Lecture Notes 4: Non-Convex Stochastic Gradient Descent and Decreasing Learning Rates

Instructor: Ashok Cutkosky

Previously, we showed that gradient descent can find critical points of smooth functions in the deterministic setting. Now, we'll consider the stochastic setting.

Algorithm 1 Stochastic Gradient Descent

Input: Initial Point $\mathbf{w}_1$, learning rates $\eta_1, \dots, \eta_T$, time horizon T.
for $t = 1 \dots T$ **do**
  Sample $z_t \sim P_z$.
  Set $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$.
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$.
end for

The classic analysis of stochastic gradient descent is as follows:

Theorem 1. *Suppose $\mathcal{L}$ is H-smooth, and $\nabla \ell(\mathbf{w}, z)$ has variance at most σ^2 for all $\mathbf{w}$ (that is for all $\mathbf{w}$, $\mathbb{E}_z[\|\nabla \ell(\mathbf{w}, z) - \nabla \mathcal{L}(\mathbf{w})\|^2] \le \sigma^2$). Let us consider SGD with a* fixed *learning rate $\eta_t = \eta$ for all t. Then so long as $\eta \le \frac{1}{H}$, Algorithm 1 guarantees:*

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{2\Delta}{\eta} + HT\eta\sigma^2$$

Further, if we set $\eta = \min\left(\frac{1}{H}, \frac{\sqrt{\Delta}}{\sigma\sqrt{HT}}\right)$, then:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le 2\Delta H + 3\sigma\sqrt{\Delta H T}$$

, then if $\hat{\mathbf{w}}$ is selected uniformly at random from $\mathbf{w}_1, \dots, \mathbf{w}_T$,

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \le \frac{\sqrt{2\Delta H}}{\sqrt{T}} + \frac{\sqrt{3\sigma\sqrt{\Delta H}}}{T^{1/4}}$$

Notice that the second statement of the Theorem bounds the expected gradient norm by a sum of two terms. The first term is identical to the bound for non-stochastic gradient descent, while the second term depends on the variance σ and has a slower $O(1/T^{1/4})$ rate of dependence on the time T.

Proof. Again, let's use our understanding of smooth losses to bound the progress made in one step of stochastic gradient descent:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) &\le \mathcal{L}(\mathbf{w}_t) + \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= \mathcal{L}(\mathbf{w}_t) - \eta \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle + \frac{H\eta^2}{2}\|\mathbf{g}_t\|^2
\end{aligned}$$

Now, in deference to the randomness of our situation, we take the expected value of both sides:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\, \mathbb{E}[\nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle] + \frac{L\eta^2}{2}\, \mathbb{E}[\|\mathbf{g}_t\|^2]$$

Now, use the fact that $\mathbb{E}[\mathbf{g}_t|\mathbf{w}_t] = \nabla\mathcal{L}(\mathbf{w}_t)$, so that:

$$= \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta^2}{2}\,\mathbb{E}[\|\mathbf{g}_t\|^2]$$

From bias variance decomposition:

$$\begin{aligned}
&\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta^2}{2}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 + \sigma^2] \\
&= \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta(1 - \eta L/2)\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta^2\sigma^2}{2}
\end{aligned}$$

Since $\eta \leq \frac{1}{L}$:

$$\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{\eta}{2}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta^2\sigma^2}{2}$$

Summing over t and telescoping:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_1)] &\leq -\sum_{t=1}^{T}\frac{\eta}{2}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta^2\sigma^2}{2} \\
\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\leq \frac{2\Delta}{\eta} + HT\eta\sigma^2
\end{aligned}$$

This proves the first part of the Theorem. Now, for the second part we consider the provided setting for η. There are two cases, either $\eta = \frac{1}{H} \leq \frac{\sqrt{\Delta}}{\sigma\sqrt{HT}}$ or not. If $\eta = \frac{1}{H}$, then:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\leq \frac{2\Delta}{\eta} + HT\eta\sigma^2 \\
&= 2\Delta H + HT\eta\sigma^2
\end{aligned}$$

Further, since $\eta \leq \frac{\sqrt{\Delta}}{\sigma\sqrt{HT}}$,

$$HT\eta\sigma^2 \leq \sigma\sqrt{\Delta HT}$$

so altogether:

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq 2\Delta H + \sigma\sqrt{\Delta HT}$$

Next, consider the case $\eta = \frac{\sqrt{\Delta}}{\sigma\sqrt{HT}}$. Then, by plugging in η, we have:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\leq \frac{2\Delta}{\eta} + HT\eta\sigma^2 \\
&= 3\sigma\sqrt{\Delta HT}
\end{aligned}$$

so overall, we have that $\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2]$ is bounded by the maximum of these quantities, which is

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq 2\Delta H + 3\sigma\sqrt{\Delta HT}$$

Now, for the final statement of the theorem, divide by T and apply Jensen's inequality:

$$\begin{aligned}
\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|] &\le \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2]} \\
&\le \sqrt{\frac{2\Delta H}{T} + \frac{3\sigma\sqrt{\Delta H}}{\sqrt{T}}} \\
&\le \frac{\sqrt{2\Delta H}}{\sqrt{T}} + \frac{\sqrt{3\sigma\sqrt{\Delta H}}}{T^{1/4}}
\end{aligned}$$

□

Theorem 1 has the problem that the learning rate η is set based on various parameters like H, σ and T, which are presumably not actually known. In practice, the common strategy is to simply guess the learning rate. That is:

1. Try several learning rates out.
2. Choose the one that resulted in best performance on a validation set.

However, it *is* possible to make some guarantees without requiring detailed settings for η. The standard approach is to set $\eta_t \propto \frac{1}{\sqrt{t}}$, and to rely on some kind of Lipschitz assumption. For example, one could assume that $\|\nabla\ell(\mathbf{w}, z)\| \le G$ always. This would be implied it $\ell(\mathbf{w}, Z)$ is G-Lipschitz as a function of $\mathbf{w}$. In fact, we can make do with a slightly weaker assumption that $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z)\|^2] \le G^2$. Note that we do not need to *know* G in order to guarantee convergence, although knowing it might allow us to set the c coefficient in η_t more optimally.

Theorem 2. *Suppose $\mathcal{L}$ is H-smooth, and $\nabla\ell(\mathbf{w}, z)$ satisfies $\mathbb{E}_z[\|\nabla\ell(\mathbf{w}, z)\|] \le G^2$ for all $\mathbf{w}$. Let $\eta_1 \ge \dots \ge \eta_T$ be an arbitrary* deterministic *and decreasing learning rate schedule. Then:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \le \frac{\Delta}{T\eta_T} + \frac{HG^2}{2T\eta_T}\sum_{t=1}^{T}\eta_t^2$$

Next, set $\eta_t = \frac{c}{\sqrt{t}}$ for some c. Then:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \le \frac{\Delta}{c\sqrt{T}} + \frac{HG^2c(1+\log(T))}{2\sqrt{T}}$$

In particular, if $\hat{\mathbf{w}}$ is randomly selected from $\mathbf{w}_1, \dots, \mathbf{w}_T$, then

$$\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|] \le \frac{\sqrt{\Delta/c + G^2cH(1+\log(T)/2)}}{T^{1/4}}$$

Proof. Again we have

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) &\le \mathcal{L}(\mathbf{w}_t) + \langle\nabla\mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle + \frac{H}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= \mathcal{L}(\mathbf{w}_t) - \eta_t\langle\mathcal{L}(\mathbf{w}_t), \mathbf{g}_t\rangle + \frac{H\eta_t^2}{2}\|\mathbf{g}_t\|^2 \\
\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] &\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta_t\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta_t^2G^2}{2}
\end{aligned}$$

Again we sum over t and telescope:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_1)] - \sum_{t=1}^{T}\eta_t\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta_t^2G^2}{2}$$

Use the fact that $\eta_T \leq \eta_t$ for all t:

$$\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_1)] - \eta_T \sum_{t=1}^T \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{HG^2}{2} \sum_{t=1}^T \eta_t^2$$

rearrange terms:

$$\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \leq \frac{\mathbb{E}[\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_{T+1})]}{T\eta_T} + \frac{HG^2}{2T\eta_T} \sum_{t=1}^T \eta_t^2$$

so that we have shown the first part of the Theorem. Now, we get to use an identity that will become useful time and time again:

$$\begin{aligned}
\sum_{t=1}^T \frac{1}{t} &= 1 + \sum_{t=2}^T \frac{1}{t} \\
&\leq 1 + \int_1^T \frac{dt}{t} \\
&= 1 + \log(T)
\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] &\leq \frac{\mathbb{E}[\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_{T+1})]}{T\eta_T} + \frac{HG^2}{2T\eta_T} \sum_{t=1}^T \eta_t^2 \\
&= \frac{\Delta}{c\sqrt{T}} + \frac{cHG^2}{2\sqrt{T}} \sum_{t=1}^T \frac{1}{\sqrt{t}} \\
&\leq \frac{\Delta}{c\sqrt{T}} + \frac{cHG^2(1+\log(T))}{2\sqrt{T}}
\end{aligned}$$

The final statement follows from taking square roots and using Jensen inequality. □

1 Minibatch SGD

While SGD is the basis for almost all the popular algorithms in use for training neural networks today, most of the time a number of different modifications are put in place. By far the most common change is the use of *minibatches* (in fact, minibatches are almost never *not* used). Minibatching is a way to leverage parallelism to speed up the total amount of time taken to train a model. The entire training set is called the "batch", and a random small subset of the training set is called a "minibatch". Using a minibatch is straightforward: any time you wish to call the stochastic gradient oracle, instead call it B times and return the average of those B vectors. B is called the minibatch size. The code for the most basic version of minibatch SGD is below:

Algorithm 2 Minibatch Stochastic Gradient Descent

Input: Initial Point $\mathbf{w}_1$, learning rates $\eta_1, \dots, \eta_T$, time horizon T, batch size B.
for $t = 1 \dots T$ **do**
  for $i = 1 \dots B$ **do**
    $\mathbf{g}_{t,i} = \nabla \ell(\mathbf{w}_t, z_{(t-1)B+i})$
  end for
  Set $\mathbf{g}_t = \frac{1}{B}\sum_{i=1}^B \mathbf{g}_{t,i}$.
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$.
end for

Let's analyze this algorithm. In order to do so, we are going to re-use the analysis of Theorem 1 by considering $\mathbf{g}_t$ as the output of a stochastic gradient oracle with variance $\frac{\sigma^2}{B}$. Specifically:

Proposition 3. *In Algorithm 2, suppose* $\mathbb{E}_{\mathbf{z}}[\|\nabla \ell(\mathbf{w}, \mathbf{z}) - \nabla \mathcal{L}(\mathbf{w})\|^2] \le \sigma^2$ *for all* $\mathbf{w}$*. Then* $\mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{\sigma^2}{B}$*.*

Proof. This is a standard property of averages: they decrease the variance by the number of averaged items.

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2] &= \mathbb{E}\left[\left\|\frac{1}{B}\left(\sum_{i=1}^{B} \mathbf{g}_{t,i} - \nabla \mathcal{L}(\mathbf{w}_t)\right)\right\|^2\right] \\
&= \frac{1}{B^2}\left(\mathbb{E}\left[\sum_{i=1}^{B} \|\mathbf{g}_{t,i} - \nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] + \mathbb{E}\left[\sum_{i \neq j} \langle \mathbf{g}_{t,i} - \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_{t,j} - \nabla \mathcal{L}(\mathbf{w}_t)\rangle\right]\right)
\end{aligned}$$

Using $\mathbb{E}[\|\mathbf{g}_{t,i} - \nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \sigma^2$ and $\mathbb{E}[\mathbf{g}_{t,i} - \nabla \mathcal{L}(\mathbf{w}_t)] = 0$,

$$\le \frac{B\sigma^2}{B^2} = \frac{\sigma^2}{B}$$

□

Now, we notice that Algorithm 2 is actually the same as Algorithm 1, with the difference that the gradient estimates $\mathbf{g}_t$ have variance decreased to σ^2/B. Therefore by Theorem 1:

Theorem 4. *Suppose* $\mathcal{L}$ *is* H*-smooth, and* $\hat{\nabla}\ell(\mathbf{w})$ *has variance at most* σ^2 *for all* $\mathbf{w}$ *(that is for all* $\mathbf{w}$*,* $\mathbb{E}[\|\hat{\nabla}\ell(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w})\|^2] \le \sigma^2$*). Let us consider a* fixed *learning rate* $\eta_t = \eta$ *for all* t*. Then so long as* $\eta \le \frac{1}{H}$*, Algorithm 2 guarantees:*

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{2\Delta}{\eta} + \frac{HT\eta\sigma^2}{B}$$

Further, if we set $\eta = \frac{1}{\max(H, \sigma\sqrt{LT/B})}$*, then if* $\hat{\mathbf{w}}$ *is selected uniformly at random from* $\mathbf{w}_1, \dots, \mathbf{w}_T$*,*

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \le \frac{\sqrt{2\Delta H}}{\sqrt{T}} + \frac{\sqrt{\sigma(1+2\Delta)}H^{1/4}}{(BT)^{1/4}}$$

Let's take a moment to appreciate how the learning rates changed when we incorporated the minibatch. For large enough T, the learning rate suggested by the theory is:

$$\eta \propto \frac{\sqrt{B}}{\sqrt{T}}$$

and the gradient size identified is:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \le O\left(\frac{1}{(BT)^{1/4}}\right)$$

When considering the *total computational cost*, we notice that when using a minibatch of size B, each iteration takes B times more compute since we need to compute B gradients. Thus the cost is $C = TB$. This is also the *oracle complexity* - the number of calls to the stochastic gradient oracle. Re-writing these results:

$$\eta \propto \frac{B}{\sqrt{C}}$$
$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \le O\left(\frac{1}{C^{1/4}}\right)$$

The first lesson here is that *the gradient size is independent of the batch size*. In fact, if we are more careful we would see that the best thing to optimize the constants hiding in this analysis is to set $B = 1$.

In one point of view, this is a bad thing: we increased the batch size, but the performance may not get any better! Fortunately, although we may not save on total compute cost, we might actually save in terms of *total time*. Specifically, the B computations $\mathbf{g}_{t,i} = \nabla \ell(\mathbf{w}_t, z_{(t-1)B+i})$ in Algorithm 2 can all be done in parallel. So, in theory if we had access to M machines and ignore a plethora of issues involving communication overheads, we might be able to have the total time spent by the algorithm equal to $\tau = \frac{C}{M}$. Thus, in time τ we have:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \leq O(\frac{1}{(M\tau)^{1/4}})$$

and so there is a clear advantage to increasing the batch size. However, there is a caveat here: all of this analysis only holds for sufficiently large T. If we have $B \geq T$, then we hit a point of diminishing returns:

Exercise 5. *Show that for $B \geq T$, the optimal value for η obtains only:*

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \leq O(\frac{1}{\sqrt{T}}) = O(\frac{\sqrt{B}}{\sqrt{C}})$$

so that increasing B may be actively harmful. Can you think of an intuitive reason why this should be expected?