Machine Learning Theory Notes

Gaotang Li

March 2, 2023

Contents

Part I

Generalization Theory

Chapter 1

Supervised Learning Framework

1.1 Basic Setups

In a supervised learning problem, we have a goal to predict a label given an input. Let S denote the dataset $\{(x_i, y_i)\}_{i=1}^n$ for

- $x_i \in \mathcal{X}$, the inputs in the input space.
- $y_i \in \mathcal{Y}$, the label associated with x_i in the label space.

We assume that the data are drawn **i.i.d.** from an unknown probability distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. We aim at learning a function mapping $h \colon \mathcal{X} \to \mathcal{Y}$ (aka *hypothesis*, *predictor*, *model*).

To evaluate the performance of h, we specify a loss function. A loss function $\ell \colon \mathcal{Y}, \mathcal{Y} \to \mathbb{R}$ measures the difference between the predicted label and the groundtruth label.

Definition 1.1.1 (population risk). The *population risk* of a *hypothesis* h is its expected loss over the data distribution $\mathcal{P}$: $L_{\mathcal{P}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)]$

Example. Examples of Loss Functions

- Classification: 0-1 loss $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$.
- Regression: squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$

It is often impossible to consider all possible function mappings from $\mathcal{X} \to \mathcal{Y}$. We usually only consider a *hypothesis class* $\mathcal{H}$.

Example. Examples of $\mathcal{H}$.

- Linear function class: $\mathcal{H} = \{h_\theta | h_\theta(x) = \theta^T x, \theta \in \mathbb{R}\}$
- General parametric function class: $\mathcal{H} = \{h_\theta | h_\theta(x) = f(x, \theta), \theta \in \mathbb{R}^p\}$

1.2 Empirical Risk Minimization

Definition 1.2.1 (Empirical Risk). The *empirical risk* of a hypothesis h is its average loss over the dataset S

$$L_s(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Empirical risk minimization (ERM) is any algorithm that minimizes the empirical risk over the hypothesis class $\mathcal{H}$. We denote a hypothesis returned by ERM as $\hat{h}_{ERM}$, *i.e.*:

$$\hat{h}_{ERM} \in \text{argmin}_{h \in \mathcal{H}} L_S(h)$$

If we assume h is independent of S, then $\mathbb{E}_S[L_S(h)] = L_P(h)$. But in reality h and S are not independent.

1.3 Questions

In the supervised learning part of this course, we are mainly interested in the following two fundamental problems:

- **Statistical**: What guarantee do we have about $L_P(\hat{h}_{ERM})$?
- **Optimization**: When may ERM be achieved efficiently?

Chapter 2

Concentration Inequality

Concentration inequalities are a mathematical tool to study the relation between population and empirical quantities. Consider the following main question: for i.i.d. random variables $X_1, \dots, X_n$, how does $\frac{1}{n}\sum_{i=1}^n X_i$ relate to $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n X_i] = \mu$?

2.1 Chebyshev's Inequality

Lemma 2.1.1 (Markov's Inequality). Let X be a non-negative random variable, then for all $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(x)}{t}$$

Proof.

$$\mathbb{E}(X) \geq \Pr(X < t) * 0 + \Pr(X > t) * t$$

■

Theorem 2.1.1 (Chebyshev's Inequality). Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $t > 0$,

$$\Pr\left[|X - \mathbb{E}(X)| \geq t\right] \leq \frac{\mathrm{Var}(X)}{t}$$

Proof.

$$\begin{aligned}\Pr\left[X - \mathbb{E}[X] \geq t\right] &= \Pr\left[(X - \mathbb{E}[X])^2 \geq t^2\right] \\ &= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{t^2} \\ &= \frac{\mathrm{Var}[X]}{t^2}\end{aligned}$$

■

Remark.

$$\Pr\left[|X - \mathbb{E}[X]| \geq t\right] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^P]}{t^P}$$

Corollary 2.1.1. Let $x_1, \dots, x_n$ be i.i.d. random variables such that $\mathbb{E}\left[x_i\right] = \mu$, $\mathrm{Var}[x_i] = \sigma^2$. Then:

$$\Pr\left[|\frac{1}{n}\sum_{i=1}^n x_i - \mu| \geq t\right] \leq \frac{\sigma^2}{nt^2}$$

2.2 Hoeffding Inequality

Lemma 2.2.1. If $X \in [0,1]$ a.s. Then,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2}{8}}$$

for all $\lambda \in \mathbb{R}$.

Proof. Let $Z = X - \mathbb{E}[X]$, then $\mathbb{E}[Z] = 0$.
Define $\psi(\lambda) := \log \mathbb{E}\left[e^{\lambda Z}\right]$.
Using the Taylor expansion to get that $\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\lambda')$ where λ' is between 0 and λ.

Here the first term $\psi(0) = \log 1 = 0$, and the second term $\lambda\psi'(0) = \mathbb{E}[Z] = 0$. The only thing we need to is to compute the third term. The idea is to bound the third term by $1/4$.

Then

$$\begin{aligned}
\psi'(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda Z} Z\right]}{e^{\lambda Z}} = \mathbb{E}[Y] \\
\psi''(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda Z} Z^2\right]}{e^{\lambda Z}} - \left(\frac{\mathbb{E}\left[e^{\lambda Z} Z\right]}{\mathbb{E}[e^{\lambda Z}]}\right)^2 \\
&= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\
&= \mathrm{Var}[Y]
\end{aligned}$$

Where we can think of Y as a reweighted version of Z, and we have that

$$dP_Y(x) = \frac{e^{\lambda x}}{\mathbb{E}\left[e^{\lambda Z}\right]} dP_Z(x)$$

not finished yet... ■

Remark. We also call such random variables **subgaussian** random variables. Another interpretation is that bounded random variables are subgaussian.

Another reminder is that the expectation is in the form of **Moment Generating Function**, where $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

Theorem 2.2.1 (Hoeffding Inequality)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables such that for each i, $X_i \in [0,1]$ a.s. Then for all $t > 0$:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq t\right] \leq e^{-2nt^2}$$
$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \leq -t\right] \leq e^{-2nt^2}$$

Proof. Let us use $\bar{X}$ to denote $\frac{1}{n}\sum_{i=1}^n X_i$. Then we have that

$$\begin{aligned}
\Pr\left[\bar{X} - \mathbb{E}[\bar{X}] \geq t\right] &= \Pr\left[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])} \geq e^{\lambda t}\right] && \lambda > 0 \\
&\leq \frac{\mathbb{E}\left[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}\right]}{e^{\lambda t}} && \text{Markov} \\
&= e^{-\lambda t} \cdot \mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}\right] \\
&= e^{-\lambda t} \cdot \mathbb{E}\left[\prod_{i=1}^n e^{\frac{\lambda}{n}(X_i - \mathbb{E}[X_i])}\right] \\
&= e^{-\lambda t} \prod_{i=1}^n \mathbb{E}\left[e^{\frac{\lambda}{n}(X_i - \mathbb{E}[X_i])}\right] && \text{Independence} \\
&\leq e^{-\lambda t}\left(e^{\frac{1}{8}(\frac{\lambda}{n})^n}\right) \\
&= e^{-\lambda t + \frac{\lambda^2}{8n}} \\
&= e^{-2nt^2} && \text{let } \lambda = 4nt
\end{aligned}$$

By symmetry, we complete the proof. ■

Remark (Equivalent Definition of Hoeffding Inequality)**.** Let $X_1, \ldots, X_n \in [0,1]$ a.s. and independent,

$$\forall \delta \in (0,1), \text{ w.p. } \geq 1-\delta: \quad \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\text{w.p. } \geq 1-\delta: \quad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] - \frac{1}{n}\sum_{i=1}^n X_i \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

2.3 Bounded Difference Concentration Inequality

We are concerned with diffeormorphism (*i.e.* change in one coordinate) and formally

$$f(X_1, \cdots, X_n) \mapsto \mathbb{E}\left[f(X_1, \cdots, X_n)\right]$$

Theorem 2.3.1 (Mcdiarmid's inequality)**.** Suppose $X_1, \ldots, X_n$ are independent random variables taking values in a set A. Let $f\colon A^n \to \mathbb{R}$ be a function that satisfies the *bounded difference* condition:

$$\exists c_1, \ldots, c_n > 0 \text{s.t.} \forall\ x_1, \ldots, x_n \in A, x_i' \in A |f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i$$

Then, for all $t > 0$,

$$\Pr\left[f(X_1, \ldots, X_n) - \mathbb{E}\left[f(X_1, \ldots, X_n)\right] \geq t\right] \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

Remark. If $f(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n x_i$ and $A = [0,1]$, then $c_i = \frac{1}{n}$ and the bound recovers the Hoeffding inequality as e^{-2nt^2}.

Chapter 3

Rademacher Complexity

3.1 Uniform Convergence

Motivation: we want to study $L(\hat{h}_{\text{ERM}})$ and compare it against $h^* \in \text{argmin}_{h\in\mathcal{H}} L(h)$. We want to bound the difference $L(\hat{h}_{\text{ERM}}) - L(h^*)$, which is also referred to as the "**excess risk**".

$$L(\hat{h}_{\text{ERM}}) - L(h^*) = \left(L(\hat{h}_{\text{ERM}}) - L_S(\hat{h}_{\text{ERM}})\right) + \left(L_S(\hat{h}_{\text{ERM}}) - L_S(h^*)\right) + \left(L_S(h^*) - L(h^*)\right)$$

where the second term is smaller or equal to 0 by definition, and the third term can be bounded using the Hoeffding inequality as h^* does not depend on S.

Consequently, our aim becomes bounding the first term and we define the following **generalization gap**:

Definition 3.1.1 (Uniform Convergence).

$$L(\hat{h}_{\text{ERM}}) - L_S(\hat{h}_{\text{ERM}}) \le \sup_{h\in\mathcal{H}} \left(L(h) - L_S(h)\right)$$

, where the bounded difference is called the generalization gap.

Theorem 3.1.1 (Generalization Bound for finite hypothesis class). If $\mathcal{H}$ is finite, then for any $\delta \in (0,1)$, we have

$$\text{w.p.} \ge 1-\delta, \sup_{h\in\mathcal{H}} \left(L(h) - L_S(h)\right) \le \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

Remark. If $n >> \log|\mathcal{H}|$, excess risk $\to 0$.

What if $\mathcal{H}$ is infinite?

– Idea: Reduce infinite case to finite case.

3.2 Rademacher Complexity

Notation: Given $\mathcal{H}$ and ℓ, define the family of loss mappings:

$$\mathcal{G} = \{g_h : (x,y) \mapsto \ell(h(x),y), h \in \mathcal{H}\}$$

where $z = (x,y) \sim P$, $z_i = (x_i, y_i)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $L(h) = \mathbb{E}_{z\sim P}\left[g_h(z)\right]$, $L_S(h) = \frac{1}{n}\sum_{i=1}^n g_h(z_i)$.

$$\sup_{h \in \mathcal{H}} (L(h) - L_S(h)) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P}[g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right)$$

Definition 3.2.1 (Empirical Rademacher Complexity)**.** Let $\mathcal{G}$ be a set of functions mapping $\mathcal{Z} \to \mathbb{R}$. Let $S = \{z_1, \cdots, z_n\} \subseteq \mathcal{Z}$.

The **empirical Rademacher complexity** of $\mathcal{G}$ with respect to the simple set S is:

$$R_S(\mathcal{G}) = \mathbb{E}_{\sigma_1, \cdots, \sigma_n} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right]$$

where $\sigma_i = \begin{cases} +1 & \text{w.p.} \frac{1}{2} \\ -1 & \text{w.p.} \frac{1}{2} \end{cases}$ i.i.d (called Rademacher random variables).

Remark. Rademacher complexity measures the ability of a function class to fit random noise

$$R_S(\mathcal{G}) = \mathbb{E}_{\vec{\sigma}} \left[\sup_{g \in \mathcal{G}} \frac{1}{n} < \vec{\sigma}, \vec{g}_s > \right]$$

Definition 3.2.2 (Rademacher Complexity)**.** Let P be a distribution over $\mathcal{Z}$.

For an integer $n \geq 1$, the **Rademacher complexity** of $\mathcal{G}$ is

$$R_n(\mathcal{G}) = \mathbb{E}_{S \sim P^n} [R_S(\mathcal{G})]$$

Theorem 3.2.1 (Generalization Bound using Rademacher Complexity)**.** Let $\mathcal{G}$ be a function class mapping $\mathcal{Z}$ to $[0, 1]$, $S = \{z_1, \ldots, z_n\} \sim P^n$. Then for any $\delta \in (0, 1)$:

$$\text{w.p.} \geq 1 - \delta, \; \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right) \leq 2R_n(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$$\text{w.p.} \geq 1 - \delta, \; \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right) \leq 2R_S(\mathcal{G}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2n}}$$

Proof. **Step 1: Relate the sup terms to the expectation of sups using Mcdiarmid's ineq**

Define $f(z_1, \cdots, z_n) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}_{z \sim P} [g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right)$.

Consider $\{z_1, \ldots, z_n\}$ and $\{z'_1, \ldots, z'_n\}$ that only differs by 1 point (*i.e.* $z_k \neq z'_k, z_i = z'_i \; \forall i \neq k$).

$$\begin{aligned}
f(z_1, \ldots, z_n) &= \sup_{g \in \mathcal{G}} \left(\mathbb{E} [g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z'_i) + \frac{1}{n} \sum_{i=1}^{n} g(z'_i) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right) \\
&\leq \sup_{g \in \mathcal{G}} \left(\mathbb{E} [g(z)] - \frac{1}{n} \sum_{i=1}^{n} g(z'_i) \right) + \sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^{n} g(z'_i) - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right) \\
&= f(z'_1, \ldots, z'_n) + \sup_{g \in \mathcal{G}} \left(\frac{1}{n} g(z'_k) - \frac{1}{n} g(z_k) \right) \\
&\leq f(z'_1, \ldots, z'_n) + \frac{1}{n}
\end{aligned}$$

Similarly, $f(z_i', \dots, z_n') - f(z_1, \dots, z_n) \le \frac{1}{n}$. Combining them we can get that $|f(z_1, \dots, z_n) - f(z_1', \dots, z_n')| \le \frac{1}{n}$.

Applying the Mcdiarmid's inequality, we can get the following bound:

$$\text{w.p.} \ge 1 - \delta, f(z_1, \dots, z_n) - \mathbb{E}\left[f(z_1, \dots, z_n)\right] \le \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Step 2: Bound $\mathbb{E}_S\left[\sup_{g\in\mathcal{G}}\left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right)\right]$ by Rademacher Complexity Draw a fresh set of n samples $S' = \{z_1', \dots, z_n'\} \sim P^n$. Fix S, we have

$$\begin{aligned}
\sup_{g\in\mathcal{G}}\left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right) &= \sup_{g\in\mathcal{G}}\left(\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n g(z_i)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right) \\
&= \sup_{g\in\mathcal{G}}\left(\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n (g(z_i') - g(z_i))\right]\right) \\
&\le \mathbb{E}_{S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n (g(z_i') - g(z_i))\right]
\end{aligned}$$

Taking expectation over S on both sides generate that

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{g\in G}\left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right)\right] &\le \mathbb{E}_{S,S'}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n (g(z_i') - g(z_i))\right] \\
&= \mathbb{E}_{S,S',\vec{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i\left(g(z_i') - g(z_i)\right)\right] \qquad ? \\
&\le \mathbb{E}_{S,S',\vec{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n \sigma_i g(z_i')\right] + \mathbb{E}_{S,S'\vec{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n}\sum_{i=1}^n -\sigma_i g(z_i)\right] \\
&= 2R_n(\mathcal{G})
\end{aligned}$$

Combining the result from **step 1** and **step 2**, we prove the first inequality in the theorem.

Step 3: Prove $R_n(\mathcal{G})$ and $R_S(\mathcal{G})$ are close Similar to step 1, we can verify that $R_S(\mathcal{G})$ satisifies the bounded difference property.

Apply Mcdiarmid's inequality, we can get that

$$\text{w.p.} \ge 1 - \delta, R_n(\mathcal{G}) \le R_S(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Combining the outputs from step 1 - 3 and replacing δ with $\delta/2$ gives the second inequality. ■

Chapter 4

VC-Dimension

In this chapter, we only consider the binary classification case with the 0-1 loss, *i.e.* $y = \{\pm 1\}$ and $\mathcal{G} = \{(x, y) \mapsto \mathbb{1}[h(x) \neq y] : h \in \mathcal{H}\}$.

4.1 Growth Function Bounds

Lemma 4.1.1. $R_n(\mathcal{G}) = \frac{1}{2} R_n(\mathcal{H})$

Proof. Given $S = \{(x_i, y_i)\}_{i=1}^n$, we have

$$\begin{aligned}
R_S(\mathcal{G}) &= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sigma_i \mathbb{1}[h(x_i) \neq y_i]\right] \\
&= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i \frac{1 - y_i h(x_i)}{2}\right] \\
&= \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i + \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i(-y_i)h(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \sigma_i h(x_i)\right] \\
&= \frac{1}{2} R_S(\mathcal{H})
\end{aligned}$$

∎

Remark. It then becomes natural to bound $R_n(\mathcal{H})$.

Definition 4.1.1 (Growth Function)**.** The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis class $\mathcal{H}$ that maps to $y = \{\pm 1\}$ is defined as

$$\Pi_{\mathcal{H}}(n) = \sup_{x_1,\ldots,x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|$$

Remark. This definition defines the set of all possible predictions on a given set of inputs.

Theorem 4.1.1 (Generalization bound using VC-dimension)**.** Let $\mathcal{H}$ be a hypothesis class taking values $y = \{\pm 1\}$. Then

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(n)}{n}}$$

Proof. Let $S = \{x_1, \dots, x_n\}$, $Q = Q_S = \{(h(x_1), \dots, h(x_n) \colon h \in \mathcal{H})\}$.

We want to show that $R_S(\mathcal{H}) \leq \sqrt{\frac{2\log|Q|}{n}}$

$$\begin{aligned} R_S(\mathcal{H}) &= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i)\right] \\ &= \mathbb{E}_{\vec{\sigma}}\left[\sup_{\vec{v}\in Q} \frac{1}{n}\sum_{i=1}^{n} \sigma_i v_i\right] && \text{Apply Hoeffding} \end{aligned}$$

Then for all $\lambda > 0$,

$$\begin{aligned} e^{\lambda R_S(\mathcal{H})} &= e^{\lambda \mathbb{E}_{\vec{\sigma}}\left[\sup_{\vec{v}\in Q} \frac{1}{n}\sum_{i=1}^{n} \sigma_i v_i\right]} \\ &\leq \mathbb{E}_{\vec{\sigma}}\left[e^{\lambda \sup_{\vec{v}\in Q} \frac{1}{n}\sum_{i=1}^{n} \sigma_i v_i}\right] && \text{Jensen's ineq} \\ &\leq \mathbb{E}_{\vec{\sigma}}\left[\sum_{\vec{v}\in Q} e^{\lambda \frac{1}{n}\sum_{i=1}^{n} \sigma_i v_i}\right] && \text{why?} \\ &= \sum_{\vec{v}\in Q} \mathbb{E}_{\vec{\sigma}}\left[e^{\lambda \frac{1}{n}\sum_{i=1}^{n} \sigma_i v_i}\right] \\ &\leq \sum_{\vec{v}\in Q} e^{\frac{\lambda^2}{2n}} && \text{by Hoeffding} \\ &= |Q| e^{\frac{\lambda^2}{2n}} \end{aligned}$$

This gives that $R_S(\mathcal{H}) \leq \frac{1}{\lambda}\log|Q| + \frac{\lambda}{2n}$

Choose $\lambda = \sqrt{2n\log|Q|}$ and we can get that $R_S(\mathcal{H}) \leq \sqrt{\frac{2\log|Q|}{n}}$ ■

Remark. Discussions about the growth function:

- When $\mathcal{H}$ is finite, we have that $\Pi_{\mathcal{H}}(n) \leq |\mathcal{H}|$

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{n}} \text{ recovers Thm 1}$$

- When $\mathcal{H}$ is "super power", $\Pi_{\mathcal{H}}(n) = 2^n$, *i.e. overfitting.*

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2\log 2^n}{n}} = \sqrt{2\log 2}$$

- What if the growth function is in-between, a polynomial function?

  Suppose $\Pi_{\mathcal{H}}(n) \leq n^d$, we have that

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2d\log n}{n}} \to 0 \text{ if } n >> d\log d$$

Definition 4.1.2 (VC-dimension)**.** The VC-dimension of a class of hypothesis function $\mathcal{H}$ is

$$\mathrm{VC}(\mathcal{H}) = \max\{n \colon \Pi_{\mathcal{H}}(n) = 2^n\}$$

Definition 4.1.3 (Shatter). $S = \{x_1, \ldots, x_n\}$ can be shattered by $\mathcal{H}$ if $\forall\ y_1, \ldots, y_n \in \{\pm 1\}, \exists h \in \mathcal{H}$s.t.$h(x_i) = y_i$ for all $i = \{1, \ldots, n\}$.

Remark. The VC-dimension is the maximum size of a sample set S that can be shattered by $\mathcal{H}$.

Example (Threshold Function). Let $\mathcal{X} = \mathbb{R}, \mathcal{H} = \{h_a : a \in \mathbb{R}\}$, $h_a \in \mathcal{H}$, $h_a(x) = \begin{cases} +1, & \text{if } x \geq a \\ -1, & \text{if } x < a \end{cases}$

$$\text{Then } \mathbf{VC-dim}(\mathcal{H}) = 1$$

Proof.

1. any input $x \in \mathbb{R}$ can be shattered

$$h_{x-1}(x) = +1, \quad h_{x+1} = -1$$

2. any inputs $x_1, x_2 \in \mathbb{R}$ cannot be shattered

$$x_1 \leq x_2, \text{ impossible to label } (+1, -1)$$

■

Theorem 4.1.2 (growth function bound). Let $\mathcal{H}$ be a hypothesis class with VC-dimension d. Then,

$$\forall\ n >> d \colon \Pi_{\mathcal{H}}(n) << \left(\frac{e^n}{d}\right)^d \leq n^d \text{ if } d \geq 3$$

Theorem 4.1.3 (Generalization Bound Using VC-Dimension). Let $\mathcal{H}$ be a hypothesis class taking values in $y = \{\pm 1\}$ and has VC-dim d. Consider the 0-1 loss.

Then, for all $\delta \in (0, 1)$,

$$\text{w.p.} \geq 1 - \delta, \sup_{h \in \mathcal{H}} (L(h) - L_S(h)) \leq \sqrt{\frac{2d \log e^n}{d}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Remark. This VC-dimension bound requires $n >> d$. In other words, it is effective when the hypothesis class is relatively less expressive.

4.2 More on VC-Dimension

First we look at more examples illustrating the concept of VC-dimension.

Example (Axis-aligned rectangles). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_{a,b,c,d} \colon a, b, c, d \in \mathbb{R}\}$, and $h \in \mathcal{H}$ be the form

$$h_{a,b,c,d}(x) = \begin{cases} 1 & \text{if } x_1 \in [a, b], x_2 \in [c, d] \\ -1 & \text{otherwise} \end{cases}$$

Then we have **Vc-dim** $\mathcal{H} = 4$.

Proof.

1. there exists 4 points that can be shattered exists or for all?
2. Any 5 points cannot be shattered
   Choose the minimum axis-aligned rectangle that contains all 5 points, then it is impossible to label the sides +1 while labeling inside one -1

■

Example (Linear Functions). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_w : w \in \mathbb{R}^d\}$, and

$$h_w(x) = \text{sign}(w^T x) = \begin{cases} 1 & \text{if } w^T x \geq 0 \\ -1 & \text{if } w^t x < 0 \end{cases}$$

Then $\textbf{Vc-dim}(\mathcal{H}) = d$.

Proof.
1. $\exists$ d points that can be shattered Same Question exists or for all?

   Choose $x_1, \ldots, x_d \in \mathbb{R}^d$ that are linearly independent.
   Then for all $y_1, \ldots, y_d \in \{\pm 1\}$, we can find a $w \in \mathbb{R}^d$ such that $w^T x_i = y_i$, for all $i = 1, \ldots, d$ by solving the set of linear equations.

2. Any $d + 1$ point cannot be shattered

   Assume for the sake of contradiction that there exists $d + 1$ points: $x_1, \ldots, x_d$ that can be shattered.
   In formal terms, $\exists$ $\alpha = (\alpha_1, \ldots, \alpha_{d+1})$ s.t. $\sum_{i=1}^{d+1} \alpha_i x_i = 0$, $\alpha \neq 0$, *i.e.* $\exists$ a coordinate $k \in \{1, \ldots, d+1\}$ s.t. $\alpha_k \neq 0$. WLOG we can assume $\alpha_k > 0$.

   For all $w \in \mathbb{R}^d$, we must have $\sum_{i=1}^{d+1} \alpha_i w^T x_i = 0$. why?

   Then let $y_i = \text{sign}(\alpha_i), i = 1, \ldots, d+1$. $\exists w \in \mathbb{R}^d$ s.t. $\text{sign}(w^T x_i) = y_i$.

   Then we find the contradiction:

$$0 = \sum_{i=1}^{d+1} \alpha_i (w^T x_i) < 0 \qquad \text{opposite sign}$$

■

Example (Sine Function). Let $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_\omega : \omega \in \mathbb{R}\}$, and $h = \text{sign}\left(\sin(\omega x)\right)$

Then $\textbf{Vc-dim}(\ \mathcal{H}) = \infty$.

Proof. It suffices to show that $\exists$ n points that can be shattered, for any n.
Consider n points, $x_i = 2^{-i}$ $(i = 1, \ldots, n)$ and any labeling $y_1, \ldots, y_n \in \{\pm 1\}$.

Define $\frac{\omega}{\pi} = \left(y'_n y'_{n-1} \ldots y'_1 1\right)_2$ in terms of binary integer, where $y'_i = \begin{cases} 0 & \text{if } y_i = 1 \\ 1 & \text{if } y_i = -1 \end{cases}$

WTS $\text{sign}\left(\sin(\omega x_i)\right) = y_i$,

which can be realized through

$$\frac{\omega x_i}{\pi} = \frac{\omega}{\pi} 2^{-i} = \left(y'_n y'_{n-1} \ldots y'_1 1\right)_2$$

Not fully understand ■

Theorem 4.2.1 (VC-dimension in finite precision). Let $\mathcal{H}$ be parametrized by p parameters, with each stored in k bits. $\mathcal{H} = \{h_\theta, \theta \in \mathbb{R}^P\}$, then $\text{VC-dim}(\mathcal{H}) \leq k \cdot p$.

Proof. There are $(2^k)^p$ choices for $\theta = (\theta_1, \ldots, \theta_p)$, and then

$$2^{\text{Vc-dim}(\mathcal{H})} \leq |\mathcal{H}| \leq 2^{kp}$$

■

Remark (Limitation of VC-dimension).

$$\begin{aligned} L(h) - L_S(h) &\leq \tilde{O}\left(\sqrt{\frac{VC - dim(\mathcal{H})}{n}}\right) \\ &\leq \tilde{O}\left(\sqrt{\frac{\#\text{params}}{n}}\right) \end{aligned}$$

If $\#$ params » $\#$ samples, the bound will become vacuous.

Chapter 5

Margin Theory

We focus on the binary classification setting where $y = \{\pm 1\}$.

5.1 Basic Setups

Definition 5.1.1 (Margin). The margin of a function $h\colon \mathcal{X} \to \mathbb{R}$ at a point $x \in \mathcal{X}$ labeled with $y \in \{\pm 1\}$ is $yh(x)$.

Remark. We have $\hat{y} = \operatorname{sign}(h(x))$; and a classification is correct when $yh(x) > 0$.

Definition 5.1.2 (Margin Loss). For any $\gamma > 0$, define γ-margin loss as

$$\ell_\gamma(y', y) = \ell_\gamma(yy') = \begin{cases} 1, & \text{if } yy' \leq 0 \\ 1 - \frac{yy'}{\gamma} & \text{if } 0 < yy' < \gamma \\ 0, & \text{if } yy' \geq \gamma \end{cases}$$

Remark. Margin Loss $\geq$ 0-1 loss (in terms of their graphs).

Definition 5.1.3 (Population & Empirical Risk for Margin Loss).

$$L_\gamma(h) = \mathbb{E}_{(x,y)\sim P}\left[\ell_\gamma\left(h(x), y\right)\right]$$

$$L_{\gamma,S}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell_\gamma\left(h(x_i), y_i\right)$$

Remark. $\ell_\gamma(\cdot)$ is $\frac{1}{\gamma}$-Lipschitz.

SideNote: We say $f\colon \mathbb{R} \to R$ is $C-$Lipschitz if $|f(x) - f(x')| \leq C|x - x'|$ for all $x, x' \in \mathbb{R}$. OR equivalently, $|f'(x)| \leq C, \forall x \in R$.

Lemma 5.1.1 (Talagrend's Lemma). Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a C-Lipschitz function. Then,

$$R_S(\phi \circ \mathcal{H}) \leq C \cdot R_S(\mathcal{H})$$

where $\phi \circ \mathcal{H} = \{z \mapsto \phi(h(z))\colon h \in \mathcal{H}\}$.

Theorem 5.1.1 (Margin-based generalization bound for binary classificaiton). Let $\mathcal{H}$ be a function class mapping $\mathcal{X} \to \mathbb{R}$. Fix $\gamma > 0$. Then, for any $\delta \in (0, 1)$, with probability $\geq 1 - \delta$ we have:

$$\sup_{h \in \mathcal{H}} (L_\gamma(h) - L_{\gamma,S}(h)) \leq \frac{2}{\gamma} R_n(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Also with probability $\geq 1 - \delta$, we have:

$$\sup_{h \in \mathcal{H}} (L_\gamma(h) - L_{\gamma,S}(h)) \leq \frac{2}{\gamma} R_S(\mathcal{H}) + 3\sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Proof.

$$\begin{aligned} G_\gamma &= \{(x, y) \mapsto \ell_\gamma(yh(x)) \colon h \in \mathcal{H}\} \\ &= \{(x, y) \mapsto \ell_\gamma(\hat{h}(x, y)) \colon \hat{h} \in \hat{\mathcal{H}}\} \\ &= \ell_\gamma \circ \hat{\mathcal{H}} \end{aligned}$$

where $\hat{\mathcal{H}} = \{(x, y) \to yh(x) \colon h \in \mathcal{H}\}$.

$$\begin{aligned} R_S(\hat{\mathcal{H}}) &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i y_i h(x_i) \right] \\ &= \mathbb{E}_{\vec{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right] \\ &= R_S(\mathcal{H}) \end{aligned}$$

By Talagrend's lemma, $R_S(G_\gamma) \leq \frac{1}{\gamma} R_S(\hat{\mathcal{H}}) = \frac{1}{\gamma} R_S(\mathcal{H})$.

Completes the proof by applying the generalization bound for G_γ

What generalization bound? ■

Part II

Optimization

Chapter 6

Gradient Descent

Please refer to appendix 1 and 2 for a basic calculus and linear algebra recap at first.

Gradient Descent is an iterative algorithm, where we are concerned with

$$\min_{x \in \mathbb{R}^d} f(x)$$

The algorithm starts at $x_0 \in \mathbb{R}^d$ and iteratively update the variable $x_1, x_2, \ldots$.

When the point is at x_t, we can do 1st-order Taylor expansion:

$$f(x_t + \Delta x) \approx f(x_t) + \langle \nabla f(x_t), \Delta x \rangle + \cdots$$

In order to decrease f as much as possible, we can choose $\Delta x // -\nabla f(x_t)$.

Remark.

$$\inf_{\|\Delta x\|_2 \leq \epsilon} \langle a, \Delta x \rangle = -\epsilon \|a\|_2$$

the optimum occurs at $\Delta x = \epsilon \frac{a}{\|a\|_2}$

This motivates **Gradient Descent**(GD).

$$x_{t+1} = x_t - \eta \nabla f(x_t), \ t = 0, 1, 2, \ldots$$

where $\eta > 0$ is called *step size* or *learning rate.*

In order for GD to do what it's supposed to do, we want the 1st-order Taylor expansion to be accurate.

Error of 1st-order Taylor:

$$\begin{aligned} f(x) - f(x_t) - \langle \nabla f(x_t), x - x_t \rangle &= \frac{1}{2}(x - x_t)^T \nabla^2 f(\xi)(x - x_t) \\ &\leq \frac{1}{2}\|\nabla^2 f(\xi)\|_2 + \|x - x_t\|_2^2 \end{aligned}$$

Definition 6.0.1 (smoothness). A differentiable function $f \colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth ($\beta > 0$) if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2, \ \forall \ x, y$$

In other words, gradient of f is β-Lipschitz.

Remark. When f is twice differentiable, $f\colon \mathbb{R}^d \to \mathbb{R}$ is equivalent to

$$\|\nabla^2 f(x)\|_2 \leq \beta,\ \forall\ x$$

Lemma 6.0.1. If f is β-smooth, then:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{\beta}{2}\|x-y\|_2^2$$

Proof.

$$\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= |\int_0^1 \langle \nabla f(x + t(y-x)), y - x\rangle - \int_0^1 \langle \nabla f(x), y-x\rangle dt| && \text{FTC}\\
&= |\int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x), y - x\rangle dt|\\
&\leq \int_0^1 \|\nabla f(x+t(y-x)) - \nabla f(x)\|_2 \|y-x\|_2 && \text{Cauchy-Schwardz}\\
&\leq \int_0^1 \beta \cdot \|t(y-x)\|_2 \cdot \|y-x\|_2 dt && \text{beta - smooth}\\
&= \beta\|y-x\|_2^2 \cdot \int_0^1 tdt\\
&= \frac{\beta}{2}\|y-x\|_2^2
\end{aligned}$$

■

Lemma 6.0.2 (Descent Lemma)**.** If f is β-smooth and $\eta \leq \frac{1}{\beta}$, then GD with step size η ($x_{t+1} = x_t - \eta\nabla f(x_t)$) satisfies

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$$

Proof.

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \langle f(x_t), t_1 - t\rangle + \frac{\beta}{2}\|t_1 - t\|_2^2 && \text{previous lemma}\\
&= f(x_t) + \langle f(x_t), -\eta\nabla f(x_t)\rangle + \frac{\beta}{2}\|-\eta\nabla f(x_t)\|_2^2\\
&= f(x_t) - \left(\eta - \frac{\beta}{2}\eta^2\right)\|f(x_t)\|_2^2\\
&\leq f(x_t) - \frac{\eta}{2}\|f(x_t)\|_2^2
\end{aligned}$$

■

Remark. Descent Lemma shows that every step in a $\beta-$smooth function f decreases the function value.

Corollary 6.0.1. If f is β-smooth, then GD with step size $\eta \leq \frac{1}{\beta}$ must satisfy:

- $\lim_{t\to\infty} f(x_t)$ exists
- $\lim_{t\to\infty} \|\nabla f(x_t)\|_2 = 0$, since function converges and $f(x_t) - f(x)$ is bounded by it.

6.1 Convex Optimization

Definition 6.1.1 (convexity)**.** We present the following definitions:

convex set: A set $X \subseteq \mathbb{R}^d$ is convex if

$$\forall x, y \in X,\ \forall \gamma \in (0,1)\colon (1-\gamma)x + \gamma y \in X$$

convex function: A function $f\colon X \to \mathbb{R}$ is convex if X is convex and

$$\forall x, y \in X,\ \forall \gamma \in (0,1)\colon f\left((1-\gamma)x + \gamma y\right) \leq (1-\gamma)f(x) + \gamma f(y)$$

Example. Common Convex Functions

- linear function
- squared norm

Example (Examples that preserve convexity)**.** E.g.

- non-negative weighted sum
- composition with affine mapping
- pointwise supreme

Example (Linear Model)**.** Given dataset $S = \{x_i, y_i\}_{i=1}^n$, $\mathcal{H} = \{x \mapsto w^T x \colon w \in \mathbb{R}^d\}$. Empirical risk $L_S(w) = \frac{1}{n}\sum_{i=1}^n \ell(w^T x_i, y_i)$

claim: If $\ell(y', y)$ is convex in its first argument (for any fixed y), then L_S is convex.

Let's see the common loss functions that are convex in first argument. $(y \in \{\pm 1\})$

- squared loss: $\ell(y', y) = (y - y')^2$ Convex
- 0-1 loss: $\ell(y', y) = \mathbb{1}\left[yy' \leq 0\right]$ not Convex
- Margine loss: not convex
- Hinge loss: convex
- logistic loss: $\ell(y', y) = \log\left(1 + e^{-yy'}\right)$ convex

Lemma 6.1.1 (first-order & second-order characterization of convex functions)**.** First, if f is differentiable, then f is convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle,\ \forall x, y$$

Second, if f is twice-continuously differentiable, then f is convex if and only if

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

Definition 6.1.2 (Local Minimum)**.** A local minimum of a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is a point $x \in \mathbb{R}$ such that $\exists\ \epsilon > 0$:

$$f(x) \leq f(y),\ \forall\ y \text{ satisfying } \|y - x\|_2 \leq \epsilon$$

Lemma 6.1.2. Every local minimum of a convex function is a global minimum.

Proof. Suppose x is a local minimum but not global minimum, *i.e.*, $\exists\ y$ s.t. $f(y) < f(x)$.

By convexity, for all $\gamma \in (0, 1)$,

$$\begin{aligned} f((1-\gamma)x + \gamma y) &\leq (1-\gamma)f(x) + \gamma f(y) \\ &\leq (1-\gamma)f(x) + \gamma f(x) \\ &= f(x) \end{aligned}$$

As we take $\gamma \to 0$, we have $\|(1-\gamma)x + \gamma y - x\|_2 \to 0$, yielding a contraction. ■

Remark. Isn't it "As we take $\gamma \to 1$, we have $\|f((1-\gamma)x + \gamma y) - f(x)\|_2 \to 0$, yielding a contraction."

Lemma 6.1.3. If $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and $\nabla f(x) = 0$ (*i.e.* x is a stationary point), then x is a global minimum.

Proof. $\forall\, y, f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle = f(x)$ ■

Lemma 6.1.4. If $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable, and x is a local minimum, then $\nabla f(x) = 0$.

Corollary 6.1.1. If f is convex and differentiable, then x is a global minimum if and only if $\nabla f(x) = 0$.

6.2 Convergence of GD for Smooth Convex Functions

Lemma 6.2.1 (contraction lemma)**.** If f is convex and β-smooth, and $\eta \leq \frac{1}{\beta}$, then:

$$\|x_{t+1} - x^*\|_2 \leq \|x_t - x^*\|_2, \ \forall\, t$$

Proof.

$$\begin{aligned} \|x_{t+1} - x^*\|_2^2 &= \|(x_t - x^*) - \eta \nabla f(x)\|_2^2 \\ &= \|x_t - x^*\|_2^2 - 2\eta \langle x_t - x^*, \nabla f(x_t) \rangle + \eta^2 \|\nabla f(x)\|_2^2 \\ &\leq \|x_t - x^*\|_2^2 - 2\eta\left(f(x_t) - f(x^*)\right) + 2\eta\left(f(x_t) - f(x^{t+1})\right) \\ &= \|x_t - x^*\|_2^2 - 2\eta\left(f(x_{t+1}) - f(x^*)\right) \\ &\leq \|x_t - x^*\|_2^2 \end{aligned}$$

Where the step from 2nd to 3rd line relies on the convexity assumption and the descent lemma. ■

Remark. In addition to the descent lemma, contraction lemma tells us that not only the function value decreases, the next step's x always gets closer to the optimum point.

Theorem 6.2.1 (GD convergence for smooth convex functions)**.** If f is convex and β-smooth, and $\eta \leq \frac{1}{\beta}$, then:

$$f(x_t) - f(x^*) \leq \frac{2\|x_0 - x^*\|_2^2}{\eta t}, \quad \forall t \geq 1$$

Proof. Let $\delta_t = f(x_t) - f(x^*)$.
By the descent lemma,

$$\delta_{t+1} \leq \delta_t - \frac{\eta}{2}\|\nabla f(x)\|_2^2 \tag{1}$$

By convexity and contraction lemma,

$$\begin{aligned}\delta_t \leq \langle \nabla f(x_t), x_t - x^* \rangle &\leq \|\nabla f(x_t)\| \|x_t - x^*\| \\ &\leq \|\nabla f(x_t)\| \|x_0 - x^*\| \end{aligned} \tag{2}$$

By putting (1) and (2) together, we have

$$\delta_{t+1} \leq \delta_t - \frac{\eta}{2} \left(\frac{\delta_t}{\|x_0 - x^*\|_2} \right)^2$$

Diving this inequality by $\delta_t \cdot \delta_{t+1}$, we can get the following:

$$\frac{1}{\delta_t} \leq \frac{1}{\delta_{t+1}} - \frac{\eta}{2\|x_0 - x^*\|_2} \cdot \frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{\eta}{\|x_0 - x^*\|_2}$$

Taking the sum over $t = 0, 1, \ldots, t-1$:

$$\sum_{s=0}^{t-1} \left(\frac{1}{\delta_s} - \frac{1}{\delta_{s+1}} \right) \leq -\frac{\eta t}{2\|x_0 - x^*\|_2}$$

which gives to the desired inequality:

$$\delta_t \leq \frac{2\|x_0 - x^*\|_2^2}{\eta t}$$

■

Remark. Let $\eta = \frac{1}{\beta}$, we get $\delta_t \leq \frac{2\beta\|x_0 - x^*\|_2^2}{t}$.

To get $\delta_t \leq \epsilon$, we need the step size $t \geq \frac{2\beta\|x_0 - x^*\|_2^2}{\epsilon}$

6.3 Convergence of GD for smooth and strongly convex functions

Definition 6.3.1 (strong convexity). $f \colon \mathbb{R}^d \to \mathbb{R}$ is α-strongly convex ($\alpha > 0$) if $f(x) - \frac{\alpha}{2}\|x\|_2^2$ is convex.

Lemma 6.3.1 (first-order & second-order characterization of strong convexity).

1. first-order: If $f \colon \mathbb{R}^d \to \mathbb{R}$ is differentiable, then f is α-SC if and only if:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2 \ \forall\ x, y$$

2. second-order: If $f \colon \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable, then f is α-SC if and only if:

$$\nabla^2 f(x) \succeq \alpha I, \ \forall x$$

Example. $f(x) = \frac{1}{2}x^T A x$ for symmetric A. Then we know $\nabla^2 f(x) = A$.

If $\lambda_{\min}(A) > 0$, then f is $\lambda_{\min}(A) -$ SC, and $\lambda_{\min} -$ smooth.

Theorem 6.3.1 (GD convergence for smooth and strongly-convex functions). If f is β-smooth and $\alpha -$ SC, and $\alpha \leq \frac{1}{\beta}$, then

1. $\|x_{t+1} - x^*\|_2^2 \le (1 - \alpha\eta) \cdot \|x_t - x^*\|_2^2$, $\forall t$.

2. $f(x_t) - f(x^*) \le \frac{\beta}{2}(1 - \alpha\eta)^t \|x_0 - x^*\|_2^2$.

Proof.

$$\begin{aligned}
\|x_{t+1} - x^*\|_2^2 &= \|(x_t - x^*) - \eta \nabla f(x_t)\|_2^2 \\
&= \|x_t - x^*\|_2^2 - 2\eta \langle x_t - x^*, \nabla f(x_t)\rangle + \eta^2 \|\nabla f(x_t)\|_2^2 \\
&\le \|x_t - x^*\|_2^2 - 2\eta \left(f(x_t) - f(x^*) + \frac{\alpha}{2}\|x_t - x^*\|_2^2 \right) + 2\eta \left(f(x_t) - f(x_{t+1}) \right) \\
&= (1 - \alpha\eta)\|x_t - x^*\|_2^2 - 2\eta \left(f(x_{t+1}) - f(x^*) \right) \\
&\le (1 - \alpha\eta)
\end{aligned}$$

which finishes (1). To see how we prove (2), we proceed as follows:

By (1), we have $\|x_t - x^*\|_2^2 \le (1 - \alpha\eta)^t \|x_0 - x^*\|_2^2$, then:

$$\begin{aligned}
f(x_t) - f(x^*) &\le \langle \nabla f(x^*), x_t - x^* \rangle + \frac{\beta}{2}\|x_t - x^*\|_2^2 \\
&\le \frac{\beta}{2}(1 - \alpha\eta)^t \|x_0 - x^*\|_2^2
\end{aligned}$$

■

Remark. If $\eta = \frac{1}{\beta}$:

$$\begin{aligned}
\|x_t - x^*\|_2^2 &\le \left(1 - \frac{\alpha}{\beta}\right) \|x_0 - x^*\|_2^2 \\
&= \left(1 - \frac{1}{\kappa}\right) \|x_0 - x^*\|_2^2 \\
&\le e^{-\frac{t}{\kappa}} \|x_0 - x^*\|_2^2
\end{aligned}$$

where we set $\kappa = \frac{\beta}{\alpha}$ to be the "condition number" and the last line comes from $1 - \frac{1}{\kappa} \le e^{-\frac{1}{\kappa}}$.

Alternatively, if we want $\|x_t - x^*\|_2^2 \le \epsilon$, we need to set the step size $t \ge \kappa \log \frac{\|x_0 - x^*\|_2^2}{\epsilon}$.

Part III

Appendix

Chapter 7

Calculus

7.1 Taylor Expansion

In 1-dimension, for a function $f : \mathbb{R} \to \mathbb{R}$, the taylor expansion of f at a point x_0 can be expressed as:

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \cdots + \frac{1}{k!}f^{(k)}(x_0)(x - x_0)^k + \cdots$$

In particular, we call the **approximation error**, or equivalently, **remainder** as:

$$f(x) - f_{k,x_0}(x) = \frac{1}{(k+1)!}f^{(k+1)}(\xi)(x - x_0)$$

where ξ is between x_0 and x.

Let's do some extension to **multivariate** functions where $f : \mathbb{R}^d \to \mathbb{R}$.

The taylor expansion at x_0 can be expressed as:

$$f(x) \approx f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}(x - x_0)^T \nabla^2 f(x_0)(x - x_0) + \cdots$$

Similarly, we can also have remainder:

$$f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0 \rangle = \frac{1}{2}(x - x_0)^T \nabla^2 f(\xi)(x - x_0)$$

where $\xi = x_0 + t(x - x_0)$, for some $t \in [0, 1]$.

Fundamental Theorem of Calculus

in 1-d:

$$f(x) - f(x_0) = \int_{x_0}^{x} f'(t)dt$$

in multivariate:

$$f(x) - f(x_0) = \int_0^1 \langle \nabla f(x_0 + t(x - x_0)), x - x_0 \rangle dt$$

7.2 Linear Algebra

Definition 7.2.1. Let $A \in \mathbb{R}^{d\times d}$

- Eigenvalue, eigenvector: $A\vec{v} = \lambda\vec{v}$, $\vec{v} \neq 0$
- If A is symmetric, it has d real eigenvalues:

$$\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A) = \lambda_{\min}(A)$$

- Spectral norm: $\|A\|_2 = \sup_{\|v\|=1} \|Av\|_2 = \sup_{\|u\|=\|v\|=1} |u^T Av|$

  If A is symmetric, $\|A\|_2 = \max_i |\lambda_i(A)| = \max\{|\lambda_1(A), \lambda_d(A)|\}$
- PSD matrix: $A \succeq 0$ if A is symmetric and $\lambda_{\min}(A) \geq 0$, or equivalently: A is symmetric and $v^T Av \geq 0$, for all $v \in \mathbb{R}^d$.

  Moreover, we write $A \succeq B$ if $A - B \succeq 0$, $\lambda_{\min}(A) \cdot I \preceq A\lambda_{\max}(A) \cdot I$