Optimal Rates in Convex Optimization

Statistical Machine Learning, Spring 2014

Ryan Tibshirani (with Larry Wasserman)

1 Optimal rates, connections to statistics

- Statisticians aren't the only ones who care about lower bounds. Optimizers do too. Here we cover lower bound on rates of convergence in convex optimization, a subject dominated by the work of Yurii Nesterov, a living legend in optimization. Our discussion is based on his book Nesterov (2004)

- There are some interesting connections, and distinctions, to minimax theory in statistics. In statistics, our concern is (roughly speaking)

$$\min_{\text{estimators}} \max_{\text{parameters}} \mathbb{E}[d(\text{estimator}, \text{parameter})],$$

  where d is a loss function, and the expectation is taken over data samples that we used to form the estimate. The minimization over estimators is usually unrestricted, but we usually restrict the maximization to be over a parameter set of interest. In optimization, our concern is (again, roughly speaking)

$$\min_{\text{algorithms}} \max_{\text{functions}} \big(\text{algorithm's guess at } k\text{th iteration} - \text{optimal function value}\big).$$

  Here we typically restrict both the minimization over algorithms and the maximization over functions to range over classes of interest. Also, there is no expectation to be computed; we are looking at the "raw" worst-case over functions, and so interestingly, the proofs of lower bounds end up looking quite different than those in statistical minimax theory. That is, to derive a lower bound, we only have to construct one properly "hard" function to minimize

- In the next two sections, we look at lower bounds on convergence rates when we restrict the optimization algorithms to be first-order methods, that use either subgradients (nonsmooth case), or gradients (smooth case). The relevant function classes are then defined appropriately

2 Nonsmooth first-order methods

- Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \text{dom}(f)$, recall that we defined a subgradient $g \in \mathbb{R}^n$ at x to be any point satisfying

$$f(y) \geq f(x) + g^T(y - x)$$

  for all $y \in \text{dom}(f)$. Subgradients always exist for convex functions f

- Recall the subgradient method for minimizing a convex function f. Written in a slightly more convenient notation for this section: we pick some initial point $y^{(0)}$, and repeat

$$y^{(k)} = y^{(k-1)} - t_k \cdot g^{(k-1)}, \quad k = 1, 2, 3, \ldots,$$

where $g^{(k-1)}$ is any subgradient of f at $y^{(k-1)}$. Since this is not a descent method, we have to keep track of the best iterate $x^{(k)}$ among $y^{(0)}, \ldots y^{(k)}$ so far, i.e., defined by

$$f(x^{(k)}) = \min_{i=0,\ldots k} f(y^{(i)})$$

- It turns out that for a particular choice of step sizes t_k, $k = 1, 2, 3, \ldots$ (e.g., what are called Polyak step sizes), we can show that when f is convex and Lipschitz, the convergence rate of the subgradient method is $O(1/\sqrt{k})$, or in other words,

$$f(x^{(k)}) - f^\star \leq \frac{C}{\sqrt{k}}, \tag{1}$$

where $f^\star$ is the minimum of f, and C is a constant depending on the Lipschitz constant of f, and the initial point $x^{(0)} = y^{(0)}$

- Another way of thinking about this is in terms of the minimum number of iterations k needed to achieve a certain distance from optimality, $f(x^{(k)}) - f^\star \leq \epsilon$. Note that this is the analogy to sample complexity in statistics, and so we might dub it *iteration complexity*. Formally, iteration complexity is defined as

$$k(\epsilon) = \min \Big\{ k = 1, 2, 3, \ldots : f(x^{(k)}) - f^\star \leq \epsilon \Big\},$$

and given the bound in (1), we can see that $k(\epsilon) = O(1/\epsilon^2)$ for the subgradient method. In general, both ways of reporting convergence rates are fine, though in this document, we will stick with the k notation, rather than the ϵ notation

- A question: is (1) optimal, among algorithms that only use subgradients like the subgradient method? Nesterov (2004) shows (see Section 3.2.1) that the answer is precisely: "yes"

- We will paraphrase and prove his result here. The setup: consider a convex function f, with $\mathrm{dom}(f) = \mathbb{R}^n$, such that a solution $x^\star$ of the unconstrained problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

exists. Further assume that f is Lipschitz continuous on the set $\{x \in \mathbb{R}^n : \|x - x^\star\|_2 \leq R\}$, for some $R > 0$

We have access to a nonsmooth first-order oracle—when queried with any point $x \in \mathbb{R}^n$, the oracle gives us an arbitrary subgradient g of f at x. This is called a *weak oracle*

Given an initial point $x^{(0)}$ with $\|x^{(0)} - x^\star\|_2 \leq R$, we are allowed to use any algorithm that produces iterates satisfying

$$x^{(k)} \in x^{(0)} + \mathrm{span}\big\{ g^{(0)}, \ldots g^{(k-1)} \big\}, \quad k = 1, 2, 3, \ldots,$$

with $g^{(k)}$ the subgradient returned by the oracle at $x^{(k)}$, for each k. We call this a *nonsmooth first-order method*

- The result: in the worst-case, we can't do any better than the subgradient method in terms of a convergence rate. That is, given any $k < n$, initial point $x^{(0)}$, and constants $G, R > 0$, there exists a convex, Lipschitz function f with Lipschitz constant G, and minimizer $x^\star$ satisfying $\|x^{(0)} - x^\star\|_2 \leq R$, such that

$$f(x^{(k)}) - f^\star \geq \frac{GR}{2(1 + \sqrt{k+1})},$$

for the iterates $x^{(k)}$, $k = 1, 2, 3, \ldots$ produced by any nonsmooth first-order method, initialized at $x^{(0)}$, under the weak oracle model

- To rephrase (which should look familiar, after having seen minimax theory): given any $k < n$, and $x^{(0)}$, and $G, R > 0$, we have

$$\min_{\text{algorithm } A \in \mathcal{N}(x^{(0)})} \max_{\substack{\text{funcion } f \in \mathcal{L}(x^{(0)}, G, R), \\ \text{subgradients from weak oracle}}} f(x^{(k)}) - f^\star \;\geq\; \frac{C}{\sqrt{k}}.$$

  Here $\mathcal{N}(x^{(0)})$ denotes the set of nonsmooth first-order algorithms initialized at the point $x^{(0)}$, and we write $x^{(k)} = x^{(k)}(A)$ for the kth iterate produced by one such algorithm A. Also, $\mathcal{L}(x^{(0)}, G, R)$ is the set of G-Lipschitz convex functions, having a minimizer $x^\star$ that satisfies $\|x^{(0)} - x^\star\|_2 \leq R$

- Proof: for simplicity we'll derive the lower bound for $x^{(0)} = 0$, and a particular R and G. All arguments can be extended to the general case, but it just obfuscates the proof a bit

  Note that it suffices to be clever and find just one "hard" convex, Lipschitz function f to optimize! It turns out that Nesterov has found it; he considers the convex function

$$f(x) = \max_{i=1,\ldots k+1} x_i + \frac{1}{2}\|x\|_2^2$$

  It is not hard to see that the minimizer and minimum of f are given by

$$x^\star = \big(\underbrace{-1/(k+1), \ldots -1/(k+1)}_{k+1 \text{ times}}, 0, \ldots 0\big) \quad \text{and} \quad f^\star = -\frac{1}{2(k+1)},$$

  respectively. (You can check this using subgradients.) Consider $R = \|0 - x^\star\|_2 = 1/\sqrt{k+1}$. Note that f is Lipschitz continuous with constant $G = 1 + 1/\sqrt{k+1}$

  Remember that we are running a nonsmooth first-order method under a weak oracle model that can return arbitrary subgradients. Suppose that, given a point x, the weak oracle will always return the particular subgradient

$$g = e_j + x$$

  where j is the smallest index such that $x_j = \max_{i=1,\ldots k+1} x_i$, and e_j denotes the jth standard basis vector. We claim that, given this setup, the ℓth iterate of any nonsmooth first-order method will satisfy

$$x_{\ell+1}^{(\ell)} = \ldots = x_n^{(\ell)} = 0, \tag{2}$$

  for any $\ell = 1, 2, \ldots k$. This can be seen by arguing inductively that

$$\text{span}\{g^{(0)}, \ldots g^{(\ell-1)}\} \subseteq \text{span}\{e_1, \ldots e_\ell\},$$

  for any $\ell = 1, 2, \ldots k$. Therefore, from (2), we get that at $\ell = k$, the achieved function value is $f(x^{(k)}) \geq 0$, and so

$$f(x^{(k)}) - f^\star \geq \frac{1}{2(k+1)} = \frac{GR}{2(1+\sqrt{k+1})},$$

  completing the proof

3 Smooth first-order methods

- Now consider a smooth (differentiable) convex function $f : \mathbb{R}^n \to \mathbb{R}$. Recall that the gradient descent algorithm for minimizing f begins with an initial point $x^{(0)}$ and repeats

$$x^{(k)} = x^{(k-1)} - t_k \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

- If the gradient ∇f is Lipschitz with constant $L > 0$, and the step sizes are either fixed at $t_k = t \leq 1/L$, $k = 1, 2, 3, \ldots$ or chosen by backtracking line search, then one can show that the convergence rate of gradient descent is $O(1/k)$. In other words, we have

$$f(x^{(k)}) - f^\star \leq \frac{C}{k}, \tag{3}$$

  where C is a constant depending on the Lipschitz constant L and the initial point $x^{(0)}$. In terms of iteration complexity, this is $O(1/\epsilon)$

- Again, we ask the question: is (3) the optimal rate, among methods that only use gradient information, like gradient descent? It turns out that the answer, again due to Nesterov (2004) (see Section 2.1.2), is now: "no"

- We will paraphrase and prove his result here. The setup: consider a smooth, convex function f, with $\mathrm{dom}(f) = \mathbb{R}^n$, such that ∇f Lipschitz continuous with constant $L > 0$. Given an initial point $x^{(0)}$, we are allowed to use any algorithm that produces iterates satisfying

$$x^{(k)} \in x^{(0)} + \mathrm{span}\big\{\nabla f(x^{(0)}), \ldots \nabla f(x^{(k-1)})\big\}, \quad k = 1, 2, 3, \ldots.$$

  We call such an algorithm a (smooth) *first-order method*

- The result: under this setup, we can lower bound the convergence rate of first-order methods by $1/k^2$. Note that this bound does not match the rate of gradient descent, which means that either (1) the lower bound is too loose, or (2) gradient descent is not rate optimal. We'll see shortly that it is the second statement that is true. Formally, given any $k \leq (n-1)/2$, initial point $x^{(0)}$, and Lipschitz constant $L > 0$, there exists a smooth, convex function f, whose gradient ∇f is L-Lipschitz continuous, such that

$$f(x^{(k)}) - f^\star \geq \frac{3L\|x^{(0)} - x^\star\|_2^2}{32(k+1)^2}$$

  for any sequence of iterates $x^{(k)}$, $k = 1, 2, 3, \ldots$ produced by a first-order method

- To rephrase (in minimax notation), given any $k \leq (n-1)/2$, initial point $x^{(0)}$, and $L > 0$

$$\min_{A \in \mathcal{S}(x^{(0)})} \max_{f \in \mathcal{L}(L)} f(x^{(k)}) - f^\star \;\geq\; \frac{C}{k^2},$$

  where $\mathcal{S}(x^{(0)})$ denotes the set of first-order algorithms initialized at $x^{(0)}$, and $x^{(k)} = x^{(k)}(A)$ denotes the kth iterate from one such algorithm A. Also, $\mathcal{L}(L)$ is the set of smooth, convex functions whose gradient is L-Lipschitz continuous

- Proof: we will assume without a loss of generality that $x^{(0)} = 0$. As before, it suffices to be clever and find one "hard" function to optimize, and Nesterov has done it for us. He considers the convex, quadratic function

$$f(x) = \frac{L}{4}\Big(\frac{1}{2}x_1^2 + \frac{1}{2}\sum_{i=1}^{2k}(x_i - x_{i+1})^2 + \frac{1}{2}x_{2k+1}^2 - x_1\Big).$$

  Expressing f in matrix notation is a bit cumbersome, but once we do this it is straightforward to see that ∇f is Lipschitz with constant L, and the minimizer and minimum of f are given by

$$x_i^\star = \begin{cases} 1 - i/(2k+2) & i = 1, \ldots 2k+1 \\ 0 & i = 2k+2, \ldots n \end{cases},$$

and

$$f^\star = \frac{L}{8}\Big(\frac{1}{2k+2} - 1\Big).$$

Although it may seem strange right now, it helps to consider another convex, quadratic function, defined as

$$g(x) = \frac{L}{4}\Big(\frac{1}{2}x_1^2 + \frac{1}{2}\sum_{i=1}^{k-1}(x_i - x_{i+1})^2 + \frac{1}{2}x_k^2 - x_1\Big),$$

and by similar arguments, the minimum of g is

$$g^\star = \frac{L}{8}\Big(\frac{1}{k+1} - 1\Big).$$

Finally, consider running any first-order method to minimize f, starting at $x^{(0)} = 0$, yielding iterates $x^{(\ell)}$, $\ell = 1, 2, 3, \ldots$. One can show using induction that

$$x_{\ell+1}^{(\ell)} = \ldots = x_n^{(\ell)} = 0,$$

for all $\ell = 1, 2, \ldots k$. Note that in particular, at $\ell = k$, this means $f(x^{(k)}) = g(x^{(k)})$, so

$$f(x^{(k)}) \geq g(x^{(k)}) \geq g^\star = \frac{L}{8}\Big(\frac{1}{k+1} - 1\Big).$$

Therefore

$$\frac{f(x^{(k)}) - f^\star}{\|x^{(0)} - x^\star\|_2^2} \geq \frac{g^\star - f^\star}{\|x^{(0)} - x^\star\|_2^2} \geq \frac{\frac{L}{8}(\frac{1}{k+1} - \frac{1}{2k+2})}{\frac{1}{3}(2k+2)} = \frac{3L}{32(k+1)^2},$$

which completes the proof

4 Achieving optimality: acceleration

- The good news is that the lower bound derived in the previous section is tight, and we can achieve it with a practical first-order method: *accelerated* gradient descent
- Acceleration was briefly described in the previous set of lectures notes on convexity and optimization, and this is just giving you a quick reminder. The idea of acceleration originated with Nesterov (1983), though here we present a variant due to Beck & Teboulle (2009). Choosing an initial point $x^{(0)}$, we repeat for $k = 1, 2, 3, \ldots$,

$$\begin{aligned} y &= x^{(k-1)} + \frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)}) \\ x^{(k)} &= y - t_k \nabla f(y). \end{aligned}$$

  In the definition of y, the second term $\frac{k-2}{k+1}(x^{(k-1)} - x^{(k-2)})$ is like a momentum term which continues to push us in the direction pointing from $x^{(k-2)}$ to $x^{(k-1)}$; the weight $(k-2)/(k+1)$ gets closer and closer to 1 as k gets larger. This term helps accelerate the convergence of the algorithm when it is close to the optimum
- The step sizes t_k, $k = 1, 2, 3, \ldots$ can be chosen in either a fixed fashion or via backtracking (see the previous set of lecture notes)

- Most importantly, if f is smooth and convex with L-Lipschitz continuous gradient, then the accelerated gradient method with a fixed step size $t_k = t \leq 1/L$, $k = 1, 2, 3, \ldots$, satisfies

$$f(x^{(k)}) - f(x^\star) \leq \frac{2\|x^{(0)} - x^\star\|_2^2}{t(k+1)^2},$$

  where C is a constant depending on $x^{(0)}$ and L, and a similar result holds for backtracking. Paired with the result in the previous section, we now see that this rate of $O(1/k^2)$ is optimal among first-order methods

- Fortuitously, the exact same acceleration ideas can be applied to proximal gradient descent, used to minimize a nonsmooth but decomposable function f, yielding the same convergence rate

References

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Nesterov, Y. (1983), 'A method of solving a convex programming problem with convergence rate $O(1/k^2)$', *Soviet Mathematics Doklady* **27**(2), 372–376.

Nesterov, Y. (2004), *Introductory Lectures on Convex Optimization: A Basic Course*, Kluwer Academic Publishers, Norwell.