CMPT 409/981: Optimization for Machine Learning

Lecture 10

Sharan Vaswani

October 8, 2024

Recap

- For minimizing $f(w) = \sum_{i=1}^{n} f_i(w)$, the SGD update is $w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$, where $i_k \in [n]$.
- SGD does not require computing the gradient of all the points in the dataset, and results in cheaper iterations compared to GD.
- Compared to GD, the rate of convergence (in terms of the number of required iterations) is slower.
- To counter the noise in the stochastic gradients, the step-size η_k needs to be decayed to ensure convergence to the minimizer.

• Two key properties we used to analyze SGD: For all w,
Unbiasedness: $\mathbb{E}_i[\nabla f_i(w)] = \nabla f(w)$; **Bounded Variance**: $\mathbb{E}_i \left\|\nabla f_i(w) - \nabla f(w)\right\|^2 \leq \sigma^2$.

• For minimizing L-smooth, but potentially non-convex functions, T iterations of SGD with $\eta_k = \frac{1}{L} \frac{1}{\sqrt{k+1}}$ result in the following suboptimality for the "best" iterate $\hat{w}$,

$$\mathbb{E}[\left\|\nabla f(\hat{w})\right\|^2] \leq \frac{2L\left[f(x_0) - f^*\right]}{\sqrt{T}} + \frac{\sigma^2\left(1 + \log(T)\right)}{\sqrt{T}}$$

Minimizing smooth, convex functions using SGD

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

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L \left\| w_0 - w^* \right\|^2}{\sqrt{T}} + \frac{\sigma^2(1 + \log(T))}{L\sqrt{T}}$$

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

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

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

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

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

(Unbiasedness)

Minimizing smooth, convex functions using SGD

Recall that $\mathbb{E}[\|w_{k+1} - w^*\|^2] = \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right]$.

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

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

$$
\begin{aligned}
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + \eta_k^2 \, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] + \eta_k^2 \sigma^2 \\
&\implies \mathbb{E}[\|w_{k+1} - w^*\|^2] \leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + 2L \eta_k^2 \, \mathbb{E}[f(w_k) - f(w^*)] + \eta_k^2 \sigma^2 \\
&\qquad\qquad (\text{Using } L\text{-smoothness of } f)
\end{aligned}
$$

Minimizing smooth, convex functions using SGD

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

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

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

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

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

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

Summing from $k = 0$ to $T - 1$,

$$\eta_{\min}\sum_{k=0}^{T-1}\mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=0}^{T-1}\mathbb{E}\left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] + \sigma^2\sum_{k=0}^{T-1}\eta_k^2$$

Minimizing smooth, convex functions using SGD

Recall $\eta_{\min} \sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=0}^{T-1} \mathbb{E}\left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] + \sigma^2 \sum_{k=0}^{T-1} \eta_k^2$.

$$\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{\mathbb{E}\left[\|w_0 - w^*\|^2 - \|w_T - w^*\|^2\right]}{\eta_{\min}} + \frac{\sigma^2}{\eta_{\min}} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)]}{T} \leq \frac{\|w_0 - w^*\|^2}{\eta_{\min}\, T} + \frac{\sigma^2}{\eta_{\min}\, T} \sum_{k=0}^{T-1} \eta_k^2 \qquad \text{(Dividing by } T\text{)}$$

Define $\bar{w}_T := \frac{\sum_{k=0}^{T-1} w_k}{T}$. Since f is convex, we can use Jensen's inequality to conclude that $\mathbb{E}[f(\bar{w}_T)] \leq \frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k)]}{T}$. Choosing $\eta_k = \frac{1}{2L} \frac{1}{\sqrt{k+1}}$,

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

Minimizing smooth, convex functions using SGD

Recall that $\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\,\|w_0 - w^*\|^2}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{T}} \sum_{k=1}^{T} \frac{1}{k}$. Since $\sum_{k=1}^{T} \frac{1}{k} \leq 1 + \log(T)$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\,\|w_0 - w^*\|^2}{\sqrt{T}} + \frac{\sigma^2(1 + \log(T))}{L\sqrt{T}}$$

- Hence, compared to GD that has an $O\left(1/T\right)$ rate of convergence, SGD has an $O\left(1/\sqrt{T}\right)$ convergence rate, but each iteration of SGD is faster.
- For GD, we proved a guarantee for the last iterate w_T; for SGD, our guarantee only holds for the average iterate $\bar{w}_T$. By using a different step-size scheme, we can get last-iterate convergence.

Lower Bound: Without additional assumptions, for smooth, convex functions, no first-order algorithm using the stochastic gradient oracle can obtain a (dimension-independent) convergence rate faster than $\Omega\left(1/\sqrt{T}\right)$.

Hence, SGD is optimal for minimizing smooth, convex functions. In the stochastic setting, using momentum or Nesterov acceleration has no provable benefit in terms of the dependence on T.

Minimizing smooth, convex functions using SGD

- Let us analyze the convergence for an alternative choice of the step-size. By following the previous proof, we have that for $\eta_k \leq \frac{1}{2L}$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\left\|w_0 - w^*\right\|^2}{\eta_{\min}\, T} + \frac{\sigma^2}{\eta_{\min}\, T} \sum_{k=1}^{T} \eta_k^2$$

- If we do not decay the step-size, and set $\eta_k = \eta = \frac{1}{2L}$, then,

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

- Hence, if we use a constant step-size for SGD, it will not converge to the minimum value but will oscillate in a *neighbourhood* around the minimum.
- Recall that if we use a mini-batch size of b, the "effective" noise is reduced to $\sigma_b^2 = \frac{n-b}{n\,b}\,\sigma^2$.
- **Common practice**: *Step-size schedules* – run SGD for some iterations (in a *stage*), decrease the step-size by a multiplicative factor and use the smaller step-size in the next stage.

Questions?

Minimizing smooth, convex functions using SGD

- If $\sigma = 0$, SGD can attain an $O(1/T)$ convergence to the minimizer using a constant step-size. If $\sigma \neq 0$, then SGD can converge to the minimizer at an $\Theta(1/\sqrt{T})$ rate using a $O(1/\sqrt{k})$ step-size.

- If σ is known, SGD with a tuned step-size can attain an $O(1/T + \sigma/\sqrt{T})$ rate i.e. convergence is slowed down only by the extent of noise [GL13, Corollary 2.2].

- Using $\eta_k = \eta \leq \frac{1}{2L}$, following the same proof,

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

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

As before, taking expectation w.r.t the randomness from iterations $i = 0$ to $k - 1$ and summing,

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

$$2\eta(1 - \eta L)\,\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{T} + \sigma^2\eta^2$$

(By dividing by T and using Jensen similar to before,)

Minimizing smooth, convex functions using SGD

Recall that $2\eta(1-\eta L)\,\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{T} + \sigma^2\eta^2$. Choosing $\eta = \min\left\{\frac{1}{2L}, \frac{1}{\sigma\sqrt{T}}\right\}$

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{T\,2\eta(1-\eta L)} + \sigma^2 \frac{\eta^2}{2\eta(1-\eta L)} \leq \frac{\|w_0 - w^*\|^2}{T\,\eta} + \sigma^2\eta$$

(For $\eta \leq \frac{1}{2L}$, $\eta \leq 2\eta - 2\eta^2 L$)

$$\leq \frac{\|w_0 - w^*\|^2}{T\,\eta} + \frac{\sigma}{\sqrt{T}} \leq \frac{\|w_0 - w^*\|^2}{T} \max\left\{2L, \sigma\sqrt{T}\right\} + \frac{\sigma}{\sqrt{T}}$$

($1/\min\{a,b\} = \max\{1/a, 1/b\}$)

$$\leq \frac{\|w_0 - w^*\|^2}{T}\left(2L + \sigma\sqrt{T}\right) + \frac{\sigma}{\sqrt{T}}$$

($\max\{a, b\} \leq a + b$ for $a, b \geq 0$)

$$\implies \mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\,\|w_0 - w^*\|^2}{T} + \sigma\left[\frac{\|w_0 - w^*\|^2 + 1}{\sqrt{T}}\right]$$

Hence, with $\eta = \min\left\{\frac{1}{2L}, \frac{1}{\sigma\sqrt{T}}\right\}$, SGD converges to the minimizer at an $O(1/T + \sigma/\sqrt{T})$ rate.

Questions?

References i

Saeed Ghadimi and Guanghui Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization **23** (2013), no. 4, 2341–2368.