CMPT 409/981: Optimization for Machine Learning

Lecture 9

Sharan Vaswani

October 17, 2022

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 slow.

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 $\hat{w}$,

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

Recap

For L-smooth, convex functions, T iterations of SGD with $\eta_k = \frac{1}{2L}\frac{1}{\sqrt{k+1}}$ result in the following suboptimality for the average iterate $\bar{w} = \frac{\sum_{k=0}^{T-1} w_k}{T}$,

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

Similar proof applies for mini-batch SGD: $w_{k+1} = w_k - \eta_k\left[\frac{1}{b}\sum_{i\in B_k}\nabla f_i(w_k)\right]$. Using a mini-batch results in the same $O(1/\sqrt{T})$ rate, but the effective noise is reduced to $\sigma_b^2 = \frac{n-b}{nb}\sigma^2$.

SGD with a constant step-size $\eta \leq \frac{1}{2L}$ results in the following convergence rate:

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

Using a smaller η slows down the convergence, but results in a smaller neighbourhood.

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.

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 a *noise-adaptive rate* of $O(1/T + \sigma/\sqrt{T})$ 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 proof from Lecture 8,

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

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

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


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.