University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

Feburary 12, 2015 Notes: Benjamin Recht

Lecture 7: Analysis of the Stochastic Gradient Method

We now turn to an analysis of the stochastic gradient method. Before we proceed, let's set up some conventions. We will assume that we are trying to minimize a convex function $f : \mathbb{R}^d \to \mathbb{R}$. Let $x_\star$ denote any optimal solution of f. We will assume we gain access at every iteration to a *stochastic function* $g(x;\xi)$ such that

$$\mathbb{E}_\xi[g(x;\xi)] = \nabla f(x)\,.$$

Here ξ is a random variable which determines what our direction looks like. We additionally assume that there exist non-negative constants L_g and B such that

$$\mathbb{E}_\xi\left[\|g(x)\|_2^2\right] \leq L_g^2\|x - x_\star\|^2 + B^2\,.$$

Note that this makes no assumption about the boundedness of g, only that it is bounded in expectation.

We will study the stochastic gradient iteration

$$x_{k+1} = x_k - \alpha_k g\left(x_k;\xi_k\right)$$

for various choices of stepsizes under different assumptions about L_g and B. Throughout, we will assume that the sequence $\{\xi_j\}$ is selected i.i.d. from some fixed distribution. One can weaken the assumptions of both independence and identical distribution, but doing so will complicate and clutter the analysis.

1 The constants L_g and B

Let's briefly discuss how the constants L_g and B manifest themselves in different problem settings.

1.1 Case 1: Bounded gradients ($L_g = 0$ and $B > 0$.)

In this case, we are asserting that the stochastic gradient function is bounded almost surely for all x. Some examples of this would be the support vector machine where

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \max(1 - x^T(z_i y_i), 0)$$

In this case, for the incremental gradient method, ξ is a random index between 1 and n and

$$g(x,i) = \begin{cases} z_i y_i & x^T(z_i y_i) < 1 \\ 0 & \text{otherwise} \end{cases}$$

and hence,

$$B = \sup_i \|z_i\|_2 \,.$$

(note that this function is technically not differentiable. But any smoothing of the hinge around zero, say by convolution with a gaussian, will still result in a problem with bounded gradients)

Note that under this assumption, f cannot be strongly convex. This is because if f is a strongly convex function with parameter m, we have

$$\|\nabla f(x)\| \geq \frac{m}{2}\|x - x_\star\|$$

for all x. But, by Jensen's inequality,

$$\|\nabla f(x)\|^2 \leq \mathbb{E}[\|g(x;\xi)\|^2]$$

which implies that the norm of the stochastic gradients cannot be bounded above.

1.2 Case 2: The randomized Kaczmarz method ($B = 0$, $L_g > 0$)

In the randomized Kaczmarz method, we assume that

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}(a_i^T x - b_i)^2 \,,$$

and there exists a point $x_\star$ such that $a_i^T x_\star = b_i$ for all i. In this case

$$f(x) = \frac{1}{2n}\sum_{i=1}^{n}(x - x_\star)^T a_i a_i^T (x - x_\star)$$

and

$$g(x;i) = a_i a_i^T (x - x_\star)$$

Computing the expected norm of the stochastic gradient yields

$$\mathbb{E}[\|g(x;i)\|^2] = \mathbb{E}[\|a_i\|^2 |a_i^T(x - x_\star)|^2] \leq \mathbb{E}[\|a_i\|^4]\|x - x_\star\|^2$$

Hence, we can use the bounds

$$L_g \leq \mathbb{E}[\|a_i\|_2^4]^{1/2} \quad \text{and} \quad B = 0 \,.$$

1.3 Case 3: additive Gaussian noise

Suppose that

$$g(x;\omega) = \nabla f(x) + \omega$$

where ω is a Gaussian random vector with mean 0 and covariance $\sigma^2 I$. In this case,

$$\mathbb{E}[g(x;\omega)] = \nabla f(x)$$

and

$$\mathbb{E}[\|g(x;\omega)\|^2] = \|\nabla f(x)\|^2 + \sigma^2 d$$

Thus, L_g is upper bounded by the Lipschitz constant of the gradient of f. We additionally have that $B \leq \sigma\sqrt{d}$.

1.4 Case 4: The incremental gradient method

For the general incremental gradient method, we have

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

and

$$g(x;i) = \nabla f_i(x)\,.$$

Assume that $\nabla f_i(x)$ has Lipschitz constant L_i. Let $x_\star^{(i)}$ denote any point where $\nabla f_i(x_\star^{(i)}) = 0$. Then we can compute

$$\begin{aligned}
\mathbb{E}[\|g(x;i)\|^2] &= \mathbb{E}[\|\nabla f_i(x)\|^2] \\
&\leq \mathbb{E}[L_i^2\|x - x_\star^{(i)}\|^2] \\
&\leq \mathbb{E}\left[2L_i^2\|x - x_\star\|^2 + 2L_i^2\|x_\star^{(i)} - x_\star\|^2\right] \\
&= \left(\frac{2}{n}\sum_{i=1}^{n} L_i^2\right)\|x - x_\star\|^2 + \frac{2}{n}\sum_{i=1}^{n} L_i^2\|x_\star^{(i)} - x_\star\|^2\,.
\end{aligned}$$

Thus, we can identify

$$L_g \leq \left(\frac{2}{n}\sum_{i=1}^{n} L_i^2\right)^{1/2} \quad \text{and} \quad B \leq \left(\frac{2}{n}\sum_{i=1}^{n} L_i^2\|x_\star^{(i)} - x_\star\|^2\right)^{1/2}\,.$$

There is a nice intuition of this parameter B. If all of the minima of f_i coincide with $x_\star$, then B will equal zero. We will see in the next section that this means that the stochastic gradient method would converge at a linear rate. When B is nonzero, we will only be able to prove convergence at a rate inversely proportional to the iteration counter. However, the smaller B, the faster the convergence.

2 Analysis

The analysis for the different settings of L_g and B are different, but they all start from the same point. We begin by expanding the distance to the optimal solution

$$\begin{aligned}
\|x_{k+1} - x_\star\|^2 &= \|x_k - \alpha_k g_k(x_k;\xi_k) - x_\star\|^2 \\
&= \|x_k - x_\star\|^2 - 2\alpha_k\langle g_k(x_k;\xi_k), x_k - x_\star\rangle + \alpha_k^2\|g_k(x_k;\xi_k)\|^2
\end{aligned}$$

We deal with each term in this expansion separately. First note that if we apply the law of iterated expectation

$$\begin{aligned}
\mathbb{E}[\langle g_k(x_k;\xi_k), x_k - x_\star\rangle] &= \mathbb{E}\left[\mathbb{E}_{\xi_k}[\langle g_k(x_k;\xi_k), x_k - x_\star\rangle \mid \xi_0,\ldots,\xi_{k-1}]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}_{\xi_k}[g_k(x_k;\xi_k) \mid \xi_0,\ldots,\xi_{k-1}], x_k - x_\star\rangle\right] \\
&= \mathbb{E}\left[\langle \nabla f(x_k), x_k - x_\star\rangle\right]\,.
\end{aligned}$$

Here, we are simply using the fact that ξ_k being independent of all of the preceding ξ_i implies that it is independent of x_k. This means that when we iterate the expectation, the stochastic gradient can be replaced by the gradient.

By a similar argument, we can bound the last term as

$$\mathbb{E}[\|g(x_k;\xi_k)\|_2^2] = \mathbb{E}\left[\mathbb{E}_{\xi_k}[\|g(x_k;\xi_k)\|_2^2 \mid \xi_0,\ldots,\xi_{k-1}]\right] \leq \mathbb{E}\left[L_g^2\|x_k - x_\star\|_2^2 + B^2\right]$$

Letting $a_k := \mathbb{E}[\|x_k - x_\star\|^2]$, this gives

$$a_{k+1} \leq (1+\alpha_k^2 L_g^2)a_k - 2\alpha_k \mathbb{E}\left[\langle \nabla f(x_k), x_k - x_\star\rangle\right] + \alpha_k^2 B^2\,. \tag{1}$$

All of our analyses follow from different manipulations of (1). We'll proceed through several cases.

2.1 Case 1: $L_g = 0$.

Let's run the algorithm on a convex f with $L_g = 0$. Let α_k be our stepsize sequence. Define

$$\lambda_k = \sum_{j=0}^{k} \alpha_j\,.$$

This is the sum of all the stepsizes up to iteration k. Also define

$$\bar{x}_k = \lambda_k^{-1} \sum_{j=0}^{k} \alpha_j x_j\,.$$

$\bar{x}_k$ is the average of the iterates weighted by the stepsize. We are going to analyze the deviation of $f(\bar{x}_k)$ from optimality.

Also let $D_0 = \|x_0 - x_\star\|^2$. D_0 is the initial distance to an optimal solution. It is not necessarily a random variable.

To proceed, we just expand the following expression:

$$\begin{aligned}
\mathbb{E}\left[f\left(\bar{x}_T\right) - f(x_\star)\right] &\leq \mathbb{E}\left[\lambda_T^{-1}\sum_{t=0}^{T}\alpha_t(f(x_t) - f(x_\star))\right] && \text{(2a)}\\
&\leq \lambda_T^{-1}\sum_{t=0}^{T}\alpha_t\mathbb{E}[\langle\nabla f(x_t), x_\star - x_t\rangle] && \text{(2b)}\\
&\leq \lambda_T^{-1}\sum_{t=0}^{T}\tfrac{1}{2}(a_t - a_{t+1}) + \tfrac{1}{2}\alpha_t^2 B^2 && \text{(2c)}\\
&= \frac{a_0 - a_{T+1} + B^2\sum_{t=0}^{T}\alpha_t^2}{2\lambda_T} && \text{(2d)}\\
&\leq \frac{D_0^2 + B^2\sum_{t=0}^{T}\alpha_t^2}{2\sum_{t=0}^{T}\alpha_t} && \text{(2e)}
\end{aligned}$$

Here, (2a) follows because f is convex, (2b) is the first order condition for convexity, and (2c) uses (1).

With this in hand, we can now easily prove the following

Proposition 1 (Nemirovski *et al* [?]) *Suppose we run the SGM on a convex f with $L_g = 0$ for T steps with stepsize α. Define*

$$\alpha_{\text{opt}} = \frac{D_0}{B\sqrt{T}}$$

and

$$\theta := \frac{\alpha}{\alpha_{\text{opt}}}$$

Then we have the bound

$$\mathbb{E}[f(\bar{x}) - f_\star] \leq \left(\tfrac{1}{2}\theta + \tfrac{1}{2}\theta^{-1}\right) \frac{BD_0}{\sqrt{T}}\,. \tag{3}$$

This proposition asserts that we pay linearly for errors in selecting the optimal constant stepsize. If we guess a constant stepsize that is two-times or one-half of the optimal choice, then we need to run for twice as many iterations.

We can prove this just by minimizing (2e). Plugging in our stepsize, we see that

$$\mathbb{E}\left[f\left(\bar{x}_T\right) - f(x_\star)\right] \leq \frac{D_0^2 + B^2 T\alpha^2}{2T\alpha} = \left(\tfrac{1}{2}\theta^{-1} + \tfrac{1}{2}\theta\right) \frac{BD_0}{\sqrt{T}}\,.$$

Other stepsizes could also be selected here, including diminishing stepsizes. But the constant stepsize, it turns out, is optimal for this upper bound.

2.2 Case 2: $B = 0$

When $B = 0$, we surprisingly get a *linear rate* of convergence provided that f is strongly convex. Assume f has strong convexity parameter m. Then we have the inequality

$$\langle \nabla f(x), x - x_\star \rangle \geq m\|x - x_\star\|^2\,.$$

Plugging this into (1) with $B = 0$ gives the recursion

$$a_{k+1} \leq (1 - 2m\alpha_k + L_g^2\alpha_k^2)a_k\,. \tag{4}$$

Choosing any constant α in the range $(0, 2m/L_g^2)$ gives a linear rate of convergence. The optimal rate is when $\alpha = \frac{m}{L_g^2}$, in which case we have

$$a_k \leq \left(1 - \frac{m^2}{L_g^2}\right)^k D_0\,,$$

and

$$T = \frac{L_g^2}{m^2} \log\left(\frac{D_0}{\epsilon}\right)$$

suffice to guarantee that $\mathbb{E}[\|x_T - x_\star\|^2] \leq \epsilon$.

2.3 Case 3: B and L_g both nonzero

We finally turn to analyzing the general form of the recursion (1) when f is strongly convex:

$$a_{k+1} \leq (1 - 2m\alpha_k + \alpha_k^2 L_g^2)a_k + \alpha_k^2 B^2$$

Constant stepsize First, consider the case of a constant stepsize. Assuming that $\alpha \in (0, \frac{m}{L_g^2})$, we can roll the recursion out to find

$$a_k \leq (1 - 2m\alpha + \alpha^2 L_g^2)^k D_0 + \frac{\alpha B^2}{2m - \alpha L_g^2}\,.$$

What we see is that no matter how long we run, we can't prove that a_k converges to 0. Indeed, even "at infinity" (taking the limit as k tends to infinity), we see that

$$\lim_{k\to\infty} a_k \leq \frac{\alpha B^2}{2m - \alpha L_g^2}$$

This is actually representative of what happens in practice. The iterates converge to a ball around the optimal solution, but bounce around inside of this ball. To get closer to the optimal solution, we can reduce the stepsize α. But then, the rate at which we converge to the optimal ball is controlled by the quantity $1 - 2m\alpha + \alpha^2 L_g^2$. This number tends to 1 as α tends to zero.

One way to balance these two effects is to use *epochs*. The idea is to run with an agressively large stepsize for T iterations. Then, we halve the stepsize and double the number of iterations. This has the effect of shrinking the radius about the optimal solution, and guarantees we get close to this radius by running for an extended period of time.

Diminishing stepsize Note that by running in epochs, we are making a piecewise approximation to the stepsize C/k. We can use such a stepsize and get direct convergence to the optimum. Suppose we choose the stepsize

$$\alpha_k = \frac{\gamma}{k_0 + k}$$

for some constants γ and k_0. We will now show that for certain choices of these constants, we can guarantee that

$$a_k \leq \frac{Q}{k_0 + k}\,.$$

One can prove by induction the following

Proposition 2 *Suppose f is strongly convex with parameter m. If we run the stochastic gradient method with stepsize*

$$\alpha_k = \frac{1}{2m(L_g^2/2m^2 + k)}$$

then

$$\mathbb{E}[\|x_k - x_\star\|^2] \leq \frac{B^2}{2m(L_g^2/2m^2 + k)}$$