CMPT 409/981: Optimization for Machine Learning
Lecture 10: Additional Notes

Sharan Vaswani *

October 20, 2022

In Lecture 10, on Slide 6, we proved an $O(1/T)$ convergence rate for SGD when minimizing smooth, strongly-convex functions. For simplicity, we assumed that the stochastic gradients are bounded i.e. there exists a G such that $\mathbb{E}\left\|\nabla f_i(w)\right\|^2 \leq G^2$ for all w.

In this note, we relax this assumption and use a proof similar to Gower et al. (2019). For this, we will use ideas from the proof for the decreasing step-size (Slide 6 in Lecture 10) and constant step-size (Slide 1 in Lecture 10). We will prove the following claim.

Claim: For L-smooth, μ-strongly convex functions, T iterations of SGD with

$$\eta_k = \frac{1}{L} \quad ; \text{For } k < k_0 \quad \text{(Phase 1)}$$
$$\eta_k = \frac{1}{\mu\,(k+1)} \quad ; \text{For } k \geq k_0 \quad \text{(Phase 2)}$$

for $k_0 := \lceil 2\kappa - 1 \rceil$ returns iterate $\bar{w}_T := \frac{\sum_{k=k_0}^{T-1} w_k}{T-k_0}$ such that for $T \geq k_0$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\mu\,\lceil 2\kappa - 1 \rceil}{T - \lceil 2\kappa - 1 \rceil}\left[\exp\left(\frac{-\lceil 2\kappa - 1 \rceil}{\kappa}\right)\left\|w_0 - w^*\right\|^2 + \frac{\sigma^2}{\mu\,L}\right] + \frac{\sigma^2\,[1 + \log(T)]}{\mu\,(T - \lceil 2\kappa - 1 \rceil)}\,.$$

Proof: For the proof, we will require that $\eta_k \leq \frac{1}{2L}$ in Phase 2, i.e. for all $k \geq k_0$

$$\implies \frac{1}{\mu\,(k+1)} \leq \frac{1}{2L} \implies k \geq 2\kappa - 1\,.$$

Since Phase 2 only starts when $k \geq k_0 = \lceil 2\kappa - 1 \rceil$, this ensures that $\eta_k \leq \frac{1}{2L}$ in Phase 2. Expanding the iterate distance to w^* similar to the previous proofs,

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

*Thanks to Reza Babanezhad for checking the proof.

Taking expectation w.r.t i_k on both sides,

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

Using μ-strong convexity, $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$ with $y = w^*$ and $x = w_k$,

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

Using L-smoothness of f,

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

Let us first analyze Phase 2. Since $\eta_k \le \frac{1}{2L}$ in Phase 2, using Eq. (1) for all $k \ge k_0$,

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

Proceeding with the proof as in Slides 7-8,

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

Taking expectation w.r.t the randomness from iterations $k = k_0$ to $T - 1$,

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

Summing from $k = k_0$ to $T - 1$ in Phase 2

$$\begin{aligned}
\sum_{k=k_0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] &\le \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2\ (1 - \mu\, \eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \sigma^2 \sum_{k=k_0}^{T-1} \eta_k \\
&\le \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2\ (1 - \mu\, \eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \sigma^2 \sum_{k=0}^{T-1} \frac{1}{\mu\, (k+1)} \\
&\le \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2\ (1 - \mu\, \eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \frac{\sigma^2\, [1 + \log(T)]}{\mu}
\end{aligned}$$

Dividing by $T - k_0$, using Jensen's inequality for the LHS, and by definition of $\bar{w}_T$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \le \frac{1}{T - k_0} \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2\ (1 - \mu\, \eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \frac{\sigma^2\, [1 + \log(T)]}{\mu\, (T - k_0)}$$

Let us now simplify the second term similar to Slide 9,

$$\frac{1}{T-k_0}\sum_{k=k_0}^{T-1}\frac{\mathbb{E}\left[\|w_k-w^*\|^2\,(1-\mu\,\eta_k)-\|w_{k+1}-w^*\|^2\right]}{\eta_k}$$

$$= \frac{1}{T-k_0}\,\mathbb{E}\left[\sum_{k=k_0+1}^{T-1}\left[\|w_k-w^*\|^2\left(\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}-\mu\right)\right]+\|w_{k_0}-w^*\|^2\left(\frac{1}{\eta_{k_0}}-\mu\right)-\frac{\|w_T-w^*\|^2}{\eta_{T-1}}\right]$$

$$\leq \frac{1}{T-k_0}\,\mathbb{E}\left[\sum_{k=k_0+1}^{T-1}\left[\|w_k-w^*\|^2\,(\mu(k+1)-\mu k-\mu)\right]+\|w_{k_0}-w^*\|^2\,(\mu(k_0+1)-\mu)\right]$$

$$\implies \frac{1}{T-k_0}\sum_{k=k_0}^{T-1}\frac{\mathbb{E}\left[\|w_k-w^*\|^2\,(1-\mu\,\eta_k)-\|w_{k+1}-w^*\|^2\right]}{\eta_k}\leq\frac{\mu k_0}{T-k_0}\mathbb{E}\left[\|w_{k_0}-w^*\|^2\right]$$

Putting everything together,

$$\mathbb{E}[f(\bar{w}_T)-f(w^*)]\leq\frac{\mu k_0}{T-k_0}\mathbb{E}\left[\|w_{k_0}-w^*\|^2\right]+\frac{\sigma^2\,[1+\log(T)]}{\mu\,(T-k_0)}$$

Since k_0 is a constant, this already implies an $O(1/T)$ rate if we can control $\|w_{k_0}-w^*\|^2$. We will analyze Phase 1 to bound this term. Proceeding with the proof in Slide 1, using Eq. (1) for $k<k_0$.

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

Since $\eta_k=\frac{1}{L}$ for all $k<k_0$,

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

Since the above inequality is true for all $k<k_0$, using it for $k=k_0-1$,

$$\mathbb{E}[\|w_{k_0}-w^*\|^2]\leq\left(1-\frac{\mu}{L}\right)\|w_{k_0-1}-w^*\|^2+\frac{\sigma^2}{L^2}$$

Taking expectation w.r.t the randomness from iterations $k=0$ to k_0-1,

$$\mathbb{E}[\|w_{k_0}-w^*\|^2]\leq\rho\,\mathbb{E}\,\|w_{k_0-1}-w^*\|^2+\frac{\sigma^2}{L^2}\qquad\text{(Denoting } \rho:=1-{}^{\mu}/_{L}\text{)}$$

Unrolling the recursion until $k=0$,

$$\mathbb{E}[\|w_{k_0}-w^*\|^2]\leq\rho^{k_0}\,\|w_0-w^*\|^2+\frac{\sigma^2}{L^2}\sum_{k=0}^{k_0-1}\rho^k\leq\rho^{k_0}\,\|w_0-w^*\|^2+\frac{\sigma^2}{L^2}\sum_{k=0}^{\infty}\rho^k$$

$$\leq\rho^{k_0}\,\|w_0-w^*\|^2+\frac{\sigma^2}{L^2}\frac{1}{1-\rho}\qquad\text{(Infinite geometric series)}$$

$$=\left(1-\frac{\mu}{L}\right)^{k_0}\|w_0-w^*\|^2+\frac{\sigma^2}{\mu\,L}$$

$$\implies\mathbb{E}[\|w_{k_0}-w^*\|^2]\leq\exp\left(\frac{-k_0}{\kappa}\right)\|w_0-w^*\|^2+\frac{\sigma^2}{\mu\,L}\qquad(1-x\leq\exp(-x))$$

Putting everything together,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \le \frac{\mu k_0}{T - k_0}\left[\exp\left(\frac{-k_0}{\kappa}\right)\|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu L}\right] + \frac{\sigma^2\left[1 + \log(T)\right]}{\mu\,(T - k_0)}$$
$$\implies \mathbb{E}[f(\bar{w}_T) - f(w^*)] \le \frac{\mu\lceil 2\kappa - 1\rceil}{T - \lceil 2\kappa - 1\rceil}\left[\exp\left(\frac{-\lceil 2\kappa - 1\rceil}{\kappa}\right)\|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu L}\right] + \frac{\sigma^2\left[1 + \log(T)\right]}{\mu\,(T - \lceil 2\kappa - 1\rceil)}$$

Hence, we have controlled $\|w_{k_0} - w^*\|^2$ term, and this gives us an overall $O(1/T)$ rate. We can do a more careful analysis of Phase 2 to get last-iterate convergence i.e. for w_T instead of $\bar{w}_T$.

References

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209. PMLR, 2019.