MATH73B: Optimization for DS & ML

Rabbittac

April 13, 2021

Abstract

Warning: This is only a piece of lecture notes written by a careless scribe. So just be careful with and tolerant of any possible typos or misunderstandings when you read. The scribe does not intend to make anyone driven by his stupidity! Also, the professor's explanation is extremely helpful as he discusses a lot about the interpretable ideas behind the dull scripts. So watch the lecture before reading this. If you have any suggestions (e.g. typos, typography, logistics), please do not hesitate contacting the scribe!

Contents

MATH173B **Spring 2021**

Lecture 1: Introduction, Review of Convexity

Lecturer: Rayan Saab *Scribes: Rabbittac*

1.1 Introduction

The task in data science / machine learning is to optimize the loss function

$$\min_w \frac{1}{N}\sum_{i=1}^{N} f(w; x_i)$$

where w is model parameters and x_i is training examples.

e.g.1. **Examples of applications**

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

In general, convex optimization tends to be easier than non-convex optimization because local minima is global minima. Examples of applications of convex optimizations: linear regression, logistic regression, support vector machine. On the other hand, non-convex optimization is hard. One important application with non-convex loss functions is deep learning.

1.2 Convex Functions

Definition 1.1 (*Convex*)

A function f is ***convex*** *if and only if* $\forall \alpha \in [0,1]$

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$$

Proposition 1.1 (*Properties of Convex Function*)

- *Every local minima of a convex function is a global minima.*
- *If $f_1(x)$ and $f_2(x)$ are convex, $\alpha_1, \alpha_2 \geq 0$, then $f(x) = \alpha_1 f_1(x) + \alpha_2 f_1(x)$ is convex.*
- *If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$, then $h(x) = f(Ax + b)$ is convex.* [1.1]

[1.1] Deep learning involves compositions of functions. However, compositions of convex functions may be non-convex.

Proposition 1.2 (*First Order Property of Convex Functions*)

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

Corollary 1.3

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

Proof: Applying the First Order Property gives $f(y) \geq f(x) + (y - x)^T \nabla f(x)$ and $f(x) \geq f(y) + (x - y)^T \nabla f(y)$. Adding the inequalities and canceling $f(x) + f(y)$ gives the result. ■

MATH173B **Spring 2021**

Lecture 2: Review of Strong Convexity, Gradient Descent

Lecturer: Rayan Saab *Scribes: Rabbittac*

Proposition 2.1 (*Second Order Property of Convex Functions*)

$$\nabla^2 f(x) \succeq 0$$

which means the Hessian is positive semidefinite.

e.g.1. Some convex functions

- x^2
- $x^T Ax + b^T x + c$
- e^x
- $\log(1 + e^x)$

2.1 Strong Convexity

Definition 2.1 (*Strong Convexity*)

*f is **strongly convex** if the function $h : h(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is convex for some $\mu > 0$.*

Proposition 2.2

A function f for which gradient exists is strongly convex if and only if

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

A function f for which Hessian exists is strongly convex if and only if [2.1]

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

e.g.2. Which of these functions is strongly convex?

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

[2.1] *Notation*: We write $A \succeq B$ if $A - B$ is positive semidefinite.

2.2 Gradient Descent

Definition 2.2 (*Gradient Descent*)

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

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

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

Proposition 2.3

GD for strong convex and smooth functions converges.

Proof: We now examine GD in the *strongly convex* and *smooth* cases. That is, we assume $\mu I \preceq \nabla^2 f(x) \preceq LI$ for $\mu > 0$ and $L > 0$, where I is the identity matrix. Think of LI as the Hessian of the function $\frac{L}{2}\|x\|^2$, μI as the Hessian of the function $\frac{\mu}{2}\|x\|^2$. Let x^* be the global optimizer of $f(x)$. We want to prove GD converges (fast) in this setting.

$x^{(t+1)} - x^{(t)} = x^{(t)} - \alpha \nabla f(x^{(t)}) - x^* = x^{(t)} - x^* - \alpha(\nabla f(x^{(t)}) - \nabla f(x^*)) = x^{(t)} - x^* - \alpha \nabla^2 f(z^{(t)})(x^{(t)} - x^*)$ by the Mean Value Theorem. Then $x^{(t+1)} - x^* = (I - \alpha \nabla^2 f(z^{(t)}))(x^{(t)} - x^*)$ implies $\|x^{(t+1)} - x^*\| \leq \|I - \nabla^2 f(z^{(t)})\| \|x^{(t)} - x^*\|$. Since $\mu I \preceq \nabla^2 f(z^{(t)}) \preceq LI$, $\|x^{(t+1)} - x^*\| \leq \|I - \nabla^2 f(z^{(t)})\| \|x^{(t)} - x^*\| \leq \max\{|1-\alpha\mu|, |1-\alpha L|\} \|x^{(t)} - x^*\|$ [2.2].

So if we choose $\alpha = \frac{2}{L+\mu}$, $\|x^{(t+1)} - x^*\| \leq \frac{L-\mu}{L+\mu} \|x^{(t)} - x^*\|$. So after N steps, $\|x^{(N)} - x^*\| \leq (\frac{L-\mu}{L+\mu})^N \|x^{(t)} - x^*\|$. ■

Today's applications actually show that GD is too computationally expensive (even for first order problems).

e.g.3. Consider $F(x) = \frac{1}{N} \sum_{i=1}^N f(w; x_i)$. We'd like to fit $a^T z + b = y$ given a bunch of points (z_i, y_i). So, we'd like to find $a \in \mathbb{R}^d, b \in \mathbb{R}$ so that $\frac{1}{N} \sum_{i=1}^N |a^T z_i + b - y_i|^2$ is minimized. In general, with $F(x) = \frac{1}{N} \sum_{i=1}^N f(w; x_i)$ using GD, we want to compute $\nabla F(x) = \frac{1}{N} \sum_{i=1}^N \nabla f(w; x_i)$. Computing this ∇F requires $O(N)$ time.

To introduce less expensive methods, we will discuss Stochastic Gradient Descent in the next lecture.

[2.2] Exercise: Prove $\|x^{(t+1)} - x^*\| \leq \max\{|1 - \alpha\mu|, |1 - \alpha L|\}$ using the strong convexity and smoothness of f.

MATH173B **Spring 2021**

Lecture 3: Probability

Lecturer: Rayan Saab *Scribes: Rabbittac*

To give a brief introduction to **Stochastic Gradient Descent (SGD)**

Theorem 3.1 (*SGD*)

SGD computes

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

At every step, choose one data and use it to compute $\nabla f(x^{(t)}; y_{i_t})$. In subsequential iterations, repeat with a different randomly chosen data point $y_{i_{t+1}}$.

The idea is that

$$\mathbb{E}[x^{(t+1)}] = \mathbb{E}[x^{(t)}] - \alpha \mathbb{E}[\nabla f(x^{(t)}; y_{i_t})] = \mathbb{E}[x^{(t)}] - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla f(x^{(t)}; y_i)$$

We need to introduce probability to understand SGD.

Definition 3.1 (*Sample Space*)

*The **sample space** Ω is the set of all possible outcomes of a random experiment.*

Definition 3.2 (*Events*)

*The **set of events** $\mathscr{F}$ a set whose elements $A \in \mathscr{F}$ are subset of Ω (event space). $\mathscr{F}$ should satisfy*

1. $\emptyset \in \mathscr{F}$
2. $A \in \mathscr{F} \implies \Omega \setminus A \in \mathscr{F}$
3. $A_1, \cdots \in \mathscr{F} \implies \bigcup_i A_i \in \mathscr{F}$

e.g.1. For any choice of Ω, $\mathscr{F} = \{\emptyset, \Omega\}$ is an acceptable (but not an intersecting).

Definition 3.3 (*Probability Measure*)

*The **probability measure** is a function $P : \mathscr{F} \to \mathbb{R}$ that satisfies the axioms of probability*

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

Definition 3.4 (*Probability Space*)

*A **probability space** is defined by the tuple $(\Omega, \mathscr{F}, P)$.*

e.g.2. $\Omega = \{1, 2, 3, 4, 5, 6\}$ which is dice rolling. $\mathscr{F} = \{\emptyset, \{1\}, \ldots, \{6\}, \{1, 2\}, \ldots, \Omega\}$. For P, we can assign a probability of $\frac{i}{6}$ to a set A with i elements in A *e.g.* $P(\{2, 4\}) = \frac{2}{6}, P(\emptyset) = 0$.

Proposition 3.2

1. *If $A \subset B$, then $P(A) \leq P(B)$.*
2. *$P(A \cap B) \leq \min\{P(A), P(B)\}$.*
3. *$P(A \cup B) \leq P(A) + P(B)$.*
4. *$P(\Omega \setminus A) = 1 - P(A)$.*
5. *Law of Total Probability: If $A_1, A_2, \ldots$ are disjoint sets with $\bigcup_i A_i = \Omega$, then $P(\bigcup_i A_i) = \sum_i P(A_i) = 1$.*

Definition 3.5 (*Conditional Probability*)

Let B be such that $P(B) \neq 0$, then we define the conditional probability of A given B as

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

Definition 3.6 (*Independence*)

*Two events A and B are **independent** if and only if $P(A \cap B) = P(A)P(B) \iff P(A|B) = P(A)$.*

Definition 3.7 (*Random Variable*)

*A **random variable** is a function $X : \Omega \to \mathbb{R}$ and we usually write $X(\omega) = x$ for $\omega \in \Omega$.*

e.g.3. Consider Ω to be the set of all length-3 sequences of heads and tails *e.g.* $\omega = (H, H, T) \in \Omega$. The random variable X is the function that counts how many heads we have. So $X(\omega) = |H|$. In our particular example, $X : \Omega \to \mathbb{N}$ is a **discrete random variable**.

For a discrete random variable

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

For a continuous random variable

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

Definition 3.8 (*Cumulative Distribution Function*)

*The **cumulative distribution function (CDF)** is $F_X(x) := P(X \leq x)$.*

The CDF allows us to calculate the probability of any event in $\mathscr{F}$.

Proposition 3.3

1. $0 \leq F_X(x) \leq 1$
2. $\lim_{x\to\infty} F(x) = 1$
3. $\lim_{x\to-\infty} F(x) = 0$
4. $x \leq y \implies F_X(x) \leq F_X(y)$

Definition 3.9 (*Probability Density Function*)

The ***probability density function (PDF)*** *is* [3.1]

$$f_X(x) := \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$$

For a PDF to exist, the CDF must be differentiable.

Proposition 3.4

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} f_X(x) = 1$
3. $\int_{x\in A} f_X(x)\,\mathrm{d}x = P(x \in A)$

e.g.4. Tossing a fair coin: $\Omega = \{H, T\}, \mathscr{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}, P(\emptyset) = 0, P(\{H\}) = P(\{T\}) = \frac{1}{2}, P(\{H, T\}) = 1$.

e.g.5. Tossing a fair coin twice: $\Omega = \{HT, TH, HH, TT\}, \mathscr{F} =$ power set of Ω. To assign the probability, $P(\{HH, HT\}) = \frac{1}{2}$. Exercise: $P(\{HH, HT, TH\}) =$?

[3.1] Note that $f_X(x) \neq P(X = x)$

MATH173B **Spring 2021**

Lecture 4: Probability

Lecturer: Rayan Saab *Scribes: Rabbittac*

This lecture includes many examples of probability. The scribe is idle so some examples are missing.

e.g.1. $\omega = \{1,2,3\}$ and $\mathscr{F} = \{\emptyset, \{1\}, \{2,3\}, \Omega\}$. To show $\mathscr{F}$ is legit:

1. $\emptyset \in \mathscr{F}$
2. $A \in \mathscr{F} \implies \Omega \setminus A \in \mathscr{F}$
3. $A_1, A_2, \cdots \in \mathscr{F} \implies \bigcup_i A_i \in \mathscr{F}$

The probability measure will be $P(\emptyset) = 0, P(\Omega) = 1, P(\{1\}) + P(\{2,3\}) = 1$.

e.g.2. (Not rigorous). Randomly choose a real number between 0 and 1. So $\Omega = [0,1]$, $\mathscr{F}$ is not considered [4.1]. $P((a,b)) = b - a$. More generally, $P(A)$ where $A \subset [0,1]$ is the length of A. $P((0,0.6)|(0.3,0.7)) = \frac{P((0,0.6)\cap(0.3,0.7))}{P((0.3,0.7))} = 0.75$.

e.g.3. Let A be the event that coin lands heads, B be the event of getting 6 on the dice. Then A and B are independent since $P(A|B) = P(A)$.

e.g.4. (Random Variable) For coin flips $\Omega = \{H,T\}$, $X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$. For rolling 2 dices, $\Omega = \{1,\ldots,6\} \times \{1,\ldots,6\}, X(\omega) = X((\omega_1,\omega_2)) = \max\{\omega_1,\omega_2\}$.

e.g.5. (Continuous Random Variable) Randomly throw a dart at a circle board. The random variable $\theta(\omega)$ is the angle with x-axis. Then Ω is the space of all possible locations the dart can land. $F_\Theta(\theta) = \frac{\theta}{2\pi}, f_\Theta(\theta) = \frac{1}{2\pi}$.

e.g.6. For a fair coin $X(\omega) = \begin{cases} 1 & \omega = H \\ 0 & \omega = T \end{cases}$. Here the probability mass function is

$$f_X(x) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{2} & x = 0 \\ 0 & \text{otherwise} \end{cases}.$$

[4.1] This requires knowledge of measure theory.

Definition 4.1 (*Expectation*)

Let X be a discrete random variable taking values in some set S with a probability mass function $P_X(x)$. Then we define ***expectation*** *of X to be*

$$\mathbb{E}[X] := \sum_{x \in S} x P_X(x)$$

If X is a continuous random variable with pdf $f_X(x)$ then we define

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f_X(x) \, \mathrm{d}x$$

e.g.7. Rolling a fair coin, $\mathbb{E}[X] = \sum_{i=1}^{6} \frac{1}{6} = 3.5$.

Proposition 4.1

$$\mathbb{E}[g(x)] := \sum_{x \in S} g(x) P_X(x)$$

$$\mathbb{E}[g(x)] := \int_{-\infty}^{\infty} g(x) f_X(x) \, \mathrm{d}x$$

e.g.8. For rolling a dice, what is $\mathbb{E}[X^2]$?

e.g.9. If X is a uniform random variable $X \sim U(0,1)$ where $f_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$.

Then $\mathbb{E}[X] = \int_0^1 x \, \mathrm{d}x = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$.

Proposition 4.2

$$\mathbb{E}[ag(x) + bf(X)] = a\mathbb{E}(g(X)) + b\mathbb{E}(f(X))$$

Proposition 4.3

If X is a discrete random variable and $\mathbb{I}_{x=k} := \begin{cases} 1 & x = k \\ 0 & \text{otherwise} \end{cases}$, then

$$\mathbb{E}[\mathbb{I}_{x=k}] = P(x = k)$$

For a continuous random variable

$$\mathbb{E}[\mathbb{I}_{x=k}] = f_X(x)$$

Definition 4.2 (*Variance*)

$$\mathrm{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Lemma 4.4

$$\mathrm{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Proof: $\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] = \mathbb{E}[X^2] - 2(\mathbb{E}[X])(\mathbb{E}[X]) + (\mathbb{E}[X])^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ ■

Proposition 4.5

1. $\mathrm{var}(c) = 0$ *for constant* c
2. $\mathrm{var}(af(X)) = a^2 \,\mathrm{var}(f(X))$

e.g.10. Calculate the variance for $X \sim U([0, 1])$.

MATH173B **Spring 2021**

Lecture 5: Random Variables, SGD

Lecturer: Rayan Saab *Scribes: Rabbittac*

Definition 5.1 (*Bernoulli Random Variable*)

The p.m.f. of a Bernoulli random variable is

$$P_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$$

Definition 5.2 (*Binomial Random Variable*)

The p.m.f. of a Binomial random variable is

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

Definition 5.3 (*Uniform Random Variable*)

The p.d.f. of a Uniform random variable is

$$P_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & otherwise \end{cases}$$

Definition 5.4 (*Gaussian/Normal Random Variable*)

The p.d.f. of a Gaussian random variable $\mathcal{N}$ is

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

Proposition 5.1

$\mathbb{E}[\mathcal{N}(\mu, \sigma^2)] = x$ *and* $\text{var}(\mathcal{N}(\mu, \sigma^2)) = \sigma^2$.

5.1 SGD

Consider the situation where we are given a set of data points $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, \ldots, N$. We want to fit a simple linear model to fit the data. So we want to solve for $a \in \mathbb{R}^d, b \in \mathbb{R}$ so that

$$a^T x_i + b_i \approx y_i$$

One way to do this is to solve

$$\min_{a\in\mathbb{R}^d, b\in\mathbb{R}} \sum_{i=1}^{N} |a^T x_i + b - y_i|^2$$

Let $F(a, b; (x_i, y_i), i = 1, \ldots, N) = \sum_{i=1}^{N} |a^T x_i + b - y_i|^2$. Notice that for arbitrary x, $a^T x + b = \begin{pmatrix} a \\ b \end{pmatrix}^T \begin{pmatrix} x \\ 1 \end{pmatrix}$. Then we can write this optimization problem as

$$\min_{w\in\mathbb{R}^{d+1}} \sum_{i=1}^{N} |w^T z_i - y_i|^2$$

where $w = \begin{pmatrix} a \\ b \end{pmatrix}, z_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix}$. Using SGD to solve this, we want to compute

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

where $\nabla F(w^{(t)}) = \sum_{i=1}^{N} 2(w^T z_i - y_i) y_i$ in our case.

What if instead, we update via $w^{(t+1)} = w^{(t)} - \alpha_t g^{(t)}$ where $g^{(t)}$ is cheap to compute and $\mathbb{E}[g^{(t)}] = \nabla F(w^{(t)})$. In our example, this is implemented by

---

Algorithm 5.1 SGD: Notice that $\mathbb{E}[g^{(t)}] = \sum_{i=1}^{N} 2(w^T z_i - y_i) y_i = \nabla F(w^{(t)})$.

---

1. Randomly select one data point (x_{i_t}, y_{i_t}) by drawing i_t randomly from $\{1, \ldots, N\}$ where each index has probability of $\frac{1}{N}$.
2. Using $g(t) = 2(w^T z_{i_t} - y_{i_t}) y_{i_t}$.

Let $f_i(w)$ denote $f(w; (z_i, y_i))$. Now think about SGD iterations: [5.1]

$$\begin{aligned} \mathbb{E}_{i_t}[w^{(t+1)}] &= \mathbb{E}_{i_t}[w^{(t)}] - \alpha_t \mathbb{E}_{i_t}[f_{i_t}(w^{(t)})] \\ &= w^{(t)} - \alpha_t \nabla F(w^{(t)}) \end{aligned}$$

This implies that, on average, SGD looks like GD.

Now we consider proving the convergence guarantee for SGD. The challenge is that since we are not moving in the direction of the negative gradient, there is now guarantee of descent at every step. To analyze SGD, we'll need some assumptions:

1. $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and ∇F is L-Lipschitz so that $\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|$.
2. F is twice differentiable and $\nabla^2 F$ satisfies $\|\nabla^2 F\| \leq L$.

Both assumptions can be used to deduce that

$$\forall x, h : F(x + h) \leq F(x) + \nabla F(x)^T h + \frac{1}{2} L \|h\|^2$$

[5.1] $\mathbb{E}_{i_t}$ denotes that the expectation is taken only with respect to i_t.

Lemma 5.2

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