CMPT 409/981: Optimization for Machine Learning

Lecture 8

Sharan Vaswani

October 13, 2022

Optimization Zoo

Function class	L-smooth	L-smooth + convex	L-smooth + μ-strongly convex
Gradient Descent	$\Theta\left({}^{1}/_{\epsilon}\right)$	$O\left({}^{1}/_{\epsilon}\right)$	$O\left(\kappa \log\left({}^{1}/_{\epsilon}\right)\right)$
Nesterov Acceleration	-	$\Theta\left({}^{1}/_{\sqrt{\epsilon}}\right)$	$\Theta\left(\sqrt{\kappa} \log\left({}^{1}/_{\epsilon}\right)\right)$

Table 1: Using the first-order oracle that returns $\nabla f(w)$

Today, we will use a stochastic first-order oracle that is less expensive, but returns a noisy estimate of the gradient.

Stochastic Gradient Descent

In machine learning, we typically care about minimizing the average of *loss functions*,

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \,.$$

i.e. our model should perform well on average across examples.

Example: In supervised learning using a dataset of n input-output pairs $\{X_i, y_i\}_{i=1}^{n}$, for linear regression, $f(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left(\langle X_i, w\rangle - y_i\right)^2$. Similarly, for logistic regression for binary classification where $y_i \in \{-1, +1\}$, $f(w) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle X_i, w\rangle\right)\right)$.

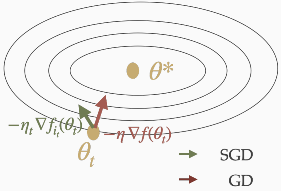
Gradient-based methods on such functions require computing $\nabla f(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w)$ which is an $O(n)$ operation. Typically, n is large in practice and hence computing the gradient across the whole datasets is expensive.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) only requires computing the gradient of one loss function in each iteration. At iteration k, SGD samples loss function i_k (uniformly) randomly:

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

Unlike GD, each iteration of SGD is cheap and does not depend on n.



θ^*
$-\eta_t \nabla f_{i_t}(\theta_t)$
$-\eta \nabla f(\theta_t)$
θ_t
SGD
GD

Unbiasedness: Since i_k is picked uniformly at random, $\nabla f_{ik}(w)$ is unbiased,

$$\mathbb{E}[\nabla f_{ik}(w)] = \sum_{i=1}^{n} \frac{1}{n} \nabla f_i(w) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)\,.$$

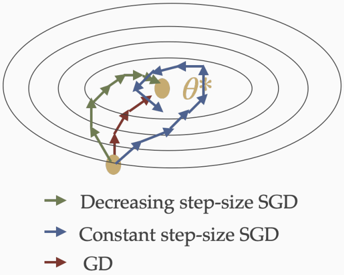
We will assume that $f(w)$ is a finite-sum of n points only for convenience. In general, all the results hold when using a *stochastic first-order oracle* that returns $\nabla f(w, \xi)$ such that $\mathbb{E}_\xi[\nabla f(w, \xi)] = \nabla f(w)$.

Stochastic Gradient Descent

Bounded variance: In order to analyze the convergence of SGD, we need to assume that the variance (*noise*) in the stochastic gradients is bounded for all w, i.e. for $\sigma^2 < \infty$,

$$\mathbb{E}_i \left\| \nabla f_i(w) - \nabla f(w) \right\|^2 \leq \sigma^2 \,.$$

For SGD to converge to the minimizer, the step-size η_k needs to decrease with k.
The schedule according to which η_k needs to decrease depends on the properties of f. For example, for smooth convex functions, $\eta_k = O\left(1/\sqrt{k}\right)$, whereas for smooth, strongly-convex functions, $\eta_k = O\left(1/k\right)$.



Optimization Zoo

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

Table 2: Using the **stochastic** first-order oracle that returns $\nabla f(w, \xi)$

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	$O\left(1/\epsilon^2\right)$	$O\left(1/\epsilon^2\right)$	$O\left(1/\epsilon\right)$

Table 3: Comparing the convergence rates of GD and SGD

Questions?

Minimizing smooth, non-convex functions using SGD

Claim: For L-smooth functions lower-bounded by f^*, T iterations of stochastic gradient descent with $\eta_k = \frac{1}{L} \frac{1}{\sqrt{k+1}}$ returns an iterate $\hat{w}$ such that,

$$\mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2L\left[f(x_0) - f^*\right]}{\sqrt{T}} + \frac{\sigma^2\left(1 + \log(T)\right)}{\sqrt{T}}$$

Proof: Using the L-smoothness of f with $x = w_k$ and $y = w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$,

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

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

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

(Since η_k is independent of i_k)

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

Minimizing smooth, non-convex functions using SGD

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

$$
\begin{aligned}
\mathbb{E}[f(w_{k+1})] &\leq f(w_k) - \eta_k \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta_k^2}{2}\,\mathbb{E}\left[\left\|\nabla f_{ik}(w_k) - \nabla f(w_k) + \nabla f(w_k)\right\|^2\right] \\
&\leq f(w_k) - \eta_k \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta_k^2}{2}\,\mathbb{E}\left[\left\|\nabla f_{ik}(w_k) - \nabla f(w_k)\right\|^2\right] + \frac{L\eta_k^2}{2}\mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] \\
&\qquad\qquad \text{(Since } \mathbb{E}[\langle \nabla f(w_k), \nabla f_{ik}(w_k) - \nabla f(w_k)\rangle] = 0\text{)} \\
&= f(w_k) - \eta_k \left\|\nabla f(w_k)\right\|^2 + \frac{L\eta_k^2}{2}\mathbb{E}\left[\left\|\nabla f(w_k)\right\|^2\right] + \frac{L\sigma^2\eta_k^2}{2} \\
&\qquad\qquad \text{(Using the bounded variance assumption)}
\end{aligned}
$$

Setting $\eta_k \leq \frac{1}{L}$ for all k,

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

Minimizing smooth, non-convex functions using SGD

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

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

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

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

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

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

$$\frac{\eta_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \leq \sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w_{k+1})] + \frac{L\sigma^2\eta_k^2}{2}$$

$$\implies \frac{\eta_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \leq \mathbb{E}[f(w_0) - f(w_T)] + \frac{L\sigma^2}{2} \sum_{k=0}^{T-1} \eta_k^2$$

Minimizing smooth, non-convex functions using SGD

Recall that $\frac{\eta_{\min}}{2} \sum_{k=0}^{T-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \leq \mathbb{E}[f(w_0) - f(w_T)] + \frac{L\sigma^2}{2} \sum_{k=0}^{T-1} \eta_k^2$. Dividing by T,

$$\frac{\eta_{\min}}{2} \frac{\sum_{k=0}^{T-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2\right]}{T} \leq \frac{\mathbb{E}[f(w_0) - f(w_T)]}{T} + \frac{L\sigma^2}{2\,T} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \min_{k=0,\ldots,T-1} \mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \leq \frac{2\,\mathbb{E}[f(w_0) - f^*]}{\eta_{\min}\; T} + \frac{L\sigma^2}{\eta_{\min}\; T} \sum_{k=0}^{T-1} \eta_k^2$$

Define $\hat{w} := \arg\min_{k \in \{0,1,\ldots,T-1\}} \mathbb{E}[\|\nabla f(w_k)\|^2]$ and choosing $\eta_k = \frac{1}{L} \frac{1}{\sqrt{k+1}}$

$$\implies \mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2\,\mathbb{E}[f(w_0) - f^*]}{\eta_{\min}\; T} + \frac{L\sigma^2}{\eta_{\min}\; T} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2L\,\mathbb{E}[f(w_0) - f^*]}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{T}} \sum_{k=1}^{T} \frac{1}{k}$$

Minimizing smooth, non-convex functions using SGD

Recall that $\mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2L\,\mathbb{E}[f(w_0) - f^*]}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{T}} \sum_{k=1}^{T} \frac{1}{k}$. Since $\sum_{k=1}^{T} \frac{1}{k} \leq 1 + \log(T)$,

$$\implies \mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2L\,[f(w_0) - f^*]}{\sqrt{T}} + \frac{\sigma^2\,(1 + \log(T))}{\sqrt{T}}$$

Hence, compared to GD that has an $O\left(1/T\right)$ rate of convergence, SGD has an $O\left(1/\sqrt{T}\right)$ convergence rate, but each iteration of SGD is faster.

Can modify the proof such that we get a guarantee for a random iterate j i.e. run SGD for T iterations, randomly sample an iterate and in expectation (over the iterations), it will have small gradient norm.

Minimizing smooth, non-convex functions using SGD

Typically in practice, we use a mini-batch of size b in the SGD update. At iteration, sample a batch B_k of examples:

$$w_{k+1} = w_k - \eta_k \left[\frac{1}{b} \sum_{i \in B_k} \nabla f_i(w_k) \right]$$

The examples in the batch can be sampled independently uniformly at random with replacement, but other sampling schemes also work. The gradients can be computed in parallel (on a GPU for example) and the resulting update is efficient.

Theoretically, the same proof works. But since we are sampling with replacement, the "effective" noise is reduced to $\sigma_b^2 = \frac{n-b}{n\,b}\,\sigma^2$. Hence, if $b = n$, $\sigma_b = 0$.

Lower Bound: Without additional assumptions, for smooth functions, no first-order algorithm using the stochastic gradient oracle can obtain a (dimension-independent) convergence rate faster than $\Omega\left(1/\sqrt{T}\right)$.

Hence, similar to the deterministic setting, SGD is optimal for minimizing general smooth, non-convex functions.

Questions?

Minimizing smooth, convex functions using SGD

Claim: For L-smooth, convex functions, T iterations of stochastic gradient descent with $\eta_k = \frac{1}{2L} \frac{1}{\sqrt{k+1}}$ returns an iterate $\bar{w}_T = \frac{\sum_{k=0}^{T-1} w_k}{T}$ such that,

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

. **Proof**: Using the SGD update, $w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$,

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

Taking expectation w.r.t i_k on both sides, and assuming η_k is independent of i_k

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

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

(Unbiasedness)

Minimizing smooth, convex functions using SGD

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

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

(Since $\mathbb{E}[\langle \nabla f(w_k), \nabla f_{ik}(w_k) - \nabla f(w_k) \rangle] = 0$)

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

(Using the bounded variance assumption)

Using convexity of f, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ with $y = w^*$ and $x = w_k$,

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

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

(Using L-smoothness of f)

Minimizing smooth, convex functions using SGD

Recall $\mathbb{E}[\|w_{k+1} - w^*\|^2] \leq \|w_k - w^*\|^2 - 2\eta_k[f(w_k) - f(w^*)] + 2L\,\eta_k^2\,\mathbb{E}[f(w_k) - f(w^*)] + \eta_k^2\,\sigma^2$.
Using $\eta_k \leq \frac{1}{2L}$ for all k,

$$\begin{aligned}\mathbb{E}[\|w_{k+1} - w^*\|^2] &\leq \|w_k - w^*\|^2 - 2\eta_k[f(w_k) - f(w^*)] + \eta_k\,\mathbb{E}[f(w_k) - f(w^*)] + \eta_k^2\,\sigma^2 \\ &= \|w_k - w^*\|^2 - \eta_k[f(w_k) - f(w^*)] + \eta_k^2\,\sigma^2\end{aligned}$$

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

$$\implies \eta_{\text{min}}[f(w_k) - f(w^*)] \leq \left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] + \eta_k^2\,\sigma^2$$

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

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

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

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

Minimizing smooth, convex functions using SGD

Recall $\eta_{\min} \sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=0}^{T-1} \mathbb{E}\left[\|w_k - w^*\|^2 - \|w_{k+1} - w^*\|^2\right] + \sigma^2 \sum_{k=0}^{T-1} \eta_k^2$.

$$\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{\mathbb{E}\left[\|w_0 - w^*\|^2 - \|w_T - w^*\|^2\right]}{\eta_{\min}} + \frac{\sigma^2}{\eta_{\min}} \sum_{k=0}^{T-1} \eta_k^2$$

$$\implies \frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)]}{T} \leq \frac{\|w_0 - w^*\|^2}{\eta_{\min}\, T} + \frac{\sigma^2}{\eta_{\min}\, T} \sum_{k=0}^{T-1} \eta_k^2 \qquad \text{(Dividing by } T\text{)}$$

Define $\bar{w}_T := \frac{\sum_{k=0}^{T-1} w_k}{T}$. Since f is convex, we can use Jensen's inequality to conclude that $\mathbb{E}[f(\bar{w}_T)] \leq \frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k)]}{T}$. Choosing $\eta_k = \frac{1}{2L} \frac{1}{\sqrt{k+1}}$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\, \|w_0 - w^*\|^2}{\sqrt{T}} + \frac{\sigma^2}{2L\sqrt{T}} \sum_{k=1}^{T} \frac{1}{k}$$

Minimizing smooth, convex functions using SGD

Recall that $\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{2L\,\|w_0 - w^*\|^2}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{T}} \sum_{k=1}^{T} \frac{1}{k}$. Since $\sum_{k=1}^{T} \frac{1}{k} \leq 1 + \log(T)$,

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

Hence, compared to GD that has an $O\left({}^{1}/_{T}\right)$ rate of convergence, SGD has an $O\left({}^{1}/_{\sqrt{T}}\right)$ convergence rate, but each iteration of SGD is faster.

For GD, we proved a guarantee for the last iterate w_T; for SGD, our guarantee only holds for the average iterate $\bar{w}_T$. By using a different step-size scheme, we can get last-iterate convergence.

Lower Bound: Without additional assumptions, for smooth, convex functions, no first-order algorithm using the stochastic gradient oracle can obtain a (dimension-independent) convergence rate faster than $\Omega\left({}^{1}/_{\sqrt{T}}\right)$.
Hence, unlike the deterministic setting, SGD is optimal for minimizing smooth, convex functions.
In the stochastic setting, using momentum or Nesterov acceleration has no provable benefit in terms of the dependence on T.

Minimizing smooth, convex functions using SGD

Let us analyze the convergence for alternative choices of the step-size. By following the previous proof, we have that for $\eta_k \leq \frac{1}{2L}$,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\|w_0 - w^*\|^2}{\eta_{\min} T} + \frac{\sigma^2}{\eta_{\min} T} \sum_{k=1}^{T} \eta_k^2$$

If we do not decay the step-size, and set $\eta_k = \eta = \frac{1}{2L}$, then,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \underbrace{\frac{2L\,\|w_0 - w^*\|^2}{T}}_{\text{bias}} + \underbrace{\frac{\sigma^2}{2L}}_{\text{neighbourhood}}$$

Hence, if we use a constant step-size for SGD, it will not converge to the minimum value but will oscillate in a *neighbourhood* around the minimum. Recall that if we use a mini-batch size of b, the "effective" noise is reduced to $\sigma_b^2 = \frac{n-b}{n\,b}\,\sigma^2$. Hence, the size of the neighbourhood in which SGD oscillates is reduced. If $b = n$, $\sigma_b^2 = 0$ and SGD with a constant step-size (same as GD) will converge to the minimizer.

Questions?