Lecture Notes 11: Momentum in Stochastic Optimization

Instructor: Ashok Cutkosky

1 Momentum as averaging

When discussing Adam, we briefly motivated the idea of momentum as related to artificially increasing the batch size rather than as a method connected to physical simulation. Now, let's try to make this connection more formal. We will consider a simple SGD with momentum algorithm:

Algorithm 1 SGD with Momentum

```
**Input:** Initial Point $\mathbf{w}_1$, learning rate $\eta$, momentum parameters $\alpha$, time horizon $T$:
Sample $z_1$.
Set $\mathbf{m}_1 = \nabla \ell(\mathbf{w}_1, z_1)$.
Set $\mathbf{w}_2 = \mathbf{w}_1 - \mathbf{m}_1$
for $t = 2 \dots T$ do
    Sample $z_t$.
    Set $\mathbf{m}_t = (1-\alpha)\mathbf{m}_{t-1} + \alpha \nabla \ell(\mathbf{w}_t, z_t)$.
    Set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{m}_t$.
end for
```

The difficulty in analyzing SGD with momentum is that it is not the case that

$$\mathbb{E}[\mathbf{m}_t] = \nabla \mathcal{L}(\mathbf{w}_t)$$

Thus, our standard strategy of writing:

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

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

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

is not valid because $\mathbb{E}[\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{m}_t \rangle] \ne \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2]$. Instead, we will need to do something a little more subtle. In order to facilitate this analysis, we'll define the quantity:

$$\epsilon_t = \mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)$$

Then, ϵ_t is measuring the amount of "error" in the estimate $\mathbf{g}_t$. Notice that $\mathbb{E}[\epsilon_t] \ne 0$.

Also, let's define:

$$r_t = \nabla \ell(\mathbf{w}_t, z_t) - \nabla \mathcal{L}(\mathbf{w}_t)$$

Thus, r_t is the error in the ordinary non-momentum gradient estimates, and has $\mathbb{E}[r_t] = 0$. We'll also assume as usual that $\mathbb{E}[\|r_t^2\|] \le \sigma^2$.

With this notation, we can produce an analog of our one-step progress lemma for SGD, now incorporating momentum:

Lemma 1. *Suppose that $\mathcal{L}$ is H-smooth. Then so long as $\eta_t \le \frac{1}{4H}$,*

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1}) - L(\mathbf{w}_t)] \le -\frac{\eta}{4}\,\mathbb{E}[\|\nabla L(\mathbf{w}_t)\|^2] + \frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2]$$

Proof. From the standard smoothness lemma:

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

taking expectation of both sides:

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

Using Young inequality $\langle x, y \rangle \le \frac{\|x\|^2}{2\lambda} + \frac{\lambda \|y\|^2}{2}$ with $\lambda = 1$:

$$\begin{aligned}
&\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{\eta}{2}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{\eta}{2}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta^2 H\,\mathbb{E}[\|\mathbf{m}_t\|^2]}{2} \\
&\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{\eta_t}{2}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{\eta}{2}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta^2 H\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t) + \epsilon_t\|^2]}{2}
\end{aligned}$$

Using $\|a + b\|^2 \le 2\|a\|^2 + 2\|b\|^2$:

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

Using $\eta_t \le \frac{1}{4H}$:

$$\begin{aligned}
&\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{\eta_t}{2}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{\eta_t}{2}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta_t\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\| + \|\epsilon_t\|^2]}{4} \\
&= \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{\eta_t}{4}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{3\eta_t}{4}\,\mathbb{E}[\|\epsilon_t\|^2]
\end{aligned}$$

□

Let's compare this result to our standard gradient descent result when $\mathbb{E}[\mathbf{m}_t] = \nabla \mathcal{L}(\mathbf{w}_t)$. This required only $\eta \le 1/H$ and achieved:

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

The dependency on $\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2]$ is similar - just the constants have changed. However, the dependency on $\|\epsilon\|^2$ has changed quite a bit: the power of η has decreased. As a result, *this lemma is not sufficient to guarantee convergence unless ϵ becomes small.*

How small does ϵ need to be? Notice that if $\|\epsilon_t\|^2 \ge \frac{1}{3}\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$, then Lemma 1 does not actually imply that $\mathcal{L}(\mathbf{w}_{t+1}) \le \mathcal{L}(\mathbf{w}_t)$ anymore. Therefore, we should expect to make progress on the function value only when $\|\epsilon_t\|^2 \le \frac{1}{3}\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$. Further, if our goal is to at least match the performance of regular non-momentum SGD, then we'd hope to make progress so long as $\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 > \Omega(1/\sqrt{T})$. This suggests that we need to get $\|\epsilon_t\|^2$ down to $O(1/\sqrt{T})$.

This is the crux of the momentum idea: if we view $\mathbf{m}_t$ as running average of gradient estimates, it is reasonable to believe that $\mathbf{m}_t$ would be a very close estimate of $\nabla \mathcal{L}(\mathbf{w}_t)$, and so the value of ϵ_t would indeed be small. Let's try to gain some intuition behind how this should happen.

Instead of considering the exponentially weighted moving average, let's just consider a simple *windowed* average:

$$\mathbf{m}_t = \frac{1}{K}\sum_{i=0}^{K-1} \nabla \ell(\mathbf{w}_{t-i}, z_{t-i})$$

What is ϵ_t in this setting?

$$\begin{aligned}\epsilon_t &= \frac{1}{K}\sum_{i=0}^{K-1}(\nabla \ell(\mathbf{w}_{t-i}, z_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_t))\\ &= \frac{1}{K}\sum_{i=0}^{K-1}(\nabla \ell(\mathbf{w}_{t-i}, z_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_{t-i})) + \frac{1}{K}\sum_{i=0}^{K-1}(\nabla \mathcal{L}(\mathbf{w}_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_t))\end{aligned}$$

Now, the first term above, $\frac{1}{K}\sum_{i=0}^{K-1}(\nabla \ell(\mathbf{w}_{t-i}, z_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_{t-i}))$, is an average of K mean-zero values, and so will be on average $O(1/\sqrt{K})$. The next term, on the other hand is a kind of "bias" that is harder to control. To gain some rough idea of how we could make it small, notice that by smoothness, $\|\nabla \mathcal{L}(\mathbf{w}_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_t)\| \le H\|\mathbf{w}_{t-i} - \mathbf{w}_t\|$. Further, if we expect $\|\mathbf{w}_{t-1} - \mathbf{w}_t\| = O(\eta)$ since the learning rate is η, we can argue:

$$\begin{aligned}\left\|\frac{1}{K}\sum_{i=0}^{K-1}(\nabla \mathcal{L}(\mathbf{w}_{t-i}) - \nabla \mathcal{L}(\mathbf{w}_t))\right\| &\le O\left(\frac{1}{K}\sum_{i=0}^{K-1}\eta H\right)\\ &= O\left(K\eta H\right)\end{aligned}$$

So overall, we have:

$$\|\epsilon_t\| = O(\frac{1}{\sqrt{K}} + \eta K)$$

Setting K optimally yields:

$$\|\epsilon_t\| = O(\eta^{1/3})$$

Now, let's try to apply Lemma 1:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t)] &\le O(-\eta\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \eta^{5/3})\\ \sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] &\le O\left(\frac{\Delta}{\eta} + T\eta^{2/3}\right)\end{aligned}$$

So, optimizing η yields:

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le O\left(T^{3/5}\right)$$

In contrast, with regular SGD we obtained $\sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le O\left(T^{1/2}\right)$, so it seems like things have actually gotten worse!

The issue here was that our bound $\|\epsilon_t\| = O(\eta^{1/3})$ was actually a bit loose. If we are more careful, we will be able to obtain $\|\epsilon_t\| = O(\eta^{1/2})$ instead. The key observation is that the bound $\|\mathbf{w}_{t-1} - \mathbf{w}_t\| = O(\eta)$ was also a bit loose: really, we have $\|\mathbf{w}_{t-1} - \mathbf{w}_t\| = O(\eta \mathbf{m}_{t-1}) = O(\eta\|\epsilon_{t-1}\| + \eta\|\nabla \mathcal{L}(\mathbf{w}_{t-1})\|)$. Thus, if ϵ_t and $\nabla \mathcal{L}(\mathbf{w}_t)$ are both getting smaller, we should expect $\|\mathbf{w}_{t-1} - \mathbf{w}_t\|$ to also be getting smaller, which in turn makes ϵ even smaller. This positive feedback cycle is what will eventually yield our improved bounds.

Next, we have the following fact about the error in the momentum estimates. This is the Lemma that will inform our choices for α and η, and will need to be modified for alternative momentum schemes.

Lemma 2. *For any $\alpha \le 1/2$, setting $\eta = \frac{\alpha}{\sqrt{48H}}$ guarantees:*

$$\frac{\sqrt{3}}{8H}\,\mathbb{E}[\|\epsilon_{t+1}\|^2 - \|\epsilon_t\|^2] \le -\frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta}{8}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + 3\sqrt{3}H\eta^2\sigma^2$$

Proof. Recall our notation:

$$r_t = \nabla \ell(\mathbf{w}_t, z_t) - \nabla \mathcal{L}(\mathbf{w}_t)$$

Note that we have the important property:

$$\mathbb{E}[r_t] = 0$$

Note however that $\mathbb{E}[\epsilon_t] \neq 0$.

Now, we derive a recursive formula for ϵ_{t+1} in terms of ϵ_t:

$$\begin{aligned}
\epsilon_{t+1} &= \mathbf{m}_{t+1} - \nabla \mathcal{L}(\mathbf{w}_{t+1}) \\
&= (1-\alpha)\mathbf{m}_t + \alpha_t \nabla \ell(\mathbf{w}_{t+1}, z_{t+1}) - \nabla \mathcal{L}(\mathbf{w}_{t+1}) \\
&= (1-\alpha)(\mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_{t+1})) + \alpha(\nabla \ell(\mathbf{w}_{t+1}, z_{t+1}) - \nabla \mathcal{L}(\mathbf{w}_{t+1})) \\
&= (1-\alpha)(\mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)) + (1-\alpha)(\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})) + \alpha r_{t+1} \\
&= (1-\alpha)\epsilon_t + (1-\alpha)(\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})) + \alpha r_{t+1}
\end{aligned}$$

Now, remember that we are actually interested in $\mathbb{E}[\|\epsilon_t\|^2]$, so let us take the norm-squared of both sides in the above, and use the fact that $\mathbb{E}[r_t] = 0$:

$$\begin{aligned}
\mathbb{E}[\|\epsilon_{t+1}\|^2] &= (1-\alpha)^2\,\mathbb{E}[\|\epsilon_t\|^2] + 2(1-\alpha)^2\,\mathbb{E}[\langle \epsilon_t, \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\rangle] + (1-\alpha)^2\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2] \\
&\quad + \alpha^2\,\mathbb{E}[\|r_{t+1}\|^2] \\
&\le (1-\alpha)^2\,\mathbb{E}[\|\epsilon_t\|^2] + 2(1-\alpha)^2\,\mathbb{E}[\langle \epsilon_t, \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\rangle] + (1-\alpha)^2\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2] \\
&\quad + \alpha^2\sigma^2
\end{aligned}$$

Notice that by H-smoothness,

$$\begin{aligned}
\|\nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\| &\le H\|\mathbf{w}_t - \mathbf{w}_{t+1}\| \\
&\le H\eta\|\mathbf{m}_t\| \\
&\le H\eta(\|\nabla \mathcal{L}(\mathbf{w}_t)\| + \|\epsilon_t\|)
\end{aligned}$$

Further, by Young inequality again, for any λ we have:

$$\langle \epsilon_t, \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})\rangle \le \frac{\|\epsilon_t\|^2}{2\lambda} + \lambda H^2\eta^2(\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \|\epsilon_t\|^2)$$

Thus, putting this together with $(1-\alpha)^2 \le 1$, we have:

$$\mathbb{E}[\|\epsilon_{t+1}\|^2] \le \left[(1-\alpha)^2 + \frac{1}{\lambda} + 2\lambda H^2\eta^2 + 2H^2\eta^2\right]\mathbb{E}[\|\epsilon_t\|^2] + \left[2\lambda H^2\eta^2 + H^2\eta^2\right]\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \alpha^2\sigma^2$$

Let's take a step back and see where we are going with this. Notice that $(1-\alpha)^2 = 1 - 2\alpha + \alpha^2$. Therefore, by setting $\lambda \approx 1/\alpha$, and then setting η in some clever way we might be able to guarantee something like $(1-\alpha)^2 + \frac{1}{\lambda} + 2\lambda L^2\eta^2 + 2L^2\eta^2 \approx 1 - \alpha$. Discarding many constants, and observing that $\lambda\eta^2 \approx \eta^2/\alpha$ is bigger η^2, this would yield something like:

$$\mathbb{E}[\|\epsilon_{t+1}\|^2] \le (1-\alpha)\,\mathbb{E}[\|\epsilon_t\|^2] + O(\eta^2/\alpha\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \alpha^2\sigma^2)$$

This suggests that $\|\epsilon_t\|^2$ will decrease until it reaches some "equilibrium" value when the above recurrence has $\|\epsilon_{t+1}\| = \|\epsilon_t\|$. Solving for this, we obtain:

$$\mathbb{E}[\eta\|\epsilon_t\|^2] \le O\left(\frac{\eta^3}{\alpha^2}\,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \alpha\eta\sigma^2\right)$$

So in order to keep the $\eta\|\epsilon_t\|^2$ terms in Lemma 1 from dominating the $-\eta\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$ terms, we should try to set $\eta = \alpha$

Let's see if this is possible. We will try to obtain $(1-\alpha)^2 + \frac{1}{\lambda} + 2\lambda H^2\eta^2 + 2H^2\eta^2 \le 1 - \frac{1}{2}\alpha$. We can expand the $(1-\alpha)^2$ to $1 - 2\alpha + \alpha^2$. Thus, we have an "extra" $-\frac{3}{2}\alpha$ term of slack, and three positive terms α^2, $\frac{1}{\lambda}$ and $2H^2(\lambda+1)\eta^2$. We will try to get all three positive terms at most $\frac{1}{2}\alpha$ to obtain the entire expression is at most $1 - \frac{1}{2}\alpha$. To accomplish this, we need:

$$\begin{aligned}
\alpha &\le \frac{1}{2}\\
\lambda &= \frac{2}{\alpha}\\
\eta &\le \frac{\alpha}{4H}
\end{aligned}$$

The first two equations deal with α^2 and $\frac{1}{\lambda}$. For the final term, notice that $\lambda > 1$, so that $2H^2(\lambda+1)\eta^2 \le 4H^2\lambda\eta^2 = \frac{8H^2\eta^2}{\alpha}$. So if we set $\eta \le \frac{\alpha}{4H}$, we will have $(1-\alpha)^2 + \frac{1}{\lambda} + 2\lambda H^2\eta^2 + 2H^2\eta^2 = 1 - \frac{1}{2}\alpha \le 1 - \frac{1}{2}\alpha$.

Thus overall we have obtained:

$$\begin{aligned}
\mathbb{E}[\|\epsilon_{t+1}\|^2] &\le (1 - \frac{\alpha}{2})\,\mathbb{E}[\|\epsilon_t\|^2] + \left[2\lambda H^2\eta^2 + H^2\eta^2\right]\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \alpha^2\sigma^2\\
&\le (1 - \frac{\alpha}{2})\,\mathbb{E}[\|\epsilon_t\|^2] + 4\lambda H^2\eta^2\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \alpha^2\sigma^2\\
&\le (1 - \frac{\alpha}{2})\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{8H^2\eta^2}{\alpha}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \alpha^2\sigma^2\\
\frac{3\eta}{2\alpha}\,\mathbb{E}[\|\epsilon_{t+1}\|^2 - \|\epsilon_t\|^2] &\le -\frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{6H^2\eta^3}{\alpha^2}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{3\eta\alpha}{4}\sigma^2
\end{aligned}$$

Now, if we set $\eta = \frac{\alpha}{\sqrt{48}H}$ (which satisfies $\eta \le \frac{\alpha}{4H}$), this simplifies to:

$$\frac{3\eta}{2\alpha}\,\mathbb{E}[\|\epsilon_{t+1}\|^2 - \|\epsilon_t\|^2] \le -\frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta}{8}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + 3\sqrt{3}H\eta^2\sigma^2$$

And the Lemma follows by observing $\frac{3\eta}{2\alpha} = \frac{\sqrt{3}}{8H}$ □

Theorem 3. *Algorithm 1 with* $\eta = \frac{\alpha}{\sqrt{48}H}$ *and* $\alpha = \min\left(\frac{1}{2}, \frac{\sqrt{32\sqrt{3}H\Delta + 12\sigma^2}}{\sigma\sqrt{T}}\right)$ *guarantees*

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

Proof. The proof works using the so-called *potential* method. We define the value

$$\Phi_t = \mathcal{L}(\mathbf{w}_t) + \frac{\sqrt{3}}{8H}\|\epsilon_t\|^2$$

We will then show that Φ_t roughly decreases with t at rate that depends on $\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2$. Specifically:

$$\mathbb{E}[\Phi_{t+1} - \Phi_t] = \mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) + \frac{\sqrt{3}}{8H}\|\epsilon_{t+1}\|^2 - \frac{\sqrt{3}}{8H}\|\epsilon_t\|^2]$$

applying Lemmas 1 and 2:

$$\begin{aligned}
&\le -\frac{\eta}{4}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2]\\
&\qquad - \frac{3\eta}{4}\,\mathbb{E}[\|\epsilon_t\|^2] + \frac{\eta}{8}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + 3\sqrt{3}H\eta^2\sigma^2
\end{aligned}$$

(Notice that the $\|\epsilon_t\|^2$ terms will magically cancel in the above: this was the point of including this weird-seeming $\frac{\sqrt{3}}{8H}\|\epsilon_t\|^2$ in the potential:

$$= -\frac{\eta}{8}\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + 3\sqrt{3}\eta^2\sigma^2$$

Now sum over t:

$$\mathbb{E}[\Phi_{T+1} - \Phi_1] \le -\frac{\eta}{8}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + 3\sqrt{3}\sigma T H\eta^2$$

rearranging...

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{8}{\eta}\,\mathbb{E}[\Phi_1 - \Phi_{T+1}] + 24\sqrt{3}T\eta H\sigma^2$$

Now, observe that

$$\begin{aligned}\mathbb{E}[\Phi_1 - \Phi_{T+1}] &= \mathbb{E}[\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_{T+1}) + \frac{\sqrt{3}}{8H}\|\epsilon_1\|^2 - \frac{\sqrt{3}}{8H}\|\epsilon_{T+1}\|^2] \\ &\le \Delta + \frac{\sqrt{3}\sigma^2}{8H}\end{aligned}$$

where in the second line we used $\mathbb{E}[\|\epsilon_1\|^2] = \mathbb{E}[\|r_1\|^2] \le \sigma^2$. Putting all this together with $\eta = \frac{\alpha}{H\sqrt{48\alpha}}$ yields:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{8}{T\eta}\,\mathbb{E}[\Phi_1 - \Phi_{T+1}] + 24\sqrt{3}\sigma^2 H\eta \\ &\le \frac{32\sqrt{3}H\Delta + 12\sigma^2}{T\alpha} + 6\sigma^2\alpha\end{aligned}$$

Now, we would like to balance these terms and set:

$$\alpha = \frac{\sqrt{32\sqrt{3}H\Delta + 12\sigma^2}}{\sigma\sqrt{T}}$$

but we also need to satisfy $\alpha \le 1/2$. So, this motivates the setting:

$$\alpha = \min\left(\frac{1}{2}, \frac{\sqrt{32\sqrt{3}H\Delta + 12\sigma^2}}{\sigma\sqrt{T}}\right)$$

Now, if $\alpha \ne 1/2$, we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{2\sqrt{32\sqrt{3}H\Delta + 12\sigma^2}\sigma\sqrt{6}}{\sqrt{T}} = O\left(\frac{\sigma\sqrt{H\Delta} + \sigma^2}{\sqrt{T}}\right)$$

On the other hand, if $\alpha = 1/2$, this implies

$$\frac{\sqrt{32\sqrt{3}H\Delta + 12\sigma^2}}{\sigma\sqrt{T}} \ge 1/2$$

so that:

$$\frac{\sigma^2}{4} \le \frac{32\sqrt{3}H\Delta + 12\sigma^2}{T}$$

so that overall:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{64\sqrt{3}H\Delta + 24\sigma^2}{T\alpha} + 3\sigma^2 \\ &\le O\left(\frac{H\Delta + \sigma^2}{T}\right)\end{aligned}$$

so that adding the two bounds for the two cases yields the desired result.

□