CMPT 409/981: Optimization for Machine Learning

Lecture 10

Sharan Vaswani

October 20, 2022

Minimizing smooth, strongly-convex functions using SGD

For smooth, strongly-convex functions, SGD with an $O(1/k)$ decreasing step-size converges to the minimizer at an $\Theta(1/T)$ rate (we will prove this later today).

Similar to the convex setting, using SGD with a constant step-size results in convergence to the neighbourhood that depends on the noise in the stochastic gradients.

Claim: For L-smooth, μ-strongly convex functions, T iterations of SGD with $\eta_k = \eta = \frac{1}{L}$ returns iterate w_T such that,

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

Hence, SGD results in an exponential convergence to the neighbourhood of the minimizer.

Unlike the convex case for which we proved a guarantee on the average iterate $\bar{w}_T$, here we have a guarantee for the last iterate w_T.

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,

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

(Assuming η_k is independent of i_k and Unbiasedness)

Minimizing smooth, strongly-convex functions using SGD

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

$$
\begin{aligned}
&\mathbb{E}[\|w_{k+1} - w^*\|^2] \\
&= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f(w_k) + \nabla f(w_k)\|^2\right] \\
&\leq \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] + \eta_k^2 \, \sigma^2 \\
&\hspace{20em} \text{(Using the bounded variance assumption)}
\end{aligned}
$$

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 $x = w_k$,

$$
\begin{aligned}
&\leq \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] - \mu\eta_k \|w_k - w^*\|^2 + \eta_k^2 \, \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] + \eta_k^2 \, \sigma^2 \\
&\hspace{20em} \text{(Eq. (1))}
\end{aligned}
$$

$$
\begin{aligned}
\implies &\mathbb{E}[\|w_{k+1} - w^*\|^2] \\
&\leq (1 - \mu\eta_k) \|w_k - w^*\|^2 - 2\eta_k [f(w_k) - f(w^*)] + 2L \, \eta_k^2 \, \mathbb{E}[f(w_k) - f(w^*)] + \eta_k^2 \, \sigma^2 \\
&\hspace{20em} \text{(Using } L\text{-smoothness of } f\text{)}
\end{aligned}
$$

Minimizing smooth, strongly-convex functions using SGD

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

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

Since the above inequality is true for all k, using it for $k = T - 1$,

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

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

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

Minimizing smooth, strongly-convex functions using SGD

Recall that $\mathbb{E}[\|w_T - w^*\|^2] \leq \rho\, \mathbb{E}\, \|w_{T-1} - w^*\|^2 + \frac{\sigma^2}{L^2}$. Unrolling the recursion until $k = 0$,

$$\begin{aligned}
\mathbb{E}[\|w_T - w^*\|^2] &\leq \rho^T \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \sum_{k=0}^{T-1} \rho^k \leq \rho^T \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \sum_{k=0}^{\infty} \rho^k \\
&\leq \rho^T \|w_0 - w^*\|^2 + \frac{\sigma^2}{L^2} \frac{1}{1-\rho} && \text{(Infinite geometric series)} \\
&= \left(1 - \frac{\mu}{L}\right)^T \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu\, L} \\
&\leq \exp\left(\frac{-T}{\kappa}\right) \|w_0 - w^*\|^2 + \frac{\sigma^2}{\mu\, L} && (1 - x \leq \exp(-x)) \\
\implies \mathbb{E}[\|w_T - w^*\|^2] &\leq \underbrace{\exp\left(\frac{-T}{\kappa}\right) \|w_0 - w^*\|^2}_{\text{bias}} + \underbrace{\frac{\sigma^2}{\mu\, L}}_{\text{neighbourhood}}
\end{aligned}$$

Questions?

Minimizing smooth, strongly-convex functions using SGD

Let us prove that SGD with an $O(1/k)$ step-size results in $O(1/T)$ convergence to the minimizer. Similar to [LJSB12], for simplicity, let us assume that the stochastic gradients are bounded in expectation, i.e. there exists a G such that $\mathbb{E} \|\nabla f_i(w)\|^2 \leq G^2$ 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,

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

Three problems – the above result (i) requires knowledge of μ, (ii) requires bounded stochastic gradients, (iii) the guarantee only holds for the average iterate and not the last iterate.

[GLQ$^+$19, Theorem 3.2] uses a constant, then $O(1/k)$ step-size. Solves (ii), (iii)

[LZO21, VDTB21] use an $O\left((1/T)^{k/T}\right)$ step-size and solves all three problems. Also prove a noise-adaptive $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\sigma^2}{T}\right)$ rate, but requires knowledge of T.

Minimizing smooth, strongly-convex functions using SGD

Proof: Following the previous proof,

$$
\begin{aligned}
&\mathbb{E}\left\|w_{k+1} - w^*\right\|^2 \\
&\leq \left\|w_k - w^*\right\|^2 - 2\eta_k[f(w_k) - f(w^*)] - \mu\eta_k \left\|w_k - w^*\right\|^2 + \eta_k^2\,\mathbb{E}\left[\left\|\nabla f_{ik}(w_k)\right\|^2\right] \\
&\leq \left\|w_k - w^*\right\|^2 - 2\eta_k[f(w_k) - f(w^*)] - \mu\eta_k \left\|w_k - w^*\right\|^2 + \eta_k^2\, G^2 \\
&\hspace{8cm}\text{(Using the boundedness of stochastic gradients)}
\end{aligned}
$$

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

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

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

Minimizing smooth, strongly-convex functions using SGD

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

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

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

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

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

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

Minimizing smooth, strongly-convex functions using SGD

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

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

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

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

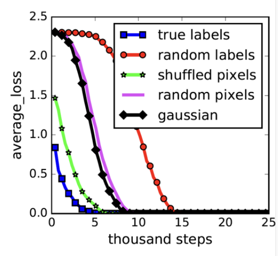
Putting everything together,

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

Questions?

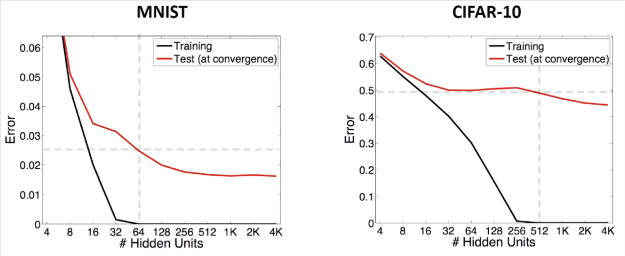
Interpolation for over-parameterized models

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!

Questions?

References

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).