CSE 547/Stat 548: Machine Learning for Big Data **Lecture**

SGD and Averaging

Instructor: Sham Kakade

1 SGD and optimality

There is a strong sense in which SGD can be made "optimal", if we perform averaging. SGD itself is really not optimal, from a statistical perspective. Analyzing these issues is subtle. However, just examining things in one dimensional already provides much of the insight.

2 Background

Stochastic gradient descent is among the most commonly used practical algorithms for large scale stochastic optimization. The seminal result of Ruppert [1988], Polyak and Juditsky [1992] formalized this effectiveness, showing that for certain (locally quadric) problems, asymptotically, stochastic gradient descent is statistically minimax optimal (provided the iterates are averaged). There are a number of more modern proofs Dieuleveut and Bach [2015], Défossez and Bach [2015], Jain et al. [2017] of this fact, which provide finite rates of convergence.

This lecture will look at a short proof of this minimax optimality for SGD, with averaging, in the one dimensional case. See Polyak and Juditsky [1992] for a self contained proof of this for the case of least squares and Polyak and Juditsky [1992] for a more general treatment.

3 The one dimensional case

The expected square loss for $w \in \mathbb{R}$ over $y \in \mathbb{R}$ sampled from a distribution $\mathcal{D}$, is:

$$L(w) = \mathbb{E}_{y\sim\mathcal{D}}[(y-w)^2]$$

The optimal weight is simply the mean, denoted by:

$$w^* := \mathbb{E}[y] := \arg\min_w L(w)\,.$$

Stochastic gradient descent proceeds as follows: at each iteration t, using an i.i.d. sample $y_t \sim \mathcal{D}$, the update of w_t is:

$$w_t = w_{t-1} + \gamma_t(y_t - w_{t-1})\,.$$

Clearly, to obtain convergence of w_t, we must decay the stepsize. If we knew a stopping time T in advance, we could set T as a function of the stopping time. For simplicity, let us consider a γ to be a fixed stepsize.

$$w_t = w_{t-1} + \gamma_t(y_t - w_{t-1})\,.$$

The statistically optimal rate. Using n samples (and for large enough n), the minimax optimal rate is achieved by the sample mean (more generally, the maximum likelihood estimator, or, equivalently, the empirical risk minimizer). Denote the variance as:

$$\sigma^2 := \mathbb{E}\left[(y - \mathbb{E}[y])^2\right] .$$

Given n i.i.d. samples $\{(y_i)\}_{i=1}^n$, the best estimator is sample mean:

$$\widehat{w}_n^{\text{SampMean}} := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(y_i - w \cdot x_i\right)^2 .$$

This optimal (among estimators) is characterized as follows:

$$\mathbb{E}[L(\widehat{w}_n^{\text{SampMean}})] - L(w^*) = \frac{\sigma^2}{n}$$

4 SGD itself isn't all that great...

This is a little surprising, that even in one dimension, SGD doesn't really get it right.

SGD with a constant learning rate. Define the noise in iteration t (of the t-th sample) as:

$$\epsilon_t := \mathbb{E}[y] - y_t,$$

which is a mean 0 quantity. The SGD rule can be written as:

$$\begin{aligned} w_t - w^* &= w_{t-1} - w^* + \gamma(y_t - w_{t-1}) \\ &= (1-\gamma)(w_{t-1} - w^*) - \gamma\epsilon_t \end{aligned}$$

.

Roughly speaking, the above shows how the process on $w_t - w^*$ consists of a contraction along with an addition of a zero mean quantity.

From recursion,

$$w_t - w^* = (1-\gamma)^t(w_0 - w^*) - \gamma\left(\epsilon_t + (1-\gamma)\epsilon_{t-1} + \cdots + (1-\gamma)^t\epsilon_1\right) .$$

Lemma 4.1. *We have that:*

$$\begin{aligned} \mathbb{E}[L(w_t)] - L(w^*) = \mathbb{E}[(w_t - w^*)^2] &= (1-\gamma)^t(w_0 - w^*)^2 + \gamma^2\sigma^2\sum_{\tau=0}^{t-1}(1-\gamma)^{2t} \\ &\le \exp\left(-\gamma t\right)(w_0 - w^*)^2 + \gamma\sigma^2 . \end{aligned}$$

Proof. That $\mathbb{E}[L(w_t) - L(w^*) = \mathbb{E}[(w_t - w^*)^2]$ is straight forward. To prove the equality, we note that the noise is mean 0, i.e. $\mathbb{E}[\epsilon_t] = 0$, and the noise is independent, so $\mathbb{E}[\epsilon_t\epsilon_{t'}] = 0$ for $t \neq t'$. Hence, we have that:

$$\mathbb{E}[(w_t - w^*)^2] = (1-\gamma)^t(w_0 - w^*) - \gamma^2\left(\sigma^2 + (1-\gamma)^2\sigma^2 + \cdots + (1-\gamma)^{2t}\sigma^2\right) . .$$

which leads to the first claim. The last step simply follows from summing the geometric series, to obtain:

$$\gamma^2\sum_{\tau=0}^{t-1}(1-\gamma)^t \le \frac{\gamma^2}{1-(1-\gamma)^2} \le \frac{\gamma^2}{1-(1-\gamma)} = \gamma$$

which completes the proof. □

There is no good choice of a learning rate. We would ideally hope for a rate that (eventually) matches σ^2/n which is the rate of the sample average.

Note that our derivation is exact (the only step with an inequality is summing the geometric series, which is looses very little).

Let us try out a few learning rates to get intuition. Let us consider $\gamma = 1/2$. (let's rule out $\gamma = 1$, as we jump to the 0 bias in one step. This does not hold for other problems, e.g. in more than one dimension such as for regression.) Here we have:

$$\mathbb{E}[(w_t - w^*)^2] \leq \exp(-t/2)(w_0 - w^*)^2 + \frac{\sigma^2}{2}\,.$$

The first term (the bias) is dropping extremely quickly (geometrically). The variance is not even going to 0, so, of course, this is extremely poor.

Setting $\eta = 1/\sqrt{t}$, we have:

$$\mathbb{E}[(w_t - w^*)^2] \leq \exp(-\sqrt{t})(w_0 - w^*)^2 + \frac{\sigma^2}{\sqrt{t}}\,.$$

which at least goes to 0. However, the bias term is dropping much more slowly than before, and the variance term, while going to 0, is much worse than the rate of the sample average!

Now, let us look at $\eta = 1/T$. Here we have:

$$\mathbb{E}[(w_t - w^*)^2] \leq \exp(-1)(w_0 - w^*)^2 + \frac{\sigma^2}{t}\,.$$

Here, the variance term is in fact dropping at the optimal rate. However, the bias term does not even go 0, so, overal, this is another extremely poor choice.

Is there really no choice of time varying learning rate?.

One may hope to instead try a decaying learning scheme, where we set γ_t and decay it over time (rather than just setting as a function of the stopping time). This does not improve things (though a decaying γ as $O(1/t)$ does improve upon the previous bounds).

The reader might not that restart scheme will work here, if one thinks this through.

More generally (if we move to regression), there is in fact no decaying learning rate scheme which is optimal (one can prove a lower bound on this); meaning that the statistical minimax rate will not be reached.

5 Iterate Averaging

Remarkably, an extremely procedure will give us the best of both worlds, where our bias drops geometrically and the variance is optimal (to within a constant). The approach is extremely effective even in non-convex settings.

The iterate averaging algorithm does not actually change SGD algorithm. Instead, you just keep track of a running average of your w_t's (say starting at some point in time) and you use that instead of using the last point w_t. Note that you still just run your usual SGD algorithm as you were doing before!

Denote the average iterate, averaged from iteration t to T, by:

$$\overline{w}_{t:T} := \frac{1}{T-t}\sum_{t'=t}^{T-1} w_{t'}\,.$$

Note there is a choice of when to start averaging. In practice, we often cycle through our dataset.

Theorem 5.1. *Suppose $\gamma < 1/2$. The risk is bounded as:*

$$\mathbb{E}[(\overline{w}_{t:T} - w^*)^2] \leq 2\exp\left(\big(-t/2\big)(w_0 - w^*)^2 + 4\frac{\sigma^2}{T-t}\right).$$

For $t = T/2$, i.e. we start our average over the second half of the samples, we have:

$$\mathbb{E}[(\overline{w}_{T/2:T} - w^*)^2] \leq 2\exp\left(\big(-T/4\big)(w_0 - w^*)^2 + 8\frac{\sigma^2}{T}\right).$$

We have that, with iterate averaging, the bias term (the first term) decays at a geometric rate, and the variance term is withint a constant factor of the optimal variance (we have not optimized this constant). Also, even if did not know T in advance, it is easy enough to maintain multiple running averages (or restart the running average).

Thus, iterative averaging gives the best of both worlds! And this phenomena seems to far more general. Empirically, it also works very well in non-convex cases.

This theorem is really a simple special case of results in the literature, e.g. see Jain et al. [2017]. It is not particularly difficult to prove in the 1-dimensional case.

References

Alexandre Défossez and Francis R. Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *AISTATS*, volume 38, 2015.

Aymeric Dieuleveut and Francis R. Bach. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 2015.

P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and V. Pillutla andA. Sidford. A markov chain theory approach to characterizing the minimax optimality of stochastic gradient descent (for least squares). In *37th Foundations of Software Technology and Theoretical Computer Science, 2017*, 2017.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, volume 30, 1992.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Tech. Report, ORIE, Cornell University*, 1988.