CMPT 409/981: Optimization for Machine Learning

Lecture 11

Sharan Vaswani

October 10, 2024

Recap

Function class	L-smooth	L-smooth + convex	L-smooth + μ-strongly convex
Gradient Descent	$O\left(1/\epsilon\right)$	$O\left(1/\epsilon\right)$	$O\left(\kappa \log\left(1/\epsilon\right)\right)$
Stochastic Gradient Descent	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$

Table 1: Comparing the convergence rates of GD and SGD

Minimizing smooth, strongly-convex functions using SGD

- Let us prove that SGD with an $O(1/k)$ decaying step-size results in an $O(1/T)$ convergence to the minimizer.

- Following [LJSB12], let us first do the proof with an additional (strong) assumption that the stochastic gradients are bounded in expectation, i.e. there exists a G such that $\mathbb{E} \left\|\nabla f_i(w)\right\|^2 \leq G^2$ for all w.

- **Claim**: For μ-strongly convex functions with the above assumption, T iterations of SGD with $\eta_k = \frac{1}{\mu\,(k+1)}$ returns iterate $\bar{w}_T = \frac{\sum_{k=0}^{T-1} w_k}{T}$ such that,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{G^2\,[1 + \log(T)]}{2\mu\,T}$$

- Three problems with the above result: (i) setting the step-size requires knowledge of μ, (ii) requires bounded stochastic gradients (not necessarily true for quadratics), (iii) the guarantee only holds for the average and not the last iterate.

Minimizing smooth, strongly-convex functions using SGD

Proof: Following a proof similar to the convex case,

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

Taking expectation w.r.t i_k on both sides,

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

(Assuming η_k is independent of i_k and Unbiasedness)

Using μ-strong convexity, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2$ with $y = w^*$ and $x = w_k$,

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

Minimizing smooth, strongly-convex functions using SGD

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

Using the boundedness of stochastic gradients, $\mathbb{E}\,\|\nabla f_i(w)\|^2 \leq G^2$ for all w,

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

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

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

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

Minimizing smooth, strongly-convex functions using SGD

Recall that $\mathbb{E}[f(w_k) - f(w^*)] \leq \frac{\mathbb{E}\left[\left\|w_k - w^*\right\|^2 (1 - \mu\,\eta_k) - \left\|w_{k+1} - w^*\right\|^2\right]}{2\eta_k} + \frac{\eta_k}{2}\,G^2$.

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

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

$$= \sum_{k=0}^{T-1} \frac{\mathbb{E}\left[\left\|w_k - w^*\right\|^2 (1 - \mu\,\eta_k) - \left\|w_{k+1} - w^*\right\|^2\right]}{2\eta_k} + \frac{G^2}{2} \sum_{k=0}^{T-1} \frac{1}{\mu\,(k+1)}$$

$$\leq \sum_{k=0}^{T-1} \frac{\mathbb{E}\left[\left\|w_k - w^*\right\|^2 (1 - \mu\,\eta_k) - \left\|w_{k+1} - w^*\right\|^2\right]}{2\eta_k} + \frac{G^2\,[1 + \log(T)]}{2\mu}$$

Dividing by T, using Jensen's inequality for the LHS, and by definition of $\bar{w}_T$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{1}{T} \sum_{k=0}^{T-1} \frac{\mathbb{E}\left[\left\|w_k - w^*\right\|^2 (1 - \mu\,\eta_k) - \left\|w_{k+1} - w^*\right\|^2\right]}{2\eta_k} + \frac{G^2\,[1 + \log(T)]}{2\mu\,T}$$

Minimizing smooth, strongly-convex functions using SGD

Recall that $\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{1}{T} \sum_{k=0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2 (1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{2\eta_k} + \frac{G^2\,[1 + \log(T)]}{2\mu\,T}$.
Simplifying the first term on the RHS,

$$\frac{1}{2T} \sum_{k=0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2\,(1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k}$$

$$= \frac{1}{2T}\,\mathbb{E}\left[\sum_{k=1}^{T-1}\left[\|w_k - w^*\|^2 \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} - \mu\right)\right] + \|w_0 - w^*\|^2 \left(\frac{1}{\eta_0} - \mu\right) - \frac{\|w_T - w^*\|^2}{\eta_{T-1}}\right]$$

$$\leq \frac{1}{2T}\,\mathbb{E}\left[\sum_{k=1}^{T-1}\left[\|w_k - w^*\|^2\,(\mu(k+1) - \mu k - \mu)\right] + \|w_0 - w^*\|^2\,(\mu - \mu)\right] = 0$$

Putting everything together,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{G^2\,[1 + \log(T)]}{2\mu\,T}$$

Questions?

Minimizing smooth, strongly-convex functions using SGD

- Next, we will adapt the proof from [GLQ+19] that does not require bounded stochastic gradients. It uses a constant followed by a $O(1/k)$ decaying step-size, and converges to the minimizer at an $O(1/T)$ rate.

Claim: For L-smooth, μ-strongly convex functions, T iterations of SGD with

$$\eta_k = \frac{1}{L} \quad (\text{For } k < k_0) \quad [\text{Phase 1}] \quad ; \quad \eta_k = \frac{1}{\mu\,(k+1)} \quad (\text{For } k \geq k_0) \quad [\text{Phase 2}]$$

for $k_0 := \lceil 2\kappa - 1 \rceil$ returns iterate $\bar{w}_T := \frac{\sum_{k=k_0}^{T-1} w_k}{T-k_0}$ such that for $T > k_0$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\mu\, k_0}{T - k_0} \left[\exp\left(\frac{-k_0}{\kappa}\right) \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu\, L} \right] + \frac{\sigma^2\, [1 + \log(T)]}{\mu\, (T - k_0)} \,.$$

- Three problems with the above result: (i) setting the step-size requires knowledge of μ, (ii) guarantee only holds for $T > k_0$ (iii) guarantee holds only for the average iterate and not the last iterate.

Minimizing smooth, strongly-convex functions using SGD

Proof: Following the same sequence of steps as before, we obtain the following inequality:

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

Using L-smoothness,

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

Phase 2: We require that $\eta_k \leq \frac{1}{2L}$ in Phase 2, i.e. for all $k \geq k_0$,

$$\implies \frac{1}{\mu\,(k+1)} \leq \frac{1}{2L} \implies k \geq 2\kappa - 1\,.$$

Since Phase 2 only starts when $k \geq k_0 = \lceil 2\kappa - 1 \rceil$, this ensures the desired condition.

Minimizing smooth, strongly-convex functions using SGD

Phase 2: Since $\eta_k \leq \frac{1}{2L}$ in Phase 2, using Eq (1) for all $k \geq k_0$ and following the previous proof,

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

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

Taking expectation w.r.t the randomness from iterations $k = k_0$ to $T - 1$,

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

Summing from $k = k_0$ to $T - 1$ in Phase 2,

$$\sum_{k=k_0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2 (1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \sigma^2 \sum_{k=k_0}^{T-1} \eta_k$$

Minimizing smooth, strongly-convex functions using SGD

$$\sum_{k=k_0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2 (1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \sum_{k=0}^{T-1} \frac{\sigma^2}{\mu\,(k+1)}$$

$$\leq \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2 (1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \frac{\sigma^2\,[1 + \log(T)]}{\mu}$$

Dividing by $T - k_0$, using Jensen's inequality for the LHS, and by definition of $\bar{w}_T$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{1}{T - k_0} \sum_{k=k_0}^{T-1} \frac{\mathbb{E}\left[\|w_k - w^*\|^2 (1 - \mu\,\eta_k) - \|w_{k+1} - w^*\|^2\right]}{\eta_k} + \frac{\sigma^2\,[1 + \log(T)]}{\mu\,(T - k_0)}$$

Following the same proof as before, we can conclude that,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\mu k_0}{T - k_0} \mathbb{E}\left[\|w_{k_0} - w^*\|^2\right] + \frac{\sigma^2\,[1 + \log(T)]}{\mu\,(T - k_0)}.$$

Minimizing smooth, strongly-convex functions using SGD

Since k_0 is a constant, the previous slide already implies an $O(1/T)$ rate if we can control $\|w_{k_0} - w^*\|^2$ in Phase 1.

Phase 1: Using Eq(1) for $k < k_0$, for which $\eta_k = \frac{1}{L}$,

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

Since the above inequality is true for all $k < k_0$, using it for $k = k_0 - 1$ and taking expectation w.r.t the randomness from iterations $k = 0$ to $k_0 - 1$,

$$\mathbb{E}[\|w_{k_0} - w^*\|^2] \leq \rho\, \mathbb{E}\, \|w_{k_0-1} - w^*\|^2 + \frac{\sigma^2}{L^2} \qquad \text{(Denoting } \rho := 1 - \mu/L\text{)}$$

$$\implies \mathbb{E}[\|w_{k_0} - w^*\|^2] \leq \rho^{k_0} \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \sum_{k=0}^{k_0-1} \rho^k \leq \rho^{k_0} \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \sum_{k=0}^{\infty} \rho^k$$

$$\leq \rho^{k_0} \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \frac{1}{1-\rho} = \left(1 - \frac{\mu}{L}\right)^{k_0} \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu\, L}$$

Minimizing smooth, strongly-convex functions using SGD

Using the result from the previous slide,

$$\mathbb{E}[\|w_{k_0} - w^*\|^2] \leq \exp\left(\frac{-k_0}{\kappa}\right) \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu L} \qquad (1 - x \leq \exp(-x))$$

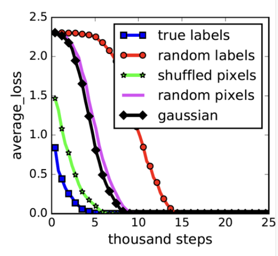
Hence, we have controlled $\|w_{k_0} - w^*\|^2$ term. Putting everything together,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\mu k_0}{T - k_0}\left[\exp\left(\frac{-k_0}{\kappa}\right) \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu L}\right] + \frac{\sigma^2 [1 + \log(T)]}{\mu (T - k_0)}$$

- By choosing a different step-size that depends on both σ^2 and μ, it is possible to prove last-iterate convergence (for $T > k_0$) for SGD [GLQ+19] The resulting rate of convergence is $O(\kappa \ln(1/\epsilon) + \sigma^2/\epsilon)$.
- [LZO21, VDTB21] use an $\eta_k = \frac{1}{2L} \left((1/T)^{k/T}\right)$ step-size, obtain a last-iterate noise-adaptive convergence rate of $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$. However, it requires knowledge of T (in practice, we can use the doubling trick).
- The resulting step-size works well in practice, and can also be combined with Nesterov acceleration to achieve an $O\left(\exp\left(\frac{-T}{\sqrt{\kappa}}\right) + \frac{\sigma^2}{T}\right)$ rate.

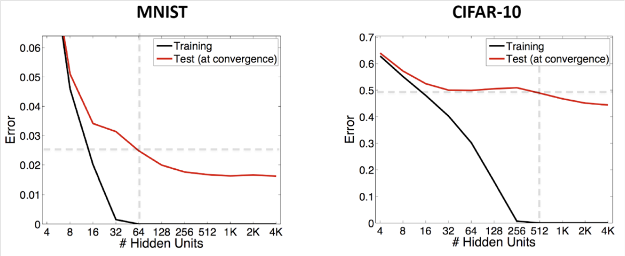
Interpolation for over-parameterized models

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



Zhang et al, "Understanding deep learning requires rethinking generalization", 2016.

Loss vs Training steps on CIFAR-10 dataset



https://www.neyshabur.net/papers/inductive_bias_poster.pdf

Error vs Network size

Formally, when minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, interpolation means that 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.

SGD under Interpolation

- Recall that SGD needs to decrease the step-size to counteract the noise (variance).

Idea: 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 the model is expressive enough such that interpolation is satisfied (for example, 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.

- In this setting, SGD matches the rate of deterministic (full-batch) GD, but compared to GD, each iteration is cheap!

- Moreover, empirical results (and theoretical results on "benign overfitting") suggest that interpolating the training dataset does not adversely affect the generalization error!

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.

- Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik, *Sgd: General analysis and improved rates*, International Conference on Machine Learning, PMLR, 2019, pp. 5200–5209.
- Simon Lacoste-Julien, Mark Schmidt, and Francis Bach, *A simpler approach to obtaining an o (1/t) convergence rate for the projected stochastic subgradient method*, arXiv preprint arXiv:1212.2002 (2012).
- Xiaoyu Li, Zhenxun Zhuang, and Francesco Orabona, *A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance*, International Conference on Machine Learning, PMLR, 2021, pp. 6553–6564.
- Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad, *Towards noise-adaptive, problem-adaptive stochastic gradient descent*, arXiv preprint arXiv:2110.11442 (2021).