CMPT 409/981: Optimization for Machine Learning

Lecture 11

Sharan Vaswani

October 24, 2022

Recap

Interpolation: Over-parameterized models (such as deep neural networks) are capable of exactly fitting the training dataset.

When minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, if $\|\nabla f(w)\| = 0$, then $\|\nabla f_i(w)\| = 0$ for all $i \in [n]$ i.e. the variance in the stochastic gradients becomes zero at a stationary point.

Under interpolation, since the noise is zero at the optimum, SGD does not need to decrease the step-size and can converge to the minimizer by using a *constant* step-size.

If f is strongly-convex and interpolation is satisfied (e.g. when using kernels or least squares with $d > n$), constant step-size SGD can converge to the minimizer at an $O(\exp(-T/\kappa))$ rate. Hence, SGD matches the rate of deterministic GD, but compared to GD, each iteration is cheap.

Minimizing smooth, strongly-convex functions using SGD under interpolation

Claim: When minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ such that (i) f is μ-strongly convex, (ii) each f_i is convex and L-smooth, (iii) interpolation is exactly satisfied i.e. $\|\nabla f_i(w^*)\| = 0$, T iterations of SGD with $\eta_k = \eta = \frac{1}{L}$ returns iterate w_T such that,

$$\mathbb{E}[\|w_T - w^*\|^2] \leq \exp\left(\frac{-T}{\kappa}\right) \|w_0 - w^*\|^2 .$$

Before analyzing the convergence of SGD, let us first study the effect of interpolation on $\sigma^2(w)$.

$$\begin{aligned}
\sigma^2(w) &:= \mathbb{E}_i \|\nabla f(w) - \nabla f_i(w)\|^2 = \|\nabla f(w)\|^2 + \mathbb{E}_i \|\nabla f_i(w)\|^2 - 2\mathbb{E}[\langle \nabla f(w), \nabla f_i(w)\rangle] \\
&= \mathbb{E}_i \|\nabla f_i(w)\|^2 + \|\nabla f(w)\|^2 - 2\|\nabla f(w)\|^2 && \text{(Unbiasedness)} \\
&\leq \mathbb{E}_i \|\nabla f_i(w)\|^2 \leq \mathbb{E}_i [2L[f_i(w) - f_i(w^*)]] \\
& && \text{(Using L-smoothness, convexity of f_i and $\nabla f_i(w^*) = 0$)} \\
\implies \sigma^2(w) &\leq 2L[f(w) - f(w^*)] && \text{(Unbiasedness)}
\end{aligned}$$

As w gets closer to the solution (in terms of the function values), the variance decreases becoming zero at w^*. Hence, under interpolation, we do not need to decrease the step-size.

Minimizing smooth, strongly-convex functions using SGD under interpolation

Proof: Following the same proof as before, we get that,

$$\begin{aligned}
\mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \\
&\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}_i \left[2L \left[f_{ik}(w_k) - f_{ik}(w^*)\right]\right] \\
&\qquad \text{(Using } L\text{-smoothness, convexity of } f_i \text{ and } \nabla f_i(w^*) = 0) \\
&= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 2L \, \eta_k^2 \, \mathbb{E}\left[f(w_k) - f(w^*)\right] \\
&\qquad \text{(Unbiasedness)} \\
&= \|w_k - w^*\|^2 (1 - \mu \eta_k) - 2\eta_k \left[f(w_k) - f(w^*)\right] + 2L \, \eta_k^2 \, \mathbb{E}\left[f(w_k) - f(w^*)\right] \\
&\qquad \text{(Strong-convexity)} \\
&= \left(1 - \frac{\mu}{L}\right) \|w_k - w^*\|^2 \qquad \text{(Since } \eta_k = \eta = \tfrac{1}{L})
\end{aligned}$$

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

$$\mathbb{E}[\|w_T - w^*\|^2] \leq \left(1 - \frac{\mu}{L}\right)^T \|w_0 - w^*\|^2 \leq \exp\left(\frac{-T}{\kappa}\right) \|w_0 - w^*\|^2$$

Minimizing smooth, strongly-convex functions using SGD under interpolation

We can modify the proof in order to get an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ where $\zeta^2 \propto \mathbb{E}_i \left\|\nabla f_i(w^*)\right\|^2$.

Moreover, as before, if we use a mini-batch of size b, the effective noise is $\zeta_b^2 \propto \frac{\mathbb{E}_i \|\nabla f_i(w^*)\|^2}{b}$. Hence, if the model is sufficiently over-parameterized so that it *almost* interpolates the data, and we are using a large batch-size, then ζ_b^2 is small, and constant step-size works well.

When minimizing convex functions under (exact) interpolation, constant step-size SGD results in $O(1/T)$ convergence, matching deterministic GD, but with much smaller per-iteration cost (Need to prove this in Assignment 3!)

Questions?

Minimizing smooth, non-convex functions using SGD under interpolation

When minimizing non-convex functions, interpolation is not enough to guarantee a fast (matching the deterministic) $O(1/T)$ rate for SGD.

Can achieve this rate under the *strong growth condition* (SGC) on the stochastic gradients. Formally, there exists a constant $\rho > 1$ such that for all w,

$$\mathbb{E}_i \left\| \nabla f_i(w) \right\|^2 \leq \rho \left\| \nabla f(w) \right\|^2$$

Hence, SGC implies that $\left\| \nabla f_i(w^*) \right\|^2 = 0$ for all i and hence interpolation.

As before, let us study the effect of SGC on the variance $\sigma^2(w)$.

$$\sigma^2(w) := \mathbb{E}_i \left\| \nabla f_i(w) - \nabla f(w) \right\|^2 = \mathbb{E}_i \left\| \nabla f_i(w) \right\|^2 - \left\| \nabla f(w) \right\|^2 \qquad \text{(Unbiasedness)}$$

$$\implies \sigma^2(w) \leq (\rho - 1) \left\| \nabla f(w) \right\|^2 \qquad \text{(SGC)}$$

Hence, SGC implies that as w gets closer to a stationary point (in terms of the gradient norm), the variance decreases and constant step-size SGD converges to a stationary point.

Minimizing smooth, non-convex functions using SGD under interpolation

Claim: For (i) L-smooth functions lower-bounded by f^*, (ii) under ρ-SGC, T iterations of SGD with $\eta_k = \frac{1}{\rho L}$ returns an iterate $\hat{w}$ such that,

$$\mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2\rho L\,[f(w_0) - f^*]}{T}$$

Proof: Similar to the proof in Lecture 8, using the L-smoothness of f with $x = w_k$ and $y = w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$,

$$f(w_{k+1}) \leq f(w_k) + \langle \nabla f(w_k), -\eta_k \nabla f_{ik}(w_k) \rangle + \frac{L}{2}\,\eta_k^2\,\|\nabla f_{ik}(w_k)\|^2$$

Taking expectation w.r.t i_k on both sides and using that η_k is independent of i_k

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

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \eta_k\,\|\nabla f(w_k)\|^2 + \frac{L\eta_k^2}{2}\,\mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \qquad \text{(Unbiasedness)}$$

Minimizing smooth, non-convex functions using SGD under interpolation

Recall $\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \eta_k \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta_k^2}{2} \mathbb{E}\left[\left\|\nabla f_{ik}(w_k)\right\|^2\right]$. Using ρ-SGC,

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \eta_k \left\|\nabla f(w_k)\right\|^2 + \frac{L\rho\eta_k^2}{2} \left\|\nabla f(w_k)\right\|^2$$

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \frac{1}{2\rho L} \left\|\nabla f(w_k)\right\|^2 \qquad \text{(Using } \eta_k = \eta = \tfrac{1}{\rho L}\text{)}$$

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

$$\sum_{k=0}^{T-1} \mathbb{E}[\left\|\nabla f(w_k)\right\|^2] \leq 2\rho L \sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w_{k+1})] \implies \frac{\sum_{k=0}^{T-1} \mathbb{E}[\left\|\nabla f(w_k)\right\|^2]}{T} \leq \frac{2\rho L\, \mathbb{E}[f(w_0) - f^*]}{T}$$

(Dividing by T)

Defining $\hat{w} := \arg\min_{k \in \{0,1,\dots,T-1\}} \mathbb{E}[\left\|\nabla f(w_k)\right\|^2]$,

$$\mathbb{E}[\left\|\nabla f(\hat{w})\right\|^2] \leq \frac{2\rho L\,[f(w_0) - f^*]}{T}$$

Questions?