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 $\sigma^2(w)$.

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

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,

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

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 $\sigma^2(w)$.

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

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

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,

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

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

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

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

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

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

(Dividing by T)

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

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

Questions?

Stochastic Line-Search

Algorithmically, convergence at a fast rate under interpolation requires a constant step-size that depends on L. We will use a *stochastic line-search* (SLS) procedure to estimate L. SLS is similar to the deterministic setting in Lecture 3, but uses only stochastic function/gradient evaluations.

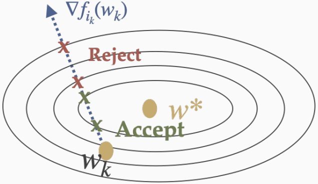
Algorithm SGD with Stochastic Line-search

```
function SGD with Stochastic Line-search (f, w_0, η_max, c ∈ (0,1), β ∈ (0,1))
for k = 0, ..., T − 1 do
    η̃_k ← η_max
    while f_ik(w_k − η̃_k ∇f_ik(w_k)) > f_ik(w_k) − c · η̃_k ||∇f_ik(w_k)||^2 do
        η̃_k ← η̃_k β
    end while
    η_k ← η̃_k
    w_{k+1} = w_k − η_k ∇f_ik(w_k)
end for
return w_T
```

Stochastic Line-Search

SLS searches for a good step-size in the wrong direction.

Since all the functions share the same minimizer (because of interpolation), SGD with SLS converges to the minimizer.



Claim: The (exact) backtracking procedure for SLS terminates and returns $\eta_k \in \left[\min\left\{\frac{2\,(1-c)}{L}, \eta_{\max}\right\}, \eta_{\max}\right]$.

Proof: Similar to the deterministic case (Lecture 3), but requires that each f_{ik} is L-smooth.

Minimizing smooth, strongly-convex functions using SGD + SLS 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 SLS (with $c = 1/2$) returns iterate w_T such that,

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

Proof: Similar to the previous proof, we get that,

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

Since η_k depends on i_k, we can not push the expectation in. Since η_k is set by SLS, it satisfies the stochastic Armijo condition. Simplifying the third term and denoting $f_{ik}^* := \min f_{ik}(w)$,

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

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Using Eq. (1) + Eq. (2),

$$\mathbb{E}[\|w_{k+1} - w^*\|^2] = \|w_k - w^*\|^2 - 2\mathbb{E}\left[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle\right] + \mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \quad (3)$$

$$\mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] = \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*) + f_{ik}(w^*) - f_{ik}^*\right)\right] \qquad (\text{Setting } c = 1/2)$$

$$= \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] + \mathbb{E}\left[2\eta_k \underbrace{\left(f_{ik}(w^*) - f_{ik}^*\right)}_{\text{Positive}}\right]$$

$$\leq \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] + 2\eta_{\max}\mathbb{E}\left[f_{ik}(w^*) - f_{ik}^*\right] \quad (\text{Since } \eta_k \leq \eta_{\max})$$

Since f_{ik} is convex and $\nabla f_{ik}(w^*) = 0$, $f_{ik}(w^*) = f_{ik}^*$.

$$\mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \leq \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] \quad (4)$$

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Using Eq. (3) + Eq. (4),

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

Since f_{ik} is convex, $f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle \leq 0$

$$\begin{aligned}&\leq \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle\right] \\ &\qquad \left(\text{Lower-bounding } \eta_k.\ \eta_{\min} := \min\left\{\tfrac{1}{L}, \eta_{\max}\right\}\right) \\ &= \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f(w_k) - f(w^*) + \langle \nabla f(w_k), w^* - w_k \rangle\right] \\ &\qquad (\text{Unbiasedness}) \\ &\leq \|w_k - w^*\|^2 + 2\eta_{\min} \left[\frac{-\mu}{2} \|w_k - w^*\|^2\right] \qquad (f \text{ is } \mu\text{-strongly convex})\end{aligned}$$

$$\implies \mathbb{E}[\|w_{k+1} - w^*\|^2] \leq (1 - \mu\, \eta_{\min})\, \|w_k - w^*\|^2$$

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Recall that $\mathbb{E}[\|w_{k+1} - w^*\|^2] \leq (1 - \mu\,\eta_{\min})\, \|w_k - w^*\|^2$. Taking expectation w.r.t the randomness from iterations $k = 0$ to $T - 1$ and recursing,

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

Hence, when minimizing smooth, strongly-convex functions under interpolation, SGD + SLS will will converge to the minimizer at an exponential rate.

Similarly, When minimizing convex functions under (exact) interpolation, SGD + SLS results in an $O(1/T)$ rate without requiring knowledge of L. (Need to prove this in Assignment 3!)

We can modify the proof in order to get an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ where $\zeta^2 := \mathbb{E}\left[f_{ik}(w^*) - f_{ik}^*\right]$.

Questions?