Lecture 14

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

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

Stochastic Gradient Methods

Today
- Stochastic Gradient Descent
- Examples
- Analysis

Before we had an exact gradient oracle
$$ x \mapsto \nabla f(x). $$

Now we have a stochastic gradient oracle
$$ x \mapsto g(x, z). $$

random variable iid at each call

Such that
$$ \mathbb{E} g(x, z) = \nabla f(x) \quad (\text{Unbiased}) $$

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

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

A natural algorithm updates

Draw $z_k$.

$$ x_{k+1} \leftarrow x_k - \alpha_k g(x_k, z_k). $$
Relevant properties of the expectation

- **Linearity**
  
  Given $X_1, \ldots, X_n$ r.v. and constants $\lambda_1, \ldots, \lambda_n$, we have
  
  $$E \left[ \sum_{i=1}^{n} \lambda_i X_i \right] = \sum_{i=1}^{n} \lambda_i E X_i. $$

- **Tower law**
  
  Given two random variables $X, Y$
  
  $$E_x \left[ E[Y \mid X] \right] = E[Y]  \quad \text{conditional expectation}$$

Examples of oracles

**Example 1:** Coordinate approach

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

Pick $i \in \{1, \ldots, d\}$ 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
\[ \mathbb{E} [ q(x) ] = \frac{1}{d} \sum d \frac{\partial f}{\partial x_i}(x) \cdot e_i \]
\[ = \sum \frac{\partial f}{\partial x_i}(x) \cdot e_i = \nabla f(x). \]

(Chech 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
\[ \mathbb{E} [ \|
\nabla f_i(x) - \frac{1}{n} \sum \nabla f_i(x) \|^2 ] \leq 2L^2 \|x\|^2. \]

**Example 3: Stochastic programming (Infinite sum)**
Suppose we want to solve
\[ \min_x \mathbb{E} [f(x, z)] \]
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 \( s \subseteq \{1, \ldots, n\} \) 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. \( X_1, \ldots, X_n \in \mathbb{R}^n \)

\[
\text{Var} \left( \frac{1}{k} \sum_{i=1}^k X_i \right) = \frac{1}{k^2} \text{Var} \left( X_i \right)
\]

Better to consider \( k > 1 \)

Idea 2: Variance reduction

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

Pick \( i \in \{1, \ldots, n\} \) uniformly at random

\[
g(x, i) = \nabla f(x) + \nabla f_i(x) - \nabla f_i(x)
\]
small when \( x - \hat{x} \)
is small and \( f \) is \( \ell \)-Lipschitz.
It is also unbiased

\[ E[g(x, i)] = \nabla f(x) + E\nabla f_i(x) - E\nabla f_i(x) \]

These two cancel out.

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

\[ E \left[ \| \nabla f(x) - \nabla f(x) + (\nabla f_i(x) - \nabla f_i(x)) \| ^2 \right] \leq 4L^2 \| x - x \| ^2 \]

\[ g(x, i) - \nabla f(x) \]

can be made small.

SVRG [Johnson, Zhang, 2013].

Analysis for non-convex functions.

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

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

Then the iterates of stochastic gradient descent with \( 0 < \alpha_k < 2L \) satisfy

\[ E \left[ \min_{k=1}^{T} \| \nabla f(x_i) \| ^2 \right] \leq \frac{(f(x_0) - \min f) + \frac{\sigma^2 L}{2} \sum_{k=0}^{T} \alpha_k^2}{\sum_{k=0}^{T} \alpha_k (1 - \frac{L \alpha_k}{2})} \]
Proof: By the Taylor Approximation Theorem

\[ f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{1}{2} \| x_{k+1} - x_k \|^2 \]

\[ = f(x_k) - \alpha_k \nabla f(x_k)^T g_k + \frac{\alpha_k^2}{2} \| g_k \|^2 \]

Conditioning on \( x_k \)

\[ E[f(x_{k+1}) | x_k] \leq f(x_k) - \alpha_k \ E[\nabla f(x_k)^T g_k | x_k] + \frac{\alpha_k^2}{2} E[\| g_k \|^2 | x_k] \]

Linearity

\[ = f(x_k) - \alpha_k \nabla f(x_k)^T E[g_k | x_k] + \frac{\alpha_k^2}{2} E[\| g_k \|^2 | x_k] \]

\[ \leq f(x_k) - \alpha_k \| \nabla f(x_k) \|^2 \]

\[ + \frac{\alpha_k^2}{2} \left[ \sigma^2 + \| \nabla f(x_k) \|^2 \right] \]

\[ = f(x_k) - (\alpha_k + \frac{\alpha_k^2}{2}) \| \nabla f(x_k) \|^2 \]

\[ + \frac{\alpha_k^2}{2} \sigma^2. \]

By Tower Law

\[ E[f(x_{k+1})] \leq E[f(x_k) - (\alpha_k + \frac{\alpha_k^2}{2}) \| \nabla f(x_k) \|^2 \]

\[ + \frac{\alpha_k^2}{2} \sigma^2. \]
By recursively applying this formula
\[ E \left[ f(x_{T+1}) \right] \leq E f(x_0) - \sum_{k=0}^{T} \left( \alpha_k - \frac{L \alpha_k^2}{2} \right) E \|\nabla f(x_k)\|^2 + \sum_{k=0}^{T} L \alpha_k \sigma^2 \]

The result follows from reordering and using the fact that
\[ E \left[ \min_{k \leq T} \|\nabla f(x_k)\|^2 \right] \sum_{k=0}^{T} \left( \alpha_k - \frac{L \alpha_k^2}{2} \right) \leq \sum_{k=0}^{T} \left( \alpha_k - \frac{L \alpha_k^2}{2} \right) E \left[ \|\nabla f(x_k)\|^2 \right] \]

Next time we will make a \( O(\frac{1}{\kappa_{14}}) \) in the general case and \( O(\frac{1}{\kappa_{12}}) \) in the convex case.