Lecture Notes 7: Adaptive Learning Rates II

Instructor: Ashok Cutkosky

Throughout these notes, we adopt the notation:

$$\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$$

to make the equations look a little simpler.

1 Finer-Grained Adaptive SGD analysis

Previously, we examined the learning rates schedule:

$$\eta_t = \frac{c}{\sqrt{\epsilon^2 + \sum_{i=1}^t \|\mathbf{g}_i\|^2}}$$

with the corresponding algorithm:

Algorithm 1 Adaptive Stochastic Gradient Descent

Input: Initial Point $\mathbf{w}_1$, learning rates scaling c, small constant ϵ (e.g. 1e-4).
for $t = 1 \dots T$ **do**
  Sample $z_t \sim P_z$.
  Set $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$.
  Set $\eta_t = \frac{c}{\sqrt{\epsilon^2 + \sum_{i=1}^t \|\mathbf{g}_i\|^2}}$
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$.
end for

And we were able to prove the following result:

Theorem 1. *Suppose $\mathcal{L}$ is H-smooth and $G_{\mathcal{L}}$ Lipschitz, and let $\Delta = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$. Suppose $\mathbb{E}[\max_{t \le T} \|\mathbf{g}_t\|] \le G_\ell$ and $\mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \sigma^2$ for all t. Define*

$$K = \frac{\Delta}{c} + \frac{Hc \log\left(\frac{T(G_{\mathcal{L}}^2 + \sigma^2)}{\epsilon^2}\right)}{2} + \frac{G_{\mathcal{L}} G_\ell}{\epsilon} = O(\log(T))$$

Then Algorithm 1 guarantees:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \|\nabla \mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{K\sqrt{2} + \sqrt{K\epsilon}}{\sqrt{T}} + \frac{\sqrt{K\sigma}}{T^{1/4}}$$

However, it turns out an even more fine-grained result is true:

Theorem 2. *Suppose $\mathcal{L}$ is H-smooth and $G_{\mathcal{L}}$ Lipschitz, and let $\Delta = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$. Suppose $\mathbb{E}[\max_{t \le T} \|\mathbf{g}_t\|] \le G_\ell$ and define $\sigma_t^2 = \mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2]$ for all t. Then*

$$K = \frac{\Delta}{c} + \frac{Hc \log\left(\frac{T(G_{\mathcal{L}}^2 + \sigma^2)}{\epsilon^2}\right)}{2} + \frac{G_{\mathcal{L}} G_\ell}{\epsilon} = O(\log(T))$$

Then Algorithm 1 guarantees:

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \leq \frac{1}{\sqrt{T}}\,\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2}\right] \leq \frac{K\sqrt{2}+\sqrt{K\epsilon}}{\sqrt{T}} + \frac{\sqrt{K\sqrt{\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\sigma_t^2\right]}}}{T^{1/4}}$$

The proof is essentially the same - the only difference is that we have not performed the last Cauchy-Schwarz argument to bound $\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \leq \frac{1}{\sqrt{T}}\,\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2}\right]$, and also at a certain point in the proof of Theorem 4, we used the bound

$$\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{g}_t-\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right]} \leq \sigma\sqrt{T}$$

and instead we should have used the more fine-grained bound:

$$\sqrt{\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{g}_t-\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right]} = \sqrt{\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\sigma_t^2\right]}\sqrt{T}$$

The result of Theorem 3 is much more appealing than that of Theorem 4 because it is written in terms of the varying variances σ_t. How might this be helpful? Consider a situation in which $\mathbb{E}[\|\nabla\ell(\mathbf{w},z)-\nabla\mathcal{L}(\mathbf{w})\|^2] \leq D\|\nabla\mathcal{L}(\mathbf{w})\|^2$ for all $\mathbf{w}$ for some d. That is, points with small gradients also have small variance. This is not entirely unlikely: in the case that it is possible to obtain $\mathcal{L}(\mathbf{w}_\star)=0$ and $\ell(\mathbf{w},z)\geq 0$ for all $\mathbf{w}$ and z, we would have $\mathbb{E}[\|\nabla\ell(\mathbf{w}_\star,z)-\nabla\mathcal{L}(\mathbf{w}_\star)\|^2]=0$. If in addition we have that $\ell(\mathbf{w},z)$ is H-smooth in $\mathbf{w}$ for all z, This would also imply that $\mathbb{E}[\|\nabla\ell(\mathbf{w},z)-\nabla\mathcal{L}(\mathbf{w})\|^2] \leq H^2\|\mathbf{w}-\mathbf{w}_\star\|^2$. Thus, if it happened to be that $\|\nabla\mathcal{L}(\mathbf{w})\| \geq \|\mathbf{w}-\mathbf{w}_\star\|$ (this would be implied by strong-convexity for example), then indeed it would hold that $\mathbb{E}[\|\nabla\ell(\mathbf{w},z)-\nabla\mathcal{L}(\mathbf{w})\|^2] \leq K\|\nabla\mathcal{L}(\mathbf{w})\|^2$ for all $\mathbf{w}$ for some D.

Regardless, under this assumption, we have:

$$\begin{aligned}\frac{1}{\sqrt{T}}\,\mathbb{E}\left[\sqrt{\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2}\right] &\leq \frac{K\sqrt{2}+\sqrt{K\epsilon}}{\sqrt{T}} + \frac{\sqrt{K\sqrt{\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\sigma_t^2\right]}}}{T^{1/4}} \\ &\leq \frac{K\sqrt{2}+\sqrt{K\epsilon}}{\sqrt{T}} + \frac{\sqrt{KD\sqrt{\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right]}}}{T^{1/4}}\end{aligned}$$

Now, we can do another self-bounding style argument. Let $X = \sqrt{\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right]}$. Then, we have:

$$X \leq \frac{K\sqrt{2}+\sqrt{K\epsilon}}{\sqrt{T}} + \frac{\sqrt{KDX}}{T^{1/4}}$$

Now, we can use the quadratic formula again to obtain:

$$\sqrt{X} \leq \frac{\sqrt{KD}/T^{1/4} + \sqrt{\frac{KD}{\sqrt{T}} + 4\frac{K\sqrt{2}+\sqrt{K\epsilon}}{\sqrt{T}}}}{2}$$

using $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$:

$$\leq \frac{\sqrt{KD}}{T^{1/4}} + \frac{\sqrt{K\sqrt{2}+\sqrt{K\epsilon}}}{T^{1/4}}$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$:

$$X \leq \frac{2KD + 2K\sqrt{2} + 2\sqrt{K\epsilon}}{\sqrt{T}}$$

using $\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|\right] \leq X$:

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|\right] \leq \frac{2KD + 2K\sqrt{2} + 2\sqrt{K\epsilon}}{\sqrt{T}}$$

So that we actually converge at an $O(1/\sqrt{T})$ rate rather than an $O(1/T^{1/4})$ rate - there is a kind of "virtuous cycle" in which making progress makes the variance smaller, which allows for faster progress.

Formally, we have proved the following result:

Theorem 3. *Suppose $\mathcal{L}$ is H-smooth and $G_{\mathcal{L}}$ Lipschitz, and let $\Delta = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$. Suppose $\mathbb{E}[\max_{t \leq T} \|\mathbf{g}_t\|] \leq G_\ell$, and suppose $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \mathcal{L}(\mathbf{w})\|^2] \leq D\|\nabla \mathcal{L}(\mathbf{w})\|^2$ for all $\mathbf{w}$ for some D. Define:*

$$K = \frac{\Delta}{c} + \frac{Hc \log\left(\frac{T(G_{\mathcal{L}}^2 + \sigma^2)}{\epsilon^2}\right)}{2} + \frac{G_{\mathcal{L}} G_\ell}{\epsilon} = O(\log(T))$$

Then Algorithm 1 guarantees:

$$\frac{1}{T}\,\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|\right] \leq \frac{2KD + 2K\sqrt{2} + 2\sqrt{K\epsilon}}{\sqrt{T}}$$

2 Tracking the variance

Let's take a look at one step of stochastic gradient descent:

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

Using bias-variance decomposition, and assuming η is independent of $\mathbf{g}_t$:

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

assuming $\eta \leq \frac{1}{H}$, and writing $\sigma_t^2 = \mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2]$:

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

Now, from this we should expect that the right learning rate *at this iterate* is something like $\eta = \min\left(\frac{1}{H}, \frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}{2H\sigma_t^2}\right)$ (to see this, differentiate the above with respect. to η). After a little case-work, this yields:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \min\left(\frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}{4H},\ \frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^4}{8H\sigma_t^2}\right)\right]$$

Following now the usual train logic, we end up with:

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} \min\left(\frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}{4H},\ \frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^4}{8H\sigma_t^2}\right)\right] \leq \frac{\Delta}{T}$$

Therefore, if $\hat{\mathbf{w}}$ is a randomly selected iterate, we have

$$\mathbb{E}\left[\min\left(\frac{\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2}{4H},\ \frac{\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^4}{8H\,\mathbb{E}_z[\|\nabla\ell(\hat{\mathbf{w}},z)-\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2}\right)\right] \leq \frac{\Delta}{T}$$

This suggests that the point $\hat{\mathbf{w}}$ has $\|\nabla\mathcal{L}(\hat{\mathbf{w}})\| = O(1/T^{1/4})$, but also that the proportionality constant is now related to the variance *at the same point* $\hat{\mathbf{w}}$ rather than a sum of variance terms over all iterates. This might enable us to obtain similar "virtuous cycles" as found in the previous section with much weaker conditions on the losses.

The key problem here of course is that we do not actually know the "instantaneous" variance σ_t, nor the true gradient norm $\|\nabla\mathcal{L}(\mathbf{w}_t)\|$. The latter sounds very difficult to obtain, but what if we were able to estimate the former? Could we still obtain an improved bound?

Let's consider learning rates:

$$\eta_t = \min\left(\frac{1}{H},\ \frac{\sqrt{\Delta}}{\sigma_t\sqrt{TH}}\right)$$

Then, following the same logic as usual, we obtain:

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_{t+1})\right] \leq \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \frac{\eta_t}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 + \frac{\Delta}{2T}\right]$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\eta_t\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \leq 3\Delta$$

From which a randomly selected $\hat{\mathbf{w}}$ obtains:

$$\mathbb{E}\left[\min\left(\frac{1}{H}, \frac{\sqrt{\Delta}}{\sqrt{\mathbb{E}_z[\|\nabla\ell(\hat{\mathbf{w}},z)-\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2]TH}}\right)\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2\right] \leq \frac{3\Delta}{T}$$

Thus, for large T we should expect

$$\mathbb{E}\left[\frac{\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2}{\mathbb{E}_z[\|\nabla\ell(\hat{\mathbf{w}},z)-\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2}\right] \leq O\left(\frac{1}{\sqrt{T}}\right)$$

This is a similar result, but we still have an issue: how did we know the instantaneous variance σ_t? A standard heuristic used in practice that might achieve this goal is the *exponentially weighted moving average*. In this scheme, we choose some parameter $\beta \in (0,1)$ and then set

$$\eta_t = \frac{c}{\sqrt{\epsilon^2 + \sum_{i=1}^{t}\beta^{t-1}\|\mathbf{g}_i\|^2}}$$

This can be computed efficiently, as demonstrated in the following pseudocode:

Algorithm 2 Adaptive Stochastic Gradient Descent with EMA

Input: Initial Point $\mathbf{w}_1$, learning rates scaling c, small constant ϵ (e.g. 1e-4), averaging factor β:
Set $A_0 = 0$.
for $t = 1 \ldots T$ **do**
  Sample $z_t \sim P_z$.
  Set $\mathbf{g}_t = \nabla\ell(\mathbf{w}_t, z_t)$.
  Set $A_t = \beta A_{t-1} + \|\mathbf{g}_t\|^2$.
  Set $\eta_t = \frac{c}{\sqrt{\epsilon^2 + A_t}}$
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\mathbf{g}_t$.
end for

Intuitively, this setting of η_t is allowed to "forget" about previous gradients, so that if σ_t happens to be large in the first iterates, this will not have as big an effect on later iterates.

Unfortunately, the algorithm as described here does not actually converge. The problem is that the learning rates η_t may increase quite rapidly during the training. Instead, as suggested by [1], we could try the following:

This algorithm has the following convergence guarantee:

Algorithm 3 Corrected Adaptive Stochastic Gradient Descent with EMA

Input: Initial Point $\mathbf{w}_1$, learning rates scaling c, small constant ϵ (e.g. 1e-4), averaging factor β:
Set $A_0 = 0$.
Set $B_0 = \epsilon^2$.
for $t = 1 \dots T$ **do**
  Sample $z_t \sim P_z$.
  Set $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$.
  Set $A_t = \beta A_{t-1} + \|\mathbf{g}_t\|^2$.
  Set $B_t = \max(B_{t-1}, tA_t)$
  Set $\eta_t = \frac{c}{\sqrt{B_t}}$
  Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$.
end for

Theorem 4. *Suppose $\mathcal{L}$ is H-smooth and $G_{\mathcal{L}}$ Lipschitz, and let $\Delta = \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)$. Suppose $\mathbb{E}[\max_{t \le T} \|\mathbf{g}_t\|] \le G_\ell$ for all t. Then Algorithm 3 guarantees:*

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] \le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \sum_{t=1}^{T} \eta_{t-1} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] + \frac{H}{2}(1 + \log(T)) + \eta_0 G_{\mathcal{L}} G_\ell$$

Moreover, it is also guaranteed that:

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \le \left(\Delta + \frac{H}{2}(1 + \log(T)) + \eta_0 G_{\mathcal{L}} G_\ell\right) \frac{\epsilon + G\sqrt{T}}{\sqrt{1-\beta}}$$

Note that this result actually doesn't imply any obvious gain over the previous results for adaptive gradient descent. While intuitively it seems reasonable that the algorithm will perform better (and in practice it frequently does), the analysis is sufficiently complicated that there is no solid result characterizing when Algorithm 3 is actually better than 1.

Proof. Recall from Theorem 1 of the previous notes 6 (which applies since η_t is now guaranteed to be decreasing), that we can write:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] \le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \sum_{t=1}^{T} \eta_{t-1} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{H}{2} \sum_{t=1}^{T} \eta_t^2 \|\mathbf{g}_t\|^2\right] + \eta_0 G_{\mathcal{L}} G_\ell$$

Now, to bound $\sum_{t=1}^{T} \eta_t^2 \|\mathbf{g}_t\|^2$, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \eta_t^2 \|\mathbf{g}_t\|^2 &\le \sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|^2}{tA_t} \\
&\le \frac{\|\mathbf{g}_t\|^2}{t\|\mathbf{g}_t\|^2} \\
&\le \frac{1}{t}
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \sum_{t=1}^{T} \eta_{t-1} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] + \frac{H}{2}(1 + \log(T)) + \eta_0 G_{\mathcal{L}} G_\ell \\
\mathbb{E}\left[\sum_{t=1}^{T} \eta_{t-1} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] &\le \Delta + \frac{H}{2}(1 + \log(T)) + \eta_0 G_{\mathcal{L}} G_\ell \\
\mathbb{E}\left[(\min_t \eta_t) \sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] &\le \Delta + \frac{H}{2}(1 + \log(T)) + \eta_0 G_{\mathcal{L}} G_\ell
\end{aligned}$$

Now, observe that

$$A_t \leq G^2 \sum_{i=0}^{t-1} \beta^i \leq \frac{G^2}{1-\beta}$$

Therefore, $\min_t \eta_t \geq \frac{1}{\epsilon+\sqrt{T \max A_t}} \geq \frac{\sqrt{1-\beta}}{\epsilon+G\sqrt{T}}$. This implies

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \leq \left(\Delta + \frac{H}{2}(1+\log(T)) + \eta_0 G_{\mathcal{L}} G_\ell\right) \frac{\epsilon + G\sqrt{T}}{\sqrt{1-\beta}}$$

□

References

[1] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. "On the Convergence of Adam and Beyond". In: *6th International Conference on Learning Representations, ICLR 2018*. 2018.