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 $\tilde{\mathbf{g}} \in \mathbb{R}^n$ is a unbiased stochastic gradient if it can be written as $\tilde{\mathbf{g}} = \underline{\mathbf{g}} + \mathbf{n}$, where $\mathbf{g}$ is the true gradient and $\mathbb{E}[\mathbf{n}] = \mathbf{0}$

  $\mathbb{E}[\tilde{\mathbf{g}}]$ $\mathbb{E}[\mathbf{n}]=0$

- $\mathbf{n}$ 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 <u>subgradient</u> 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.

  $\Uparrow$ $\not\Downarrow$

  convex $f(\cdot)$ convex

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 $\lim_{n\to\infty} F_n(X_n) = F(X)$, 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 $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\{|X_n - X| > \varepsilon\} = 0 .$$

Denote 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 the r-th abs. moments

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

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

Denote as $X_n \xrightarrow{L^r} X$

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

strongest

$$\xrightarrow{a.s.} \;\Rightarrow\; \xrightarrow{p} \;\Rightarrow\; \xrightarrow{D}$$

$$\Uparrow$$

$$\xrightarrow{L^s} \;\Rightarrow\; \xrightarrow{L^{r\,(=1)}}$$

if $s > r \geq 1$

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,

$$\mathbb{E}\left[\|Z_1 + \cdots + Z_n\|^2\right] \leq \mathbb{E}\left[\|Z_1\|^2 + \cdots + \|Z_n\|^2\right] \quad \left(\text{implies } \mathbb{E}[\|X_i\|^2] < \infty\right).$$

* For r.v. $Z_1 \cdots Z_n$ that are not necc. indep.

$$\mathbb{E}\left[\|Z_1 + \cdots + Z_n\|^2\right] \leq n\, \mathbb{E}\left[\|Z_1\|^2 + \cdots + \|Z_n\|^2\right]$$

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

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

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

Convergence of SGD (Convex)

Asymptotic

- 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

(Convex)

Thm: If $\mathbb{E}[\|\tilde{g}_k\|_2] \le G$, $\forall k$, $\mathbb{E}\{\|x_1 - x^*\|\} \le R$ and step-sizes $\{s_k\}_{k=1}^{\infty}$ satisfy: $s_k > 0$, $\forall k$, $\sum_{k=1}^{\infty} s_k^2 < \infty$, $\sum_{k=1}^{\infty} s_k \to \infty$.

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

Proof. Consider cond. expectation square Euclidean dist:

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

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

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

Note: $f(x^*) \ge f(x_k) + \mathbb{E}\{\tilde{g}_k | x_k\}^T(x^* - x_k)$

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

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

Note: from SGD dynamic, x_{k+1} only dep. x_k ($x_{k+1} = x_k - s_k\tilde{g}_k$) and indep. of $x_{k-1} \cdots x_1$.

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

Take expectation over joint distr. of $\{x_k \cdots x_1\}$ yields

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

Apply this process recursively:

$$\mathbb{E}\left[\|\underline{x}_{k+1}-\underline{x}^*\|^2\right] \le \underbrace{\mathbb{E}\left[\|\underline{x}_1-\underline{x}^*\|^2\right]}_{\le R^2} - 2\sum_{i=1}^{k} s_i\Big(\underbrace{\mathbb{E}[f(\underline{x}_i)-f^*]}_{\underbrace{\ge \min_{i=1\cdots k} f(\underline{x}_i)-f^*}}\Big)$$

$$+ \; \underbrace{G^2\sum_{i=1}^{k} s_k^2}_{\le B}$$

$$\Rightarrow \min_{i=1,\cdots k}\left\{\mathbb{E}[f(\underline{x}_i)]-f^*\right\} \le \frac{R^2+G^2B}{\underbrace{2\sum_{i=1}^{k} s_i}_{\to\infty}} \to 0 \text{ as } k\to\infty.$$

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

Then Jensen's ineq.:

* If f is convex: $f(\mathbb{E}X) \le \mathbb{E}f(X)$
* — — concave: $\ge$

$$\mathbb{E}[f^{(k)}_{best}] = \mathbb{E}\Big[\underbrace{\min_{i=1\cdots k} f(\underline{x}_i)}_{\text{concave}}\Big] \underset{\text{Jensen's}}{\le} \min_{i=1\cdots k}\mathbb{E}[f(\underline{x}_i)] \underset{\text{done earlier}}{\longrightarrow} f^*$$

∎

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

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

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

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

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

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

Convergence Rate (Convex)

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

$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\cdots$ (Nicole Oresme)

$=1+\frac{1}{2}+\left(\frac{1}{3}+\frac{1}{4}\right)+\left(\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\right)+\cdots$

$>1+\frac{1}{2}+\left(\frac{1}{4}+\frac{1}{4}\right)+\left(\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}\right)+\cdots$

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

$\Theta(\log(t))$

- 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 $O(1/\log(t))$

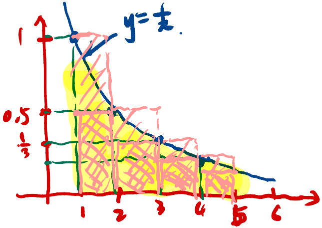
$\int_N^\infty f(x)\,dx \le \sum_{n=N}^{\infty} f(n) \le f(N) + \int_N^\infty f(x)\,dx$ $\quad\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots=\frac{\pi^2}{6}$ $\quad$ integral test

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

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

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



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 $\mathbf{x}^*$

strongly convex:

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

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

Add together: $[\underbrace{\nabla f(y)}_{\mathbb{E}[g_k|x_k] \to \nabla f(x_k)} - \underbrace{\cancel{\nabla f(x)}}_{\nabla f(x^*)=0}]^T (\overset{x_k}{y} - \overset{x^*}{x}) \geq \mu \|y-x\|^2$.

Recall: $\mathbb{E}\left[\|x_{k+1}-x^*\|\,\middle|\,x_k\right] \leq \|x_k - x^*\|^2 + s_k^2\,\mathbb{E}\left[\|\tilde{g}_k\|^2\,\middle|\,x_k\right]$
$$-2s_k\,\mathbb{E}\left[\tilde{g}_k\,\middle|\,x_k\right]^T (x_k - x^*)$$

Taking full expectation:

$$\mathbb{E}\left[\|x_{k+1}-x^*\|^2\right] \leq \mathbb{E}\left[\|x_k - x^*\|^2\right] + s_k^2\,\underbrace{\mathbb{E}\left[\|\tilde{g}_k\|^2\right]}_{\leq \sigma^2} - 2\mu s_k\,\mathbb{E}\left[\|x_k - x^*\|^2\right]$$

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

$$= (1-2\mu s_k)\,\mathbb{E}\left[\|x_k - x^*\|^2\right] + s_k^2\sigma^2 \qquad (1). \qquad s < \frac{2}{\mu}$$

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

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

(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
  - ∇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)

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

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

"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

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

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

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

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

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

$$x_{k+1} = x_k - s_k \nabla f_{i_k}(x_k).$$

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

Recall the descent lemma in GD.

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

Plugging in SGD iteration yields:

$$f(x_{k+1}) \le f(x_k) - s_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{Ls_k^2}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Take expectation w.r.t. i_k

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}\left[f(x_k) - s_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{Ls_k^2}{2}\|\nabla f_{i_k}(x_k)\|^2\right]$$

$$= \mathbb{E}[f(x_k)] - s_k\|\nabla f(x_k)\|^2 + \frac{Ls_k^2}{2}\underbrace{\mathbb{E}\left[\|\nabla f_{i_k}(x_k)\|^2\right]}_{\le \sigma^2}$$

$$\le \mathbb{E}[f(x_k)] \underbrace{- s_k\|\nabla f(x_k)\|^2}_{\text{good}} + \underbrace{\frac{Ls_k^2}{2}\sigma^2}_{\text{bad}}.$$

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

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

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

$$\sum_{k=1}^{t-1} s_{k-1} \underbrace{\mathbb{E}[\|\nabla f(x_{k-1})\|^2]}_{\geq \min_{k=0,\dots,t-1}\{\mathbb{E}[\|\nabla f(x_{k-1})\|^2]\}} \leq \underbrace{\sum_{k=1}^{t} [\mathbb{E}[f(x_{k-1})] - \mathbb{E}[f(x_k)]]}_{\text{telescope.}} + \frac{L\sigma^2}{2}\sum_{k=0}^{t-1} s_k^2$$

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

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

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

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:

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

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

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

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

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

Young's Ineq. $ab \le \frac{a^p}{p} + \frac{b^q}{q}$, with $\frac{1}{p} + \frac{1}{q} = 1$.

By picking "a" $= \sqrt{a}$, "b" $= \sqrt{b}$, $p = q = 2$, $a + b \ge 2\sqrt{ab}$ "=" iff $a = b$.

By forcing $\frac{f(x_0) - f^*}{c} = \frac{L\sigma^2 c}{2} \Rightarrow c = \sqrt{\frac{2(f(x_0) - f^*)}{L\sigma^2}}$

Then, the stated result follows. ∎

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)
  - $\mathbb{E}_{\xi}[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$ and $\mathbb{E}_{\xi}[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2] \leq \sigma^2$

- 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

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

$\leq \frac{\epsilon^2}{2}$

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

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

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

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

Note: $g(x) = \frac{1}{B}\sum_{i=1}^{B} \nabla f_i(x, \xi_i)$, where $\xi_1 \cdots \sim \xi_B$ are i.i.d. sampled from D.

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

expand indep.

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

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

$$B = \left\lceil \frac{2\sigma^2}{\varepsilon^2} \right\rceil$$

Recall "Descent lemma":

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

$\uparrow \pm g_k$.

add & subtract g_k

$$= f(x_k) + g_k^T (\underbrace{x_{k+1} - x_k}_{s_k g_k}) + (\nabla f(x_k) - g_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$\le \|\nabla f(x_k) - g_k\|^2 + \frac{1}{4 s_k}\|x_{k+1} - x_k\|^2 \quad \left(\begin{array}{l}\text{by picking } \alpha = \frac{1}{2} \\ \text{in F-Y ineq}\end{array}\right)$$

$x_{k+1} - x_k \,/\!/\, s_k g_k$.

$$= f(x_k) - s_k\|g_k\|^2 + s_k \underline{(\nabla f(x_k) - g_k)^T (x_{k+1} - x_k)} + \frac{L}{2} s_k^2 \|g_k\|^2 \quad (1)$$

Fenchel–Young's Ineq: $\underline{a}^T \underline{b} \le \frac{1}{2\alpha}\|a\|^2 + \frac{\alpha}{2}\|b\|^2$

"Convex Conjugate":

Let X be topo. space, and X^* be dual space

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

Def (Convex Conjugate): For a fn $f: X \to \mathbb{R} \cup \{-\infty, +\infty\}$, its convex conjugate is the fn: $f^*: X^* \to \mathbb{R} \cup \{-\infty, +\infty\}$, where at $x^* \in X^*$, we have:

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

$$\Rightarrow x \in X,\ p \in X^*,\ \underset{\text{inner}}{\langle p, x \rangle} \le \underset{\|\cdot\|_2}{f(x)} + \underset{\|\cdot\|_2}{f^*(p)}.$$

$$(1) \Rightarrow f(x_{k+1}) \le f(x_k) - s_k\|g_k\|^2 + s_k\|\nabla f(x_k) - g_k\|^2 + \left(\frac{1}{4 s_k} + \frac{L}{2}\right) s_k^2 \|g_k\|^2.$$

$$= f(x_k) - s_k\left[1 - \left(\frac{1}{4} + \frac{L s_k}{2}\right)\right]\|g_k\|^2 + s_k\|\nabla f(x_k) - g_k\|^2 \quad (2)$$

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

$$\Rightarrow -\left(\frac{1}{4} + \frac{L s_k}{2}\right) \ge -\frac{1}{2} \Rightarrow -\left[1 - \left(\frac{1}{4} + \frac{L s_k}{2}\right)\right] \le -\frac{1}{2}$$

Then, $(2) \Rightarrow f(x_{k+1}) \le f(x_k) \underbrace{- \frac{s_k}{2}\|g_k\|^2}_{\text{good}} + \underbrace{s_k\|\nabla f(x_k) - g_k\|^2}_{\text{bad}}$

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

$\pm \nabla f(x_k)$

add & subtract

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

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

$$= f(x_k) - \frac{s_k}{2}\|\nabla f(x_k)\|^2 + \underbrace{\frac{s_k}{2}\mathbb{E}\left[\overbrace{\|g_k - \nabla f(x_k)\|^2 \,\middle|\, x_k}^{\le \frac{\varepsilon^2}{2}}\right]}_{\le \frac{s_k \varepsilon^2}{4}} \qquad (3).$$

Taking full expectation on both sides, choosing $s_k \le \frac{1}{2L}$, summing (3) for $k = 0, \cdots, t-1$, we have:

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

$$= \frac{4L}{t}\left(f(x_0) - \underbrace{\mathbb{E}[f(x_t)]}_{\le -f^*}\right) + \frac{\varepsilon^2}{2}$$

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

Finally, choosing output $\hat{x}$ uniformly at random from $\{x_0 \cdots x_{t-1}\}$ we have $\mathbb{E}\left[\|\nabla f(\hat{x}_t)\|^2\right] \le \frac{4L(f(x_0) - f^*)}{t} + \frac{\varepsilon^2}{2}$ ∎

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 $B_{k+1} = B_k +$ const.

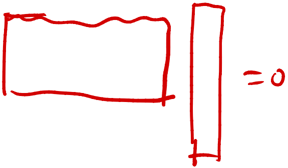
Mini-Batching SGD as Gradient Descent with Error

- 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