Math 173B - Lecture 1: Introduction and Review

Introduction

In data science and machine learning, we often need to optimize functions of the form:

$$\min_x \sum_{i=1}^{N} f(x; y_i),$$

where x represents the model parameters, y_i are the training data (or training examples), and f is the loss function. Examples include loss functions such as:

- Linear regression
- Logistic regression
- Least squares fit to a model
- Principal Component Analysis (PCA)
- Neural network loss functions

Convex vs. Non-Convex Optimization

- Applications such as linear regression, logistic regression, and support vector machines (SVMs) involve convex optimization. Convex optimization is generally easier than non-convex optimization because in convex optimization, local minima are also global minima.
- Non-convex optimization tends to be hard due to the presence of many local minima. Many non-convex problems are NP-hard. Deep learning models typically involve non-convex optimization.

Refresher on Convex Functions

Definition of Convexity

A function f is convex if:

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathbb{R}^n, \lambda \in [0,1].$$

This inequality ensures that the line segment between any two points on the graph of the function lies above the graph itself.

Examples (Single-variable case)

- $f(x) = x^2$
- $f(x) = e^x$
- $f(x) = |x|$
- $f(x) = \log(1 + e^x)$

Exercise: Show that the above functions are convex.

Properties of Convex Functions

- Every local minimum is a global minimum. This is a very useful property because it implies the optimization algorithm only needs to find a local minimum, and it is guaranteed to be global.
- If f is convex, then:
  - $af(x) + bg(x)$ is convex if f, g are convex and $a, b \geq 0$.
  - If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then $f(Ax + b)$ is convex for any matrix A and vector b.

Compositions of Functions

- Compositions of convex functions are not necessarily convex.
- Deep learning involves composition of functions. However, compositions of convex functions are, in general, not convex. For example, if f and g are convex, $h(x) = f(g(x))$ is not necessarily convex.

First-Order and Second-Order Conditions

First-Order Condition

- A function is convex if its linear approximation underestimates the function everywhere:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x).$$

- $(x - y)^T (\nabla f(x) - \nabla f(y)) \geq 0$.

  Proof. Add the two inequalities derived from the first-order convexity condition:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x),$$
$$f(x) \geq f(y) + \nabla f(y)^T (x - y),$$

  and cancel terms and simplify to obtain the result.

Second-Order Condition

- A twice-differentiable function f is convex if its Hessian $\nabla^2 f(x)$ is positive semidefinite for all x.

Examples Using Second-Order Condition

- $f(x) = x^2$:
$$\nabla f(x) = 2x, \quad \nabla^2 f(x) = 2 \geq 0,$$
so f is convex.
- Quadratic form $f(x) = x^T Ax + b^T x + c$:
$$\nabla f(x) = 2Ax + b, \quad \nabla^2 f(x) = 2A.$$
If A is positive semidefinite (PSD), then f is convex.
- Exponential function $f(x) = e^x$:
$$\nabla f(x) = e^x, \quad \nabla^2 f(x) = e^x \geq 0,$$
so f is convex.
- $f(x) = \log(1 + e^x)$:
$$\nabla f(x) = \frac{e^x}{1 + e^x}, \quad \nabla^2 f(x) = \frac{e^x}{(1 + e^x)^2} \geq 0,$$
so f is convex.

Strong Convexity

Strongly convex functions are easier to optimize and allow for faster convergence guarantees.

- $f(x)$ is strongly convex, if there is a $\mu > 0$, such that the function $h(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex.
- A differentiable function f is strongly convex if there exists a constant $\mu > 0$ such that:
$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$
- A function for which the Hessian exists is strongly convex if:
$$\nabla^2 f(x) \succeq \mu I,$$
where $\mu > 0$ is a constant. This means that the smallest eigenvalue of the Hessian is at least μ.

Refresher on Gradient Descent

To optimize f using gradient descent, we run the iteration:
$$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)}),$$
where η is the step size, and iterations continue until some stopping criterion is met.

Proof of Convergence for Strongly Convex Smooth Functions

Let's examine gradient descent when the function is not only convex, but also strongly convex and smooth. That is, assume:

$$\mu I \preceq \nabla^2 f(x) \preceq LI,$$

where L is the Lipschitz constant of the gradient and $\mu > 0$ is the strong convexity constant.

We can think of LI as an upper bound on the Hessian $\nabla^2 f(x)$ and μI as a lower bound.

Recall that $\nabla f(x^*) = 0$. From the gradient descent update rule, and the mean value theorem, there is a $\zeta^{(t)}$ on the line segment between $x^{(t)}$ and x^*, with

$$\left(x^{(t+1)} - x^*\right) = \left(x^{(t)} - x^*\right) - \eta\left(\nabla f(x^{(t)}) - \nabla f(x^*)\right) \tag{1}$$

$$= \left(x^{(t)} - x^*\right) - \eta\nabla^2 f(\zeta^{(t)})\left(x^{(t)} - x^*\right) \tag{2}$$

$$= \left(I - \eta\nabla^2 f(\zeta^{(t)})\right)\left(x^{(t)} - x^*\right). \tag{3}$$

Thus, using the matrix norm inequality $\|Az\|_2 \leq \|A\|\|z\|_2$:

$$\|x^{(t+1)} - x^*\|_2 \leq \left\|I - \eta\nabla^2 f(\zeta^{(t)})\right\|\|x^{(t)} - x^*\|_2 \tag{4}$$

$$\leq \max\left(|1 - \eta\mu|, |1 - \eta L|\right)\|x^{(t)} - x^*\|_2. \tag{5}$$

Exercise: prove (5) using the definition of PSD matrices and the properties of the Hessian.

So, if we choose $\eta = \frac{2}{L+\mu}$, we have

$$\|x^{(t+1)} - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right) \cdot \|x^{(t)} - x^*\|_2, \tag{6}$$

and at step, say N,

$$\|x^{(N)} - x^*\|_2 \leq \left(\frac{L-\mu}{L+\mu}\right)^N \cdot \|x^{(0)} - x^*\|_2, \tag{7}$$

which goes to 0 very fast as N grows.

Math 173B - Lecture 2: Convexity and Stochastic Gradient Descent

1 Second order conditions

Proposition 1. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a second order differentiable function. Then f is convex if and only if $\nabla^2 f \succeq 0$.*

Remark 1. For a symmetric matrix $A \in \mathbb{R}^{n\times n}$, $A \succeq 0$ if and only if $\forall x \in \mathbb{R}^n$, $y^\top Ay \geq 0$. This is also equivalent to "all the eigenvalues of A are non-negative".

Exercises

- x^2.
- $x^\top Ax + b^\top x + c$, where A is an $n \times n$ positive semidefinite matrix.
- e^x.
- $\log(1 + e^x)$.

2 Strong convexity

Definition 2.1. $f : \mathbb{R}^n \to \mathbb{R}$ is strongly convex if and only if $h(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex for some $\mu > 0$.

Proposition 2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a first order differentiable function. Then f is strongly convex if and only if there exists $\mu > 0$, such that the following holds for all x, $y \in \mathbb{R}^n$.*

$$(x - y)^\top (\nabla f(x) - \nabla f(y)) \geq \mu\|x - y\|^2.$$

Proposition 3. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a second order differentiable function. Then f is strongly convex if and only if there exits $\mu > 0$, such that the following holds for all $x \in \mathbb{R}^n$.*

$$\nabla^2 f(x) \succeq \mu I.$$

Remark 2. For two symmetric matrices A, $B \in \mathbb{R}^{n\times n}$, $A \succeq B$ if and only if $A - B \succeq 0$.

Strongly convex functions are in some sense "easy" to minimize and also easy to prove theorems about.
Exercise: Which of the following functions is strongly convex? ($f : \mathbb{R} \to \mathbb{R}$)

- e^x.
- $\log(1 + e^x)$.
- x.
- $|x|$.
- $x^2 + 7x + 5$.

3 Review of Gradient descent

To optimize $f(x)$, we run the iteration

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla f(x^{(t)})$$

for $t = 0, 1, ...$ until some stopping criterion is met.

Let us examine gradient descent when the function is strongly convex and smooth. That is, we assume

$$\mu I \preceq \nabla^2 f(x) \preceq LI,$$

where $0 < \mu < L$, and I represents the identity matrix. In other words, we assume

$$\nabla^2 f(x) - \mu I \succeq 0,\ LI - \nabla^2 f(x) \preceq 0.$$

Think of LI as the Hessian of the function $\frac{L}{2}\|x\|^2$ and μI as the Hessian of the function $\frac{\mu}{2}\|x\|^2$. Let's suppose that x^* is the global minimizer of $f(x)$.

Proof that gradient descent converges (fast) in this setting. We have

$$\begin{aligned} x^{(t+1)} - x^* &= x^{(t)} - \alpha \nabla f(x^{(t)}) - x^* \\ &= x^{(t)} - x^* - \alpha \left(\nabla f(x^{(t)}) - \nabla f(x^*) \right) \\ &= \left(x^{(t)} - x^* \right) - \alpha \nabla^2 f(z^{(t)}) \left(x^{(t)} - x^* \right), \end{aligned}$$

where in the last line, we used the mean-value theorem. Then

$$x^{(t+1)} - x^* = \left[I - \alpha \nabla^2 f(z^{(t)}) \right] \left(x^{(t)} - x^* \right),$$

which implies

$$\|x^{(t+1)} - x^*\| \leq \left\| I - \alpha \nabla^2 f(z^{(t)}) \right\| \cdot \left\| x^{(t)} - x^* \right\|.$$

Here, we used the norm inequality for matrix-vector multiplication $\|Ax\| \leq \|A\| \cdot \|x\|$. Next we use the fact (left as exercise) that

$$\left\| I - \alpha \nabla^2 f(z^{(t)}) \right\| \leq \max \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\}.$$

Thus we deduce

$$\|x^{(t+1)} - x^*\| \leq \max \left\{ |1 - \alpha\mu|, |1 - \alpha L| \right\} \cdot \left\| x^{(t)} - x^* \right\|.$$

If we pick step size α to be $\frac{2}{L+\mu}$, we have

$$\|x^{(t+1)} - x^*\| \leq \frac{L - \mu}{L + \mu} \cdot \left\| x^{(t)} - x^* \right\|,$$

where the coefficient $\frac{L-\mu}{L+\mu} < 1$. As a result, after N steps,

$$\left\| x^{(N)} - x^* \right\| \leq \left(\frac{L - \mu}{L + \mu} \right)^N \left\| x^{(0)} - x^* \right\|,$$

which goes to 0 exponentially fast as N goes the infinity. □

Today's data science problems often imply that even a first order algorithm like gradient descent is "too expensive".

Consider

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f(x; y_i).$$

For example, we'd like to fit

$$a^\top z + b = y$$

given a bunch of points $\{z_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}_{i=1}^N$. So, here, we'd like to find $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ so that

$$\frac{1}{N}\sum_{i=1}^{N} |a^\top z_i + b - y_i|^2$$

is minimized.

In general with $F(x) = \frac{1}{N}\sum_{i=1}^{N} f(x; y_i)$ using gradient descent, we'd need to compute $\nabla_x F(x) = \frac{1}{N}\sum_{i=1}^{N} \nabla_x f(x; y_i)$. Computing this $\nabla_x F(x)$ requires $O(N)$ time (N evaluations of $\nabla_x f$).

So in "Big Data" type applications, when N is huge, gradient descent can be too expensive. So how do we come up with less expensive methods?

Math 173B - Lecture 3: A Brief Introduction of Stochastic Gradient Descent

1 Stochastic Gradient Descent

Recall in machine learning tasks, we usually want to minimize functions of the following form,

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f(x; y_i).$$

Here N represents the size of the data set, which is usually large. We've mentioned that gradient descent in this context might be inefficient.

To resolve the issue, Stochastic Gradient Descent (SGD) is introduced. SGD computes

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla f(x^{(t)}; y_{i_t}),$$

where $i_t \in \{1, 2, ..., N\}$. Specifically,

- At every step of SGD, select only one data point and use it to compute $\nabla f(x^{(t)}; y_{i_t})$.
- In subsequent iteations, repeat with a different **randomly chosen** data pount $y_{i_{t+1}}$.

The idea behind this is that the expectation

$$\begin{aligned}\mathbb{E}x^{(t+1)} &= \mathbb{E}x^{(t)} - \alpha_t \mathbb{E}\nabla f(x^{(t)}; y_{i_t}) \\ &= \mathbb{E}x^{(t)} - \alpha_t \frac{1}{N}\sum_{i=1}^{N} \nabla f(x^{(t)}; y_i).\end{aligned}$$

So we are going to need some probability to understand SGD.

2 Probability basics

Definition 2.1. A *probability space* is the triple $(\Omega, \mathcal{F}, \mathbb{P})$. Here,

- Ω is called the *sample space*, which is the set of all possible outcomes of a random experiment.
- $\mathcal{F}$ is called the *event space*, which is a set whose elements $A \in \mathcal{F}$ are subsets of Ω, i.e. $\mathcal{F} \subseteq 2^{\Omega}$. Further, $\mathcal{F}$ should satisfy
  1. $\emptyset \in \mathcal{F}$, i.e., the empty set is an element of $\mathcal{F}$.
  2. $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$, i.e., $\mathcal{F}$ is closed under complements.
  3. If $A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$, i.e., $\mathcal{F}$ is closed under countable unions.
- $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called the probability measure, which should satisfy the axioms of probability:

1. $\mathbb{P}(A) \geq 0$ for $\forall A \in \mathcal{F}$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_1, A_2, ...$ are disjoint, so that $A_i \cap A_j = \emptyset$ for $\forall i \neq j$, then

$$\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i).$$

Example 1. Examples of event spaces:

- For any choice of sample space Ω, $\mathcal{F} = \{\emptyset, \Omega\}$ is an acceptable (but not an interesting) event space.
- For finite sample space Ω, power set 2^Ω also works.

Example 2. Example of a probability sapce: Suppose we throw a die. Then we can pick $\Omega = \{1, 2, 3, 4, 5, 6\}$. $\mathcal{F}$ is chosen to be the power set $2^\Omega = \{\emptyset, \{1\}, ..., \{6\}, \{1, 2\}, ..., \Omega\}$. For the probability measure, for $A \in \mathcal{F}$, we let

$$\mathbb{P}(A) = \frac{|A|}{6},$$

where $|A|$ is the cardinality of A. For example, $\mathbb{P}(\{2, 3\}) = 2/6$, $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\{1, 2, 3, 4, 5\}) = 5/6$.

Proposition 1 (Properties of probability measure)**.** *The following statements hold true.*

- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(A \cap B) \leq \min\{\mathbb{P}(A), \mathbb{P}(B)\}$.
- $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.
- $\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$.
- *If $A_1, A_2, ..., A_N$ are disjoint sets with $\cup_i A_i = \Omega$, then*

$$\mathbb{P}(\cup_i A_i) = \sum_i \mathbb{P}(A_i) = 1.$$

This is called the law of total probability.

Definition 2.2 (Conditional probability)**.** Let B be an event such that $\mathbb{P}(B) \neq 0$. Then we define the conditional probability of A given B as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Definition 2.3 (Independence)**.** Two events A and B are independent if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B),$$

or equivalently,

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

Definition 2.4 (Random variables)**.** A random variable X is a measurable function from Ω, to (for example) the reals $\mathbb{R}$,

$$X : \Omega \to \mathbb{R}.$$

Example 3. Suppose we toss a fair coin. Let Ω be the set of all length 3 sequences of heads and tails, so that $\omega = (H, H, T) \in \Omega$. The randoms variable X is the function that counts how many heads we have:

$$X(\omega) = \#\text{ of heads in } \omega.$$

In our particular example, $X : \Omega \to \mathbb{N}$. So we call it a *discrete random variable*. Here

$$\mathbb{P}(X = k) := \mathbb{P}(\{\omega : X(\omega) = k\}).$$

Apart from discrete random variables, we can think about continuous random variables, where we instead consider

$$\mathbb{P}(a \leq X \leq b) := \mathbb{P}(\{\omega : a \leq X(\omega) \leq b\}).$$

Definition 2.5 (Cumulative distribution function)**.** Suppose we have a real valued random variable X defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then the cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \leq x).$$

The CDF allows us to calculate the probability of $\{X \in A\}$ for any Borel measurable sets $A \subseteq \mathbb{R}$.

Proposition 2 (Properties of CDFs)**.** *The following statements hold true.*

- $0 \leq F_X(x) \leq 1$.
- $\lim_{x\to\infty} F_X(x) = 1$.
- $\lim_{x\to-\infty} F_X(x) = 0$.
- $x \leq y \Rightarrow F_X(x) \leq F_X(y)$.
- $F_X(x)$ *is right continuous.*

Definition 2.6 (Probability density function)**.** Suppose we have a real valued random variable X defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If the derivative of the CDF $F_X(x)$ exists, then we define the probability density functuon (pdf) $f_X(x)$ of X to be the derivative of $F_X(x)$. Namely,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Remark 1. Rougly speaking, $\mathbb{P}(x \leq X \leq x + \delta) \approx f_X(x) \cdot \delta$. But importantly, $f_X(x) \neq \mathbb{P}(X = x)$.

Proposition 3 (Properties of pdfs)**.** *The following statements hold true.*

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x) = 1$.
- $\int_{x\in A} f_X(x)\, dx = \mathbb{P}(X \in A)$ *for any Borel measurable set* $A \subseteq \mathbb{R}$.

Example 4. Suppose we toss a fair coin. We have the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \{H, T\}$, $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ and

$$\mathbb{P}(\emptyset) = 0,\ \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 1/2,\ \mathbb{P}(\{H, T\}) = 1.$$

Example 5. Suppose we toss a fair coin twice. We have the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\Omega = \{HH, HT, TH, TT\}$, $\mathcal{F} = 2^{\Omega}$. Suppose we want to assign the probability that the first coin lands heads. Then

$$\mathbb{P}(\{HH, HT\}) = 1/2.$$

Exercise: $\mathbb{P}(\{HH, HT, TH\}) =?$.

Math 173B - Lecture 5 and 6: Stochastic Gradient Descent

1 Random Variables Continued

Here we give some examples of random variables you might encounter. We start with discrete random variables.

- Bernoulli(p), where $p \in [0, 1]$. The probability mass function is

$$p_X(x) = \begin{cases} p, & \text{if } x = 1, \\ 1 - p, & \text{if } x = 0. \end{cases}$$

- Binomial(n, p). The probability mass function is given by

$$p_X(x) = \binom{n}{x} p^x (1-p)^x.$$

For continuous random variables, below are some examples.

- Uniform random variable $U(a, b)$. The probability density function is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b], \\ 0, & \text{Otherwise.} \end{cases}$$

- Normal/Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$. The probability density function is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x-\mu)^2/2\sigma^2).$$

  Here we have $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$. Those two are left as exercises.

2 Stochastic Gradient Descent (SGD)

2.1 Introduction

Consider the situation where we are given a set of data points

$$(w_i, z_i) \in \mathbb{R}^d \times \mathbb{R}, \ i \in 1, ..., N.$$

And we'd like to fit a simple linear model to the data. So we want to solve for $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ so that

$$a^\top w_i + b \approx z_i$$

for each i.

One way to do this is to solve:

$$\min_{a\in\mathbb{R}^d, b\in\mathbb{R}} \underbrace{\frac{1}{N}\sum_{i=1}^{N} |a^\top w_i + b - z_i|^2}_{F(a,b;(w_i,z_i)_{i=1}^N)}.$$

Notice that for arbitrary w,

$$a^\top w + b = \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}^\top}_{x} \underbrace{\begin{pmatrix} w \\ 1 \end{pmatrix}}_{y}.$$

Here column vectors $x \in \mathbb{R}^{d+1}$ and $y \in \mathbb{R}^{d+1}$. So now our optimization problem is

$$\min_{x\in\mathbb{R}^{d+1}} \frac{1}{N}\sum_{i=1}^{N} |x^\top y_i - z_i|^2.$$

And we can use, for example, gradient descent to solve it.

This would require computing

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla F(x^{(t)}).$$

In our example,

$$\nabla F(x) = \frac{1}{N}\sum_{i=1}^{N} 2(x^\top y_i - z_i) y_i.$$

Notice the sum range is over all data points, you have to make a pass over all the data.

What if instead, we update via

$$x^{(t+1)} = x^{(t)} - \alpha_t g^{(t)},$$

where $g^{(t)}$ is

- easy/inexpensive to compute,
- $\mathbb{E}g^{(t)} = \nabla F(x^{(t)})$.

So in our example, this might be implemented by

- randomly selecting one data point (y_{i_t}, z_{i_t}) by drawing i_t randomly from $\{1, ..., N\}$ where each index has probability $1/N$.
- using $g^{(t)} = 2(x^{(t)\top} y_{i_t} - z_{i_t}) y_{i_t}$. Notice that $\mathbb{E}g^{(t)} = \sum_{i=1}^{N} \frac{1}{N} 2(x^{(t)\top} y_i - z_i) y_i = \nabla F(x^{(t)})$.

More generally,

$$\begin{aligned} F(x) &= \frac{1}{N}\sum_{i=1}^{N} f(x; y_i) \\ &=: \frac{1}{N}\sum_{i=1}^{N} f_i(x). \end{aligned}$$

2.2 Thoery

Let's start thinking about SGD iterates:

$$\underbrace{\mathbb{E}_{i_t} x^{(t+1)}}_{\text{Expectation on } i_t} = \mathbb{E}_{i_t} \underbrace{x^{(t)}}_{\text{not dependent on } i_t} - \alpha_t \mathbb{E}_{i_t} \underbrace{f_{i_t}}_{\text{random index}} (x^{(t)}).$$

Thus

$$\mathbb{E}_{i_t} x^{(t+1)} = \underbrace{x^{(t)} - \alpha_t \nabla F(x^{(t)})}_{\text{on average SGD looks like GD}}.$$

Here comes the question: can we have good convergence guarantees for SGD? The challenge is that since we are not moving in the direction of the negative gradient, there is no guarantee of descent at every step. To analyze SGD, we will need some assumptions.

Assumption 2.1. $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and ∇F is $L-$lipschitz so that

$$\| \underbrace{\nabla F(x_1)}_{\in \mathbb{R}^d} - \underbrace{\nabla F(x_2)}_{\in \mathbb{R}^d} \| \leq L \|x_1 - x_2\|.$$

Assumption 2.2. F is twice differentiable and $\nabla^2 F$ satisfies

$$\| \underbrace{\nabla^2 F}_{\text{Hessian: } d \times d \text{ matrix}} \| \leq L$$

Here recall

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\| \leq 1} \|Ax\|.$$

Both assumptions can be used to deduce that $\forall x, h$, we have

$$F(x + h) \leq F(x) + \nabla F(x)^\top h + \frac{1}{2} L \|h\|^2. \tag{1}$$

Exercise. Prove that Assumption 2.2 implies (1). Hint: Taylor expansion is all you need.

Lemma 2.3. *We have*

$$\begin{aligned} &\mathbb{E}_{i_t} \left[F(x^{(t+1)}) \right] - F(x^{(t)}) \\ &\leq -\alpha_t \nabla F(x^{(t)})^\top \mathbb{E}_{i_t} \left[\nabla f_{i_t}(x^{(t)}) \right] + \frac{1}{2} \alpha_t^2 L \mathbb{E}_{i_t} \left[\|\nabla f_{i_t}(x^{(t)})\|^2 \right]. \end{aligned}$$

Proof. Since $x^{(t+1)} = x^{(t)} - \alpha_t \nabla f_{i_t}(x^{(t)})$. Using Taylor expansion, there exists some z such that

$$\begin{aligned} F(x^{(t+1)}) &= F(x^{(t)} - \alpha_t \nabla f_{i_t}(x^{(t)})) \\ &= F(x^{(t)}) - \alpha_t \nabla F(x^{(t)})^\top \nabla f_{i_t}(x^{(t)}) + \frac{1}{2} \alpha_t^2 \nabla f_{i_t}(x^{(t)})^\top \nabla^2 F(z) \nabla f_{i_t}(x^{(t)}). \end{aligned}$$

Then

$$\begin{aligned} F(x^{(t+1)}) &= F(x^{(t)}) - \alpha_t \nabla F(x^{(t)})^\top \nabla f_{i_t}(x^{(t)}) + \frac{1}{2} \alpha_t^2 \nabla f_{i_t}(x^{(t)})^\top \nabla^2 F(z) \nabla f_{i_t}(x^{(t)}) \\ &\leq F(x^{(t)}) - \alpha_t \nabla F(x^{(t)})^\top \nabla f_{i_t}(x^{(t)}) + \frac{1}{2} \alpha_t^2 L \|\nabla f_{i_t}(x^{(t)})\|^2. \end{aligned}$$

This further implies

$$\mathbb{E}_{i_t} \left[F(x^{(t+1)}) \right] - F(x^{(t)}) \leq -\alpha_t \nabla F(x^{(t)})^\top \mathbb{E}_{i_t} \left[\nabla f_{i_t}(x^{(t)}) \right] + \frac{1}{2} \alpha_t^2 L \mathbb{E}_{i_t} \left[\|\nabla f_{i_t}(x^{(t)})\|^2 \right].$$

This completes the proof. □

Remark 1 (Notes on the lemma)**.** We want $-\alpha_t \nabla F(x^{(t)})^\top \mathbb{E}_{i_t} \nabla f_{i_t}(x^{(t)})$ to be negative because $\frac{1}{2}\alpha_t^2 L \mathbb{E}_{i_t} \|\nabla f_{i_t}(x^{(t)})\|^2$ is non-negative.

Corollary 1. *In our case,* $\mathbb{E}_{i_t} \nabla f_{i_t}(x^{(t)}) = \nabla F(x^{(t)})$ *(why?). So*

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) \leq -\alpha_t \|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 L \mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right].$$

We will soon prove that SGD iterates end up somewhere close to optimal provided that the RHS of the inequality above is bounded in a way that guarantees descent as t grows.

To get this bound, we will need another assumption.

Assumption 2.4. We assume

$$\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\right] \|^2 \leq M + M_G \|\nabla F(x^{(t)})\|^2,$$

where $M_G \geq 0$.

Lemma 2.5. *Under assumptions 2.1, 2.2 and 2.4, we have*

$$\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) \leq -\left[\left(1 - \frac{1}{2}\alpha_t L M_G\right)\alpha_t\right] \|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM.$$

Proof. We have

$$\begin{aligned}\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) &\leq -\alpha_t \|\nabla F(x^{(t)})\|^2 + \frac{1}{2}L\alpha_t^2 (M + M_G \|\nabla F(x^{(t)})\|^2) \\ &= -\left[\left(1 - \frac{1}{2}\alpha_t L M_G\right)\alpha_t\right] \|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM.\end{aligned}$$

□

We make some observations here. If α_t is small, then $\left(1 - \frac{1}{2}\alpha_t L M_G\right)\alpha_t$ is negative. So now, we would like to ensure that $\alpha_t^2 LM$ is not too big because it is non-negative.

Assumption 2.6 (Strong Convexity)**.** There exists $c > 0$ such that $\forall x, y \in \mathbb{R}^d$, we have

$$F(y) \geq F(x) + \nabla F(x)^\top (y - x) + \frac{1}{2}c\|y - x\|^2.$$

This assumption guarantees a unique minimizer x^* (why? check Discussion 1). It also guarantees (why? check 173A)

$$F(x) - F(x^*) \leq \frac{\|\nabla F(x)\|^2}{2c}.$$

We are almost ready to complete the ingredient for our first result on SGD, but we first mention some notation, and details. Notice that $x^{(t+1)}$ depends on $x^{(t)}$ and i_t, but $x^{(t)}$ itself depends on i_{t-1} and $x^{(t-1)}$. So when we write $\underbrace{\mathbb{E}F(x^{(t)})}_{\text{no subscript}}$ as opposed to, say, $\mathbb{E}_{i_t} F(x^{(t)})$, that means the expectation $\mathbb{E}F(x^{(t)})$ is taken with respect to all the random variables $i_1, i_2, ..., t_{t-1}$. In other words,

$$\mathbb{E}\left[F(x^{(t)})\right] = \mathbb{E}_{i_1}\mathbb{E}_{i_2}...\mathbb{E}_{i_{t-1}}\left[F(x^{(t)})\right].$$

Theorem 2.7. *If F satisfies:*

- $\|\nabla F(x)\| \leq L$ *for all x.*

- $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right] \leq M + M_G\|\nabla F(x^{(t)})\|^2$.
- *F is strongly convex with constant c.*

and if we choose α so that

$$0 < \alpha \leq \frac{1}{LM_G} \text{ and } \alpha \leq \frac{1}{C},$$

then SGD with step-size α and uniform random choice of i_t from the set $\{1, 2, ..., N\}$ guarantees

$$\mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{\alpha LM}{2c} + (1 - \alpha c)^{t-1}\left[F(x^{(1)}) - F(x^*) - \frac{\alpha LM}{2c}\right].$$

Further, when $t \to \infty$, we have

$$\lim_{t\to\infty} \mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{\alpha LM}{2c}.$$

Math 173B - Lecture 7: Stochastic Gradient Descent

1 Main Theorem

Theorem 1.1. *If F satisfies:*

- $\|\nabla F(x)\| \leq L$ *for all* x.
- $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right] \leq M + M_G\|\nabla F(x^{(t)})\|^2$.
- F *is strongly convex with constant* c.

and if we choose α *so that*

$$0 < \alpha \leq \frac{1}{LM_G} \text{ and } \alpha \leq \frac{1}{c},$$

then SGD with step-size α *and uniform random choice of* i_t *from the set* $\{1, 2, ..., N\}$ *guarantees*

$$\mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{\alpha LM}{2c} + (1-\alpha c)^{t-1}\left[F(x^{(1)}) - F(x^*) - \frac{\alpha LM}{2c}\right].$$

Further, when $t \to \infty$, *we have*

$$\lim_{t\to\infty} \mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{\alpha LM}{2c}.$$

Proof. We begin by examining the difference between function values at successive iterations:

$$\begin{aligned}\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) &\leq -\left(1 - \frac{\alpha LM_G}{2}\right)\alpha\|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha^2 LM \\ &\leq -\frac{\alpha}{2}\|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha^2 LM\end{aligned}$$

Using the strong convexity assumption $f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x})^T(\mathbf{x} - \mathbf{x}^*) + \frac{c}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$, we can derive the following inequality:

$$\mathbb{E}[\|\nabla f(x^{(t)})\|^2] \geq 2c\big(\mathbb{E}[f(x^{(t)})] - f(\mathbf{x}^*)\big).$$

Substituting this back, we obtain:

$$\begin{aligned}&\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) \\ \leq &-\alpha c(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\alpha^2 LM,\end{aligned}$$

Subtract $F(x^*)$ from both sides and rearrange to obtain

$$\mathbb{E}\left[F(x^{(t+1)})\right] - F(x^*) \leq (1-\alpha c)(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\alpha^2 LM,$$

where the expectation on the left hand side is the expectation over $i_1, i_2, ..., i_t$.

Now subtract $\frac{\alpha LM}{2c}$ from both sides to get

$$\begin{aligned}
&\mathbb{E}\left[F(x^{(t+1)}) - F(x^*)\right] - \frac{\alpha LM}{2c} \\
\leq &(1-\alpha c)E\left(F(x^{(t)}) - F(x^*)\right) + \frac{1}{2}\alpha^2 LM - \frac{\alpha LM}{2c} \\
= &\underbrace{(1-\alpha c)}_{<1}\left[E\left(F(x^{(t)}) - F(x^*)\right) - \frac{\alpha LM}{2c}\right]
\end{aligned}$$

Finally, we arrive at our conclusion

$$\mathbb{E}\left[F(x^{(t+1)}) - F(x^*)\right] - \frac{\alpha LM}{2c} \leq (1-\alpha c)^t\left[E\left(F(x^{(1)}) - F(x^*)\right) - \frac{\alpha LM}{2c}\right].$$

□

2 Guarantee for Convergence

So far, we showed that under assumptions

- $\|\nabla F(x)\| \leq L$ for all x.
- $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x^{(t)})\|^2\right] \leq M + M_G\|\nabla F(x^{(t)})\|^2$.
- F is strongly convex with constant c.
- $0 < \alpha \leq \frac{1}{LM_G}$ and $\alpha \leq \frac{1}{c}$.

, we have

$$\lim_{t\to\infty}\mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{\alpha LM}{2c}.$$

Smaller step sizes α lead to smaller error bounds but require longer running times. To balance this trade-off, we can let α vary with t to get the error to 0.

We recall from last time

$$\begin{aligned}
\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) &\leq -\left(1 - \frac{\alpha_t LM_G}{2}\right)\alpha_t\|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM \\
&\leq -\frac{\alpha_t}{2}\|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM
\end{aligned}$$

The expectation over all the random variables will give us

$$\mathbb{E}\left[F(x^{(t+1)}) - F(x^{(t)})\right] \leq -\frac{\alpha_t}{2}\mathbb{E}\left[\|\nabla F(x^{(t)})\|^2\right] + \frac{1}{2}\alpha_t^2 LM$$

The left hand side forms a telescoping sum where we can easily get $\mathbb{E}\left[F(x^{(t+1)}) - F(x^{(0)})\right]$ when summing up from 0 till $T-1$ for any fixed T.

By summing over the right hand side, we can get the following inequality

$$\mathbb{E}\left[F(x^{(T)}) - F(x^{(0)})\right] \leq -\sum_{t=0}^{T-1}\frac{\alpha_t}{2}\mathbb{E}\left[\|\nabla F(x^{(t)})\|^2\right] + \sum_{t=0}^{T-1}\frac{1}{2}\alpha_t^2 LM.$$

Equivalently,

$$\sum_{t=0}^{T-1} \frac{\alpha_t}{2} \mathbb{E}\left[\|\nabla F(x^{(t)})\|^2\right] \leq \sum_{t=0}^{T-1} \frac{1}{2}\alpha_t^2 LM + \mathbb{E}\left[F(x^{(T)}) - F(x^{(0)})\right] \tag{1}$$

We want to get to a point where $\|\nabla F(x^{(t)})\| \approx 0$. Let's do something crazy by randomly selecting one of the points $x^{(0)}, x^{(1)}, ..., x^{(T-1)}$ with the following probability

$$\mathbb{P}(\tau = t) = \frac{\alpha_t}{\sum_{i=0}^{T-1} \alpha_i}.$$

Now let us take expectation w.r.t the entire trajectory and this newly built τ.

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(x^{(\tau)})\|^2\right] =& \sum_{t=0}^{T-1} \frac{\alpha_t}{\sum_{i=0}^{T-1} \alpha_i} \mathbb{E}\left[\|\nabla F(x^{(t)})\|^2\right] \\
=& \frac{2}{\sum_{i=0}^{T-1} \alpha_i} \left(\sum_{t=0}^{T-1} \frac{\alpha_t}{2} \mathbb{E}\left[\|\nabla F(x^{(t)})\|^2\right] \right) \\
\leq& \frac{2}{\sum_{i=0}^{T-1} \alpha_i} \left(\sum_{t=0}^{T-1} \frac{1}{2}\alpha_t^2 LM + \mathbb{E}\left[F(x^{(T)}) - F(x^{(0)})\right] \right)
\end{aligned}$$

Here in the last inequality we used (1).

We will choose α_t's so that $\sum_{i=0}^{T-1} \alpha_i \to \infty$ and $\frac{\sum_{i=0}^{T-1} \alpha_i^2}{\sum_{i=0}^{T-1} \alpha_i} \to 0$ and one choice is choosing $\alpha_t = \frac{1}{L\sqrt{t+1}}$.

Thus we have $\sum_{i=0}^{T-1} \alpha_i \geq \frac{2\sqrt{T-1}-1}{L}$ and $\sum_{i=0}^{T-1} \alpha_i^2 \leq \frac{\log T+1}{L^2}$ (Exercise) and we can plug in them to the above inequalities.

Now we can show

$$\mathbb{E}\left[\|\nabla F\|^2\right] = O(\frac{\log T}{\sqrt{T}}) \to 0 \text{ as } T \to \infty.$$

Here no convexity assumptions were used.

Next lecture we will try to do a bit better by assuming more about the function and choosing α_t differently.

Math 173B - Lecture 8: More on SGD

1 More refined analysis of SGD

We first recall SGD algorithm to optimize

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

reads

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla f_{i_t}(x^{(t)}),$$

where i_t is a random variable drawn uniformly from $\{1, ..., N\}$.

We showed (under some assumptions like strong convexity, Lipschitz gradient,...) that

$$\mathbb{E}\left[F(x^{(t)})\right] - F(x^*) \leq \underbrace{\frac{\alpha LM}{2c}}_{\text{constant that depends on } \alpha} + \underbrace{(1-\alpha c)^{t-1}\left[F(x^{(0)}) - F(x^*) - \frac{\alpha LM}{2c}\right]}_{\to 0 \text{ as } t \to \infty}.$$

Recall also, we had "Idea 1" to improve convergence guarantees:

- Used α_t (not fixed):

$$\alpha_t = \frac{1}{L\sqrt{t+1}}.$$

- Randomly selecting a point $x^{(t)}$ from our sequence of iterates (with non-equal probability) proportional to

$$\frac{\alpha_t}{\sum_{s=1}^{T}\alpha_s}$$

to show

$$\mathbb{E}\left[\|\nabla F\|^2\right] = O(\frac{\log T}{\sqrt{T}}) \to 0 \text{ as } T \to 0.$$

Here no convexity assumptions were used.

Now let us try to do a bit better by assuming more about the function and choosing α_t differently. Recall we have

$$\mathbb{E}\left[F(x^{(t)})\right] - F(x^*) \leq \underbrace{\frac{\alpha LM}{2c}}_{\text{constant that depends on } \alpha} + \underbrace{(1-\alpha c)^{t-1}\left[F(x^{(0)}) - F(x^*) - \frac{\alpha LM}{2c}\right]}_{\to 0 \text{ as } t \to \infty}.$$

A practical strategy is

- Run SGD with fixed α, say $\alpha = \alpha_1$ for $t = 1, 2, ...$ until $F(x^{(t)})$ stops decreasing very much, so now the upper bound is

$$\lesssim \frac{2\alpha_1 LM}{2c}.$$

- Suppose this happens at $t = T_1$.
- Resume the algorithm staring $x^{(T_1)}$ but pick $\alpha_2 = \frac{\alpha_1}{2}$ (for example).
- Do until algorithm stalls, say $t = T_2$.
- Repeat: ..., $\alpha_r = 2^{-r+1}\alpha_1$ then

$$\mathbb{E}\left[F(x^t)\right] - F(x^*) \lesssim 2^{-r}\alpha_1 \frac{LM}{2c}$$

for $t \geq T_r$. It turns out that $T_2 \approx 2T_1, T_3 \approx 2T_2$.

More rigorously: Recall

$$\begin{aligned}
\mathbb{E}_{i_t}\left[F(x^{(t+1)})\right] - F(x^{(t)}) &\leq -\left(1 - \frac{\alpha_t LM_G}{2}\right)\alpha_t \|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM \\
&\leq -\frac{\alpha_t}{2}\|\nabla F(x^{(t)})\|^2 + \frac{1}{2}\alpha_t^2 LM \\
&\leq -\alpha_t c(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\alpha_t^2 LM,
\end{aligned}$$

where the last step is due to the strong convexity. Now we subtract $F(x^*)$ from both sides to obtain

$$\mathbb{E}\left[F(x^{(t+1)})\right] - F(x^*) \leq (1 - \alpha_t c)(F(x^{(t)}) - F(x^*)) + \frac{1}{2}\alpha_t^2 LM,$$

where the expectation on the left hand side is the expectation over $i_1, i_2, ..., i_t$.

One can show if we choose $\alpha_t = \frac{\beta}{\gamma + t}$ with $\beta > 1/c$ and $\gamma > 0$ then

$$\mathbb{E}\left[F(x^{(t)}) - F(x^*)\right] \leq \frac{v}{\gamma + t},$$

where

$$v = \max\left\{\frac{\beta^2 LM}{2(\beta c - 1)}, (\gamma + 1)(F(x^{(1)}) - F(x^*))\right\}.$$

Proof is by induction, we won't do it here. We have the following "theorem".

Theorem 1.1. *If F is such that*

- $\|\nabla F\| \leq L$.
- $\mathbb{E}_{i_t}\|\nabla f_{i_t}(x^{(t)})\|^2 \leq M + M_G\|\nabla F(x^{(t)})\|^2$.
- *F is strongly convex, with constant c.*
- *$\alpha_t = \frac{\beta}{\gamma + t}$, where $\beta > 1/c$, $\gamma > 0$ and $\alpha_1 < \frac{1}{LM_G}$, then*

$$\mathbb{E}\left[F(x^{(t)})\right] - F(x^*) \leq \frac{v}{\gamma + t} \to 0 \text{ as } t \to \infty,$$

where

$$v = \max\left\{\frac{\beta^2 LM}{2(\beta c - 1)}, (\gamma + 1)(F(x^{(1)}) - F(x^*))\right\}.$$

2 Informal comparison with GD

2.1 Gradient descent (173A)

- $F(x^{(t)}) - F(x^*) = O(\frac{1}{\sqrt{t}})$ when α_t chosen well. (assumes only convexity)
- $F(x^{(t)}) - F(x^*) = O(\frac{1}{t})$ with $\alpha_t = \alpha$ (fixed). (assumes convexity, Lipschitz gradient $\|\nabla^2 F\| \leq L$)
- $F(x^{(t)}) - F(x^*) = O(e^{-ct/2})$. (assumes strong convexity with constant c and Lipschitz gradient)

2.2 Stochastic gradient descent (173B)

- Converging to $F(x^*)$ only up to a constant error term when α is fixed.
- $\alpha_t \sim 1/\sqrt{t}$ implies $\mathbb{E}\|\nabla F(x^{(t)})\|^2 \to 0$. (extra randomness, no convexity)
- Even with strong convexity and careful choice of α_t, we only got

$$\mathbb{E}\left[F(x^{(t)})\right] - F(x^*) = O(\frac{1}{\gamma + t}).$$

While this looks bad for SGD, it doesn't take into account that for

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x),$$

a ∇F calculation requires computing N gradients, whereas an SGD iteration requires 1.

3 Variations on stochastic gradient descent

3.1 Mini-Batching

To optimize $F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, the idea is to update

$$x^{(t+1)} = x^{(t)} - \alpha_t g(x^{(t)}),$$

where $g(x) = \frac{1}{m}\sum_{k=1}^{m} \nabla f_{i_k}(x)$. Here i_k is randomly drawn independently, with equal probability from $\{1, ..., N\}$.

Here again, as in "vanilla" SGD:

$$\mathbb{E}g(x) = \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \nabla f_{i_k}(x)\right] = \frac{1}{m}\sum_{k=1}^{m} \mathbb{E}\nabla f_{i_k} = \frac{1}{m} \cdot m \cdot \mathbb{E}\nabla f_{i_k} = \frac{1}{N}\sum_{i=1}^{N} f_i(x) = \nabla F(x).$$

In other words, we have an unbiased estimate of $\nabla F(x)$ through $g(x)$ since $\mathbb{E}g(x) = \nabla F(x)$.

So, same expectation as SGD, but what about variance? We will now see that $var(g(x))$ decreases with m.

3.2 Warm-up

We flip m coins and associate H with $x = 1$, and T with $x = 0$ for each coin. Suppose we want to understand

$$S_m = \frac{\#\text{ Heads}}{\#\text{ Total}}.$$

We know

$$S_m = \frac{\sum_{i=1}^{m} x_i}{m}.$$

Note that

$$\mathbb{E}S_m = \frac{1}{m}\mathbb{E}\sum_{i=1}^{m} x_i = \frac{1}{m} \cdot m\mathbb{E}x = \frac{1}{2}.$$

What about the variance of S_m. We know

$$Var(S_m) = \mathbb{E}\left[S_m^2\right] - \left(\mathbb{E}\left[S_m\right]\right)^2.$$

Now

$$\begin{aligned}
\mathbb{E}\left[S_m^2\right] &= \mathbb{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^2\right] \\
&= \frac{1}{m^2}\mathbb{E}\left[\left(\sum_{i=1}^{m} x_i\right)^2\right] \\
&= \frac{1}{m^2}\mathbb{E}\left[\left(\sum_{i=1}^{m} x_i\right) \cdot \left(\sum_{i=1}^{m} x_i\right)\right] \\
&= \frac{1}{m^2}\mathbb{E}\left[\sum_{i=1}^{m} x_i^2 + \sum_{i\neq j} x_i x_j\right] \\
&= \frac{1}{m^2}\mathbb{E}\left[\sum_{i=1}^{m} x_i^2\right] + \frac{1}{m^2}\left[\sum_{i\neq j} x_i x_j\right] \\
&= \frac{1}{m}\mathbb{E}\left[x_i^2\right] + \frac{1}{m^2} \cdot (m^2 - m)\underbrace{\mathbb{E}\left[x_i\right]}_{1/2}\underbrace{\mathbb{E}\left[x_j\right]}_{1/2} \\
&= \frac{1}{m} \cdot \frac{1}{2} + \frac{m^2 - m}{4m^2}.
\end{aligned}$$

Thus

$$\mathbb{E}\left[S_m^2\right] = \frac{1}{4m + 4}.$$

This implies

$$Var(S_m) = \frac{1}{4m}.$$

Math 173B - Lecture 9: More on SGD

1 Returning to SGD with mini-batching

Consider the loss function

$$F(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x),$$

where $x \in \mathbb{R}^d$. Let us recall,

- Vanilla SGD:

$$x^{(t+1)} = x^{(t)} - \alpha \nabla f_{i_t}(x).$$

- SGD with mini-batching:

$$x^{(t+1)} = x^{(t)} - \alpha g(x),$$

where

$$g(x) = \frac{1}{m}\sum_{k=1}^{m} \nabla f_{i_k}(x).$$

To compute the variance of $g(x)$, notice

$$g(x) = (g_1(x), ..., g_d(x))$$

and

$$\nabla F(x) = \left(\frac{\partial F}{\partial x_1}(x), ..., \frac{\partial F}{\partial x_d}(x)\right).$$

Since $\mathbb{E}g(x) = \nabla F(x)$, we have

$$\mathbb{E}g_j = \frac{\partial F}{\partial x_j}(x).$$

Then we can compute

$$\begin{aligned} Var(g_j(x)) &= \mathbb{E}\left[(g_j(x) - \mathbb{E}g_j(x))^2\right] \\ &= \mathbb{E}\left[g_j(x)^2\right] - \left(\mathbb{E}\left[g_j(x)\right]\right)^2 \\ &= \mathbb{E}\left[g_j(x)^2\right] - \left(\frac{\partial F}{\partial x_j}(x)\right)^2, \end{aligned}$$

where recall that each

$$g_j(x) = \frac{1}{m}\left[\sum_{k=1}^{m} \nabla f_{i_k}(x)\right]_j.$$

Here the subscript j denotes the j-th entry. Thus

$$\begin{aligned}
\mathbb{E}\left[g_j(x)^2\right] &= \frac{1}{m^2}\mathbb{E}\left[\left(\sum_{k=1}^{m} \nabla f_{i_k}(x)\right)_j \left(\sum_{k=1}^{m} \nabla f_{i_k}(x)\right)_j\right] \\
&= \frac{1}{m^2}\mathbb{E}\left[\left(\sum_{k=1}^{m} \left(\frac{\partial f_{i_k}}{\partial x_j}(x)\right)^2\right) + \left(\sum_{l\neq k} \frac{\partial f_{i_k}}{\partial x_j}(x)\frac{\partial f_{i_l}}{\partial x_j}(x)\right)\right] \\
&= \frac{1}{m^2}\left[m \cdot \mathbb{E}\left[\left(\frac{\partial f_{i_k}}{\partial x_j}(x)\right)^2\right] + m(m-1)\mathbb{E}\left[\frac{\partial f_{i_k}}{\partial x_j}(x)\right]^2\right] \\
&= \frac{1}{m}\mathbb{E}\left[\left(\frac{\partial f_{i_k}}{\partial x_j}(x)\right)^2\right] + \frac{m-1}{m}\left(\frac{\partial F}{\partial x_j}(x)\right)^2.
\end{aligned}$$

Then we deduce

$$\begin{aligned}
Var(g_j(x)) &= \mathbb{E}\left[g_j(x)^2\right] - \left(\frac{\partial F}{\partial x_j}(x)\right)^2 \\
&= \frac{1}{m}\mathbb{E}\left[\left(\frac{\partial f_{i_k}}{\partial x_j}(x)\right)^2\right] + \frac{m-1}{m}\left(\frac{\partial F}{\partial x_j}(x)\right)^2 - \left(\frac{\partial F}{\partial x_j}(x)\right)^2 \\
&= \frac{1}{m}\mathbb{E}\left[\left(\frac{\partial f_{i_k}}{\partial x_j}(x)\right)^2\right] - \frac{1}{m}\left(\frac{\partial F}{\partial x_j}(x)\right)^2 \\
&= \frac{1}{m}\cdot\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f_i}{\partial x_j}(x)\right)^2 - \frac{1}{m}\left(\frac{\partial F}{\partial x_j}(x)\right)^2.
\end{aligned}$$

So the variance of each coordinate of $g(x)$ decreases as m increases. In summary, how does mini-batching help:

- Each step has $1/m$ times the variance but costs m times more to calculate.
- Average error analysis shows no advantage over standard SGD (we expect this because $g(x)$ has the same expectation as $\frac{f_{i_k}}{\partial x}$)
- However, we can calculate each summand in

$$g(x) = \frac{1}{m}\sum_{k=1}^{m} \nabla f_{i_k}(x)$$

  separately (on a different computational node). This leads to potential parallel computing.

It is used for training neural networks for these reasons.

2 Final Thoughts on SGD

We can improve SGD in similar ways to how we improved GD in math 173A. For example, recall we have the following GD with momentum:

$$x^{(t+1)} = x^{(t)} - \alpha_t \nabla F(x^{(t)}) + \beta_t(x^{(t)} - x^{(t-1)}).$$

You would just replace $\nabla F(x^{(t)})$ by its SGD estimate to get the SGD counterpart. Also, you can do SGD with Nesterov's acceleration.