ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-4: Stochastic Gradient Descent

Jia (Kevin) Liu

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

Spring 2022

Outline

In this lecture:

- Noisy unbiased gradient
- Stochastic gradient method
- Convergence results

Unbiased Stochastic Gradient

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

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

  holds almost surely.

Stochastic Gradient Descent Method

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

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

- 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 $f_{\text{best}}^{(k)} \triangleq \min_{i=1,\ldots,k}\{f(\mathbf{x}_i)\}$ and $\|\nabla f_{\text{best}}^{(k)}\| \triangleq \min_{i=1,\ldots,k}\{\|\nabla f(\mathbf{x}_i)\|\}$

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 θ. 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 distr. to X if $\lim_{n\to\infty} F_n(X_n) = F(X)$, when F_n and F are cdf of X_n and X, resp. Denoted as : $X_n \xrightarrow{D} X$.

2. Convergence in prob. to r.v. :

$\{X_n\}$ converges in prob. to a r.v. X if $\forall \varepsilon > 0$, $\lim_{n\to\infty} \Pr\{|X_n - X| > \varepsilon\} = 0$. Denoted as : $X_n \xrightarrow{P} X$.

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 if $\Pr\{\lim_{n\to\infty} X_n = X\} = 1$. Denoted as $X_n \xrightarrow{a.s.} X$.

4. Convergence in expectation : Given $r \geq 1$. $\{X_n\}$ converges in r-th mean to r.v. X if r-th absolute moments $\mathbb{E}\{|X_n|^r\}$ and $\mathbb{E}\{|X|^r\}$ exist, and $\lim_{n\to\infty} \mathbb{E}\{|X_n - X|^r\} = 0$. Denoted as $X_n \xrightarrow{L^r} X$.

* $r=1$: X_n converges in mean to X.
* $r=2$: — — — - - - mean square to X.

strongest

$$\begin{array}{ccccc} \xrightarrow{a.s.} & \Rightarrow & \xrightarrow{P} & \Rightarrow & \xrightarrow{D} \\ & & \Uparrow & & \\ \xrightarrow{L^s} & \Rightarrow & \xrightarrow{L^r} & & \end{array}$$

$s > r \geq 1$

Markov ineq:

X: non-neg. r.v. For some $a > 0$

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

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

$$\mathbb{E}\left[\|z_1 + \cdots + z_n\|^2\right] \leq \mathbb{E}\left[\|z_1\|^2 + \cdots + \|z_n\|^2\right]$$

* – – – – – – – – not necc. indep., we have

$$\mathbb{E}\left[\|z_1 + \cdots + z_n\|^2\right] \leq n\,\mathbb{E}\left[\|z_1\|^2 + \cdots + \|z_n\|^2\right]$$

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

not needed.

$$s_k > 0,\ \forall k, \quad \boxed{\sum_{k=1}^{\infty} s_k^2 < \infty,} \quad \sum_{k=1}^{\infty} s_k = \infty$$

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

Convergence of SGD (Convex)

- Convergence in expectation:

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

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

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

- Almost sure convergence

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

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

Thm: If $\mathbb{E}\{\|\tilde{g}_k\|\} \le G$, $\forall k$ $\mathbb{E}\{\|x_1 - x^*\|\} \le R$, and step-sizes $\{s_k\}_{k=1}^{\infty}$ satisfy: $s_k > 0, \forall k$, $\sum_{k=1}^{\infty} s_k^2 = B < \infty$, $\sum_{k=1}^{\infty} s_k \to \infty$, then:

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

Proof. Consider the conditional expected square Eclidean dist:

$$\mathbb{E}\{\|x_{k+1} - x^*\|_2^2 \,|\, x_k\} = \mathbb{E}\{\|x_k - s_k\tilde{g}_k - x^*\|^2 \,|\, x_k\}$$

$$= \mathbb{E}\{\|x_k - x^*\|^2 + s_k^2\|\tilde{g}_k\|^2 - 2s_k\tilde{g}_k^T(x_k - x^*) \,|\, x_k\}$$

$$= \|x_k - x^*\|^2 + s_k^2\,\mathbb{E}[\|\tilde{g}_k\|^2 \,|\, x_k] \underbrace{- 2s_k\,\mathbb{E}\{\tilde{g}_k \,|\, x_k\}^T(x_k - x^*)}_{(*)} \quad (1).$$

$$\mathbb{E}\{\tilde{g}_k \,|\, x_k\} = \nabla f(x_k)$$

By convexity: $f(x^*) \ge f(x_k) + \underbrace{\mathbb{E}\{\tilde{g}_k \,|\, x_k\}^T}_{\nabla f(x_k)}(x^* - x_k)$

$$\Rightarrow -\mathbb{E}\{\tilde{g}_k \,|\, x_k\}^T(x_k - x^*) \le -(f(x_k) - f^*)$$

Therefore: $(1) \le \|x_k - x^*\|^2 + s_k^2\,\mathbb{E}[\|\tilde{g}_k\|^2 \,|\, x_k] - 2s_k(f(x_k) - f^*)$.

Note: x_{k+1} only dep. x_k and cond. indep. of $x_{k-1}, \dots, x_1$

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 \,|\, x_k] = \mathbb{E}[\|x_{k+1} - x^*\|^2 \,|\, x_k \dots x_1]$$

Take expectation over joint distr. of $\{x_k, \dots, x_1\}$. yields:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \mathbb{E}[\|x_k - x^*\|^2] - 2s_k[\mathbb{E}[f(x_k) - f^*]] + s_k^2\,\mathbb{E}[\|\tilde{g}_k\|^2]$$

Apply the process recursively. noting $\mathbb{E}\{\|\tilde{g}_k\|_2^2\} \le G^2$:

$$\mathbb{E}[\|x_{k+1} - x^*\|^2] \le \mathbb{E}[\|x_1 - x^*\|^2] - 2\sum_{i=1}^{k} s_i(\underbrace{\mathbb{E}[f(x_i)]}_{\ge \min_{i=1,\dots,k}\mathbb{E}[f(x_i)]} - f^*) + G^2\sum_{k=1}^{\infty} s_k^2$$

$$\Rightarrow \min_{i=1,\dots,k}\{\mathbb{E}[f(x_i)] - f^*\} \le \frac{R^2 + G^2 B}{2\sum_{i=1}^{k} s_i} \to 0 \text{ as } k \to \infty.$$

∞

Claim: The fn $g(y) \triangleq \min_{i=1,\cdots,k} \{y_i\}$ is concave. $\forall y \in \mathbb{R}^k$ (HW).

Thus, by Jensen's ineq:

$$\mathbb{E}[f^{(k)}_{best}] = \mathbb{E}\Big\{\underbrace{\min_{i=1,\cdots,k} f(x_i)}_{\text{concave,}}\Big\} \underset{\text{Jensen}}{\le} \min_{i=1,\cdots,k} \mathbb{E}\{f(x_i)\} \to f^*.$$

i.e., convergence in $\mathbb{E}$ is done. $\blacksquare$

Convergence in Expectation and Probability (Convex)

Proof Sketch:

- Key quantity: Expected squared Euclidean distance to the optimal set. Let $\mathbf{x}^*$ be any minimzer 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

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

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

Convergence in Expectation and Probability (Convex)

- Jensen's inequality and concavity of minimum yields

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

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

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

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

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

Convergence Rate (Convex)

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

$\leftarrow O(1)$ $\quad$ $\leftarrow \log(t)$

$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \to \infty$ (Nicole Oresme).

$1 + \frac{1}{2} + \underbrace{\frac{1}{3} + \frac{1}{4}} + \underbrace{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}} + \cdots$

$> 1 + \frac{1}{2} + \underbrace{\frac{1}{4} + \frac{1}{4}} + \underbrace{\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}} + \cdots$

$= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots \to \infty$

$H_n = \sum_{k=1}^{n} \frac{1}{k}$

$\gamma + \log n < H_n < \gamma + \log(n+1)$ $\quad$ γ: Euler–Mascheroni const. $\gamma \approx 0.57721\cdots$

- Classical diminishing step-sizes $s_k = \alpha/k$ for some $\alpha > 0$: $\sum_{k=1}^{t} s_k = O(\log(t))$ and $\sum_k s_k^2 = O(1)$. So convergence rate is $O(1/\log(t))$

  $\Theta(\log(t))$ $\quad$ $1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots = \frac{\pi^2}{6}$ $\quad$ Basel problem. (3Blue1Brown).

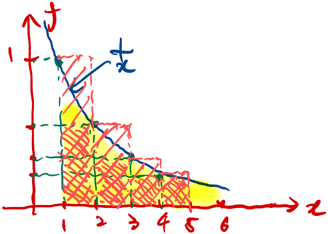
- Diminishing step-sizes $s_k = \alpha/\sqrt{k}$ for some $\alpha > 0$: $\sum_k s_k = O(\sqrt{t})$ and $\sum_k s_k^2 = O(\log(t))$. So convergence rate is $O(\log(t)/\sqrt{t}) = \tilde{O}(1/\sqrt{t})$

- 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 $O(1/t) + O(\alpha)$

integral test: $\sum f(n) \sim \int f(x)\,dx$

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

$\int_1^n \frac{1}{x}\,dx = \log(n) \to \infty$ as $n \to \infty$.



Convergence Rate (Strongly Convex)

Theorem 1 (Optimality Gap)

If $f(\cdot)$ is μ-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 $\mathbf{x}^*$

Proof:

Strong Convexity:

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

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

Add together $\Rightarrow$ $[\nabla f(y) - \nabla f(x)]^T (y-x) \geq \mu \|y-x\|^2$.

Recall: $\mathbb{E}[\|x_{k+1}-x^*\|^2 | x_k] = \|x_k - x^*\|^2 + s_k^2 \mathbb{E}[\|g_k\|^2 | x_k]$

$$\underline{- 2 s_k \mathbb{E}[g_k | x_k]^T (x_k - x^*)}.$$

$$\leq 2 s_k \mu \mathbb{E}[\|x_k - x^*\|^2 | x_k].$$

Taking full expectation:

$$\mathbb{E}[\|x_{k+1}-x^*\|^2] \leq \mathbb{E}[\|x_k - x^*\|^2] + s_k^2 \underbrace{\mathbb{E}[\|g_k\|^2]}_{\leq \sigma^2} - 2 s_k \mu \mathbb{E}[\|x_k - x^*\|^2]$$

$$= (1 - 2 s_k \mu) \mathbb{E}[\|x_k - x^*\|^2] + s_k^2 \sigma^2 \qquad (1)$$

Applying (1) recursively from $k-1$ down to 1, letting $s_k = s < \frac{2}{\mu}$, $\forall k$, using the bounding of geometric series:

$$\mathbb{E}[\|x_k - x^*\|^2] \leq (1 - 2\mu s)^k \|x_0 - x^*\| + \frac{s\sigma^2}{2\mu}.$$

1° if $\sigma^2 = 0$, linear convergence of GD (to x^*).

2° if $\sigma^2 > 0$, $\cdots$ ———————— of SGD to a $\frac{s\sigma^2}{2\mu}$ – neighborhood.

Convergence Rate (Nonconvex) – Finite Sum

$f_i(\underline{\underline{W}}) := (y_i - \underline{\underline{W}}_L \sigma(\underline{\underline{W}}_{L-1} \sigma(\cdots \sigma(\underline{\underline{W}}_2 \sigma(\underline{\underline{W}}_1 \underline{\underline{x}}_i)) \cdots))^2.$

$(\underline{\underline{x}}_i, y_i)$

- Consider the following finite-sum minimization

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

\# of samples

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 g: one sample at a time
  - ∇f is differentiable with L-Lipschitz continuous gradients (L-smooth)
  - $\mathbb{E}[\|\nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$ for some σ^2 and all $\mathbf{x}$ (bounded gradient, can be relaxed)

relaxed to: $\mathbb{E}\left[\|\nabla f_i(\underline{x}_k) - \nabla f(\underline{x}_k)\|^2\right] \leq \sigma^2$

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

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

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

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

Remark:

Proof: Randomly select i_k from $\{1, 2, \cdots, N\}$.

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

With $P_r(i_k = i) = \frac{1}{N}$ (uniformly sample at random), then.

SG is unbiased estimation of grad.

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

Recall the descent lemma:

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

Plug in SGD iteration: $\underline{x}_{k+1} - \underline{x}_k = -s_k \nabla f_{i_k}(\underline{x}_k)$.

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

Now, take expectation w.r.t. i_k assuming $P_r(i_k = i) = \frac{1}{N}$.

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

As in GD, re-arrange to get grad norm on LHS:

$$s_k \|\nabla f(x_k)\|^2 \le \mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})] + \frac{L s_k^2}{2}\sigma^2. \qquad (1).$$

Sum (1) from 1 to t, and use iterative expectation to get:

$$\sum_{k=1}^{t} s_{k-1} \underbrace{\mathbb{E}\left[\|\nabla f(x_{k-1})\|^2\right]}_{\ge \min\limits_{k=0,\cdots,t-1} \mathbb{E}[\|\nabla f(x_k)\|^2]} \le \underbrace{\sum_{k=1}^{t}\left[\mathbb{E}[f(x_{k-1})] - \mathbb{E}[f(x_k)]\right]}_{\text{telescoping}} + \sum_{k=1}^{t} s_{k-1}^2 \frac{L\sigma^2}{2}$$

$$\Rightarrow \min_{k=0,\cdots,t-1} \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \sum_{k=0}^{t-1} s_k \le f(x_0) - \underbrace{\mathbb{E}[f(x_t)]}_{\ge f^*} + \frac{L\sigma^2}{2}\sum_{k=0}^{t-1} s_{k-1}^2$$

$$\Rightarrow \min_{k=0,\cdots,t-1} \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \le \frac{f(x_0) - f^*}{\sum_{k=0}^{t-1} s_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} s_k^2}{\sum_{k=0}^{t-1} s_k}. \qquad \blacksquare$$

Convergence Rate (Nonconvex) - Finite Sum+Time Oracle

Theorem 3 ([Ghadimi & Lan '13])

Suppose $f(\cdot)$ is L-smooth and has σ-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

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

Theorem 3 ([Ghadimi & Lan '13])

Suppose $f(\cdot)$ is L-smooth and has σ-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

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

Proof.

We have shown:

$$\min_{k=0,\dots,t-1} \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \le \frac{f(x_0) - f^*}{\sum_{k=0}^{t-1} s_k} + \frac{L\sigma^2}{2} \frac{\sum_{k=0}^{t-1} s_k^2}{\sum_{k=0}^{t-1} s_k}. \quad (1).$$

$$s_k = \frac{c}{\sqrt{T}} \Rightarrow \sum_{k=0}^{T-1} s_k = T \cdot \frac{c}{\sqrt{T}} = c\sqrt{T}.$$

$$\sum_{k=0}^{T-1} s_k^2 = T \cdot \frac{c^2}{T} = c^2$$

$$(1) \Rightarrow \min_{k=0,\dots,T-1} \mathbb{E}\left[\|\nabla f(x_k)\|^2\right] \le \frac{f(x_0) - f^*}{c\sqrt{T}} + \frac{L\sigma^2}{2} \frac{c^{\cancel{2}}}{\cancel{c}\sqrt{T}}$$

$$= \frac{1}{\sqrt{T}}\left(\frac{f(x_0) - f^*}{c} + \frac{L\sigma^2 c}{2}\right). \quad (2).$$

Young's Ineq: $ab \le \frac{a^p}{p} + \frac{b^q}{q}$, for $\frac{1}{p} + \frac{1}{q} = 1$. (pick $\sqrt{a}$, $\sqrt{b}$ as "a" & "b", $p=q=2$).

$2\sqrt{ab} \le a+b$, w/ equality holding for $a=b$.

Minimizing const in (2). $\Rightarrow$ $c = \sqrt{\frac{2(f(x_0)-f^*)}{L\sigma^2}}$

The stated result immediately follows after plugging c in (2). ∎

Recall UB:

$$\frac{1}{\sqrt{T}}\left(\frac{\overbrace{f(x_0)-f^*}^{\Delta f}}{c} + \frac{L\sigma^2 c}{2}\right) = \frac{2\Delta f + L\sigma^2 \overbrace{c^2}^{\sim O(1)}}{\underbrace{2c}_{\sim O\left(\frac{1}{\sqrt{L}}\right)}}$$

$$\approx O(\sqrt{L}).$$

as long as $L = o(\sqrt{T})$

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 ξ is a random vaiable with distribution $\mathcal{D}$.

- Consider using SGD to solve this problem under the following assumptions:
  - $f(\cdot)$ is nonconvex and bounded from below
  - ∇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 $\xi_1, \ldots, \xi_{B_k}$ 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

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

where $\hat{\mathbf{x}}_t$ is chosen uniformly at random from $\mathbf{x}_0, \ldots, \mathbf{x}_{t-1}$. 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)

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 $s_k \le \frac{1}{2L}$, $\forall k$, satisfies

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

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

Proof: ① WTS: when $B_k = B = \max\{1, \frac{2\sigma^2}{\varepsilon^2}\}$, we have:

$$\mathbb{E}\left[\|g(x) - \nabla f(x)\|_2^2 \,\middle|\, x\right] \le \frac{\varepsilon^2}{2}.$$

From def, $g(x) = \frac{1}{B}\sum_{i=1}^{B} \nabla f_i(x, \xi_i)$, where $\xi_1, \ldots, \xi_B$ are i.i.d. sampled from $\mathcal{D}$. Thus:

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

$$\mathbb{E}\left[\|g(x) - \nabla f(x)\|^2 \,\middle|\, x\right] = \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} \nabla f_i(x,\xi_i) - \nabla f(x)\right\|^2 \,\middle|\, x\right]$$

$$= \frac{1}{B^2}\sum_{i=1}^{B} \mathbb{E}_{\xi_i}\Big[\underbrace{\|\nabla f(x,\xi_i) - \nabla f(x)\|^2}_{\le \sigma^2} \,\Big|\, x\Big] \le \frac{\sigma^2}{B} \le \frac{\varepsilon^2}{2}$$

for $B = \max\{1, \frac{2\sigma^2}{\varepsilon^2}\}$.

②. Consider descent lemma:

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2. \qquad (1).$$

Use Fenchel-Young's Ineq:

$$\underline{a}^T \underline{b} \le \frac{1}{2\alpha} \|\underline{a}\|^2 + \frac{\alpha}{2} \|\underline{b}\|^2$$

Let X be some real topo. space, and X^* be its dual space.

$\langle \cdot, \cdot \rangle : X^* \times X \to \mathbb{R}$.

Convex Conjugate: For a fn: $X \to \mathbb{R} \cup \{-\infty, +\infty\}$, its convex conjugate is the fn: $f^* : X^* \to \mathbb{R} \cup \{-\infty, +\infty\}$, where value at $\underline{x}^* \in X^*$ is defined as:

$$f^*(\underline{x}^*) \triangleq \sup \{ \langle \underline{x}^*, \underline{x} \rangle - f(\underline{x}) : \underline{x} \in X \}.$$

$$(\text{or } f^*(\underline{x}^*) \triangleq - \inf \{ f(\underline{x}) - \langle \underline{x}^*, \underline{x} \rangle : \underline{x} \in X \}.$$

$\Rightarrow \underline{x} \in X,\ p \in X^*,\ \langle p, \underline{x} \rangle \le f(\underline{x}) + f^*(p)$.

Let $f(\cdot) = \|\cdot\|_2$, $f^*(\cdot) = \|\cdot\|_2$

add & subtract g

$$(1) \Rightarrow f(\underline{x}_{k+1}) \le f(\underline{x}_k) + g_k^T(\underline{x}_{k+1} - \underline{x}_k) + (\nabla f(\underline{x}_k) - g_k)^T (\underline{x}_{k+1} - \underline{x}_k) + \frac{L}{2} \|\underline{x}_{k+1} - \underline{x}_k\|^2$$

by Fench-Young with $\alpha = \frac{1}{2s_k}$

$$\le f(\underline{x}_k) - s_k \|g_k\|^2 + s_k \|\nabla f(\underline{x}_k) - g_k\|^2 + \left(\frac{1}{4s_k} + \frac{L}{2}\right) \|\underline{x}_{k+1} - \underline{x}_k\|^2.$$

$$= f(\underline{x}_k) - s_k \left[1 - \left(\frac{1}{4} + \frac{Ls_k}{2}\right)\right] \|g_k\|^2 + s_k \|\nabla f(\underline{x}_k) - g_k\|^2. \qquad (2)$$

Since $s_k \le \frac{1}{2L} \Rightarrow Ls_k \le \frac{1}{2} \Rightarrow \frac{Ls_k}{2} \le \frac{1}{4} \Rightarrow \frac{1}{4} + \frac{Ls_k}{2} \le \frac{1}{2}$.

$$\Rightarrow -\left(\frac{1}{4} + \frac{Ls_k}{2}\right) \ge -\frac{1}{2} \Rightarrow \left(1 - \left(\frac{1}{4} + \frac{Ls_k}{2}\right)\right) \ge \frac{1}{2}$$

Thus, (2) $\Rightarrow$ $$f(x_{k+1}) \le f(x_k) - \frac{s_k}{2}\|g_k\|^2 + s_k\|\nabla f(x_k) - g_k\|^2. \quad (3).$$

③: Take cond. expectation on both sides:

$$\mathbb{E}\left[f(x_{k+1}) \,\middle|\, x_k\right] \le f(x_k) - \frac{s_k}{2}\mathbb{E}\left[\|g_k\|^2 \,\middle|\, x_k\right] + s_k \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \,\middle|\, x_k\right]$$

add & subtract $\nabla f(x_k)$. using unbiasedness of g_k.

$$\overset{\text{add \& subtract } \nabla f(x_k)}{\underset{\text{using unbiasedness of } g_k}{=}} f(x_k) - \frac{s_k}{2}\left[\|\nabla f(x_k)\|_2^2 + \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \,\middle|\, x_k\right]\right]$$

$$+ s_k \mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \,\middle|\, x_k\right]$$

$$= f(x_k) \underbrace{- \frac{s_k}{2}\|\nabla f(x_k)\|^2}_{\text{good}} + \underbrace{\frac{s_k}{2}\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2 \,\middle|\, x_k\right]}_{\text{bad, but ctrl'ble by batching.}} \quad (4).$$

Take full expectation on both sides, choosing $s_k = \frac{1}{2L}$, and summing (4) for $k = 0, \cdots, t-1$, we have:

$$\frac{1}{t}\sum_{k=0}^{t-1}\mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \le \frac{4L}{t}\sum_{k=0}^{t-1}\left(\mathbb{E}[f(x_k)] - \mathbb{E}[f(x_{k+1})]\right) + \frac{\varepsilon^2}{2}.$$

$$= \frac{4L}{t}\left(f(x_0) - f^*\right) + \frac{\varepsilon^2}{2}.$$

Finally, choose output $\hat{x}$ uniformly at random from $\{x_0 \cdots x_{t-1}\}$.

we have the stated result. $\blacksquare$

Mini-Batching SGD as Gradient Descent with Error

- SGD with mini-batcch:

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

- 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 σ^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-batcch:

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

- For a fixed B_k: sublinear convergence rate
  - Fixed step-size: sublinear convergence to an error ball around a stationary point
  - Diminishing step-size: sublienar 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