# Lecture 14

HW 3 due an hour ago
Midterm release tomorrow at 7 am.

**Last time**

- Black-box convex optimization
- Things that break
- Analysis

**Today**

- Stochastic Gradient Descent
- Examples
- Analysis

## Stochastic Gradient Methods

Before we had an exact gradient oracle

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

Now we have an stochastic gradient oracle

$$x \longmapsto g(x, z).$$

random variable iid at each call

Such that

$$\mathbb{E}_z\, g(x, z) = \nabla f(x)$$ (Unbiased)

$$\mathbb{E}\left(\|g(x,z) - \nabla f(x)\|^2\right) \leq \sigma^2$$ (Bounded variance)

$$\mathbb{E}\left[\|g(x,z)\|^2\right] \overset{?}{=} \|\nabla f(x)\|^2$$

A natural algorithm updates

Draw $z_k$.

$$x_{k+1} \leftarrow x_k - \alpha_k \overbrace{g(x_k, z_k)}^{g_k}.$$

## Relevant properties of the expectation

- Linearity

Given $X_1, \dots, X_n$ r.v. and constants $\lambda_1, \dots, \lambda_n$, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i X_i\right] = \sum_i \lambda_i \mathbb{E} X_i .$$

- Tower law

Given two random variables $X, Y$

$$\mathbb{E}_X \left[ \underset{\uparrow}{\mathbb{E}[Y \mid X]} \right] = \mathbb{E}[Y]$$

conditional expectation

## Examples of oracles

**Example 1:** Coordinate approach

We want to solve $\min f(x)$ with $f: \mathbb{R}^d \to \mathbb{R}$.

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

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

Let's check that it is unbiased

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

$$= \sum \frac{\partial f}{\partial x_i}(x) \cdot e_i = \nabla f(x).$$

(Check that $\sigma$ depends on the dim).

## Example 2: Finite sum

Suppose we want to minimize

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

← We have seen many examples

Then

$$g(x, i) = \nabla f_i(x)$$

yields an unbiased gradient oracle.

One can prove that if $\nabla f_i$ L-Lips

$$\mathbb{E}\left[\|\nabla f_i(x) - \tfrac{1}{n}\sum \nabla f_i(x)\|^2\right] \le 2L^2 \|x\|^2.$$

## Example 3: Stochastic programming (Infinite sum)

Suppose we want to solve

$$\min_x \mathop{\mathbb{E}}_z f(x, z)$$

and we only have access to samples $z$.

Then

$$g(x, z) = \nabla f(x, z).$$

This is unbiased by definition.

# Example 4: Improved oracles for finite sums

**Idea 1:** Look at batches/minibatches of samples.

Pick $S \subseteq \{1, \dots, n\}$ with $|S| = k$ uniformly at random with or without replacement.

Take

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

which is clearly unbiased.

**Intuition**

Consider i.i.d. r.v $X_1, \dots, X_n \in \mathbb{R}^n$

$$\mathrm{Var}\left(\frac{1}{k}\sum_{i=1}^{k} X_i\right) = \frac{1}{k^2} \mathrm{var}(X_i)$$

← Better to consider $k > 1$

**Idea 2:** Variance reduction

Compute full gradients every now and then $\nabla f(\tilde{x}) = \frac{1}{n} \sum \nabla f_i(\tilde{x})$.

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

$$g(x, i) = \nabla f(\tilde{x}) + \underbrace{\nabla f_i(x) - \nabla f_i(\tilde{x})}_{\text{small when } x - \tilde{x} \text{ is small and } f \text{ is } L\text{-Lipschitz.}}$$

It is also unbiased

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

These two cancel out.

One can show that when $\nabla f_i$ L-Lips, then

$$\mathbb{E}\Big[\|\underbrace{\nabla f(\tilde{x}) - \nabla f(x) + (\nabla f_i(x) - \nabla f_i(\tilde{x})}_{g(x,i) - \nabla f(x)})\|^2\Big] \leq 4L^2\|x - \tilde{x}\|^2$$

can be made small.

SVRG [Johnson, Zhang, 2013].

## Analysis for nonconvex functions.

**Theorem** Suppose $f: \mathbb{R}^d \to \mathbb{R}$ is L-smooth and $g(x,z)$ is an unbiased estimator such that

$$\mathbb{E}[\|g(x,z) - \nabla f(x)\|^2] \leq \sigma^2 \quad \forall x.$$

Then the iterates of stochastic gradient descent with $0 < \alpha_k < 2/L$ satisfy

$$\mathbb{E}\left[\min_{k \leq T} \|\nabla f(x_i)\|_2^2\right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2 L}{2}\sum_{k=0}^{T}\alpha_k^2}{\sum_{k=0}^{T}\alpha_k\left(1 - \frac{L\alpha_k}{2}\right)}$$

**Proof:** By the Taylor Approximation Theorem

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

Conditioning on $x_k$

random because of $z_k$

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

Linearity

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

By Tower Law

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

By recursively applying this formula

$$\mathbb{E}[f(x_{T+1})] \leq \mathbb{E} f(x_0) - \sum_{k=0}^{T} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \mathbb{E}\|\nabla f(x_k)\|^2 + \sum_{k=0}^{T} \frac{L\alpha_k \sigma^2}{2}$$

The result follows from reordering and using the fact that

$$\mathbb{E}\left[\min_{k \leq T} \|\nabla f(x_k)\|^2\right] \sum_{k=0}^{T} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \leq \sum_{k=0}^{T} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \mathbb{E}\left[\|\nabla f(x_k)\|^2\right]. \quad \square$$

Next time we will make a $O(1/k^{1/4})$ in the general case and $O(1/k^{1/2})$ in the convex case.