Stochastic Gradient Methods

My dog "Moreau"

$f_{\alpha}(\bar{x}) = \min f(x) + \frac{1}{2\alpha}\|x-\bar{x}\|^2$

Moreau Envelope

Previously gradient oracle

$$x \mapsto \nabla f(x)$$

Now we have stochastic gradient oracle

$$x \mapsto \underbrace{g(x)}_{\text{random variable/vector}}$$

"randomized independently each time we call"

such that

$$\mathbb{E}[g(x)] = \nabla f(x)$$ "unbiased estimator"

$$\operatorname{var}[g(x)] \le \sigma^2$$

$$:= \mathbb{E}[\|g(x) - \nabla f(x)\|_2^2]$$

$$= \mathbb{E}[\|g(x)\|^2] - \|\underbrace{\mathbb{E} g(x)}_{= \nabla f(x)}\|^2$$

Brief Probability Review:

Linearity of Expectation: For random variables $X_1, \dots, X_n$ and constants $\lambda_1, \dots, \lambda_n$

$$\mathbb{E}[\lambda_1 X_1 + \dots + \lambda_n X_n] = \lambda_1 \mathbb{E}[X_1] + \dots + \lambda_n \mathbb{E}[X_n]$$

Law of Total Expectation:

$$\mathbb{E}_{x_k}\Big[\mathbb{E}[f(x_{k+1}) \mid x_k]\Big] = \mathbb{E}[f(x_{k+1})]$$

Outline

1. Example of Stochastic Gradient Oracles
2. Nonconvex Guarantees : $\mathbb{E}\,\|\nabla f(x_k)\|^2$ small
3. Convex Guarantees : $\mathbb{E}\, f(x_k) - f(x^*)$ small
4. Improvements: Coordinate Methods,
   Acceleration,
   ~~Variable~~ Variance Reduction.

1. Examples

Example 1 "Coordinate Approach", $\min f(x)$, $f : \mathbb{R}^d \to \mathbb{R}$

Pick $i \in \{1, \dots, d\}$ uniformly at random

Set $g(x) = d \cdot \frac{\partial f}{\partial x_i}(x) \cdot e_i$

i^{th} basis vector

(Check unbiased oracle

$$\mathbb{E}[g(x)] = \frac{1}{d} \sum_i d \cdot \frac{\partial f}{\partial x_i}(x)\, e_i$$

$$= \sum_i \frac{\partial f}{\partial x_i}(x)\, e_i = \nabla f(x).)$$

Example 2 "Finite Sum"

$$\min f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

Pick $i \in \{1 \dots n\}$ uniformly

Set $g(x) = \nabla f_i(x)$

(Check unbiased oracle).

$$\min \frac{1}{n} \|Ax-b\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (a_i^T x - b_i)^2$$

Each (a_i, b_i) is a data point

a_i feature vector pictures of dogs or not

b_i measurement.

Example 3 (or 2.1). Infinite Sum / Expectation

There is a distribution of data points

$(a_i, b_i) \sim D$

$$\min \mathbb{E}_{(a_i,b_i)} \left[(a_i^T x - b_i)^2\right]$$

(Linearity Expections $\Rightarrow \mathbb{E} \nabla f(x, a_i, b_i) = \nabla \mathbb{E} f(x, a_i, b_i)$)

Example 4 (or 2.2) Better (Lower Variance) Oracles for finite sums.

Idea 1: Look at batches/minibatches of samples at each step

Pick $S \subseteq \{1, \dots, n\}$ with $|S| = K$

uniformly at random with or without replacement

$$g(x) = \frac{1}{K} \sum_{i \in S} \nabla f_i(x).$$

Idea 2: Variance Reduction, given $\tilde{x}$

Compute the full gradient $\nabla f(\tilde{x}) = \frac{1}{n} \sum \nabla f_i(\tilde{x})$

Pick $i \in \{1, \dots, n\}$ uniformly

$$g(x) = \nabla f(\tilde{x}) + \underbrace{\nabla f_i(x) - \nabla f_i(\tilde{x})}_{\text{small when } x - \tilde{x} \text{ small}}$$

(Check unbiased:

$$\mathbb{E}[g(x)] = \nabla f(\tilde{x}) + \underbrace{\mathbb{E}[\nabla f_i(x)]}_{= \nabla f(x)} - \underbrace{\mathbb{E}[\nabla f_i(\tilde{x})]}_{= \nabla f(\tilde{x})}$$

cancel

$$= \nabla f(x).\ )$$

SVRG [Johnson, Zhang, 2013].

2. Nonconvex Stochastic Gradient Method Analysis

Consider $x_{k+1} = x_k - \alpha_k g(x_k)$

independent at each iteration.

Aside: Not a descent method. For example,

$$g(x) = \begin{cases} 3 \nabla f(x) & \text{with prob } 1/2 \\ -\nabla f(x) & \text{with prob } 1/2 \end{cases}$$

Theorem Suppose f has L-Lipschitz gradient and $g(x)$ is unbiased estimator of $\nabla f(x)$ with variance σ^2. Then for any $0 \le \alpha_k \le 2/L$,

$$\mathbb{E}\left[\min_{i \le k} \|\nabla f(x_i)\|^2\right] \le \frac{f(x_0) - \min f + \frac{\sigma^2 L}{2} \sum_{i=1}^{k} \alpha_i^2}{\sum_{i=1}^{k} \alpha_i \left(1 - \frac{L\alpha_i}{2}\right)}.$$

Proof. Our Taylor Approximation Theorem ensures

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

Fixing x_k (just consider randomness in step k),

Linearity

$$\mathbb{E}[f(x_{k+1}) \mid x_k] \le f(x_k) - \alpha_k \mathbb{E}[\nabla f(x_k)^T g(x_k) \mid x_k] + \frac{L\alpha_k^2}{2} \mathbb{E}[\|g(x_k)\|^2 \mid x_k]$$

Linearity

$$= f(x_k) - \alpha_k \nabla f(x_k)^T \mathbb{E}[g(x_k) \mid x_k] + \frac{L\alpha_k^2}{2} \mathbb{E}[\|g(x_k)\|^2 \mid x_k]$$

by unbiased and variance bound

$$\le f(x_k) - \alpha_k \|\nabla f(x_k)\|_2^2 + \frac{L\alpha_k^2}{2} \left(\sigma^2 + \|\nabla f(x_k)\|^2\right)$$

$$= f(x_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2 \sigma^2}{2}$$

By law of total expectation, we have on the overall result of this stochastic process.

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

Induction on this ensures

$$\min f(x) \le \mathbb{E}[f(x_{k+1})] \le \underbrace{\mathbb{E}[f(x_0)]}_{\overset{\shortparallel}{f(x_0)}} - \sum_{i=0}^{k} (\alpha_i - \tfrac{L\alpha_i^2}{2}) \mathbb{E}\|\nabla f(x_i)\|^2 + \sum_{i=0}^{k} \frac{L\alpha_i^2\sigma^2}{2}.$$

$$\overset{\text{Linearity}}{\Rightarrow} \mathbb{E}\Big[\underbrace{\sum_i \frac{\alpha_i(1-\frac{L\alpha_i}{2})\|\nabla f(x_i)\|^2}{\sum_j \alpha_j(1-\frac{L\alpha_j}{2})}}_{\substack{\text{a weighted average}\\ \text{of the gradient}\\ \text{norm squared}}}\Big] \le \frac{f(x_0) - \min f + \sum \frac{L\alpha_i^2\sigma^2}{2}}{\sum \alpha_i(1-\frac{L\alpha_i}{2})}$$

$$\Rightarrow \mathbb{E}\Big[\min_{i\le k} \|\nabla f(x_i)\|^2\Big] \le \frac{f(x_0) - \min f + \sum \frac{L\alpha_i^2\sigma^2}{2}}{\sum \alpha_i(1-\frac{L\alpha_i}{2})}. \quad \square$$

Next time this is a $O(\frac{1}{\sqrt{k}})$ rate.