Lecture Notes 6: Adaptive Learning Rates I

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 Adaptive Gradient Descent

Previously, we saw that the "optimal" learning rate for SGD is roughly $\min\left(\frac{1}{H}, \frac{\sqrt{\Delta}}{\sigma\sqrt{TH}}\right)$. This is nice in theory, but in reality you probably don't actually know the values for H or Δ or σ or even T. We saw that it's possible to avoid needing to know T via a varying learning rate of $\eta_t = O(1/\sqrt{t})$ at the cost of a logarithmic factor, but how would we be able to deal with the other unknown values?

This is objective of *adaptive* algorithms. Adaptive algorithms "adapt" to the problem at hand, achieving faster convergence on "easier" problems while gracefully degrading on harder problems. Concretely, we can think of adaptive algorithms as figuring out some parameters of the problem (such as σ of H) "on-the-fly".

In the *convex* setting, this area has been heavily studied, and there are very sophisticated adaptive algorithms available that can adapt to all sorts of different types of "easy" problems. See [1, 2, 3, 4] for some examples of such algorithms. In the non-convex setting, however, things are a little less developed. Today we will discuss a way to adapt well to σ, and to some extent H. There is still no good way to adapt to Δ.

1.1 Intuition for Adaptive Learning Rates

Let's take a look at the learning rate setting we saw previously:

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

A simple way to try to design an adaptive algorithm is to "guess" this value of η on the fly. To start, let's make a simplification: for large enough T, we should always expect the $1/\sqrt{T}$ part to be dominant, so let's focus on estimating $\sigma\sqrt{T}$.

Ignoring the other constants for now, we want to make a learning rate like

$$\eta_t = \frac{c}{\sigma\sqrt{t}}$$

So now we need only estimate σ. How can we estimate the variance? The simplest method might be to try some form of empirical estimate. Since the variance of a random quantity Z is $\mathbb{E}[\|Z\|^2 - \|\mathbb{E}[Z]\|^2]$, we could try to estimate σ^2 by:

$$\sigma^2 \approx \frac{1}{t}\sum_{i=1}^t \|\mathbf{g}_i\|^2 - \left\|\frac{1}{t}\sum_{i=1}^t \mathbf{g}_i\right\|^2$$

Or, we could use the slightly more advance *unbiased* estimate:

$$\overline{\mathbf{g}} = \frac{1}{t}\sum_{i=1}^t \mathbf{g}_i \sigma^2 \approx \frac{1}{t-1}\sum_{i=1}^t \|\mathbf{g}_i - \overline{\mathbf{g}}\|^2$$

where here $\mathbf{g}_i = \nabla \ell(\mathbf{w}_i, z_i)$, the ith output of the stochastic gradient oracle.

However, it turns out that it will be much simpler algebraically to analyze instead the following estimate:

$$\sigma^2 \approx \frac{1}{t}\sum_{i=1}^{t} \|\mathbf{g}_i\|^2$$

This is a kind of "optimistic" guess for σ^2: if the gradients $\nabla \mathcal{L}(\mathbf{w}_t)$ are getting smaller, we might expect $\frac{1}{t}\sum_{i=1}^{t} \mathbf{g}_i$ to also be small, and so these two estimates are nearly the same. This leaves us with the learning rate:

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

This is almost identical to the popular AdaGrad learning rate [2] that we will discuss more later.

Strangely, while the above learning rate can analyzed without too much struggle in the *convex* setting, in the non-convex setting there is a technical issue in the algebra that makes it surprisingly difficult to work with. In order to appreciate this technical issue, let us start our performing the analysis and see what might go wrong.

We might think to start with our standard bounds on smooth losses (e.g Lemma 6 in Notes 3), which tell us that:

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

Now, in the past we have simplified the term $-\mathbb{E}[\eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t\rangle]$ as $-\eta_t \mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t\rangle] = -\eta_t \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2]$. However, this was only possible because η_t was a deterministic quantity. In our current setting, η_t actually depends on $\mathbf{g}_t$, and so we cannot simply pull it out of the expectation. In fact, [5] shows that is might actually be that $-\mathbb{E}[\eta_t \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t\rangle] > 0$, which suggests that the gradient step might actually make *negative* progress!

There are roughly two ways to deal with this. Probably the simplest way, as suggested in [5], is to simply change the learning rate to:

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

Note the two major differences here: the presence of ϵ, and the fact that the sum now only goes up to $t-1$. The ϵ is a *small* constant that is now necessary because otherwise $\eta_1 = \infty$. The sum going only up to $t-1$ fixes the problem with expectations, because now η_t is independent of z_t. As a result, we would have:

$$\begin{aligned}\mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), \eta_t \nabla \ell(\mathbf{w}_t, z_t)\rangle] &= \mathbb{E}[\eta_t \mathop{\mathbb{E}}_{z_t}[\langle \nabla \mathcal{L}(\mathbf{w}_t), \nabla \ell(\mathbf{w}_t, z_t)\rangle | z_t]] \\ &= \mathbb{E}[\eta_t \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2]\end{aligned}$$

However, this is not the learning rate used in practice, so we will instead consider the following rate:

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

Notice that while the ϵ is still there, the sum now again goes up to t, so that the above argument with expectations is *invalid*. In the convex setting, the ϵ is not needed, but we will see that it actually plays a role in our analysis to get around the expectation issues. In practice, the ϵ is usually added to avoid numerical instability issues: when computing η_t, it might be that the first gradient is zero or very small simply by chance. As a result, we would have $\eta_t = \infty$. Now, technically, we always have:

$$\|\eta_t \mathbf{g}_t\| \leq c$$

so that it should not matter if $\mathbf{g}_1 = 0$, but when actually implementing this we will run into numerical issues (in the worst case, you will get a divide by zero error), so the small ϵ is added to the denominator to avoid this. It just happens as a happy coincidence to help in the theory as well.

2 Adaptive SGD Algorithm

So, to summarize, we are considering the following 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

We are able to prove the following general Theorem. Notice that this result applies to *any* sequence of learning rates, not just the ones in Algorithm 1:

Theorem 1. *Suppose $\mathcal{L}$ is an H-smooth function. Defining $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$, suppose that $\mathbb{E}[\max_{t \le T} \|\mathbf{g}_t\|] \le G_\ell$ (note this is true if $\|\nabla \ell(\mathbf{w}_t, z_t)\| \le G_\ell$ with probability 1, but would be also implied by more technical conditions such as subgaussianity). Further, suppose that $\mathcal{L}$ is $G_{\mathcal{L}}$-Lipschitz. Finally, let $\eta_1, \dots, \eta_T$ be any sequence of learning rates such that (1) $\eta_t \ge 0$ for all t, (2) $\eta_1 \ge \eta_2 \ge \dots \ge \eta_T$ and also (3) the sequence is "causal" in the sense that η_t is not allowed to depend on $\mathbf{g}_{t+1}, \mathbf{g}_{t+2}, \dots, \mathbf{g}_T$. Let η_0 be a deterministic quantity such that $\eta_0 \ge \eta_1$. Consider the SGD update:*

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{g}_t$$

Then we have:

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

This Theorem is a somewhat technical-looking result, so before we prove it, let's see what it implies about the adaptive learning rate scheme we are using in Algorithm 1. To start with, notice that this learning rate scheme satisfies the conditions of Theorem 1: the learning rates are monotonically decreasing, and η_t does not depend on $\mathbf{g}_{t+1}$ or later values of $\mathbf{g}_t$. Further, we can set $\eta_0 = \frac{c}{\epsilon}$. For simplicity, let us suppose that $\|\mathbf{g}_t\| \le G$ with probability 1 and just take $G_{\mathcal{L}} = G_\ell = G$. Thus, Theorem 1 tells us that:

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

Now, let us do some back-of-the-envelope calculations to gain intuition for what this means. Note that these are *NOT CORRECT*, but will be "nearly correct", and we will make the arguments rigorous later.

First, let's do some rearrangement and use the fact $\eta_T \le \eta_t$ for all t:

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

Next, let's consider the sum

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

We will soon see that for roughly the same reasons that $\sum_{t=1}^T \frac{1}{t} \le O(\log(T))$, this sum is $O(c^2 \log(T/\epsilon^2))$. Let's just pretend it is actually bounded by $c^2 \log(T/\epsilon^2)$ for now. Then, we perform another technically-illegal operation: let's

divide by η_T inside the expectations:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \leq \mathbb{E}\left[\frac{\Delta}{\eta_T}+\frac{c^2\log(T/\epsilon^2)}{\eta_T}+\frac{cG^2}{\epsilon\eta_T}\right]$$
$$= \mathbb{E}\left[\left(\frac{\Delta}{c}+c\log(T/\epsilon^2)+\frac{G^2}{\epsilon}\right)\sqrt{\epsilon^2+\sum_{t=1}^{T}\|\mathbf{g}_t\|^2}\right]$$

Now, let's again make a leap of faith and assume that it will actually hold that

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

where $\sigma^2 \geq \mathbb{E}[\|\mathbf{g}_t-\nabla\mathcal{L}(\mathbf{w}_t)\|^2]$ for all t. Note that this is definitely not logically okay - we would need to prove this. But just to get some intuition, let's make this assumption and at least see that we don't hit a contradiction. Now, we have (using the bias-variance decomposition Proposition 4):

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

where we've assume T is sufficiently large that $\sigma\log(T)\sqrt{T} \leq \sigma^2T$. Further, by Jensen inequality:

$$\mathbb{E}\left[\sqrt{\epsilon^2+\sum_{t=1}^{T}\|\mathbf{g}_t\|^2}\right] \leq \sqrt{\epsilon^2+\mathbb{E}\left[\sum_{t=1}^{T}\|\mathbf{g}_t\|^2\right]}$$
$$\leq \epsilon+O\left(\sigma\sqrt{T}\right)$$

from which we are able to recover our assumption $\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] = O(\sigma\log(T)\sqrt{T})$ by absorbing Δ, G and ϵ into the big-Oh notation.

Now that we have this sketch for how to use the Theorem, let's proceed to prove Theorem 1.

Proof of Theorem 1. As usual, be we begin by using our key result about smooth functions:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t)-\eta_t\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+\frac{H}{2}\eta_t^2\|\mathbf{g}_t\|^2$$

now, when we take expectations, the $\eta_t\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle$ will cause trouble since η_t depends on $\mathbf{g}_t$. So let's instead write in terms of $\eta_{t-1}\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle$:

$$= \mathcal{L}(\mathbf{w}_t)-\eta_{t-1}\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+(\eta_{t-1}-\eta_t)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+\frac{H}{2}\eta_t^2\|\mathbf{g}_t\|^2$$

Iterating for all t:

$$\mathcal{L}(\mathbf{w}_{T+1}) \leq \mathcal{L}(\mathbf{w}_1)-\sum_{t=1}^{T}\eta_{t-1}\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+\sum_{t=1}^{T}(\eta_{t-1}-\eta_t)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+\sum_{t=1}^{T}\frac{H}{2}\eta_t^2\|\mathbf{g}_t\|^2$$

using the fact that η_t is decreasing:

$$\leq \mathcal{L}(\mathbf{w}_1)-\sum_{t=1}^{T}\eta_{t-1}\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle+\max_{t\leq T}|\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle|\sum_{t=1}^{T}(\eta_{t-1}-\eta_t)+\sum_{t=1}^{T}\frac{H}{2}\eta_t^2\|\mathbf{g}_t\|^2$$

Using Cauchy-Schwarz:

$$\leq \mathcal{L}(\mathbf{w}_1) - \sum_{t=1}^{T} \eta_{t-1} \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle + \max_{t \leq T} \|\nabla \mathcal{L}(\mathbf{w}_t)\| \max_{t \leq T} \|\mathbf{g}_t\| \sum_{t=1}^{T} (\eta_{t-1} - \eta_t) + \sum_{t=1}^{T} \frac{H}{2} \eta_t^2 \|\mathbf{g}_t\|^2$$

$$\leq \mathcal{L}(\mathbf{w}_1) - \sum_{t=1}^{T} \eta_{t-1} \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle + \max_{t \leq T} \|\nabla \mathcal{L}(\mathbf{w}_t)\| \max_{t \leq T} \|\mathbf{g}_t\| \eta_0 + \sum_{t=1}^{T} \frac{H}{2} \eta_t^2 \|\mathbf{g}_t\|^2$$

Taking expectations:

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

Finally, we can pull $G_{\mathcal{L}} G_\ell \eta_0$ out of the expectation since this is actually a constant. □

2.1 Convergence of Algorithm 1

Now that we have proven the technical result of Theorem 1, we can turn to formalizing the intuitive argument for how this implies a convergence result for Algorithm 1. In order to do this, we will need the following important technical Lemma:

Lemma 2. *Suppose $x_0, \ldots, x_T$ are arbitrary non-negative values. And let $f(x) : \mathbb{R} \to \mathbb{R}$ be an arbitary* decreasing *function. Then:*

$$\sum_{t=1}^{T} x_t f\left(\sum_{i=0}^{t} x_i\right) \leq \int_{x_0}^{\sum_{i=0}^{T} x_i} f(x)\, dx$$

Proof. Notice that since f is decreasing, for all intervals $[a, b]$ we have:

$$(b-a) f(b) \leq \int_a^b f(x)\, dx$$

In particular:

$$x_t f\left(\sum_{i=0}^{t} x_i\right) \leq \int_{\sum_{i=0}^{t-1} x_i}^{\sum_{i=0}^{t} x_i} f(x)\, dx$$

$$\sum_{t=1}^{T} x_t f\left(\sum_{i=0}^{t} x_i\right) \leq \sum_{t=1}^{T} \int_{\sum_{i=0}^{t-1} x_i}^{\sum_{i=0}^{t} x_i} f(x)\, dx$$

$$= \int_{x_0}^{\sum_{i=0}^{T} x_i} f(x)\, dx$$

□

As an immediate corollary of this Theorem, we have:

$$\sum_{t=1}^{T} \frac{\|\mathbf{g}_t\|^2}{\epsilon^2 + \sum_{i=1}^{t} \|\mathbf{g}_t\|^2} \leq \int_{\epsilon^2}^{\sum_{t=1}^{T} \|\mathbf{g}_t\|^2} \frac{dx}{x}$$

$$= \log\left(\frac{\sum_{t=1}^{T} \|\mathbf{g}_t\|^2}{\epsilon^2}\right)$$

Now, we are ready to state and prove the following:

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 $\mathbb{E}[\max_{t \le T} \|\mathbf{g}_t\|] \le G_\ell$ 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:

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

Proof. First, note that that by the bias-variance decomposition (Proposition 4), we have:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_t\|^2] &\le \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \mathbb{E}[\|\mathbf{g}_t - \nabla\mathcal{L}(\mathbf{w}_t)\|^2] \\
&\le \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \sigma^2 \\
&\le G_{\mathcal{L}}^2 + \sigma^2
\end{aligned} \tag{1}$$

Applying Theorem 1, with $\eta_0 = \frac{c}{\epsilon}$, we have:

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

Now, applying the definition of η_t and Lemma 2, we can see that

$$\begin{aligned}
\sum_{t=1}^{T}\eta_t^2\|\mathbf{g}_t\|^2 &= c^2\sum_{t=1}^{T}\frac{\|\mathbf{g}_t\|^2}{\epsilon^2 + \sum_{i=1}^{t}\|\mathbf{g}_t\|^2} \\
&\le c^2 \log\left(\frac{\sum_{t=1}^{T}\|\mathbf{g}_t\|^2}{\epsilon^2}\right)
\end{aligned}$$

Taking expectations and using Jensen inequality:

$$\mathbb{E}\left[\sum_{t=1}^{T}\eta_t^2\|\mathbf{g}_t\|^2\right] \le c^2\log\left(\frac{\mathbb{E}[\sum_{t=1}^{T}\|\mathbf{g}_t\|^2]}{\epsilon^2}\right)$$

Using equation (1):

$$\le c^2\log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)$$

Thus, we have:

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

rearranging:

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

using $\eta_T \le \eta_t$ for all t:

$$\mathbb{E}\left[\sum_{t=1}^{T}\eta_T\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] \le \Delta + \frac{Hc^2\log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)}{2} + \frac{cG_{\mathcal{L}}G_\ell}{\epsilon}$$

Now, we come to a primary difficulty in the proof. In our sketch before, we just divided by η_T at this point. In previous arguments this was fine, because in all our previous algorithms, η_T was just a deterministic constant. Now, however, η_T is unfortunately a *random variable*, and so it is not allowed to divide it out from both sides. We will have to be a bit more tricky. The key idea is to use Cauchy-Schwarz for random variables (see Proposition 6). In particular, if we define random variables

$$A^2 = \sum_{t=1}^{T} \eta_T \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$$
$$B^2 = \frac{1}{\eta_T}$$

Then by Proposition 6, we have

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

This is a somewhat magical trick: by doing a little manipulation with Proposition 6, we are able to get an expression in terms of $\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}\right]$, which *does not* have the η_T inside the expectation! So, continuing:

$$\frac{\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}\right]^2}{\mathbb{E}[\eta_T^{-1}]} \leq \Delta + \frac{Hc^2 \log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)}{2} + \frac{cG_{\mathcal{L}}G_\ell}{\epsilon}$$
$$\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}\right]^2 \leq \left(\Delta + \frac{Hc^2 \log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)}{2} + \frac{cG_{\mathcal{L}}G_\ell}{\epsilon}\right) \mathbb{E}[\eta_T^{-1}]$$
$$= \left(\frac{\Delta}{c} + \frac{Hc \log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)}{2} + \frac{G_{\mathcal{L}}G_\ell}{\epsilon}\right) \mathbb{E}\left[\sqrt{\epsilon^2 + \sum_{t=1}^{T} \|\mathbf{g}_t\|^2}\right]$$

Now, to avoid carrying around too much algebra, let's just define $K = \frac{\Delta}{c} + \frac{Hc \log\left(\frac{T(G_{\mathcal{L}}^2+\sigma^2)}{\epsilon^2}\right)}{2} + \frac{G_{\mathcal{L}}G_\ell}{\epsilon}$. Then, we can write this as:

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

The next part of the proof is an important trick that frequently arises when analyzing adaptive algorithms: we are going to use a *self-bounding argument*. The idea is that if we define $X = \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}\right]$, which is the quantity that we are interested in bounding, we will try to obtain an expression of the form:

$$X^2 \leq f(X)$$

for some function f. Then, we will use the shape of the function f to argue that if $X^2 \leq f(X)$, X cannot be too large. Specifically, $f(X)$ will eventually be a linear function, so we will have $X^2 \leq aX + b$, from which we can use the

quadratic formula to bound X. So far, we have obtained:

$$X^2 \leq K \, \mathbb{E}\left[\sqrt{\epsilon^2 + \sum_{t=1}^{T} \|\mathbf{g}_t\|^2}\right] \tag{2}$$

So, now we need to find some way to see X in the RHS of this expression. Our starting point is the identity $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. To see this identity, observe $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$, and then apply Young inequality (Proposition 5) to bound $2\langle a, b\rangle \leq \|a\|^2 + \|b\|^2$. From this, we have:

$$\|\mathbf{g}_t\|^2 = \|\mathbf{g}_t - \nabla\mathcal{L}(\mathbf{w}_t) + \nabla\mathcal{L}(\mathbf{w}_t)\|^2 \leq 2\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 + 2\|\mathbf{g}_t - \nabla\mathcal{L}(\mathbf{w}_t)\|^2$$

Next, we use the identity $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. This can be seen by simply squaring both sides. Thus, we have:

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

using Jensen inequality:

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

So, now going back to equation (2), we have:

$$X^2 \leq K\sqrt{\epsilon^2 + T\sigma^2} + K\sqrt{2}X$$

Now, by quadratic formula, we have:

$$\begin{aligned}
X &\leq \frac{K\sqrt{2} + \sqrt{2K^2 + 4K\sqrt{\epsilon^2 + T\sigma^2}}}{2} \\
&\leq K\sqrt{2} + \sqrt{K}\left(\epsilon^2 + T\sigma^2\right)^{1/4} \\
&\leq K\sqrt{2} + \sqrt{K\epsilon} + \sqrt{K\sigma}T^{1/4}
\end{aligned}$$

Now, finally, recall our definition of X and use Cauchy-Schwarz to conclude:

$$\begin{aligned}
X &= \mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|\nabla\mathcal{L}(\mathbf{w}_t)\|^2}\right] \\
&\geq \frac{1}{\sqrt{T}}\,\mathbb{E}\left[\sum_{t=1}^{T} \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right]
\end{aligned}$$

From which we can conclude:

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

□

2.2 More refined analysis

The analysis we've presented has a somwhat poor dependence on ϵ: it is $O\left(\frac{\sqrt{\sigma}}{\sqrt{\epsilon}T^{1/4}}\right)$, which could be quite bad. However, it is possible to obtain a more refined bound that does not have this poor dependence on ϵ. See [6] for a proof that achieves this. Note their proof is along a slightly different lines: instead of bounding the $(\eta_{t-1}-\eta_t)\langle\nabla\mathcal{L}(\mathbf{w}_t),\mathbf{g}_t\rangle$ terms by using decreasing learning rates, as we did in Theorem 1, instead they carefully bound the gap $|\eta_{t-1}-\eta_t|$ to show that each term is quite small.

3 Auxiliary technical results (also in notes 0)

Proposition 4 (Bias-variance Decomposition)**.** *Suppose $X \in \mathbb{R}^d$ is some random variable. Then for any deterministic value Y:*

$$\mathbb{E}[\|X-Y\|^2] = \|Y-\mathbb{E}[X]\|^2 + \mathbb{E}[\|X-\mathbb{E}[X]\|^2]$$

Proof. Expanding the square we have:

$$\begin{aligned}
\mathbb{E}[\|X-Y\|^2] &= \mathbb{E}[\|X-\mathbb{E}[X]+\mathbb{E}[X]-Y\|^2] \\
&= \mathbb{E}[\|Y-\mathbb{E}[X]\|^2] + \mathbb{E}[\|X-\mathbb{E}[X]\|^2] + 2\,\mathbb{E}[\langle X-\mathbb{E}[X],\mathbb{E}[X]-Y\rangle] \\
&= \|Y-\mathbb{E}[X]\|^2 + \mathbb{E}[\|X-\mathbb{E}[X]\|^2] + 2\langle\mathbb{E}[X-\mathbb{E}[X]],\mathbb{E}[X]-Y\rangle \\
&= \|Y-\mathbb{E}[X]\|^2 + \mathbb{E}[\|X-\mathbb{E}[X]\|^2]
\end{aligned}$$

where □

Another useful inequality is the Young inequality:

Proposition 5 (Young inequality)**.** *Suppose X and Y are arbitrary vectors. Then for any $\lambda > 0$ and any non-negative real numbers p and q satisfying $\frac{1}{p}+\frac{1}{q}=1$,*

$$\langle X,Y\rangle \le \frac{\|X\|^p}{p\lambda^p} + \frac{\lambda^q\|Y\|^q}{q}$$

Proof. Notice that we must have both p and q at least 1 (otherwise $1/p+1/q>1$, which is not allowed). Now, differentiating the RHS with respect to λ yields:

$$-\frac{\|X\|^p}{\lambda^{p+1}} + \lambda^{q-1}\|Y\|^q$$

Differentiating again yields:

$$(p+1)\frac{\|X\|^p}{\lambda^{p+2}} + (q-1)\lambda^{q-2}\|Y\|^q \ge 0$$

where we have used $q \ge 1$. Thus, we can find a minimum value for the RHS by setting the first derivative to 0. Solving $-\frac{\|X\|^p}{\lambda^{p+1}} + \lambda^{q-1}\|Y\|^q = 0$ for λ then gives $\lambda = \frac{\|X\|^{p/(p+q)}}{\|Y\|^{q/(p+q)}}$, which yields a minimum value of

$$\frac{\|X\|^{p-\frac{p^2}{p+q}}\|Y\|^{\frac{pq}{p+q}}}{p} + \frac{\|Y\|^{q-\frac{q^2}{p+q}}\|X\|^{\frac{pq}{p+q}}}{q} = \frac{\|X\|^{\frac{pq}{p+q}}\|Y\|^{\frac{pq}{p+q}}}{p} + \frac{\|Y\|^{\frac{pq}{p+q}}\|X\|^{\frac{pq}{p+q}}}{q}$$

Now, notice that

$$1 = \frac{1}{p}+\frac{1}{q} = \frac{p+q}{pq}$$

So that the minimum of the RHS simplifies to:

$$\frac{\|X\|\|Y\|}{p} + \frac{\|Y\|\|X\|}{q} = \|X\|\|Y\|$$

Now to conclude observe that $\langle X,Y\rangle \le \|X\|\|Y\|$ by Cauchy-Schwarz. □

Finally, we will also often need the following generalization of Cauchy-Schwarz:

Proposition 6 (Cauchy-Schwarz for random variables)**.** *Suppose $A \in \mathbb{R}$ and $B \in \mathbb{R}$ are arbitrary random variables. Then:*

$$\mathbb{E}[AB] \leq \sqrt{\mathbb{E}[A^2]\,\mathbb{E}[B^2]}$$

Proof. By Young inequality (Proposition 5), we have

$$\mathbb{E}[AB] \leq \frac{\mathbb{E}[A^2]}{2\lambda^2} + \frac{\lambda^2\,\mathbb{E}[B^2]}{2}$$

for any $\lambda > 0$. Now, set $\lambda = \frac{\mathbb{E}[A^2]}{\mathbb{E}[B^2]}$ to conclude the desired result. □

References

[1] Elad Hazan, Alexander Rakhlin, and Peter L Bartlett. "Adaptive online gradient descent". In: *Advances in Neural Information Processing Systems*. 2008, pp. 65–72.

[2] J. Duchi, E. Hazan, and Y. Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In: *Conference on Learning Theory (COLT)*. 2010.

[3] Tim van Erven and Wouter M Koolen. "MetaGrad: Multiple Learning Rates in Online Learning". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 3666–3674.

[4] Ashok Cutkosky and Francesco Orabona. "Black-Box Reductions for Parameter-free Online Learning in Banach Spaces". In: *Conference On Learning Theory*. 2018, pp. 1493–1529.

[5] Xiaoyu Li and Francesco Orabona. "On the convergence of stochastic gradient descent with adaptive stepsizes". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR. 2019, pp. 983–992.

[6] Rachel Ward, Xiaoxia Wu, and Leon Bottou. "AdaGrad stepsizes: Sharp convergence over nonconvex landscapes". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 6677–6686.