CMPT 409/981: Optimization for Machine Learning
Lecture 8: Additional Notes

Sharan Vaswani *

December 8, 2022

In Lecture 8, on Slide 17, we proved an $O(1/T + \sigma^2)$ convergence rate for constant step-size SGD when minimizing smooth, convex functions. For this result, we assumed that the variance is bounded i.e $\mathbb{E}_i \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \sigma^2$ and used a step-size $\eta = \frac{1}{2L}$ where L is the smoothness of f. However, in Assignment 3, we saw that this scheme could result in poor empirical performance because the resulting step-size is too large.

Though the proof we did is correct, it is quite loose and in this note, we will provide a better proof with a weaker notion of variance. In order for this note to be self-contained, let us first repeat the old proof from Lecture 8.

Claim: For L-smooth, convex functions, T iterations of stochastic gradient descent with $\eta_k = \frac{1}{2L}$ returns an iterate $\bar{w}_T = \frac{\sum_{k=0}^{T-1} w_k}{T}$ such that,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \underbrace{\frac{2L\,\|w_0 - w^*\|^2}{T}}_{\text{bias}} + \underbrace{\frac{\sigma^2}{2L}}_{\text{neighbourhood}}$$

Proof. Using the SGD update, $w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$,

$$\begin{aligned}\|w_{k+1} - w^*\|^2 &= \|w_k - \eta_k \nabla f_{ik}(w_k) - w^*\|^2 \\ &= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \eta_k^2 \|\nabla f_{ik}(w_k)\|^2\end{aligned}$$

Taking expectation w.r.t i_k on both sides, and assuming η_k is independent of i_k

$$\begin{aligned}\mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\mathbb{E}\left[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle\right] + \mathbb{E}\left[\eta_k^2 \|\nabla f_{ik}(w_k)\|^2\right] \\ &= \|w_k - w^*\|^2 - 2\eta_k \langle \mathbb{E}\left[\nabla f_{ik}(w_k)\right], w_k - w^* \rangle + \eta_k^2\, \mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \\ \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2\, \mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \qquad \text{(Unbiasedness)}\end{aligned}$$

Now we need to control the $\mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right]$ term.

$$\begin{aligned}&\mathbb{E}[\|w_{k+1} - w^*\|^2] \\ &\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2\, \mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f(w_k) + \nabla f(w_k)\|^2\right] \\ &= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2\, \mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f(w_k)\|^2\right] + \eta_k^2\, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \\ &\qquad \text{(Since } \mathbb{E}[\langle \nabla f(w_k), \nabla f_{ik}(w_k) - \nabla f(w_k) \rangle] = 0\text{)} \\ &\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2\, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] + \eta_k^2 \sigma^2 \\ &\qquad \text{(Using the bounded variance assumption)}\end{aligned}$$

*Thanks to Reza Babanezhad for checking the proof.

Using convexity of f, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ with $y = w^*$ and $x = w_k$,

$$\leq \|w_k - w^*\|^2 - 2\eta_k[f(w_k) - f(w^*)] + \eta_k^2 \, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] + \eta_k^2 \, \sigma^2$$

$$\mathbb{E}\,\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta_k[f(w_k) - f(w^*)] + 2L\,\eta_k^2 \, \mathbb{E}[f(w_k) - f(w^*)] + \eta_k^2 \, \sigma^2$$

(Using L-smoothness of f)

Since $\eta_k \leq \frac{1}{2L}$, $2L\,\eta_k \leq 1$,

$$\mathbb{E}\,\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - \eta_k[f(w_k) - f(w^*)] + \eta_k^2 \, \sigma^2$$

$$\implies \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{1}{\eta_k}\left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] + \eta_k \, \sigma^2$$

$$= 2L\left[\|w_k - w^*\|^2 - \mathbb{E}\,\|w_{k+1} - w^*\|^2\right] + \frac{\sigma^2}{2L} \qquad \text{(Since } \eta_k = \tfrac{1}{2L}\text{)}$$

Summing from $k = 0$ to $k = T - 1$, telescoping the first term on the RHS and dividing by T

$$\frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)]}{T} \leq \frac{2L\,\|w_0 - w^*\|^2}{T} + + \frac{\sigma^2}{2L}$$

Using Jensen's inequality on the LHS, and the definition of $\bar{w}_T$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\,\|w_0 - w^*\|^2}{T} + \frac{\sigma^2}{2L}$$

□

Let us now prove a bound that will make a weaker assumption on the variance, and the resulting algorithm will result in better empirical performance.

For this, we consider minimizing need the additional assumption that each f_i is L_i-smooth and define $L_{\max} := \max_i L_i$. We will use a step-size of $\eta = \frac{1}{4L_{\max}}$ and prove an $O(1/T + \zeta^2)$ convergence where $\zeta^2 := \mathbb{E}_i\,\|\nabla f_i(w^*)\|^2 = \mathbb{E}_i\,\|\nabla f_i(w^*) - \nabla f(w^*)\|^2$ i.e. we need the variance to be bounded only at the minimizer (instead of each iterate like in the definition of σ^2). Moreover, since $L_{\max} \geq L$, the resulting step-size will be smaller and result in better empirical performance. Let us prove the following claim:

Claim: When minimizing the function$f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ where f is convex and each f_i is L_i-smooth such that $L_{\max} = \max_i L_i$, T iterations of stochastic gradient descent with $\eta_k = \frac{1}{4L_{\max}}$ returns an iterate $\bar{w}_T = \frac{\sum_{k=0}^{T-1} w_k}{T}$ such that,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \underbrace{\frac{4L_{\max}\,\|w_0 - w^*\|^2}{T}}_{\text{bias}} + \underbrace{\frac{\zeta^2}{2L_{\max}}}_{\text{neighbourhood}}$$

Proof. Using the same initial steps as before, we reach the following inequality,

$$
\begin{aligned}
&\mathbb{E}[\|w_{k+1} - w^*\|^2] \\
&\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \\
&= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*) + \nabla f_{ik}(w^*)\|^2\right] \\
&\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 2\eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2 + 2\eta_k^2 \, \mathbb{E}\,\|\nabla f_{ik}(w^*)\|^2\right] \\
&\hspace{25em} (\text{Since } \|a+b\|^2 \leq 2\,\|a\|^2 + \|b\|^2) \\
&= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 2\eta_k^2 \, \mathbb{E}\,\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2 + 2\eta_k^2 \, \zeta^2 \\
&\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 4\eta_k^2 \, L_{\max} \mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w^*), w^* - w_k \rangle\right] + 2\eta_k^2 \, \zeta^2 \\
&\hspace{20em} (\text{Since each } f_{ik} \text{ is } L_i \text{ and hence } L_{\max}\text{-smooth}) \\
&= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 4\eta_k^2 \, L_{\max} \left[f(w_k) - f(w^*) + \langle \nabla f(w^*), w^* - w_k \rangle\right] + 2\eta_k^2 \, \zeta^2 \\
&\hspace{30em} (\text{Unbiasedness})
\end{aligned}
$$

$$
\implies \mathbb{E}\,\|w_{k+1} - w^*\|^2 \leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 4\eta_k^2 \, L_{\max} \left[f(w_k) - f(w^*)\right] + 2\eta_k^2 \, \zeta^2 \qquad (\nabla f(w^*) = 0)
$$

Using convexity of f to simplify the second term, and since $\eta_k \leq \frac{1}{4L_{\max}}$, $4L_{\max}\,\eta_k \leq 1$,

$$
\begin{aligned}
\mathbb{E}\,\|w_{k+1} - w^*\|^2 &\leq \|w_k - w^*\|^2 - \eta_k \left[f(w_k) - f(w^*)\right] + 2\eta_k^2 \, \zeta^2 \\
\implies \mathbb{E}[f(w_k) - f(w^*)] &\leq \frac{1}{\eta_k} \left[\|w_k - w^*\|^2 - \mathbb{E}\,\|w_{k+1} - w^*\|^2\right] + 2\eta_k \, \zeta^2 \\
&= 4L_{\max} \left[\|w_k - w^*\|^2 - \mathbb{E}\,\|w_{k+1} - w^*\|^2\right] + \frac{\zeta^2}{2L_{\max}} \qquad (\text{Since } \eta_k = \tfrac{1}{4L_{\max}})
\end{aligned}
$$

Summing from $k = 0$ to $k = T - 1$, telescoping the first term on the RHS and dividing by T

$$
\frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)]}{T} \leq \frac{4L_{\max}\,\|w_0 - w^*\|^2}{T} + \frac{\zeta^2}{2L_{\max}}
$$

Using Jensen's inequality on the LHS, and the definition of $\bar{w}_T$,

$$
\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{4L_{\max}\,\|w_0 - w^*\|^2}{T} + \frac{\zeta^2}{2L_{\max}}
$$

□

We can do a similar analysis for the decreasing $O(1/\sqrt{k})$ step-size, and obtain a dependence on ζ^2 (instead of σ^2).