

CMPT 409/981: Optimization for Machine Learning

Lecture 9

Sharan Vaswani

October 3, 2024

Dealing with Constrained Domains

- We have characterized the convergence of algorithms on smooth, (strongly)-convex functions when the domain was \mathbb{R}^d i.e. the optimization was “unconstrained”.

Numerous applications require optimizing functions over constrained domains. *Examples:*

- In reinforcement learning, finding the optimal policy in an MDP is equivalent to a linear programming with “flow” constraints.
- In supervised machine learning or operations research, the model parameters need to be optimized such that the resulting function is convex or monotonic in the input.
- The experts problem in online learning is used for forecasting, and involves optimizing over the probability simplex.

Projected GD: Modify GD to solve problems such as $\min_{w \in \mathcal{C}} f(w)$ where f is a convex function and \mathcal{C} is a convex set.

$$w_{k+1} = \Pi_{\mathcal{C}} [w_k - \eta \nabla f(w_k)] ,$$

where, $\Pi_{\mathcal{C}}[x] = \arg \min_{w \in \mathcal{C}} \frac{1}{2} \|w - x\|^2$ is the Euclidean projection onto the convex set \mathcal{C} .

Dealing with Constrained Domains

Q: (i) Is $\Pi_C[x]$ unique for convex sets? (ii) For non-convex sets?

Q: For $x \in \mathbb{R}^d$, compute the Euclidean projection onto the ℓ_2 -ball: $\mathcal{B}(0, 1) = \{w \mid \|w\|_2^2 \leq 1\}$?

Dealing with Constrained Domains

- For convex optimization over unconstrained domains, we know that the minimizer can be characterized by its gradient norm i.e. if w^* is a minimizer, then, $\nabla f(w^*) = 0$.

Optimality conditions: For constrained convex domains, if f is convex and $w^* \in \arg \min_{w \in \mathcal{C}} f(w)$, then $\forall w \in \mathcal{C}$,

$$\langle \nabla f(w^*), w - w^* \rangle \geq 0$$

i.e. if we are at the optimal, either the gradient is zero (if w^* is inside \mathcal{C}) or moving in the negative direction of the gradient will push us out of \mathcal{C} (if w^* is at the boundary of \mathcal{C}).

- For the Euclidean projection, if $y := \Pi_{\mathcal{C}}[x] = \arg \min_{w \in \mathcal{C}} \frac{1}{2} \|w - x\|^2$, then, using the optimality conditions above, $\forall w \in \mathcal{C}$,

$$\langle x - y, w - y \rangle \leq 0$$

i.e. the angle between the rays $y \rightarrow x$ and $y \rightarrow w$ for all $w \in \mathcal{C}$ is greater than 90° .

Q: For convex set \mathcal{C} , if $w^* = \arg \min_{w \in \mathcal{C}} f(w)$, what is $\Pi_{\mathcal{C}}[w^*]$?

Dealing with Constrained Domains

Claim: Projections onto a convex set are non-expansive operations i.e. for all x_1, x_2 , if $y_1 := \Pi_{\mathcal{C}}[x_1]$ and $y_2 := \Pi_{\mathcal{C}}[x_2]$, then, $\|y_1 - y_2\| \leq \|x_1 - x_2\|$.

Proof: Recall from the last slide, that for the Euclidean projection, $y = \Pi_{\mathcal{C}}[x]$, $\langle x - y, w - y \rangle \leq 0$ for all $w \in \mathcal{C}$. Hence,

$$\langle x_1 - y_1, w - y_1 \rangle \leq 0 \implies \langle x_1 - y_1, y_2 - y_1 \rangle \leq 0 \quad (\text{Set } w = y_2)$$

$$\langle x_2 - y_2, w - y_2 \rangle \leq 0 \implies \langle x_2 - y_2, y_1 - y_2 \rangle \leq 0 \quad (\text{Set } w = y_1)$$

Adding the two equations,

$$\begin{aligned} \langle x_2 - y_2, y_1 - y_2 \rangle + \langle x_1 - y_1, y_2 - y_1 \rangle &\leq 0 \implies \langle x_2 - x_1 + y_1 - y_2, y_1 - y_2 \rangle \leq 0 \\ \implies \langle y_1 - y_2, y_1 - y_2 \rangle &\leq \langle x_1 - x_2, y_1 - y_2 \rangle \implies \|y_1 - y_2\|^2 \leq \|x_1 - x_2\| \|y_1 - y_2\| \\ &\quad (\text{Cauchy Schwartz}) \end{aligned}$$

$$\implies \|y_1 - y_2\| \leq \|x_1 - x_2\|$$

Projected GD for Smooth, Strongly-Convex Functions

- Consider using projected GD: $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta \nabla f(w_k)]$ to solve the problem: $\min_{w \in \mathcal{C}} f(w)$, where f is an L -smooth, μ -strongly convex function and \mathcal{C} is a convex set.
- In Assignment 2, you need to prove that: w^* is a fixed point of the projected GD update i.e, for any $\eta \geq 0$, $w^* = \Pi_{\mathcal{C}}[w^* - \eta \nabla f(w^*)]$.
- Using this property and the non-expansiveness of projections with $x_1 = w^* - \eta \nabla f(w^*)$, $x_2 = w_k - \eta \nabla f(w_k)$, $y_1 = w^*$, $y_2 = w_{k+1}$,

$$\|w_{k+1} - w^*\|^2 \leq \|w_k - \eta \nabla f(w_k) - w^* + \eta \nabla f(w^*)\|^2$$

With this change, the proof proceeds as before. Using the optimality condition for w^* , smoothness and strong-convexity (similar to Lecture 4), we can derive the same linear rate (Need to prove in Assignment 2) .

- We can also redo the proof for smooth, convex functions and get the same $O(1/\tau)$ convergence rate. Hence, projected GD is a good option for minimizing convex functions over convex sets when the projection operation is computationally cheap.

Questions?

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

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.

Examples: In supervised learning using a dataset of n input-output pairs $\{X_i, y_i\}_{i=1}^n$,

$$f(w) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\langle X_i, w \rangle - y_i)^2 \quad (\text{Linear Regression})$$

$$f(w) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \langle X_i, w \rangle)) \quad (\text{Logistic Regression})$$

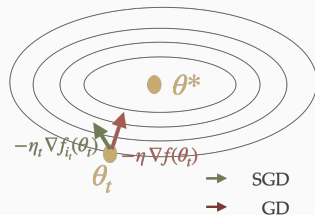
- Gradient-based methods on such functions require computing $\nabla f(w) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w)$ which is an $O(n)$ operation. Typically, n is large in practice and hence computing the gradient across the whole datasets is expensive.

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 (typically uniformly) randomly:

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

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



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

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

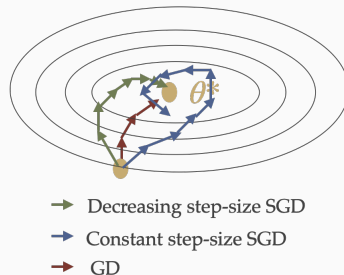
- We will assume that $f(w)$ is a finite-sum of n points only for convenience. In general, all the results hold when using a *stochastic first-order oracle* that returns $\nabla f(w, \xi)$ such that $\mathbb{E}_{\xi}[\nabla f(w, \xi)] = \nabla f(w)$.

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 (technically, this is the trace of the covariance matrix of the stochastic gradients) is bounded for all w , i.e. for $\sigma^2 < \infty$,

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

- 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 .
- *Example:* For smooth convex functions, $\eta_k = O(1/\sqrt{k})$, whereas for smooth, strongly-convex functions, $\eta_k = O(1/k)$.



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

Table 2: 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^* and with bounded noise σ^2 , T iterations of stochastic gradient descent with $\eta_k = \frac{1}{L} \frac{1}{\sqrt{k+1}}$ returns an iterate \hat{w} such that,

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

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

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

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

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

(Since η_k is independent of i_k)

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

Minimizing smooth, non-convex functions using SGD

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

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

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

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

Minimizing smooth, non-convex functions using SGD

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

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

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 $T - 1$,

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

Minimizing smooth, non-convex functions using SGD

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

$$\begin{aligned} \frac{\eta_{\min}}{2} \frac{\sum_{k=0}^{T-1} \mathbb{E} [\|\nabla f(w_k)\|^2]}{T} &\leq \frac{\mathbb{E}[f(w_0) - f(w_T)]}{T} + \frac{L\sigma^2}{2T} \sum_{k=0}^{T-1} \eta_k^2 \\ \Rightarrow \min_{k=0, \dots, T-1} \mathbb{E} [\|\nabla f(w_k)\|^2] &\leq \frac{2 \mathbb{E}[f(w_0) - f^*]}{\eta_{\min} T} + \frac{L\sigma^2}{\eta_{\min} T} \sum_{k=0}^{T-1} \eta_k^2 \end{aligned}$$

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

$$\begin{aligned} \Rightarrow \mathbb{E} [\|\nabla f(\hat{w})\|^2] &\leq \frac{2 \mathbb{E}[f(w_0) - f^*]}{\eta_{\min} T} + \frac{L\sigma^2}{\eta_{\min} T} \sum_{k=0}^{T-1} \eta_k^2 \\ \Rightarrow \mathbb{E} [\|\nabla f(\hat{w})\|^2] &\leq \frac{2L \mathbb{E}[f(w_0) - f^*]}{\sqrt{T}} + \frac{2\sigma^2}{\sqrt{T}} \sum_{k=1}^T \frac{1}{k} \end{aligned}$$

Minimizing smooth, non-convex functions using SGD

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

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

- Hence, compared to GD that has an $O(1/T)$ rate of convergence, SGD has an $O(1/\sqrt{T})$ convergence rate, but each iteration of SGD is n times 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 in expectation (over the randomness in each iteration).

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 without replacement, but other sampling schemes also work.
- The gradients can be computed in parallel (e.g. on a GPU) and the resulting update is efficient.
- Theoretically, the same proof works, but the “effective” noise is reduced to $\sigma_b^2 = \frac{n-b}{n} \sigma^2$.

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 $\Omega(1/\sqrt{T})$.

Hence, SGD is optimal for minimizing general smooth, non-convex functions.

Questions?

Minimizing smooth, convex functions using SGD

Claim: For L -smooth, convex functions with bounded noise σ^2 , 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,

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

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

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

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

$$\begin{aligned} \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\mathbb{E}[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle] + \mathbb{E}[\eta_k^2 \|\nabla f_{ik}(w_k)\|^2] \\ &= \|w_k - w^*\|^2 - 2\eta_k \langle \mathbb{E}[\nabla f_{ik}(w_k)], w_k - w^* \rangle + \eta_k^2 \mathbb{E}[\|\nabla f_{ik}(w_k)\|^2] \\ \implies \mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\eta_k \langle \nabla f(w_k), w_k - w^* \rangle + \eta_k^2 \mathbb{E}[\|\nabla f_{ik}(w_k)\|^2] \\ &\hspace{20em} \text{(Unbiasedness)} \end{aligned}$$

Minimizing smooth, convex functions using SGD

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

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

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

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

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 ,

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

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

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

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

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

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

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

Minimizing smooth, convex functions using SGD

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

$$\begin{aligned} \sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)] &\leq \frac{\mathbb{E} \left[\|w_0 - w^*\|^2 - \|w_T - w^*\|^2 \right]}{\eta_{\min}} + \frac{\sigma^2}{\eta_{\min}} \sum_{k=0}^{T-1} \eta_k^2 \\ \Rightarrow \frac{\sum_{k=0}^{T-1} \mathbb{E}[f(w_k) - f(w^*)]}{T} &\leq \frac{\|w_0 - w^*\|^2}{\eta_{\min} T} + \frac{\sigma^2}{\eta_{\min} T} \sum_{k=0}^{T-1} \eta_k^2 \quad (\text{Dividing by } T) \end{aligned}$$

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 $\eta_k = \frac{1}{2L} \frac{1}{\sqrt{k+1}}$,

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

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 $\sum_{k=1}^T \frac{1}{k} \leq 1 + \log(T)$,

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

- Hence, compared to GD that has an $O(1/T)$ rate of convergence, SGD has an $O(1/\sqrt{T})$ 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 $\Omega(1/\sqrt{T})$.

Hence, 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 an alternative choice 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 $\sigma_b^2 = \frac{n-b}{nb} \sigma^2$. 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?