Lecture Notes 15: Normalized Gradient Descent and Non-Convex Variance Reduction

Instructor: Ashok Cutkosky

We have just seen how to perform variance reduction for finite-sum convex problems. It turns out that variance reduction is in some sense even more powerful for non-convex problems. In the convex case, we are only able to see gains over SGD in the special case that $\mathcal{L}(\mathbf{w})$ is has the form of a finite sum. In contrast, for non-convex problems we will be able to obtain improved convergence to critical points without this extra restriction. For reference, recall that SGD obtained the guarantee:

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

Thus, by selecting an iterate $\hat{\mathbf{w}}$ at random from $\mathbf{w}_1,\dots,\mathbf{w}_T$, we obtain:

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

By using variance reduction, we will be able to significantly improve this to:

$$\mathbb{E}[\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|] \le O\left(\frac{1}{T^{1/3}}\right)$$

This rate discovered simultaneously by two different groups in 2018 [1, 2], and in 2020 this was shown to be the optimal rate [3].

Our presentation of the results will look slightly more similar to [1], but somewhat more streamlined borrowing ideas from [4].

In order to derive the algorithm with a good balance of intuition we will need to consider *normalized* updates for SGD. To start, let's look at the following scheme:

$$\begin{aligned}
\mathbf{m}_1 &= \nabla\ell(\mathbf{w}_1, z_1)\\
\mathbf{m}_t &= (1-\alpha)\mathbf{m}_{t-1} + \alpha\nabla\ell(\mathbf{w}_t, z_t)\\
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta\frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}
\end{aligned}$$

We'll call these updates *normalized gradient descent with momentum*.

This is almost identical to our previously studied momentum methods, but now instead of writing $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{m}_t$, we normalized the momentum term in the update. This makes the following important identity true:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1\| = \eta \text{ for all } t$$

This identity is extremely useful for analysis. Recall that when we previously analyzed momentum in the non-convex setting, we tried to view momentum as a form of averaging, and the primary difficulty was trading off some bias caused by the fact that $\mathbf{w}_t$ is changing over time. Accuratly measuring this bias was very technically challenging because there was a complicated relationship between the speed that $\mathbf{w}_t$ is changing and the amount of bias. In the end, we never actually really quantified how much this bias was, but we were able to sidestep the problem through a tricky use of a potential function. With normalized updates, we are going to be able to completely avoid all of these difficulties.

In particular, we have the following result:

Lemma 1. *Suppose that $\mathcal{L}$ is an H-smooth function and $\mathbb{E}[\|\nabla\ell(\mathbf{w},z)-\nabla\mathcal{L}(\mathbf{w})\|^2]\le\sigma^2$ for all $\mathbf{w}$. Then using the normalized gradient descent with momentum updates, we have:*

$$\mathbb{E}[\|\mathbf{m}_t-\nabla\mathcal{L}(\mathbf{w}_t)\|]\le(1-\alpha)^t\sigma+\sigma\sqrt{\alpha}+\frac{H\eta}{\alpha}$$

Proof. Let's start by obtaining an expanded expression for $\mathbf{m}_t$. To compactify the notation, set $\mathbf{g}_t=\nabla\ell(\mathbf{w}_t,z_t)$. Further, let's define:

$$\begin{aligned}\epsilon_t&=\mathbf{m}_t-\nabla\mathcal{L}(\mathbf{w}_t)\\ r_t&=\mathbf{g}_t-\nabla\mathcal{L}(\mathbf{w}_t)\end{aligned}$$

Notice that $\mathbb{E}[r_t]=0$ and $\mathbb{E}[\|r_t\|^2]\le\sigma^2$, so that by Jensen inequality, $\mathbb{E}[\|r_t\|]\le\sigma$.

Then we have:

$$\begin{aligned}\mathbf{m}_t&=(1-\alpha)\mathbf{m}_{t-1}+\alpha\mathbf{g}_t\\ \epsilon_t&=(1-\alpha)(\mathbf{m}_{t-1}-\nabla\mathcal{L}(\mathbf{w}_t))+\alpha(\mathbf{g}_t-\nabla\mathcal{L}(\mathbf{w}_t))\\ &=(1-\alpha)(\mathbf{m}_{t-1}-\nabla\mathcal{L}(\mathbf{w}_{t-1}))+(1-\alpha)(\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t))+\alpha r_t\\ &=(1-\alpha)\epsilon_{t-1}+(1-\alpha)(\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t))+\alpha r_t\end{aligned}$$

Now, we have generated a recursive expression for ϵ_t. Notice that the third term, αr_t, is zero in expectation, so we might hope that it has a small contribution to ϵ_t. The second term, is bounded by:

$$\|\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t)\|\le H\|\mathbf{w}_{t-1}-\mathbf{w}_t\|=H\eta$$

So we can control it by setting η small. Notice that by using normalized updates, we have a very tight control over the difference of the gradients because we know *exactly* how big $\|\mathbf{w}_{t-1}-\mathbf{w}_t\|$ is.

Let's continue expanding the recursive expression for ϵ_t to see how we can leverage these intuitions:

$$\begin{aligned}\epsilon_t&=(1-\alpha)\epsilon_{t-1}+(1-\alpha)(\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t))+\alpha r_t\\ &=(1-\alpha)^2\epsilon_{t-2}+(1-\alpha)^2(\nabla\mathcal{L}(\mathbf{w}_{t-2})-\nabla\mathcal{L}(\mathbf{w}_{t-1}))+\alpha(1-\alpha)r_{t_1}+(1-\alpha)(\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t))+\alpha r_t\end{aligned}$$

unrolling for t iterations:

$$=(1-\alpha)^{t-1}\epsilon_1+\alpha(1-\alpha)^{t-2}r_2+\dots+\alpha r_t+(1-\alpha)^{t-1}(\nabla\mathcal{L}(\mathbf{w}_1)-\nabla\mathcal{L}(\mathbf{w}_2))+\dots+(1-\alpha)(\nabla\mathcal{L}(\mathbf{w}_{t-1})-\nabla\mathcal{L}(\mathbf{w}_t))$$

recall that $\mathbf{m}_1=\mathbf{g}_1$ so that $\epsilon_1=r_1$:

$$\begin{aligned}&=(1-\alpha)^{t-1}r_1+\alpha(1-\alpha)^{t-2}r_2+\dots+\alpha(1-\alpha)r_t+\sum_{\tau=1}^{t-1}(1-\alpha)^{t-\tau}(\nabla\mathcal{L}(\mathbf{w}_\tau)-\nabla\mathcal{L}(\mathbf{w}_{\tau+1}))\\ &=(1-\alpha)^t r_1+\alpha(1-\alpha)^{t-1}r_1+\dots+\alpha(1-\alpha)r_t+\sum_{\tau=1}^{t-1}(1-\alpha)^{t-\tau}(\nabla\mathcal{L}(\mathbf{w}_\tau)-\nabla\mathcal{L}(\mathbf{w}_{\tau+1}))\\ &=(1-\alpha)^t r_1+\alpha\sum_{\tau=1}^{t}(1-\alpha)^{t-\tau}r_\tau+\sum_{\tau=1}^{t-1}(1-\alpha)^{t-\tau}(\nabla\mathcal{L}(\mathbf{w}_\tau)-\nabla\mathcal{L}(\mathbf{w}_{\tau+1}))\end{aligned}$$

do a little reindexing to make the geometric series in the sums clearer:

$$=(1-\alpha)^t r_1+\alpha\sum_{\tau=0}^{t}(1-\alpha)^{\tau}r_{t-\tau}+\sum_{\tau=1}^{t-1}(1-\alpha)^{\tau}(\nabla\mathcal{L}(\mathbf{w}_{t-\tau})-\nabla\mathcal{L}(\mathbf{w}_{t-\tau+1}))$$

Now, observe that all of these terms are expected to be small: the first term is of course geometrically decaying in t, and the other terms involve goemetric series of $(1-\alpha)$. Let's make this concrete by taking expectations:

$$\begin{aligned}
\mathbb{E}[\|\epsilon_t\|] &\le (1-\alpha)^t\,\mathbb{E}[\|r_1\|] + \alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] + \sum_{\tau=1}^{t-1}(1-\alpha)^\tau\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_{t-\tau}) - \nabla\mathcal{L}(\mathbf{w}_{t-\tau+1})\|]\\
&= (1-\alpha)^t\,\mathbb{E}[\|r_1\|] + \alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] + \sum_{\tau=1}^{t-1}(1-\alpha)^\tau H\eta\\
&\le (1-\alpha)^t\,\mathbb{E}[\|r_1\|] + \alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] + \sum_{\tau=0}^{\infty}(1-\alpha)^\tau H\eta\\
&= (1-\alpha)^t\,\mathbb{E}[\|r_1\|] + \alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] + \frac{H\eta}{\alpha}\\
&\le (1-\alpha)^t\sigma + \alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] + \frac{H\eta}{\alpha}
\end{aligned}$$

Now, by Jensen inequality:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] &\le \sqrt{\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|^2\right]}\\
&\le \sqrt{\mathbb{E}\left[\sum_{\tau=0}^{t}\sum_{\tau'=0}^{t}(1-\alpha)^{\tau\tau'}\langle r_{t-\tau}, r_{t-\tau'}\rangle\right]}
\end{aligned}$$

since $\mathbb{E}[\langle r_t, r_t'\rangle] = 0$ for $t \neq t'$ and $\mathbb{E}[\|r_t\|^2] \le \sigma^2$:

$$\begin{aligned}
&\le \sqrt{\sum_{\tau=0}^{t}(1-\alpha)^{2\tau}\sigma^2}\\
&\le \sigma\sqrt{\sum_{\tau=0}^{t}(1-\alpha)^{\tau}}\\
&\le \sigma\sqrt{\sum_{\tau=0}^{\infty}(1-\alpha)^{\tau}}\\
&= \frac{\sigma}{\sqrt{\alpha}}
\end{aligned}$$

Thus, $\alpha\,\mathbb{E}\left[\left\|\sum_{\tau=0}^{t}(1-\alpha)^\tau r_{t-\tau}\right\|\right] \le \sigma\sqrt{\alpha}$. So, putting all this together:

$$\mathbb{E}[\|\epsilon_t\|] \le (1-\alpha)^t\sigma + \sigma\sqrt{\alpha} + \frac{H\eta}{\alpha}$$

$\square$

This Lemma tells us that, by setting α and η appropriately, we will be able to ensure that $\mathbf{m}_t \approx \nabla\mathcal{L}(\mathbf{w}_t)$ in expectation. Now, it remains to see how we can use this property. To do this, we'll need a variation on the lemma for biased gradient descent we established when analyzing SGD with momentum:

Lemma 2. *Define $\epsilon_t = \mathbf{m}_t - \nabla\mathcal{L}(\mathbf{w}_t)$. Then we have:*

$$\mathcal{L}(\mathbf{w}_{t+1}) \le \mathcal{L}(\mathbf{w}_t) - \frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{13\eta}{6}\|\epsilon_t\| + \frac{H\eta^2}{2}$$

Proof. By the smoothness property:

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

Now, let's consider two cases, either $\|\epsilon_t\| \ge \frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|$ or not. If $\|\epsilon_t\| \ge \frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|$, then:

$$\begin{aligned}-\left\langle\nabla\mathcal{L}(\mathbf{w}_t), \frac{\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t}{\|\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t\|}\right\rangle &\le \|\nabla\mathcal{L}(\mathbf{w}_t)\|\\ &\le 2\|\epsilon_t\|\\ &\le -\frac{\|\nabla\mathcal{L}(\mathbf{w}_t)\|}{3} + \frac{13}{6}\|\epsilon_t\|\end{aligned}$$

Alternatively, if $\|\epsilon_t\| \le \frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|$:

$$\begin{aligned}\|\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t\| &\le \frac{3}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\\ -\langle\nabla\mathcal{L}(\mathbf{w}_t), \epsilon_t\rangle &\le \frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\\ -\left\langle\nabla\mathcal{L}(\mathbf{w}_t), \frac{\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t}{\|\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t\|}\right\rangle &= -\frac{\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2 + \langle\nabla\mathcal{L}(\mathbf{w}_t), \epsilon_t\rangle}{\|\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t\|}\\ &\le -\frac{\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2/2}{3\|\nabla\mathcal{L}(\mathbf{w}_2)\|/2}\\ &= -\frac{\|\nabla\mathcal{L}(\mathbf{w}_t)\|}{3}\end{aligned}$$

Therefore, either way we have:

$$-\eta\left\langle\nabla\mathcal{L}(\mathbf{w}_t), \frac{\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t}{\|\nabla\mathcal{L}(\mathbf{w}_t)+\epsilon_t\|}\right\rangle \le -\frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{13\eta}{6}\|\epsilon_t\|$$

from which the result follows. ☐

Now, we're ready to put everything together and analyze this new version of momentum:

Theorem 3. *Define* $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$. *Suppose* $\mathcal{L}$ *is* H*-smooth and* $\mathbf{g}_t$ *has variance at most* σ^2. *Then with* $\alpha = \min\left(1, \frac{\sqrt{\Delta H}}{\sigma\sqrt{T}}\right) = O(1/\sqrt{T})$ *and* $\eta = \frac{\sqrt{\Delta\alpha}}{\sqrt{HT}} = O(1/T^{3/4})$,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] &\le 24\sqrt{\Delta HT} + \frac{35(\Delta HT^3\sigma^2)^{1/4}}{2} + \frac{13\sqrt{T}}{2\sqrt{\Delta H}}\\ &\le O(T^{3/4})\end{aligned}$$

Proof. Applying Lemma 2 followed by Lemma 1, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] &\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{13\eta}{6}\|\epsilon_t\| + \frac{H\eta^2}{2}]\\ &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{H\eta^2}{2} + \frac{13\eta}{6}\left((1-\alpha)^t\sigma + \sigma\sqrt{\alpha} + \frac{H\eta}{\alpha}\right)\right]\end{aligned}$$

telescoping over t:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta}{6}\left(T\sigma\sqrt{\alpha} + \frac{HT\eta}{\alpha} + \sum_{t=1}^T (1-\alpha)^t\sigma\right)\right]\\
&\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta}{6}\left(T\sigma\sqrt{\alpha} + \frac{HT\eta}{\alpha} + \frac{\sigma}{\alpha}\right)\right]\\
&= \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta T\sigma\sqrt{\alpha}}{6} + \frac{13HT\eta^2}{6\alpha} + \frac{13\eta\sigma}{6\alpha}\right]\\
&\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{8HT\eta^2}{3\alpha} + \frac{13\eta T\sigma\sqrt{\alpha}}{6} + \frac{13\eta\sigma}{6\alpha}\right]
\end{aligned}$$

Now let's define $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$ and rearrange:

$$\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{3\Delta}{\eta} + \frac{8HT\eta}{\alpha} + \frac{13T\sigma\sqrt{\alpha}}{2} + \frac{13\sigma}{2\alpha}$$

Now, all that remains is to set α and η appropriately. This is a somewhat tricky task. To start, notice that the optimal value for η should balance the $\frac{3\Delta}{\eta}$ and the $\frac{8HT\eta}{\alpha}$ terms. From this, we can get (ignoring the constant factors) $\eta = \frac{\sqrt{\Delta\alpha}}{\sqrt{HT}}$ so that:

$$\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{11\sqrt{\Delta HT}}{\sqrt{\alpha}} + \frac{13T\sigma\sqrt{\alpha}}{2} + \frac{13\sigma}{2\alpha}$$

Now, observe that unless $\alpha \le \frac{1}{T^{2/3}}$, we should expect the $T\sqrt{\alpha}$ term to be larger than the $1/\alpha$ term. Then, to balance the first and second terms, we can set $\alpha = \frac{\sqrt{\Delta H}}{\sigma\sqrt{T}}$. This would yield:

$$\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{35(\Delta HT^3\sigma^2)^{1/4}}{2} + \frac{13\sqrt{T}}{2\sqrt{\Delta H}}$$

However, there is a subtlety: this value of α may not be allowed because we must have $\alpha \le 1$. If it is not allowed, then $\frac{\sqrt{\Delta H}}{\sigma\sqrt{T}} \ge 1$, so that $\sigma \le \frac{\sqrt{\Delta H}}{\sqrt{T}}$, and we we set $\alpha = 1$ to obtain:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] &\le 11\sqrt{\Delta HT} + \frac{13T\sigma}{2} + \frac{13\sigma}{2}\\
&\le 11\sqrt{\Delta HT} + 13\sqrt{\Delta HT}\\
&\le 24\sqrt{\Delta HT}
\end{aligned}$$

Thus, with $\alpha = \min\left(1, \frac{\sqrt{\Delta H}}{\sigma\sqrt{T}}\right)$, we obtain:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^T \|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] &\le 24\sqrt{\Delta HT} + \frac{35(\Delta HT^3\sigma^2)^{1/4}}{2} + \frac{13\sqrt{T}}{2\sqrt{\Delta H}}\\
&\le O(T^{3/4})
\end{aligned}$$

□

Now, this just recovers the standard SGD rate we've seen before. However, it turns out that a small tweak to formula will enable us to get the improved variance-reduction rate without too much extra work in the analysis.

1 Adding the Variance Reduction

The variance reduction scheme we will describe now is different than SVRG algorithm we saw earlier: we will not assume that the $\mathcal{L}$ has a finite-sum form, and we will never have to evaluate a full batch (this is good, because if $\mathcal{L}$ is not a finite-sum form it's not even possible to evaluate a full batch!). However, we will make the assumption the $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, z)]$ where $\ell(\mathbf{w}, z)$ is H-smooth in $\mathbf{w}$ for all z.

Now, our new variance-reduced momentum scheme will be the following:

$$\begin{aligned}
\mathbf{m}_1 &= \nabla \ell(\mathbf{w}_1, z_1)\\
\mathbf{m}_t &= (1-\alpha)(\mathbf{m}_{t-1} + \nabla \ell(\mathbf{w}_t, z_t) - \nabla \ell(\mathbf{w}_{t-1}, z_t)) + \alpha \nabla \ell(\mathbf{w}_t, z_t)\\
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta \frac{\mathbf{m}_t}{\|\mathbf{m}_t\|}
\end{aligned}$$

Let's call this normalized gradient descent with variance-reduced momentum.

This is almost the same as what we had previously, but now there is an extra $\nabla \ell(\mathbf{w}_t, z_t) - \nabla \ell(\mathbf{w}_{t-1}, z_t)$ added into the momentum update. Intuitively, this term is correcting some bias: since $\mathbf{m}_{t-1}$ is an estimate for the gradient at $\nabla \mathcal{L}(\mathbf{w}_{t-1})$ rather than an $\nabla \mathcal{L}(\mathbf{w}_t)$, we picked up some bias terms when analyzing $\|\epsilon_t\|$ in Lemma 1. Since $\mathbb{E}[\nabla \ell(\mathbf{w}_t, z_t) - \nabla \ell(\mathbf{w}_{t-1}, z_t)] = \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t-1})$, adding this term to the $\mathbf{m}_{t-1}$ is attempting to "de-bias" the momentum to mitigate this effect.

Let's see an analog of Lemma 1 for this new update:

Lemma 4. *Suppose that $\ell(\mathbf{w}, z)$ is an H-smooth function for all z and $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \mathcal{L}(\mathbf{w})\|^2] \le \sigma^2$ for all $\mathbf{w}$. Then using the normalized gradient descent with variance-reduced momentum updates, we have:*

$$\mathbb{E}[\|\mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|] \le (1-\alpha)^t \sigma + \sigma\sqrt{\alpha} + \frac{H\eta}{\sqrt{\alpha}}$$

Proof. The proof is extremely similar to Lemma 1. Set $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$. Define:

$$\begin{aligned}
\epsilon_t &= \mathbf{m}_t - \nabla \mathcal{L}(\mathbf{w}_t)\\
r_t &= \mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\\
\delta_t &= \mathbf{w}_t - \mathbf{w}_{t-1}
\end{aligned}$$

Now, we have:

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

Now, notice the critical difference from the proof of Lemma 1: the middle term here is now also zero in expectation! Let's define

$$s_t = \nabla \ell(\mathbf{w}_{t+1}, z_{t+1}) - \nabla \ell(\mathbf{w}_t, z_{t+1}) + \nabla \mathcal{L}(\mathbf{w}_t) - \nabla \mathcal{L}(\mathbf{w}_{t+1})$$

then we have $\mathbb{E}[s_t] = 0$, and

$$\begin{aligned}
\mathbb{E}[\|s_t\|^2] &\le \mathbb{E}[\|\nabla \ell(\mathbf{w}_{t+1}, z_{t+1}) - \nabla \ell(\mathbf{w}_t, z_{t+1})\|^2]\\
&\le H^2 \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2\\
&= H^2 \eta^2
\end{aligned}$$

Now, let's proceed to unroll the recursion once again:

$$\epsilon_t = (1-\alpha)^{t-1}\epsilon_1 + \alpha(1-\alpha)^{t-2} r_2 + \cdots + \alpha r_t + (1-\alpha)^{t-1} s_1 + \cdots + (1-\alpha) s_{t-1}$$

recall that $\mathbf{m}_1 = \mathbf{g}_1$ so that $\epsilon_1 = r_1$:

$$= (1-\alpha)^{t-1}r_1 + \alpha(1-\alpha)^{t-1}r_1 + \dots + \alpha(1-\alpha)r_t + \sum_{\tau=1}^{t-1}(1-\alpha)^{t-\tau}s_\tau$$
$$= (1-\alpha)^t r_1 + \alpha\sum_{\tau=1}^{t}(1-\alpha)^{t-\tau}r_\tau + \sum_{\tau=1}^{t-1}(1-\alpha)^{t-\tau}s_\tau$$

do a little reindexing to make the geometric series in the sums clearer:

$$= (1-\alpha)^t r_1 + \alpha\sum_{\tau=0}^{t}(1-\alpha)^{\tau}r_{t-\tau} + \sum_{\tau=1}^{t-1}(1-\alpha)^{\tau}s_\tau$$

Now, let's take norms and expectations. The first two terms are bounded identically to in the proof of Lemma 1.

$$\mathbb{E}[\|\epsilon_t\|] \le (1-\alpha)^t\sigma + \sigma\sqrt{\alpha} + \mathbb{E}\left[\left\|\sum_{\tau=1}^{t-1}(1-\alpha)^\tau s_\tau\right\|\right]$$

Now, for this last term the argument is again familiar:

$$\mathbb{E}\left[\left\|\sum_{\tau=1}^{t-1}(1-\alpha)^\tau s_\tau\right\|\right] \le \sqrt{\mathbb{E}\left[\left\|\sum_{\tau=1}^{t-1}(1-\alpha)^\tau s_\tau\right\|^2\right]}$$

using $\mathbb{E}[s_t] = 0$:

$$\le \sqrt{\sum_{\tau=1}^{t-1}(1-\alpha)^{2\tau}\,\mathbb{E}[\|s_\tau\|^2]}$$

using $\mathbb{E}[\|s_t\|^2] \le H^2\eta^2$:

$$\le H\eta\sqrt{\sum_{\tau=1}^{t-1}(1-\alpha)^{2\tau}}$$
$$\le \frac{H\eta}{\sqrt{\alpha}}$$

So over all we have obtained:

$$\mathbb{E}[\|\epsilon_t\|] \le (1-\alpha)^t\sigma + \sigma\sqrt{\alpha} + \frac{H\eta}{\sqrt{\alpha}}$$

□

Compare this result with Lemma 1: notice that the $\frac{\eta}{\alpha}$ term has improved to $\frac{\eta}{\sqrt{\alpha}}$.

Now, look back to the proof of Lemma 2: this Lemma actually made zero assumptions whatsoever about how $\mathbf{m}_t$ was generated. Thus, it applies equally well with our new improved way to generate $\mathbf{m}_t$ and so we can applying directly analogously to the proof of Theorem 3 to show:

Theorem 5. *Define $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$. Suppose $\ell(\mathbf{w}, z)$ is H-smooth for all z and $\nabla\ell(\mathbf{w}, z)$ has variance at most σ^2. Then with $\alpha = 1/T^{2/3}$ and $\eta = \frac{\sqrt{\Delta\sqrt{\alpha}}}{\sqrt{HT}} = O(1/T^{2/3})$,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le 11\sqrt{\Delta H}T^{2/3} + 13\sigma T^{2/3}$$
$$\le O(T^{2/3})$$

Proof. Applying Lemma 2 followed by Lemma 1, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] &\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{13\eta}{6}\|\epsilon_t\| + \frac{H\eta^2}{2}] \\ &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{3}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{H\eta^2}{2} + \frac{13\eta}{6}\left((1-\alpha)^t\sigma + \sigma\sqrt{\alpha} + \frac{H\eta}{\sqrt{\alpha}}\right)\right]\end{aligned}$$

telescoping over t:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1})] &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta}{6}\left(T\sigma\sqrt{\alpha} + \frac{HT\eta}{\sqrt{\alpha}} + \sum_{t=1}^{T}(1-\alpha)^t\sigma\right)\right] \\ &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta}{6}\left(T\sigma\sqrt{\alpha} + \frac{HT\eta}{\sqrt{\alpha}} + \frac{\sigma}{\alpha}\right)\right] \\ &= \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{HT\eta^2}{2} + \frac{13\eta T\sigma\sqrt{\alpha}}{6} + \frac{13HT\eta^2}{6\sqrt{\alpha}} + \frac{13\eta\sigma}{6\alpha}\right] \\ &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \frac{\eta}{3}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\| + \frac{8HT\eta^2}{3\sqrt{\alpha}} + \frac{13\eta T\sigma\sqrt{\alpha}}{6} + \frac{13\eta\sigma}{6\alpha}\right]\end{aligned}$$

Rearranging:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{3\Delta}{\eta} + \frac{8HT\eta}{\sqrt{\alpha}} + \frac{13T\sigma\sqrt{\alpha}}{2} + \frac{13\sigma}{2\alpha}$$

Now, again we need only to choose the values for η and α. Balancing the first two terms with $\eta = \frac{\sqrt{\Delta\sqrt{\alpha}}}{\sqrt{HT}}$ yields:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le 11\frac{\sqrt{\Delta HT}}{\alpha^{1/4}} + \frac{13T\sigma\sqrt{\alpha}}{2} + \frac{13\sigma}{2\alpha}$$

Now, set $\alpha = \frac{1}{T^{2/3}}$ to obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le 11\sqrt{\Delta H}T^{2/3} + 13\sigma T^{2/3}$$

If you want to be a little more careful, we can set $\alpha = \min\left(1, \frac{(\Delta H)^{2/3}}{\sigma^{4/3}T^{2/3}}\right)$. Then, if $\alpha = 1$ we have $\sigma \le$

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le 34\sqrt{\Delta HT}$$

while otherwise we have:

$$\mathbb{E}\left[\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|\right] \le \frac{35(\Delta H\sigma)^{1/3}T^{2/3}}{2} + \frac{13\sigma^{7/3}T^{2/3}}{2(\Delta H)^{2/3}}$$

□

References

[1] C Fang et al. "SPIDER: Near-optimal non-convex optimization via stochastic path integrated differential estimator". In: *Advances in Neural Information Processing Systems* 31 (2018), p. 689.

[2] Dongruo Zhou, Pan Xu, and Quanquan Gu. "Stochastic Nested Variance Reduction for Nonconvex Optimization". In: *Advances in Neural Information Processing Systems* 31 (2018), pp. 3921–3932.

[3] Yossi Arjevani et al. "Lower bounds for non-convex stochastic optimization". In: *arXiv preprint arXiv:1912.02365* (2019).

[4] Ashok Cutkosky and Harsh Mehta. "Momentum improves normalized sgd". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2260–2268.