10-425/625: Introduction to Convex Optimization (Fall 2023)

Lecture 13: SGD

Instructor:[1] *Matt Gormley* *October 9, 2023*

Today we'll continue our discussion of the stochastic gradient descent (SGD) algorithm for unconstrained optimization.

13.1 Some Examples of SGD Algorithms

Recall from a previous lecture

Last time we considered a variety of examples of SGD-like algorithms:

1. Noisy gradients
2. Incremental gradient method (IGM) (not an SGD algorithm)
3. Randomized IGM
4. IGM with Random Permutations (not quite an SGD algorithm)
5. SGD with momentum (not an SGD algorithm)

13.1.1 Mini-Batch SGD

In the ERM setting, or in the incremental gradient setting we are not restricted to using a single sample (or single function) to compute our stochastic gradient. Often in practice (due to various communication, data-manipulation bottlenecks) it will be faster to choose subsets $I_t \subset \{1, \ldots, n\}$ of size m (say), and compute:

$$x^{t+1} = x^t - \eta_t \frac{1}{m} \sum_{i \in I_t} \nabla f_i(x^t).$$

If the subsets are chosen uniformly at random from $\{1, \ldots, n\}$ then this is a valid stochastic gradient. It has a variance which is a factor of m smaller

[1] These notes were originally written by Siva Balakrishnan for 10-725 Spring 2023 (original version: here) and were edited and adapted for 10-425/625.

(but can be m times more expensive to compute). In practice, m is a hyper-parameter which needs to be tuned carefully.

13.2 A Warm-Up Example

We'd like to develop an understanding of the rates of convergence of the SGD algorithm, and perhaps some insights on step-size choices, and some insights on the role of the variance (at least intuitively, it should be the case that the variance of the stochastic gradients affects how fast the algorithm converges).

Example 13.1 (Incremental Gradient Method)**.** Suppose our goal is to optimize a very simple quadratic objective:

$$\min_x \frac{1}{2n} \sum_{i=1}^{n} \|X_i - x\|_2^2.$$

Suppose we start at $x^0 = 0$. Now, the incremental gradient algorithm would use the updates for $t = \{0, \ldots, n-1\}$.

$$x^{t+1} = x^t - \eta_t(x^t - X_{t+1}) = (1 - \eta_t)x^t + \eta_t X_{t+1}.$$

If we use the step-size $\eta_t = \frac{1}{t+1}$, then we have that,

$$x^{t+1} = \frac{tx^t + X_{t+1}}{t+1}.$$

After n iterations the incremental gradient algorithm would converge to the optimal solution (just the average of $X_1, \ldots, X_n$). One maybe shouldn't take too much away from this example (it's not even an SGD algorithm) but notice that even in this extremely favorable case (smooth, strongly convex objective) we needed our step-sizes to decay at the rate $1/(t+1)$.

Example 13.2 (One-pass SGD)**.**

One-pass SGD is a bit more interesting to study. Suppose we are interested in optimizing the population objective:

$$\min_x \frac{1}{2} \mathbb{E}_{X \sim P} \|X - x\|_2^2.$$

We obtain samples $X_1, \ldots, X_n$ from P. Lets suppose that P has mean μ and variance σ^2. From each sample, we can compute a stochastic gradient $g(x^t, X_i) = X_i - x^t$, and use this in an SGD algorithm. Suppose we use step-sizes $\eta_t = 1/(t+1)$, and $x^0 = 0$ as above. In this case, after n iterations we obtain the solution,

$$\widehat{x} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Now, we can evaluate the quality of $\widehat{x}$ via its objective value,

$$\frac{1}{2}\mathbb{E}_{X \sim P} \|X - \widehat{x}\|_2^2 = \frac{\sigma^2}{2} + \frac{\sigma^2}{2n}.$$

On the other hand the optimal solution x^* is the population mean, which achieves the objective value,

$$\frac{1}{2}\mathbb{E}_{X \sim P} \|X - \mu\|_2^2 = \frac{\sigma^2}{2}.$$

So we see that,

$$f(\widehat{x}) - f(x^*) = \frac{\sigma^2}{2n}.$$

Notice that:

1. Even in this favorable case (smooth, strongly convex objective) we obtain $1/k$-type rates of convergence.
2. Furthermore, we know that this cannot be improved in this case. Standard statistical lower bounds will tell us that the sample mean is the best possible estimator here, and that it's excess error scales exactly like $\sigma^2/2n$.
3. In this case, the SGD algorithm which processes a single sample at a time, and makes a step after each sample, is as good as any estimator which uses all of the samples $X_1, \ldots, X_n$ at once.

13.3 SGD for Lipschitz Convex Functions

Background: (Tower Property of Conditional Expectations) Given two random variables X and Y, the *tower property of conditional expectations* states that:

$$\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$$
$$\sum_x xp(x) = \sum_y p(y) \sum_x xp(x|y)$$

We will now turn our attention to some formal results for the SGD algorithm. We'll analyze SGD for non-smooth functions, and here the hypothesis will be that,

$$\mathbb{E}_\xi[g(x,\xi)] \in \partial f(x).$$

Theorem 13.3. *Suppose that f is convex, our initialization satisfies $\|x^0 - x^*\|_2 \leq R$ (for some, not necessarily unique, minimizer x^* which is fixed throughout the proof), and the stochastic gradients satisfy,*

$$\mathbb{E}\|g(x,\xi)\|_2^2 \leq G^2 \quad \textit{for all } x,$$

then if we choose $\eta = \frac{R}{G\sqrt{k}}$, we have the guarantee that,

$$\mathbb{E}f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{RG}{\sqrt{k}}.$$

Notice, the main differences between our earlier result for the subgradient method and this result:

1. We obtain a guarantee that holds in expectation, and we obtain a guarantee for the averaged iterate (similar bounds hold in high-probability and for the last iterate but are a bit more difficult to prove).
2. We make a different hypothesis, essentially that the stochastic gradients are bounded. This in some sense bounds the variance of the stochastic gradients (as well as the magnitude of the actual gradients).

3. The SGD algorithm here can be much faster than the subgradient method (at least in the ERM type problems we discussed earlier). It achieves the same rate of convergence as a function of k but each iteration of SGD can be much faster than a corresponding iteration of the sub-gradient method.

Proof: The proof is very similar to that of the subgradient method, except that we use expectations (and conditional expectations) at various points. We're using a fixed step-size across iterations. As usual we have that,

$$\|x^{t+1} - x^*\|_2^2 = \|x^t - x^*\|_2^2 + \eta^2 \|g(x^t, \xi)\|_2^2 - 2\eta(x^t - x^*)^T g(x^t, \xi).$$

Now, take a conditional expectation of both sides, $\mathbb{E}_\xi[\cdot \mid x^t]$:

$$\begin{aligned}\mathbb{E}_\xi\left[\|x^{t+1} - x^*\|_2^2 | x^t\right] &= \|x^t - x^*\|_2^2 + \eta^2 \mathbb{E}_\xi\left[\|g(x^t, \xi)\|_2^2 | x^t\right] - 2\eta(x^t - x^*)^T \mathbb{E}_\xi[g(x^t, \xi)|x^t] \\ &\leq \|x^t - x^*\|_2^2 + \eta^2 G^2 - 2\eta(x^t - x^*)^T g_{x^t},\end{aligned}$$

where $g_{x^t} \in \partial f(x^t)$. Now, we use convexity on the last term to obtain that,

$$\mathbb{E}_\xi\left[\|x^{t+1} - x^*\|_2^2 | x^t\right] \leq \|x^t - x^*\|_2^2 + \eta^2 G^2 + 2\eta(f(x^*) - f(x^t)),$$

and therefore taking an expectation under x_t of both sides, $\mathbb{E}_{x^t}[\cdot]$ and by the tower property of conditional expectations, $\mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$,

$$\mathbb{E}_\xi\left[\|x^{t+1} - x^*\|_2^2\right] \leq \mathbb{E}_{x^t}\left[\|x^t - x^*\|_2^2\right] + \eta^2 G^2 + 2\eta(f(x^*) - \mathbb{E}f(x^t)).$$

Re-arranging and telescoping the sum we obtain that,

$$\frac{1}{k}\sum_{t=1}^{k} \mathbb{E}f(x^t) - f(x^*) \leq \frac{G^2\eta}{2} + \frac{\|x^0 - x^*\|_2^2}{2k\eta}.$$

Now, by convexity we know that,

$$f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) \leq \frac{1}{k}\sum_{t=1}^{k} f(x^t), \tag{13.1}$$

so we obtain that,

$$\mathbb{E}f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{G^2\eta}{2} + \frac{\|x^0 - x^*\|_2^2}{2k\eta},$$

and using our choice of step-size $\eta = \frac{R}{G\sqrt{k}}$ this gives the desired result.

■

13.4 SGD for Strongly Convex Functions

The key takeaway from this section is that for strongly convex functions, SGD does not achieve a linear rate of convergence (and additionally assuming smoothness makes no difference). This is primarily due to the variance of the stochastic gradients, and in some later lecture we might discuss tools for variance reduction in SGD (which do in some cases yield algorithms with linear convergence rates for structured smooth and strongly convex functions).

Theorem 13.4. *Suppose f is α-strongly convex, and the stochastic gradients satisfy,*

$$\mathbb{E}\|g(x,\xi)\|_2^2 \leq G^2 \quad \textit{for all } x.$$

Then,

1. *For a fixed step-size $\eta < 1/\alpha$, we obtain,*

$$\mathbb{E}\|x^k - x^*\|_2^2 \leq (1-\alpha\eta)^k \|x^0 - x^*\|_2^2 + \frac{\eta G^2}{\alpha}.$$

2. *For $\eta_t = \frac{1}{\alpha(t+1)}$,*

$$\mathbb{E}f\left(\frac{1}{k}\sum_{t=1}^{k} x^t\right) - f(x^*) \leq \frac{G^2(1+\log k)}{2\alpha k}.$$

It is worth noticing:

1. The first result suggests that SGD iterates with a fixed step-size, will converge rapidly to some fixed ball around x^* and then bounce around there. This in turn suggests a very common practical epoch-based heuristic for SGD step-sizes – run it with some fixed step-size, when it seems like the iterates are bouncing around (or you stop making progress in function value), then decay it by some factor and continue running it.
2. In the second case, one can remove the extra log factor with some work – for instance, if you use an SGD variant where rather than average all the iterates you only average the last half the log factor can be eliminated.

Proof: Suppose we follow our earlier proof to obtain that,

$$\begin{aligned}\mathbb{E}\left[\|x^{t+1}-x^*\|_2^2|x^t\right] &= \|x^t-x^*\|_2^2+\eta_t^2\mathbb{E}\left[\|g(x^t,\xi)\|_2^2|x^t\right]-2\eta_t(x^t-x^*)^T\mathbb{E}[g(x^t,\xi)|x^t] \\ &\leq \|x^t-x^*\|_2^2+\eta_t^2G^2-2\eta_t(x^t-x^*)^T\nabla f(x^t).\end{aligned}$$

The key point to notice here is that previously we would have used the descent lemma (a consequence of smoothness) to bound the squared norm of the gradient. However, in the current stochastic gradient setup, the expected squared norm of the gradient includes two contributions: one which is roughly the squared norm of the expected gradient which we could hope to control by smoothness, and the second which is the variance of the stochastic gradients. This latter term, we should not in general expect to decrease as we get close to the optimum.

Now, using strong convexity on the last term we obtain that,

$$\mathbb{E}\left[\|x^{t+1}-x^*\|_2^2|x^t\right] \leq \|x^t-x^*\|_2^2+\eta_t^2G^2-\alpha\eta_t\|x^t-x^*\|_2^2+2\eta_t(f(x^*)-f(x^t)). \tag{13.2}$$

Proof of Claim 1: Now, to prove the first claim we use a fixed step-size η and see that,

$$\mathbb{E}\left[\|x^{t+1}-x^*\|_2^2\right] \leq (1-\alpha\eta)\mathbb{E}[\|x^t-x^*\|_2^2]+\eta^2G^2,$$

and so provided that $\alpha\eta<1$ we can unroll this recursion to obtain,

$$\mathbb{E}\left[\|x^k-x^*\|_2^2\right] \leq (1-\alpha\eta)^k\|x^0-x^*\|_2^2+\frac{\eta G^2}{\alpha}.$$

Proof of Claim 2: Rearranging (13.2), and using the tower property, we see that,

$$\mathbb{E}f(x^t)-f(x^*) \leq \frac{\mathbb{E}\left[\|x^t-x^*\|_2^2\right]-\mathbb{E}[\|x^{t+1}-x^*\|_2^2]}{2\eta_t}+\frac{\eta_tG^2}{2}-\frac{\alpha}{2}\mathbb{E}[\|x^t-x^*\|_2^2].$$

Now, one can verify that with our choice of step-sizes $\eta_t=1/\alpha(t+1)$ the first two and last terms together telescope, and we are left with $-\alpha k\|x^{k+1}-x^*\|_2^2$ which is negative and can be dropped. Thus we obtain the bound,

$$\sum_{t=0}^{k}\left[\mathbb{E}f(x^t)-f(x^*)\right] \leq \frac{G^2}{2\alpha}\sum_{t=0}^{k}\eta_t \leq \frac{G^2(1+\log k)}{2\alpha}.$$

Using the same idea as in (13.1) we obtain the final bound.

■