

# ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-4: Stochastic Gradient Descent

Jia (Kevin) Liu

Associate Professor  
Department of Electrical and Computer Engineering  
The Ohio State University, Columbus, OH, USA

Autumn 2024

# Outline

In this lecture:

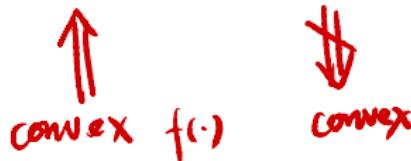
- Noisy unbiased gradient
- Stochastic gradient method
- Convergence results

# Unbiased Stochastic Gradient

- Random vector  $\tilde{g} \in \mathbb{R}^n$  is a unbiased stochastic gradient if it can be written  $\tilde{g} = g + \mathbf{n}$ , where  $g$  is the true gradient and  $\mathbb{E}[\mathbf{n}] = \mathbf{0}$
- $\mathbf{n}$  can be interpreted as error in computing  $g$ , measurement noise, Monte Carlo sampling errors, etc.
- If  $f(\cdot)$  is non-smooth,  $\tilde{g}$  is a noisy subgradient at  $x$  if

$$f(\mathbf{z}) \geq f(\mathbf{x}) + (\mathbb{E}[\tilde{g}|\mathbf{x}])^\top (\mathbf{z} - \mathbf{x}), \quad \forall \mathbf{z}$$

holds almost surely.



# Stochastic Gradient Descent Method

- Consider  $\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$ . Following standard GD, we should do:

$$\cancel{-\nabla f(\mathbf{x}_k)}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbb{E}[\tilde{\mathbf{g}}_k | \mathbf{x}_k]$$

- However,  $\mathbb{E}[\tilde{\mathbf{g}}_k | \mathbf{x}_k]$  is **difficult** to compute: Unknown distribution, too costly to sample at each iteration  $k$ , etc.
- Idea:** Simply use a noisy unbiased subgradient to replace  $\mathbb{E}[\tilde{\mathbf{g}}_k | \mathbf{x}_k]$
- The **stochastic subgradient** method works as follows:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \tilde{\mathbf{g}}_k$$

- $\mathbf{x}_k$  is the  $k$ -th iterate
- $\tilde{\mathbf{g}}_k$  is any noisy gradient of at  $\mathbf{x}_k$ , i.e.,  $\mathbb{E}[\tilde{\mathbf{g}}_k | \mathbf{x}_k] = \nabla f(\mathbf{x}_k)$
- $s_k$  is the step size
- Let  $f_{\text{best}}^{(k)} \triangleq \min_{i=1,\dots,k} \{f(\mathbf{x}_i)\}$  and  $\|\nabla f_{\text{best}}^{(k)}\| \triangleq \min_{i=1,\dots,k} \{\|\nabla f(\mathbf{x}_i)\|\}$

# Historical Perspective

- Also referred to as **stochastic approximation** in the literature, first introduced by [Robbins, Monro '51] and [Keifer, Wolfowitz '52]
- The original work [Robbins, Monro '51] is motivated by finding a root of a continuous function:

*vector-valued*  
✓  
$$f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x}, \theta)] = 0,$$

where  $F(\cdot, \cdot)$  is **unknown** and depends on a random variable  $\theta$ . But the experimenter can take random samples (noisy measurements) of  $F(\mathbf{x}, \theta)$



Herbert Robbins



Sutton Monro

# Historical Perspective

- **Robbins-Monro:**  $\mathbf{x}_{k+1} = \mathbf{x}_k + s_k Y(\mathbf{x}_k, \theta)$ , where:
  - ▶  $\mathbb{E}[Y(\mathbf{x}, \theta) | \mathbf{x} = \mathbf{x}_k] = f(\mathbf{x}_k)$  is an unbiased estimator of  $f(\mathbf{x}_k)$
  - ▶ Robbins-Monro originally showed convergence in  $L^2$  and in probability
  - ▶ Blum later prove convergence is actually w.p.1. (almost surely)
  - ▶ Key idea: Diminishing step-size provides **implicit averaging** of the observations
- Robbins-Monro's scheme can also be used in **stochastic optimization** of the form  $f(\mathbf{x}^*) = \min_{\mathbf{x}} \mathbb{E}[F(\mathbf{x}, \theta)]$  (equivalent to solving  $\nabla f(\mathbf{x}^*) = 0$ )
- Stochastic approximation, or more generally, stochastic gradient has found applications in many areas
  - ▶ Adaptive signal processing
  - ▶ Dynamic network control and optimization
  - ▶ Statistical machine learning
  - ▶ Workhorse algorithm for training **deep neural networks**

## Convergence of R.V.

1. Convergence in Distr. (weak convergence)

A seq. of (real-valued) r.v.  $\{X_n\}$  converges in dist. to  $X$  if  $\lim_{n \rightarrow \infty} F_n(X_n) = F(X)$ , where  $F_n$  and  $F$  are cdf of  $X_n$  and  $X$ , resp. Denote as  $X_n \xrightarrow{D} X$ .

2. Convergence in prob. to a r.v. ("stronger").

$\{X_n\}$  converges in prob. to a r.v.  $X$  if  $\forall \varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr\{|X_n - X| > \varepsilon\} = 0. \quad \text{Denote as: } X_n \xrightarrow{\text{P}} X.$$

3. Almost sure convergence (pt.-wise convergence in Real Analysis)

$\{X_n\}$  converges a.s. (a.e. or w.p.1, or strongly) to  $X$ .

$\nexists \Pr\{\lim_{n \rightarrow \infty} X_n = X\} = 1. \quad \text{Denoted as } X_n \xrightarrow{\text{a.s.}} X.$

4. Convergence in expectation = Given  $r \geq 1$ .  $\{X_n\}$  converges

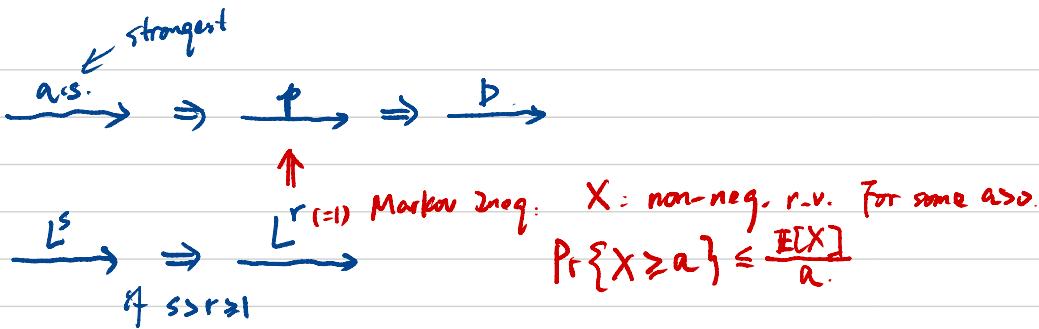
in  $r$ -th mean to r.v.  $X$  if the  $r$ -th abs. moments

$\mathbb{E}[|X_n|^r]$  and  $\mathbb{E}[|X|^r]$  exist, and

$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^r] = 0. \quad \text{Denote as } X_n \xrightarrow{L^r} X$

\*  $r=1$ :  $X_n$  converges in mean to  $X$

\*  $r=2$ :  $\underline{\hspace{1cm}}$   $\overline{\hspace{1cm}}$  mean square to  $X$ .



\* For r.v.  $z_1, \dots, z_n$  that are indep. with mean 0.

$$E[\|z_1 + \dots + z_n\|^2] \leq E[\|z_1\|^2 + \dots + \|z_n\|^2] \quad (\text{improves if } E[\|z_i\|] < 0)$$

\* For r.v.  $z_1, \dots, z_n$  that are not nec. indep.

$$E[\|z_1 + \dots + z_n\|^2] \leq n E[\|z_1\|^2 + \dots + \|z_n\|^2]$$

# Assumptions and Step Size Rules

- $f^* = \inf_x f(\mathbf{x}_k) > -\infty$ , with  $f(\mathbf{x}^*) = f^*$
- $\mathbb{E}[\|\tilde{\mathbf{g}}_k\|_2^2] \leq G^2$ , for all  $k$
- $\mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] \leq R^2$

Commonly used step-size strategies:

- Constant step-size:  $s_k = s, \forall k$
- Step-size is square summable, but not summable

$$s_k > 0, \forall k,$$

$$\boxed{\sum_{k=1}^{\infty} s_k^2 < \infty,}$$

$$\sum_{k=1}^{\infty} s_k = \infty$$

*not needed*

Note: This is stronger than needed, but just to simplify proof

# Convergence of SGD (Convex)

*Asymptotic*

- Convergence in expectation:

$$\lim_{k \rightarrow \infty} \mathbb{E}[f_{\text{best}}^{(k)}] = f^*$$

- Convergence in probability: for any  $\epsilon > 0$ ,

$$\lim_{k \rightarrow \infty} \Pr\{|f_{\text{best}}^{(k)} - f^*| > \epsilon\} = 0$$

- Almost sure convergence

$$\Pr\left\{\lim_{k \rightarrow \infty} f_{\text{best}}^{(k)} = f^*\right\} = 1$$

- See [Kushner, Yin '97] for a complete treatment on convergence analysis

(Convex)

Thm: If  $E[\|\tilde{g}_k\|_2] \leq G$ ,  $\forall k$ ,  $E\{\|\bar{x}_i - \bar{x}^*\|\} \leq R$  and

step-sizes  $\{s_k\}_{k=1}^{\infty}$  satisfy:  $s_k > 0, \forall k, \sum_{k=1}^{\infty} s_k^2 < \infty, \sum_{k=1}^{\infty} s_k \rightarrow \infty$ .

then:

$$\lim_{k \rightarrow \infty} E\{f_{\text{best}}^{(k)}\} = f^* \text{ and } \lim_{k \rightarrow \infty} \{ |f_{\text{best}}^{(k)} - f^*| > \varepsilon \} = 0, \forall \varepsilon > 0.$$

Proof. Consider cond. expectation square Euclidean dist.

$$E[\|\bar{x}_{k+1} - \bar{x}^*\|^2 | \bar{x}_k] = E[\|\underbrace{\bar{x}_k - s_k \tilde{g}_k - \bar{x}^*}_{}|^2 | \bar{x}_k]$$

$$= E[\|\bar{x}_k - \bar{x}^*\|^2 + s_k^2 \|\tilde{g}_k\|^2 - 2s_k \tilde{g}_k^T (\bar{x}_k - \bar{x}^*) | \bar{x}_k]$$

$$= \|\bar{x}_k - \bar{x}^*\|^2 + s_k^2 E[\|\tilde{g}_k\|^2 | \bar{x}_k] - 2s_k \underbrace{E[\tilde{g}_k | \bar{x}_k]}_{=\nabla f(\bar{x}_k)}^T (\bar{x}_k - \bar{x}^*) \quad (*)$$

$$\text{Note: } f(\bar{x}^*) \geq f(\bar{x}_k) + E\{\tilde{g}_k | \bar{x}_k\}^T (\bar{x}^* - \bar{x}_k)$$

$$\Rightarrow -E[\tilde{g}_k | \bar{x}_k]^T (\bar{x}_k - \bar{x}^*) \leq - (f(\bar{x}_k) - f^*)$$

$$(*) \leq \|\bar{x}_k - \bar{x}^*\|^2 + s_k^2 E[\|\tilde{g}_k\|^2 | \bar{x}_k] - 2s_k (f(\bar{x}_k) - f^*).$$

Note: from SGD dynamic,  $\bar{x}_{k+1}$  only dep.  $\bar{x}_k$  ( $\bar{x}_{k+1} = \bar{x}_k - s_k \tilde{g}_k$ )  
and indep. of  $\bar{x}_{k-1}, \dots, \bar{x}_1$ .

$$E[\|\bar{x}_{k+1} - \bar{x}^*\|^2 | \bar{x}_k] = E[\|\bar{x}_k - \bar{x}^*\|^2 | \bar{x}_k, \dots, \bar{x}_1]$$

Take expectation over joint dists. of  $\{\bar{x}_k, \dots, \bar{x}_1\}$  yields

$$E[\|\bar{x}_{k+1} - \bar{x}^*\|^2] \leq E[\|\bar{x}_k - \bar{x}^*\|^2] - 2s_k [E[f(\bar{x}_k)] - f^*]$$

$$+ s_k^2 E[\|\tilde{g}_k\|^2] \leq G$$

Apply this process recursively:

$$\mathbb{E}[||\mathbf{x}_{(k+1)} - \mathbf{x}^*||^2] \leq \mathbb{E}[||\mathbf{x}_1 - \mathbf{x}^*||^2] - 2 \sum_{i=1}^k s_i (\mathbb{E}[f(\mathbf{x}_i)] - f^*)$$

$\stackrel{\mathbf{x}_1 \in R^2}{\leq}$

$\geq \min_{i=1 \dots k} f(\mathbf{x}_i) - f^*$

$$+ G \sum_{k=1}^k s_k^2 \stackrel{\mathbf{x}_k \in B}{\leq}$$

$$\rightarrow \min_{i=1 \dots k} \{ \mathbb{E}[f(\mathbf{x}_i)] - f^* \} \leq \frac{R^2 + G^2 B}{2 \sum_{i=1}^k s_i} \xrightarrow[G \rightarrow \infty]{R^2 \rightarrow 0} 0 \text{ as } k \rightarrow \infty.$$

Claim: The fn  $g(y) \triangleq \min_{i=1 \dots k} \{y_i\}$  is concave. (HW)

Then Jensen's says:

\* If  $f$  is convex:  $f(\mathbb{E}X) \leq \mathbb{E}f(X)$

\* --- concave:  $\geq$

$$\mathbb{E}[f_{\text{best}}^{(k)}] = \mathbb{E}\left[\min_{i=1 \dots k} f(\mathbf{x}_i)\right] \stackrel{\text{concave}}{\leq} \min_{i=1 \dots k} \mathbb{E}[f(\mathbf{x}_i)] \stackrel{\text{Jensen's}}{\geq} \stackrel{\text{done earlier}}{\rightarrow} f^*$$



# Convergence in Expectation and Probability (Convex)

*Proof Sketch:*

- Key quantity: Expected squared Euclidean distance to the optimal set. Let  $\mathbf{x}^*$  be any minimizer of  $f$ . We can show that

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 | \mathbf{x}_k] \leq \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - 2s_k(f(\mathbf{x}_k) - f^*) + s_k^2 \mathbb{E}[\|\tilde{\mathbf{g}}_k\|_2^2 | \mathbf{x}_k]$$

- which can further lead to

$$\min_{i=1,\dots,k} \left\{ \mathbb{E}[f(\mathbf{x}_i)] - f^* \right\} \leq \frac{R^2 + G^2 \|s\|^2}{2 \sum_{i=1}^k s_i}$$

- The result  $\min_{i=1,\dots,k} \mathbb{E}[f(\mathbf{x}_i)] \rightarrow f^*$  simply follows from the divergent step-size series rule

# Convergence in Expectation and Probability (Convex)

- Jensen's inequality and concavity of minimum yields

$$\mathbb{E}[f_{\text{best}}^{(k)}] = \mathbb{E}\left[\min_{i=1,\dots,k} f(\mathbf{x}_i)\right] \leq \min_{i=1,\dots,k} \mathbb{E}[f(\mathbf{x}_i)]$$

Therefore,  $\mathbb{E}[f_{\text{best}}^{(k)}] \rightarrow f^*$  (convergence in expectation)

- Convergence in expectation also implies convergence in probability: By Markov's inequality, for any  $\epsilon > 0$ ,

$$\Pr\{f_{\text{best}}^{(k)} - f^* \geq \epsilon\} \leq \frac{\mathbb{E}[f_{\text{best}}^{(k)} - f^*]}{\epsilon},$$

i.e., RHS goes to 0, which proves convergence in probability.

□

## Convergence Rate (Convex)

$$\lim_{k \rightarrow \infty} \mathbb{E} \left\{ \min_{i=1, \dots, k} f(x_i) - f^* \right\} \leq \frac{\frac{R^2 + G}{2} \sum_{k=1}^{\infty} s_k}{2 \sum_{k=1}^{\infty} s_k}$$

$$\begin{aligned} & 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \dots \quad (\text{Nicole's}) \\ & = 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \dots \\ & > 1 + \frac{1}{2} + \left(\frac{1}{4} + \frac{1}{4}\right) + \left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right) + \dots \\ & = 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \dots = \infty \end{aligned}$$

$\Theta(\log(t))$ .

- Classical diminishing step-sizes  $s_k = \alpha/k$  for some  $\alpha > 0$ :

$$\int_0^n \frac{1}{x} dx \leq \sum_{k=1}^n \frac{1}{k} \leq f(n) \leq f(1) + \int_1^n \frac{1}{x} dx \quad \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \dots = \frac{\pi^2}{6}$$

integral test

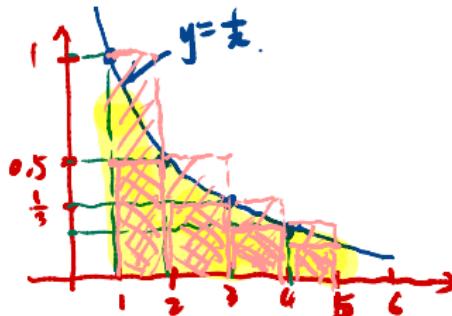
- Diminishing step-sizes  $s_k = \alpha/\sqrt{k}$  for some  $\alpha > 0$ :  $\sum_k s_k = O(\sqrt{t})$  and

$$\sum_k s_k^2 = O(\log(t)). \text{ So convergence rate is } O(\log(t)/\sqrt{t}) = \tilde{O}(1/\sqrt{t})$$

- Constant step-sizes  $s_k = \alpha$  for some  $\alpha > 0$ :  $\sum_k s_k = k\alpha$  and  $\sum_k s_k^2 = k\alpha^2$ .  
So convergence rate is  $O(1/t) + O(\alpha)$

$$\int_1^n \frac{1}{x} dx = \log(n)$$

$$\sum_{k=1}^n \frac{1}{k+1} \leq \int_1^n \frac{1}{x} dx \leq \sum_{k=1}^n \frac{1}{k}$$



# Convergence Rate (Strongly Convex)

## Theorem 1 (Optimality Gap)

If  $f(\cdot)$  is  $\mu$ -strongly convex, then the SGD method with a constant step-size  $s_k = s < 2/\mu$  satisfies:

$$\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \leq (1 - 2s\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{s\sigma^2}{2\mu}$$

Remark:

- If  $\sigma^2 = 0$  (GD), constant step-size  $s \Rightarrow$  linear convergence to  $\mathbf{x}^*$ .
- If  $\sigma^2 > 0$ , SGD with constant step-size  $s \Rightarrow$  linear convergence to  $\frac{s\sigma^2}{2\mu}$ -neighborhood of  $\mathbf{x}^*$

strongly convex:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2$$

$$f(x) \geq f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2} \|y - x\|^2$$

$$\text{Add together: } [\cancel{\nabla f(y)} - \cancel{\nabla f(x)}]^T (y - x) \geq \mu \|y - x\|^2$$

$$\text{Recall: } \mathbb{E}\left[\|\bar{x}_{k+1} - \bar{x}^*\| \mid \bar{x}_k\right] \leq \|\bar{x}_k - \bar{x}^*\|^2 + s_k^2 \mathbb{E}\left[\|\tilde{g}_k\|^2 \mid \bar{x}_k\right] \\ - 2s_k \mathbb{E}\left[\tilde{g}_k^T \mid \bar{x}_k\right]^T (\bar{x}_k - \bar{x}^*)$$

Taking full expectation:

$$\mathbb{E}\left[\|\bar{x}_{k+1} - \bar{x}^*\|^2\right] \leq \mathbb{E}\left[\|\bar{x}_k - \bar{x}^*\|^2\right] + s_k^2 \mathbb{E}\left[\|\tilde{g}_k\|^2\right] - 2\mu s_k \mathbb{E}\left[\|\bar{x}_k - \bar{x}^*\|\right] \leq s^2$$

$$\mathbb{E}\left[\tilde{g}_k^T (\bar{x}_k - \bar{x}^*) \mid \bar{x}_k\right] \geq \mu \|\bar{x}_k - \bar{x}^*\|^2$$

$$= (-2\mu s_k) \mathbb{E}\left[\|\bar{x}_k - \bar{x}^*\|^2\right] + s_k^2 \sigma^2 \quad (1) \quad s < \frac{2}{\mu}$$

Applying (1) recursively from  $k-1$  down to 1, letting  $s_k = s$  with

$$\mathbb{E}\left[\|\bar{x}_k - \bar{x}^*\|^2\right] \leq (1-2\mu s)^k \|\bar{x}_0 - \bar{x}^*\|^2 + \frac{s\sigma^2}{2\mu}. \quad \blacksquare$$

(HW): What about diminishing step-size?

# Convergence Rate (Nonconvex) – Finite Sum

- Consider the following finite-sum minimization

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) = \min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x})$$

where  $N$  is typically large, e.g., empirical risk minimization (ERM) in ML

- Consider using SGD to solve this problem under the following assumptions:
  - $f(\cdot)$  is nonconvex and bounded from below
  - $\nabla f$  is differentiable with  $L$ -Lipschitz continuous gradients ( $L$ -smooth)
  - $\mathbb{E}[\|\nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$  for some  $\sigma^2$  and all  $\mathbf{x}$  (bounded gradient, can be relaxed)

can be relaxed:  $\mathbb{E}[\|\nabla f_i(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2] \leq \sigma^2$

can be further relaxed:  $\mathbb{E}[\|\nabla f_i(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2] \leq \gamma \|\nabla f(\mathbf{x}_k)\|^2$   
"SNR  $\geq \frac{1}{\gamma}$ "

# Convergence Rate (Nonconvex) – Finite Sum

## Theorem 2 (Stationarity Gap)

If the finite-sum problem  $f(\cdot)$  is nonconvex, differentiable, and  $L$ -smooth, then the SGD method with step-sizes  $\{s_k\}$  satisfies

$$\min_{k=0,1,\dots,t-1} \{\|\nabla f(\mathbf{x}_k)\|_2^2\} \leq \frac{f(\mathbf{x}_0) - f^*}{\sum_{k=0}^{t-1} s_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} s_k^2}{\sum_{k=0}^{t-1} s_k}.$$

Remark:

- If  $\sigma^2 = 0$ , then a constant step-size yields an  $O(1/t)$  rate.
- Classical diminishing step-sizes  $s_k = \alpha/k$  for some  $\alpha > 0$ :  
 $\sum_k s_k = O(\log(t))$  and  $\sum_k s_k^2 = O(1)$ . So convergence rate is  $O(1/\log(t))$
- Diminishing step-sizes  $s_k = \alpha/\sqrt{k}$  for some  $\alpha > 0$ :  $\sum_k s_k = O(\sqrt{t})$  and  $\sum_k s_k^2 = O(\log(t))$ . So convergence rate is  $O(\log(t)/\sqrt{t}) = \tilde{O}(1/\sqrt{t})$
- Constant step-sizes  $s_k = \alpha$  for some  $\alpha > 0$ :  $\sum_k s_k = k\alpha$  and  $\sum_k s_k^2 = k\alpha^2$ .  
So convergence rate is  $O(1/t) + O(\alpha)$

## Theorem 2 (Stationarity Gap)

If the finite-sum problem  $f(\cdot)$  is nonconvex, differentiable, and  $L$ -smooth, then the SGD method with step-sizes  $\{s_k\}$  satisfies

$$\min_{k=0,1,\dots,t-1} \{\|\nabla f(\mathbf{x}_k)\|_2^2\} \leq \frac{f(\mathbf{x}_0) - f^*}{\sum_{k=0}^{t-1} s_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} s_k^2}{\sum_{k=0}^{t-1} s_k}.$$

Proof: Consider uniform sampling  $i_k \in \{1, \dots, N\}$  with

$$\Pr(i_k = i) = \frac{1}{N}.$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \nabla f_{i_k}(\mathbf{x}_k).$$

$$\mathbb{E}[\nabla f_{i_k}(\mathbf{x}_k)] = \sum_{i=1}^N \Pr(i_k = i) \nabla f_{i_k}(\mathbf{x}_k) = \frac{1}{N} \sum_{i=1}^N \nabla f_{i_k}(\mathbf{x}_k) = \nabla f(\mathbf{x}_k).$$

Recall the descent lemma in GD.

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T (\mathbf{x}_{k+1} - \mathbf{x}_k) + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$$

Plugging in SGD iteration yields:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - s_k \nabla f(\mathbf{x}_k)^T \nabla f_{i_k}(\mathbf{x}_k) + \frac{L s_k^2}{2} \|\nabla f_{i_k}(\mathbf{x}_k)\|^2$$

Take expectation w.r.t.  $i_k$

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{k+1})] &\leq \mathbb{E}\left[f(\mathbf{x}_k) - s_k \nabla f(\mathbf{x}_k)^T \nabla f_{i_k}(\mathbf{x}_k) + \frac{L s_k^2}{2} \|\nabla f_{i_k}(\mathbf{x}_k)\|^2\right] \\ &= \mathbb{E}[f(\mathbf{x}_k)] - s_k \|\nabla f(\mathbf{x}_k)\|^2 + \frac{L s_k^2}{2} \mathbb{E}\left[\|\nabla f_{i_k}(\mathbf{x}_k)\|^2\right] \\ &\leq \mathbb{E}[f(\mathbf{x}_k)] - \underbrace{s_k \|\nabla f(\mathbf{x}_k)\|^2}_{\text{good}} + \underbrace{\frac{L s_k^2}{2} \sigma^2}_{\text{bad}}. \end{aligned}$$

As in GD: rearrange to get the grad norm on LHS.

$$s_k \|\nabla f(\mathbf{z}_k)\|^2 \leq \mathbb{E}[f(\mathbf{z}_k)] - \mathbb{E}[f(\mathbf{z}_{k+1})] + \frac{Ls_k^2}{2} \sigma^2 \quad (2)$$

Summing (2) from 1 to  $t$  & use iterated expectation to get.

$$\sum_{k=1}^{t-1} s_{k+1} \mathbb{E}[\|\nabla f(\mathbf{z}_{k+1})\|^2] \leq \sum_{k=1}^t [\mathbb{E}[f(\mathbf{z}_k)] - \mathbb{E}[f(\mathbf{z}_{k+1})]] + \frac{L\sigma^2}{2} \sum_{k=0}^{t-1} s_k^2$$

$\geq \min_{k=0 \dots t-1} \{\mathbb{E}[\|\nabla f(\mathbf{z}_{k+1})\|^2]\}$  telescope.

$$\Rightarrow \min_{k=0 \dots t-1} \{\mathbb{E}[\|\nabla f(\mathbf{z}_k)\|^2]\} \leq \frac{f(\mathbf{z}_0) - f(\mathbf{z}^*)}{\sum_{k=0}^{t-1} s_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} s_k^2}{\sum_{k=0}^{t-1} s_k}. \quad \blacksquare$$

# Convergence Rate (Nonconvex) - Finite Sum+Time Oracle

## Theorem 3 ([Ghadimi & Lan '13])

Suppose  $f(\cdot)$  is  $L$ -smooth and has  $\sigma$ -bounded gradients and it is known a priori that the SGD algorithm will be executed for  $T$  iterations. Let  $s_k = c/\sqrt{T}$ , where

$$c = \sqrt{\frac{2(f(\mathbf{x}_0) - f^*)}{L\sigma^2}}.$$

Then, the iterates of SGD satisfy

$$\min_{0 \leq t \leq T-1} \mathbb{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq \sqrt{\frac{2(f(\mathbf{x}_0) - f^*)L}{T}}\sigma.$$

# Convergence Rate (Nonconvex) - General Expectation Minimization with Batching

- Consider the following general expectation minimization problem

$$f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)],$$

where  $\xi$  is a random variable with distribution  $\mathcal{D}$ .

- Consider using SGD to solve this problem under the following assumptions:
  - $f(\cdot)$  is nonconvex and bounded from below
  - $\nabla f$  is differentiable with  $L$ -Lipschitz continuous gradients ( $L$ -smooth)
  - $\mathbb{E}_\xi[f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$  and  $\mathbb{E}_\xi[\|f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2] \leq \sigma^2$
- A common approach in SGD: Rather than choosing one training sample randomly at a time, use a **larger random mini-batch of samples**  $\mathcal{B}_k$ , with  $|\mathcal{B}_k| = B_k$ . Then,  $\mathbf{g}_k = \frac{1}{B_k} \sum_{i=1}^{B_k} \nabla f(\mathbf{x}, \xi_i)$ . SGD becomes:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{g}_k = \mathbf{x}_k - \frac{s_k}{B_k} \sum_{i=1}^{B_k} \nabla f(\mathbf{x}, \xi_i),$$

where  $\xi_1, \dots, \xi_{B_k}$  are i.i.d. sampled from  $\mathcal{D}$

# Convergence Rate (Nonconvex) - General Expectation Minimization with Batching

## Theorem 4 (Stationarity Gap)

In the expectation minimization problem, supposed that  $f(\cdot)$  is nonconvex, differentiable, and  $L$ -smooth. For any given  $\epsilon > 0$ , then the SGD method with mini-batch size  $B_k = B = \max\{1, \frac{2\sigma^2}{\epsilon^2}\}$ ,  $\forall k$ , and step-sizes  $s_k \leq \frac{1}{2L}$ ,  $\forall k$ , satisfies

$$\mathbb{E}[\|\nabla f(\hat{\mathbf{x}}_t)\|_2^2] \leq \frac{4L(f(\mathbf{x}_0) - f^*)}{t} + \frac{\epsilon^2}{2}, \quad (1)$$

where  $\hat{\mathbf{x}}_t$  is chosen uniformly at random from  $\mathbf{x}_0, \dots, \mathbf{x}_{t-1}$ . Thus, Eq. (1) implies that taking  $t = \lceil \frac{8L(f(\mathbf{x}_0) - f^*)}{\epsilon^2} \rceil$  yields  $\mathbb{E}[\|\nabla f(\hat{\mathbf{x}}_t)\|_2^2] \leq \epsilon^2$ .

Sample Complexity Bound:

$$\sum_{k=0}^{t-1} B_k = \frac{2\sigma^2}{\epsilon^2} t = \left\lceil \frac{16L(f(\mathbf{x}_0) - f^*)\sigma^2}{\epsilon^4} \right\rceil = O(\epsilon^{-4})$$

- Optimal up to constant factors (see [Arjevani et al. 2019] for lower bound)

# Mini-Batching SGD as Gradient Descent with Error

- SGD with mini-batch:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s_k}{B_k} \sum_{i=1}^{B_k} \nabla f(\mathbf{x}, \xi_i)$$

- This can be viewed as a “gradient descent with error”

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k (\nabla f(\mathbf{x}_k) + \mathbf{e}_k)$$

, where  $\mathbf{e}_k$  is the difference between approximation and true gradient

- By setting  $s_k = 1/L$ , it follows from descent lemma that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \underbrace{\frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2}_{\text{good}} + \underbrace{\frac{1}{2L} \|\mathbf{e}_k\|^2}_{\text{bad}}$$

# Mini-Batching SGD as Gradient Descent with Error

- SGD progress bound with  $s_k = 1/L$  and error is:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \underbrace{\frac{1}{2L} \|\nabla f(\mathbf{x}_k)\|^2}_{\text{good}} + \underbrace{\frac{1}{2L} \|\mathbf{e}_k\|^2}_{\text{bad}}$$

- Relationship between “error-free” rate and “with error” rate:
  - If “error-free” rate is  $O(1/k)$ , you maintain this rate if  $\|\mathbf{e}_k\|^2 = O(1/k)$
  - If “error-free” rate is  $O(\rho^k)$ , you maintain this rate if  $\|\mathbf{e}_k\|^2 = O(\rho^k)$
  - If error goes to zero more slowly, error vanishing rate is the “bottleneck”
- So, need to know how batch-size  $B_k$  affects  $\|\mathbf{e}_k\|^2$

# Mini-Batching SGD as Gradient Descent with Error

- Sample with replacement:

$$\mathbb{E}[\|\mathbf{e}_k\|^2] = \frac{1}{B_k}\sigma^2,$$

where  $\sigma^2$  is the variance of the stochastic gradient norm (i.e., doubling the batch-size cuts the error in half)

- Sample without replacement (from a dataset of size  $N$ ):

$$\mathbb{E}[\|\mathbf{e}_k\|^2] = \frac{N - B_k}{N - 1} \frac{1}{B_k}\sigma^2,$$

i.e., driving error to zero as batch size approaches  $N$

- Growing batch-size:

- ▶ For  $O(\rho^k)$  linear convergence: need  $B_{k+1} = B_k/\rho$
- ▶ For  $O(1/k)$  sublinear convergence: need  $B_{k+1} = B_k + \text{const.}$

# Mini-Batching SGD as Gradient Descent with Error

- SGD with mini-batch:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s_k}{B_k} \sum_{i=1}^{B_k} \nabla f(\mathbf{x}, \xi_i)$$

- For a fixed  $B_k$ : sublinear convergence rate
  - Fixed step-size: sublinear convergence to an error ball around a stationary point
  - Diminishing step-size: sublinear convergence to a stationary point
- Can grow  $B_k$  to achieve faster rate:
  - Early iterations: cheap SG iterations
  - Later iterations: Use larger batch-sizes (no need to play with step-sizes)

## Next Class

### Variance-Reduced First-Order Methods