is linear in terms of the bit precision, this rate of convergence is termed *linear*. We now move to prove these rates.

3.4 Convergence rate strongly convex functions

Theorem 3.3. *Assume $f\colon \Omega \to \mathbb{R}$ is α-*strongly convex *and L-*Lipschitz*. Let x^* be an optimizer of f, and let x_s be the updated point at step s using projected gradient descent. Let the max number of iterations be t with an adaptive step size $\eta_s = \frac{2}{\alpha(s+1)}$, then*

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leqslant \frac{2L^2}{\alpha(t+1)}$$

The theorem implies the convergence rate of projected gradient descent for α-strongly convex functions is similar to that of β-smooth functions with a bound on error $\epsilon \leqslant O(1/t)$. In order to prove Theorem 3.3, we need the following proposition.

Proposition 3.4 (Jensen's inequality)**.** *Assume $f\colon \Omega \to \mathbb{R}$ is a convex function and $x_1, x_2, ..., , x_n, \sum_{i=1}^n \gamma_i x_i / \sum_{i=1}^n \gamma_i \in \Omega$ with weights $\gamma_i > 0$, then*

$$f\left(\frac{\sum_{i=1}^n \gamma_i x_i}{\sum_{i=1}^n \gamma_i}\right) \leqslant \frac{\sum_{i=1}^n \gamma_i f(x_i)}{\sum_{i=1}^n \gamma_i}$$

For a graphical "proof" follow this link.

Proof of Theorem 3.3. Recall the two steps update rule of projected gradient descent

$$\begin{aligned} y_{s+1} &= x_s - \eta_s \nabla f(x_s) \\ x_{s+1} &= \Pi_\Omega(y_{s+1}) \end{aligned}$$

First, the proof begins by exploring an upper bound of difference between function

values $f(x_s)$ and $f(x^*)$.

$$\begin{aligned}
f(x_s) - f(x^*) &\leqslant \nabla f(x_s)^\top (x_s - x^*) - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
&= \frac{1}{\eta_s}(x_s - y_{s+1})^\top (x_s - x^*) - \frac{\alpha}{2}\|x_s - x^*\|^2 && \text{(by update rule)} \\
&= \frac{1}{2\eta_s}(\|x_s - x^*\|^2 + \|x_s - y_{s+1}\|^2 - \|y_{s+1} - x^*\|^2) - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
& && \text{(by "Fundamental Theorem of Optimization")} \\
&= \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2) + \frac{\eta_s}{2}\|\nabla f(x_s)\|^2 - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
& && \text{(by update rule)} \\
&\leqslant \frac{1}{2\eta_s}(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2) + \frac{\eta_s}{2}\|\nabla f(x_s)\|^2 - \frac{\alpha}{2}\|x_s - x^*\|^2 \\
& && \text{(by Lemma \textbf{??})} \\
&\leqslant (\frac{1}{2\eta_s} - \frac{\alpha}{2})\|x_s - x^*\|^2 - \frac{1}{2\eta_s}\|x_{s+1} - x^*\|^2 + \frac{\eta_s L^2}{2} && \text{(by Lipschitzness)}
\end{aligned}$$

By multiplying s on both sides and substituting the step size η_s by $\frac{2}{\alpha(s+1)}$, we get

$$s(f(x_s) - f(x^*)) \leqslant \frac{L^2}{\alpha} + \frac{\alpha}{4}(s(s-1)\|x_s - x^*\|^2 - s(s+1)\|x_{s+1} - x^*\|^2)$$

Finally, we can find the upper bound of the function value shown in Theorem 3.3 obtained using t steps projected gradient descent

$$\begin{aligned}
f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) &\leqslant \sum_{s=1}^{t} \frac{2s}{t(t+1)} f(x_s) && \text{(by Proposition 3.4)} \\
&\leqslant \frac{2}{t(t+1)} \sum_{s=1}^{t} \left(s f(x^*) + \frac{L^2}{\alpha} + \frac{\alpha}{4}(s(s-1)\|x_s - x^*\|^2 - s(s+1)\|x_{s+1} - x^*\|^2)\right) \\
&= \frac{2}{t(t+1)} \sum_{s=1}^{t} s f(x^*) + \frac{2L^2}{\alpha(t+1)} - \frac{\alpha}{2}\|x_{t+1} - x^*\|^2 \\
& && \text{(by telescoping sum)} \\
&\leqslant f(x^*) + \frac{2L^2}{\alpha(t+1)}
\end{aligned}$$

This concludes that solving an optimization problem with a strongly convex objective function with projected gradient descent has a convergence rate is of the order $\frac{1}{t+1}$, which is faster compared to the case purely with Lipschitzness. ■

3.5 Convergence rate for smooth and strongly convex functions

Theorem 3.5. *Assume $f\colon \mathbb{R}^n \to \mathbb{R}$ is α-*strongly convex *and β-*smooth*. Let x^* be an optimizer of f, and let x_t be the updated point at step t using gradient descent with a constant*

step size $\frac{1}{\beta}$*, i.e. using the update rule* $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$*. Then,*

$$\|x_{t+1} - x^*\|^2 \leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2$$

In order to prove Theorem 3.5, we require use of the following lemma.

Lemma 3.6. *Assume* f *as in Theorem 3.5. Then* $\forall x, y \in \mathbb{R}^n$ *and an update of the form* $x^+ = x - \frac{1}{\beta}\nabla f(x)$*,*

$$f(x^+) - f(y) \leqslant \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2$$

Proof of Lemma 3.6.

$$\begin{aligned}
f(x^+) - f(x) + f(x) - f(y) &\leqslant \nabla f(x)^\top (x^+ - x) + \frac{\beta}{2}\|x^+ - x\|^2 && \text{(Smoothness)} \\
&\quad + \nabla f(x)^\top (x - y) - \frac{\alpha}{2}\|x - y\|^2 && \text{(Strong convexity)} \\
&= \nabla f(x)^\top (x^+ - y) + \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\
& && \text{(Definition of } x^+\text{)} \\
&= \nabla f(x)^\top (x - y) - \frac{1}{2\beta}\|\nabla f(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\
& && \text{(Definition of } x^+\text{)}
\end{aligned}$$

■

Now with Lemma 3.6 we are able to prove Theorem 3.5.

Proof of Theorem 3.5.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|x_t - \frac{1}{\beta}\nabla f(x_t) - x^*\|^2 \\
&= \|x_t - x^*\|^2 - \frac{2}{\beta}\nabla f(x_t)^\top (x_t - x^*) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\
&\leqslant (1 - \frac{\alpha}{\beta})\|x_t - x^*\|^2 && \text{(Use of Lemma 3.6 with } y = x^*, x = x_t\text{)} \\
&\leqslant (1 - \frac{\alpha}{\beta})^t\|x_1 - x^*\|^2 \\
&\leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2
\end{aligned}$$

■

We can also prove the same result for the constrained case using projected gradient descent.

Theorem 3.7. *Assume $f\colon \Omega \to \mathbb{R}$ is α-*strongly convex *and β-*smooth*. Let x^* be an optimizer of f, and let x_t be the updated point at step t using projected gradient descent with a constant step size $\frac{1}{\beta}$, i.e. using the update rule $x_{t+1} = \Pi_\Omega(x_t - \frac{1}{\beta}\nabla f(x_t))$ where Π_Ω is the projection operator. Then,*

$$\|x_{t+1} - x^*\|^2 \leqslant \exp\left(-t\frac{\alpha}{\beta}\right)\|x_1 - x^*\|^2$$

As in Theorem 3.5, we will require the use of the following Lemma in order to prove Theorem 3.7.

Lemma 3.8. *Assume f as in Theorem 3.5. Then $\forall x, y \in \Omega$, define $x^+ \in \Omega$ as $x^+ = \Pi_\Omega(x - \frac{1}{\beta}\nabla f(x))$ and the function $g\colon \Omega \to \mathbb{R}$ as $g(x) = \beta(x - x^+)$. Then*

$$f(x^+) - f(y) \leqslant g(x)^\top(x - y) - \frac{1}{2\beta}\|g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2$$

Proof of Lemma 3.8. The following is given by the Projection Lemma, for all x, x^+, y defined as in Theorem 3.7.

$$\nabla f(x)^\top(x^+ - y) \leqslant g(x)^\top(x^+ - y)$$

Therefore, following the form of the proof of Lemma 3.6,

$$\begin{aligned} f(x^+) - f(x) + f(x) - f(y) &\leqslant \nabla f(x)^\top(x^+ - y) + \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &\leqslant \nabla g(x)^\top(x^+ - y) + \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \\ &= \nabla g(x)^\top(x - y) - \frac{1}{2\beta}\|\nabla g(x)\|^2 - \frac{\alpha}{2}\|x - y\|^2 \qquad \blacksquare \end{aligned}$$

The proof of Theorem 3.7 is exactly as in Theorem 3.5 after substituting the appropriate projected gradient descent update in place of the standard gradient descent update, with Lemma 3.8 used in place of Lemma 3.6.

References

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.