CMPT 409/981: Optimization for Machine Learning

Lecture 12

Sharan Vaswani

October 17, 2024

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.

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,

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

Minimizing smooth, strongly-convex functions using SGD under interpolation

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

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

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

(Using L-smoothness, convexity of f_i and $\nabla f_i(w^*) = 0$)

$$= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + 2L\,\eta_k^2 \, \mathbb{E}\left[f(w_k) - f(w^*)\right]$$

(Unbiasedness)

$$= \|w_k - w^*\|^2 (1 - \mu\eta_k) - 2\eta_k \left[f(w_k) - f(w^*)\right] + 2L\,\eta_k^2 \, \mathbb{E}\left[f(w_k) - f(w^*)\right]$$

(Strong-convexity)

$$= \left(1 - \frac{\mu}{L}\right) \|w_k - w^*\|^2$$

(Since $\eta_k = \eta = \frac{1}{L}$)

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 $\sigma^2(w)$.

$$\begin{aligned} & \sigma^2(w) := \mathbb{E}_i \left\|\nabla f_i(w) - \nabla f(w)\right\|^2 = \mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 - \left\|\nabla f(w)\right\|^2 && \text{(Unbiasedness)} \\ \implies & \sigma^2(w) \leq (\rho - 1) \left\|\nabla f(w)\right\|^2 && \text{(SGC)} \end{aligned}$$

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 $\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \eta_k \|\nabla f(w_k)\|^2 + \frac{L\eta_k^2}{2}\mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right]$. 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 under interpolation requires knowledge of L. We will use a *stochastic line-search* (SLS) procedure [VML$^+$19] to estimate L. SLS is similar to the deterministic variant in Lecture 3, but uses only stochastic function/gradient evaluations.

Algorithm SGD with Stochastic Line-search

---

function SGD with Stochastic Line-search $(f,\ w_0,\ \eta_{\max},\ c \in (0,1),\ \beta \in (0,1))$

for $k = 0, \dots, T-1$ **do**

    $\tilde{\eta}_k \leftarrow \eta_{\max}$

    while $f_{ik}(w_k - \tilde{\eta}_k \nabla f_{ik}(w_k)) > f_{ik}(w_k) - c \cdot \tilde{\eta}_k \left\|\nabla f_{ik}(w_k)\right\|^2$ **do**

        $\tilde{\eta}_k \leftarrow \tilde{\eta}_k \beta$

    end while

    $\eta_k \leftarrow \tilde{\eta}_k$

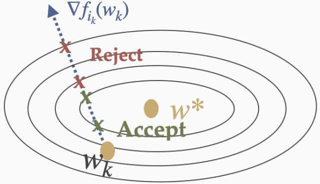
    $w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$

end for

return w_T

Stochastic Line-Search

- SLS searches for a good step-size in the wrong direction.
- Since all f_i have zero gradient at w^* and the noise decreases as we get closer to the solution (because of interpolation), SGD with SLS converges to the minimizer.



Claim: If each f_i is L-smooth, then 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_i 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. 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),

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

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

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

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

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),

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

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

$$\leq \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle\right]$$

(Lower-bounding η_k. $\eta_{\min} := \min\left\{\frac{1}{L}, \eta_{\max}\right\}$)

$$= \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f(w_k) - f(w^*) + \langle \nabla f(w_k), w^* - w_k \rangle\right]$$

(Unbiasedness)

$$\leq \|w_k - w^*\|^2 + 2\eta_{\min} \left[\frac{-\mu}{2} \|w_k - w^*\|^2\right]$$

(f is μ-strongly convex)

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

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,

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

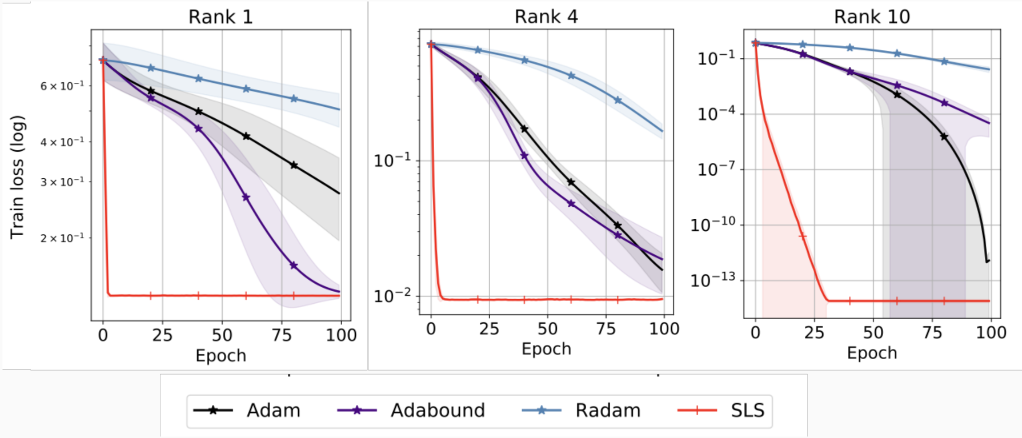
$$\implies \mathbb{E}[\|w_T - w^*\|^2] \leq \exp\left(-\mu\, T\, \min\left\{\frac{1}{L}, \eta_{\max}\right\}\right) \|w_0 - w^*\|^2$$

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

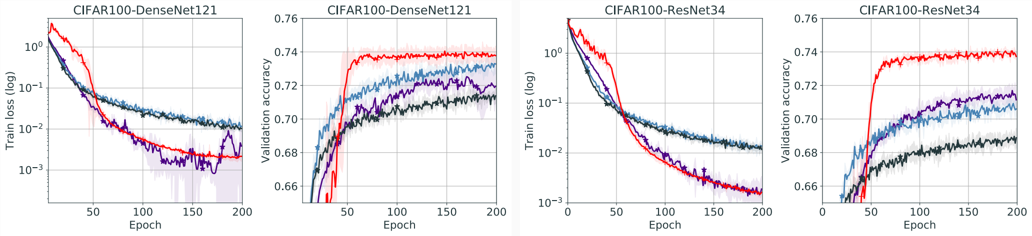
- If interpolation is not exactly satisfied, we can modify the proof to get an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate, where $\zeta^2 := \mathbb{E}\left[f_{ik}(w^*) - f_{ik}^*\right]$.
- When minimizing convex functions under (exact) interpolation, SGD + SLS results in an $O(1/T)$ rate without requiring knowledge of L. (Need to prove this in Assignment 3!)
- Do not have strong theoretical results for SGD + SLS on smooth, non-convex problems.

Stochastic Line-Search and Effect of Over-parametrization

Objective: $\min_{\theta_1,\theta_2} \frac{1}{2n} \sum_{i=1}^{n} \|\theta_2\, \theta_1 x_i - y_i\|^2$; **Parameterization**: $\theta_1 \in \mathbb{R}^{k\times 6}$, $\theta_2 \in \mathbb{R}^{10\times k}$.



Task: Multi-class classification with logistic loss.



Stochastic Polyak Step-size

- When interpolation is (approximately) satisfied, we can use SGD with the *stochastic Polyak step-size* (SPS) [LVLLJ21]: At iteration k, for hyper-parameter $c \in (0, 1)$ and $f_{ik}^* := \min_w f_{ik}(w)$,

$$\eta_k = \frac{f_{ik}(w_k) - f_{ik}^*}{c \left\|\nabla f_{ik}(w_k)\right\|^2} .$$

Common machine learning losses (squared loss, logistic loss, exponential loss) are lower-bounded by zero. Algorithmically, we can set $f_{ik}^* = 0$.

- SPS matches the SLS rates on smooth, (strongly) convex functions. E.g. SPS with $c = 1/2$ achieves the $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate for smooth, strongly-convex functions.
- Much simpler and computationally inexpensive to implement compared to SLS.
- Unlike SLS, SPS can be used for minimizing non-smooth, convex functions.
- Results in large step-sizes and requires some additional heuristics for stabilizing the method.
- For neural networks, generalization for SGD + SPS was typically worse than for SGD + SLS.
- Requires access to f_{ik}^* which might be difficult to compute for more general problems.

Adaptivity for SGD

Noise-adaptivity: When minimizing smooth, strongly-convex functions, with T iterations of SGD with $\eta_k := \frac{1}{L}\left(\frac{1}{T}\right)^{\frac{k}{T}}$, we can obtain an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\zeta^2}{T}\right)$ rate, where $\zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*]$. Adaptive to the extent of interpolation, but requires L to set the step-size.

Problem-adaptivity: SGD with the step-size set according to SLS/SPS is adaptive to L, but results in an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate.

- [VDTB21] attempts to combine the above ideas to obtain both noise and problem adaptivity i.e. use SLS to set $\gamma_k \approx \frac{1}{L}$ and use $\eta_k = \gamma_k \left(\frac{1}{T}\right)^{\frac{k}{T}}$. Either not guaranteed to converge to the minimizer or will converge to the minimizer at a slower (than optimal) rate.

- For smooth, strongly-convex problems, we do not (yet) know how to make SGD problem and noise-adaptive, and achieve the optimal rate.

- For smooth, convex problems, AdaGrad is both problem and noise-adaptive.

Questions?

References i

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien, *Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1306–1314.

Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad, *Towards noise-adaptive, problem-adaptive stochastic gradient descent*, arXiv preprint arXiv:2110.11442 (2021).

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien, *Painless stochastic gradient: Interpolation, line-search, and convergence rates*, Advances in neural information processing systems **32** (2019).