Machine Learning Theory Notes

Gaotang Li

March 31, 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.*:

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

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 $X_1, \ldots, X_n$, 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 $x_1, \ldots, x_n$ 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 $\psi(\lambda) \coloneqq \log \mathbb{E}\left[e^{\lambda Z}\right]$.
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 $t > 0$:

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

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 $f: A^n \to \mathbb{R}$ 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**".

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

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)**.**

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

, 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

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

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:

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

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

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

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

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

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 $\delta\in(0,1)$:

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

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

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

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

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

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

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

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

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

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 $S' = \{z_1', \ldots, z_n'\} \sim P^n$. Fix S, we have

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

Taking expectation over S on both sides generate that

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

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

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

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 $\Pi_{\mathcal{H}} \colon \mathbb{N} \to \mathbb{N}$ 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 $R_S(\mathcal{H}) \le \sqrt{\frac{2\log|Q|}{n}}$

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

Then for all $\lambda > 0$,

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

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

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

Remark. Discussions about the growth function:

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

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

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

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

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

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

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

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,

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

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 $\mathcal{X} = \mathbb{R}$, $\mathcal{H} = \{h_{a,b,c,d} : a, b, c, d \in \mathbb{R}\}$, 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 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

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

Remark. Margin Loss $\ge$ 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 $|f(x) - f(x')| \le C|x - x'|$ for all $x, x' \in \mathbb{R}$. OR equivalently, $|f'(x)| \le C, \forall x \in R$.

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

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

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:

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

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

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

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? ■

Remark.

$$L_{0-1}(h) \leq L_\gamma(h) \leq L_{\gamma,S}(h) + \frac{2}{\gamma} R_n(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Tradeoff: As γ increases, $L_{\gamma,S}(h)$ increases but $\frac{2}{\gamma} R_n(\mathcal{H})$ decreases.

Assume that h perfectly classifies the dataset S:

$$y_i h(x_i) > 0, \ \forall i = 1, \ldots, n \qquad L_{0-1,S}(h) = 0$$

Let $\gamma_{S,h} = \min_{i \in [n]} y_i h(x_i)$, which is the maximum γ such that $L_{\gamma,S}(h) = 0$. Then:

$$L_{0-1}(h) \leq L_\gamma(h) \leq \frac{2}{\gamma_{S,h}} R_n(\mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

5.2 Margin Bound for Linear functions

Theorem 5.2.1 (Rademacher complexity of linear functions with bounded weights)**.** Let $\mathcal{H} = \{x \mapsto w^T x \colon w \in \mathbb{R}^d, \|w\|_2 \leq B\}$ and

Chapter 6

Generalization bounds via covering numbers

6.1 Setups

Definition 6.1.1 (metric space). (M, ρ) is a metric space if M is a set and ρ: $M \times M \to \mathbb{R}$ is a distance function that satisfies:

- $\rho(x, x) = 0, \ \forall x \in M.$
- $\rho(x, y) > 0, \ \forall x, y \in M, x \neq y$
- $\rho(x, y) = \rho(y, x), \ \forall x, y \in M$
- $\rho(x, z) \leq \rho(x, y) + \rho(y, z), \ \forall x, y, z \in M$

Example. $(\mathbb{R}^d, \|\cdot\|_2), \rho(x, y) = \|x - y\|_2$.

Definition 6.1.2 (ϵ-cover/ϵ-net). Let (M, ρ) be a metric space and $A \subseteq M$. We say that $C \subseteq M$ is an ϵ-cover for A if:

$$\forall x \in A, \exists \ x' \in C \text{ s.t. } \rho(x, x') \leq \epsilon$$

Definition 6.1.3 (covering number). Covering number is the minimum size of an ϵ-cover:

$$N(\epsilon, A, \rho) = \min\{|C| \colon C \text{ is an } \epsilon\text{-cover for } A \text{ under distance } \rho\}$$

Example. Given $(\mathbb{R}^d, \|\cdot\|_2), A = \{x \in \mathbb{R}^d \colon \|x\|_2 \leq 1\}$, then for any $\epsilon \in (0, 1)$,

$$\left(\frac{1}{\epsilon}\right)^d \leq N(\epsilon, A, \|\cdot\|_2) \leq \left(\frac{3}{\epsilon}\right)^d$$

Proof. For the lower bound, Let C be an ϵ-cover, then

$$\text{Vol}(A) \leq |C| \cdot \text{Vol}(\epsilon\text{-ball})$$

Leading to that $|C| \geq \frac{\text{Vol(1-ball)}}{\text{Vol}(\epsilon\text{-ball})} = \left(\frac{1}{\epsilon}\right)^d$.

For the Upper bound, we will construct an ϵ-cover:

repeatedly adding points that are not covered by existing ϵ-balls.

$C = \{x_1, x_2, \cdots, x_n\}$: ϵ-cover, $\|x_i - x_j\|_2 > \epsilon \ \forall i \neq j$

We have the property: all $\frac{\epsilon}{2}$-balls don't intersect, and the following:

$$\begin{aligned} |C| &\leq \frac{\text{Vol}\left((1+\frac{\epsilon}{2})\text{-ball}\right)}{\text{Vol}\left(\frac{\epsilon}{2}\text{-ball}\right)} \\ &= \left(\frac{1+\frac{\epsilon}{2}}{\frac{\epsilon}{2}}\right)^d \\ &\leq \left(\frac{3}{\epsilon}\right)^d \end{aligned}$$

■

Remark.

$$d \log \frac{1}{\epsilon} \leq \log N(\epsilon, A, \|\cdot\|_2) \leq d \log \frac{3}{\epsilon}$$

6.2 Relation to Rademacher Complexity

We fix a sample set $S = \{x_1, \cdots, x_n\}$ and bound its empirical Rad complexity:

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

Define $Q = \{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}$, and $R_S(\mathcal{H}) = \mathbb{E}_{\vec{\sigma}} \left[\sup_{\vec{v} \in Q} \frac{1}{n} \langle \vec{\sigma}, \vec{v} \rangle\right]$.

Lemma 6.2.1 (Massart's lemma). If Q is finite and $\sup_{\vec{v} \in Q} \frac{1}{\sqrt{n}} \|\vec{v}\|_2 \leq M$, then

$$R_S(\mathcal{H}) \leq \sqrt{\frac{2M^2 \log |Q|}{n}}$$

Remark. We prove this earlier in the Rad complexity chapter.

Theorem 6.2.1 (Discretization theorem for Rad complexity). Let $\mathcal{H}$ be a function class taking values in $[-M, M]$, then:

$$R_S(\mathcal{H}) \leq \inf_{\epsilon > 0} \left(\epsilon + \sqrt{\frac{2M^2 \log N(\epsilon, Q, \frac{1}{\sqrt{n}} \|\cdot\|_2)}{n}} \right)$$

Proof. Let C be an ϵ-cover for Q in $(\mathbb{R}^n,)$ ■

Part II

Optimization

Chapter 7

Gradient descent and Convex Optimization

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.

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

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:

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

Definition 7.0.1 (smoothness). A differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ is β-smooth ($\beta > 0$) if

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

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

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

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

Lemma 7.0.1. 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.

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

■

Lemma 7.0.2 (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 7.0.1. 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.

7.1 Convex Optimization

Definition 7.1.1 (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: $\ell(y', y) = \mathbb{1}[yy' \leq 0]$ not Convex
- Margine loss: not convex
- Hinge loss: convex
- logistic loss: $\ell(y', y) = \log\left(1 + e^{-yy'}\right)$ convex

Lemma 7.1.1 (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 7.1.2 (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 7.1.2. 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 $\gamma \in (0,1)$,

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

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

Lemma 7.1.4. If $f\colon \mathbb{R}^d \to \mathbb{R}$ is differentiable, and x is a local minimum, then $\nabla f(x)=0$.

Corollary 7.1.1. If f is convex and differentiable, then x is a global minimum if and only if $\nabla f(x)=0$.

7.2 Convergence of GD for Smooth Convex Functions

Lemma 7.2.1 (contraction lemma). If f is convex and β-smooth, and $\eta \le \frac{1}{\beta}$, then:

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

Proof.

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

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

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

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

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

By convexity and contraction lemma,

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

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

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

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

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

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

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

which gives to the desired inequality:

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

■

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

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

Chapter 8

Convergence of GD Under Certain Conditions

8.1 For Smooth and Convex Functions

Definition 8.1.1 (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 8.1.1 (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 $\lambda_{\min}(A) - \text{SC}$, and $\lambda_{\min} - \text{smooth}$.

Theorem 8.1.1 (GD convergence for smooth and strongly-convex functions). If f is β-smooth and $\alpha - \text{SC}$, and $\alpha \leq \frac{1}{\beta}$, then

1. $\|x_{t+1} - x^*\|_2^2 \leq (1 - \alpha\eta) \cdot \|x_t - x^*\|_2^2,\ \forall t.$
2. $f(x_t) - f(x^*) \leq \frac{\beta}{2}(1 - \alpha\eta)^t \|x_0 - x^*\|_2^2.$

Proof.

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

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

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

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

■

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

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

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

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

Example (Linear Regression). $x_1, ..., x_n \in \mathbb{R}^d$, $y_1, ..., y_n \in \mathbb{R}$, $\mathcal{H} = \{x \mapsto w^T x : w \in \mathbb{R}^d\}$

ERM minimizes:

$$\begin{aligned} L(w) &= \frac{1}{2n} \sum_{i=1}^{n} (w^T x_i - y)^2 \\ &= \frac{1}{2n} \|XW - y\|_2^2 \\ &= \frac{1}{2n} (w^T X^T X w - 2y^T X w + y^T y) \end{aligned}$$

where $X = [x_1^T \dots x_n^T]^T \in \mathbb{R}^{n \times d}$, $y = [y_1 \dots y_n]^T \in \mathbb{R}^n$.

We can calculate gradient and Hessian of $L(w)$ as follows:

$$\begin{aligned} \nabla L(w) &= \frac{1}{2n} \left(2X^T X W - 2y^T X + y^T y\right) \\ \nabla^2 L(w) &= \frac{1}{n} X^T X \succeq 0 \end{aligned}$$

Let $\beta = \lambda_{\max}(\frac{1}{n} X^T X)$ and $\alpha = \lambda_{\min}(\frac{1}{n} X^T X)$

Then L is $\beta-$smooth and $\alpha-$SC (if $\alpha > 0$).

Example (Logistic Regression). $y_i \in \{\pm 1\}$

$$L(w) = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-y_i w^T x_i}\right)$$

We calculate the gradient and Hessian as above:

$$\nabla L(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-y_i w^T x_i}(-y_i x_i)}{1 + e^{-y_i w^T x_i}} \in \mathbb{R}^d$$

Note that $(\frac{e^z}{1-e^z})' = \frac{e^z}{(1+e^z)^2}$

$$\begin{aligned}\nabla^2 L(w) &= \frac{1}{n}\sum_{i=1}^{n} \frac{e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2}(-y_i x_i)(-y_i x_i)^T \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} x_i x_i^T\end{aligned}$$

It turns out that $\frac{e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} \leq \frac{1}{4}$. Therefore,

$$\nabla^2 L(w) \preceq \frac{1}{n}\sum_{i=1}^{n} \frac{1}{4} x_i x_i^T = \frac{1}{4n} X^T X$$

and we conclude that L is β-smooth with $\beta = \lambda_{max}(\frac{1}{4n} X^T X)$.

What about strong convexity? As $\|w\|_2 \to \infty$ and $w^T x_i \neq 0$, then $\frac{e^{-y_i w^T x_i}}{(1+e^{-y_i w^T x_i})^2} \to 0$. Therefore, $\exists\ w$ s.t. $\nabla^2 L(w) \prec \alpha I, \forall \alpha > 0$

8.2 Linear Convergence under PL condition

Definition 8.2.1 (Poly-Lojasiewicz(PL) condition)**.** A differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ satisfies $\alpha - PL$ condition $(\alpha > 0)$ if

$$\|\nabla f(x)\|_2^2 \geq 2\alpha\left(f(x) - f(x^*)\right),\ \forall\ x \in \mathbb{R}^d$$

where $x^* = \text{argmin}_{x\in\mathbb{R}^d} f(x)$.

Lemma 8.2.1 ($\alpha - SC \Rightarrow \alpha - PL$)**.** If f is differentiable and $\alpha - SC$, then f is $\alpha - PL$.

Proof. By the 1st-order characterization of $\alpha - SC$:

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y-x\|_2^2, \forall\ x, y \\ \min_y\{f(y)\} &\geq \min_y\{f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y-x\|_2^2\} \\ f(x^*) &\geq f(x) + \langle \nabla f(x), -\frac{1}{\alpha}\nabla f(x)\rangle + \frac{\alpha}{2}\cdot\frac{1}{\alpha^2}\|\nabla f(x)\|_2^2 \\ &= f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \end{aligned}$$

where the second last line comes from the fact that optimal $y = x - \frac{1}{\alpha}\nabla f(x)$.

why?

■

Remark. $\alpha - PL$ is a weaker condition than $\alpha - SC$, where the latter is a necessary condition for the former.

Theorem 8.2.1 (Convergence of GD for smooth and PL functions)**.** If f is $\beta-$smooth and $\alpha-$PL, and

$\eta \leq \frac{1}{\beta}$, then:

$$f(x_t) - f(x^*) \leq (1 - \alpha\eta)^t \left(f(x_0) - f(x^n)\right)$$

Proof.

$$\begin{aligned} f(x_{t+1}) - f(x^*) &\leq f(x_t) - f(x^*) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 && \text{Descent Lemma} \\ &\leq f(x_t) - f(x^*) - \frac{\eta}{2} \cdot 2\alpha(f(x_t) - f(x^*)) && \text{alpha-PL} \\ &= (1 - \eta\alpha)(f(x_t) - f(x^*)) \end{aligned}$$

■

Properties of PL functions

1. Not necessarily convex, *e.g.*, $f(x) = x^2 + 3\sin^2 x$
2. Can have multiple global minima (This is on the contrary to $\alpha - SC$ functions, where its global minima is unique)
3. All stationary points are global minima

$$\begin{aligned} \nabla f(x) = 0 \Rightarrow 0 = \|\nabla f(x)\|_2^2 &\geq 2\alpha \left(f(x) - f(x^*)\right) \\ \Rightarrow f(x) &\leq f(x^*) \\ \Rightarrow f(x) &= f(x^*) \end{aligned}$$

Example (Overparametrized Linear Regression). We start off in same setting as the previous linear regression example (that is, the same symbols, gradient, and Hessian), etc. However, here $n << d$ – there are many more parameters than observations.

In order for L to be SC, we need $\lambda_{\min}(X^T X) > 0$, where $X^T X$ is a $d \times d$ matrix.

But if $n < d$, $X^T X$ is not full-rank – we have $\text{rank}(XX^T) \leq n < d$. As non-full-rank matrices have at least one zero singular value, this implies that $\lambda_{\min} = 0$ and thus L is not SC.

Fortunately, we can still show the PL condition in this case.

Assume $\text{rank}(X) = n$, ($x_1, \ldots, x_n$ are linearly independent), then

$$\exists\ w^* \in \mathbb{R}^d \text{ s.t. } (w^*)^T x_i = y_i, \forall i$$

In fact, infinite w^* of this form exist and lie in the same subspace of $\mathbb{R}^d$. We will show, where $XX^T \in \mathbb{R}^{n \times n}$, that:

$$\|\nabla L(W)\|_2^2 \geq 2\lambda_{\min}(\frac{1}{n} XX^T) \cdot (L(w) - L(w^*))$$

To prove this,

$$\begin{aligned} \|\nabla L(W)\|_2^2 &= \frac{1}{n}\left\|X^T(Xw - y)\right\|_2^2 \\ &= \frac{1}{n^2}(Xw - y)^T XX^T (Xw - y) \\ &\geq \frac{1}{n^2}\lambda_{\min}(XX^T) \cdot \|Xw - y\|_2^2 \\ &= \frac{2}{n}\lambda_{\min}(XX^T) \cdot (L(w) - L(w^*)) \end{aligned}$$

and so L satisfies the PL condition.

> the second last line why?

Chapter 9

Non-Convex Optimization

9.1 Basics

We focus on unconstrained optimization problems

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

Assume: f is β-smooth and twice continuously differentiable.

Definition 9.1.1 (stationary point). x is a stationary point of f if $\nabla f(x) = 0$. x is an ϵ-stationary point if $\|\nabla f(x)\|_2 \leq \epsilon$.

Theorem 9.1.1 (Convergence of GD to an ϵ-stationary point). If f is β-smooth, then for any $\epsilon > 0$, GD with $\eta \leq \frac{1}{\beta}$ finds an $\epsilon-$stationary point within $T = \frac{2(f(x_0)-f(x^*))}{\eta\epsilon^2}$, *i.e.*:

$$\min_{0 \leq t \leq T} \|\nabla f(x_t)\|_2 \leq \epsilon$$

Proof. By descent lemma:

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

Sum over $t = 0, 1, \ldots, T-1$:

$$f(x_0) - f(x^*) \geq f(x_0) - f(x_T) \geq \frac{\eta}{2}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2$$

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|_2^2 \leq \frac{2}{\eta T}(f(x_0) - f(x^*)) = \epsilon^2$$

$$\min_{0 \leq t \leq T-1} \|\nabla f(x_t)\|_2^2 \leq \epsilon$$

and this gives us our result. Note when $\eta = \frac{1}{\beta}$, we have that $T = \frac{2\beta(f(x_0)-f(x^*))}{\epsilon^2}$. ■

Remark. What does this mean for non-convex functions?

- Gradient descent is unlikely to go to a local maximum.
- Local minima are likely the best we can hope for in general.

Definition 9.1.2 (saddle point). A saddle point is a stationary point but not a local minimum or local maximum. Equivalently, x is a saddle point of f if $\nabla f(x) = 0$, and $\forall\ \epsilon > 0$, $\exists y, z$ s.t. $\|y - x\|_2 \leq \epsilon$, $\|z - x\|_2 \leq \epsilon$ and $f(y) < f(x) < f(z)$.

Definition 9.1.3 (strict saddle point). A saddle point x of f is strict if $\lambda_{\min}(\nabla^2 f(x)) < 0$

Example. We have the following two functions:

$$f_1(x_1, x_2) = x_1^2 - x_2^2$$
$$f_2(x_1, x_2) = x_1^2 - x_2^3$$

With some calculations, we get that

$$\nabla f_1(x_1, x_2) = (2x_1 \quad -2x_2)^T$$
$$\nabla^2 f_1(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

At $(0, 0)$, the gradient is 0, but $\lambda_{min}(\nabla^2 f_1(0, 0)) < 0$. So $(0, 0)$ is a strict saddle point of f_1.

But with f_2,

$$\nabla f_2(x_1, x_2) = (2x_1 \quad -3x_2^2)^T$$
$$\nabla^2 f_2(x_1, x_2) = \begin{bmatrix} 2 & 0 \\ 0 & -6x_2 \end{bmatrix}$$

At $(0, 0)$, the gradient is 0, but $\lambda_{min}(\nabla^2 f_2(0, 0)) = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \succeq 0$. So $(0, 0)$ is a non-strict saddle point of f_2.

9.2 Second-order stationary point

Definition 9.2.1 (second-order stationary point). x is a second-order stationary point (SOSP) of f if $\nabla f(x) = 0$ and $\nabla^2 f(x) \succeq 0$. x is an (ϵ_1, ϵ_2)-SOSP if $\|\nabla f(x)\|_2 \leq \epsilon_1$, $\nabla^2 f(x) \succeq -\epsilon_2 I$

Definition 9.2.2 (Hessian-Lipschitzness). f is $\rho-$Hessian-Lipschitz if

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_2 \leq \rho \|x - y\|_2,\ \forall\ x, y \in \mathbb{R}^d$$

Equivalently,

$$|f(y) - \left(\langle \nabla f(x), y - x\rangle + \frac{1}{2}(y - x)^T \nabla^2 f(x)(y - x)\right)| \leq \frac{\rho}{6}\|x - y\|_2^3,\ \forall\ x, y \in \mathbb{R}^d$$

Finding $\epsilon-$SOSP using Hessian information

Suppose $\|\nabla f(x_t)\|_2 < \epsilon$ and $\lambda_{\min}(\nabla^2 f(x_t)) < -\epsilon_2$

$$f(x_t + v) = f(x_t) + \langle \nabla f(x_t), v \rangle + \frac{1}{2} v^T \nabla^2 f(x_t) v \pm \frac{\rho}{6} \|v\|_2^3$$

We want to find v such that $v^T \nabla^2 f(x_t) v$ is minimized. An intuitive solution is to use the bottom eigenvector.

$$\text{Recap: } \sup_{\|v\|_2=1} v^T A v = \lambda_{\max}(A), \text{ achieved by top eigenvector}$$
$$\inf_{\|v\|_2=1} v^T A v = \lambda_{\min}(A), \text{ achieved by bottom eigenvector}$$

Theorem 9.2.1 (GD + eigenvector finds SOSP). Assume that f is β-smooth and $\rho-$Hessian Lipschitz. Fix $\epsilon > 0$. Let $\eta = \frac{1}{\beta}$ and $\gamma = \sqrt{\frac{\epsilon}{\rho}}$.

Then within $T = \frac{3\beta(f(x_0) - f(x^*))}{\epsilon^2}$ iterations, at least one of the iterates x_t $0 \le t \le T$ is an $(\epsilon, \sqrt{\rho\epsilon})$-SOSP.

Proof. Proof by contradiction. Assume that all iterates $x_0, \dots, x_T$ are not $(\epsilon, \sqrt{\rho\epsilon})$-SOSP. Then we want to show we will have a sufficient decrease of $f(x_t)$ in every iteration, and reach a contradiction.

Case 1: $\|\nabla f(x_t)\|_2 > \epsilon$.

By descent lemma: $f(x_{t+1}) - f(x_t) \le -\frac{\eta}{2}\|\nabla f(x_t)\|_2^2 < -\frac{1}{2}\epsilon^2 = \frac{1}{2\beta}\epsilon^2$

Case 2: $\|\nabla f(x_t)\|_2 \le \epsilon$.

By our assumption that x_t is not an $(\epsilon, \sqrt{\rho\epsilon})$-SOSP, we know $\lambda_{\min}(\nabla^2 f(x_t)) < -\sqrt{\rho\epsilon}$

By choice of v_t: $v_t^T \nabla^2 f(x_t) v_t = \lambda_{\min}(\nabla^2 f(x_t)) < -\sqrt{\rho\epsilon}$.

$$\begin{aligned}
f(x_{t+1}) &= \min\{f(x_t + \gamma v_t), f(x_t - \gamma v_t)\} \\
&\le f(x_t) + \min\{\langle \nabla f(x_t), \gamma v_t \rangle, \langle \nabla f(x_t), -\gamma v_t \rangle\} + \frac{1}{2}(\gamma v_t)^T \nabla^2 f(x_t)(\gamma v_t) + \frac{\rho}{6}\|\gamma v_t\|_2^3 \\
&\le f(x_t) + 0 + \frac{\gamma^2}{2}(-\sqrt{\rho\epsilon}) + \frac{\rho}{6}\gamma^3 \\
&= f(x_0) - \frac{1}{3}\sqrt{\frac{\epsilon^3}{\rho}} && \text{let gamma = sqrt(ep/rho)} \\
&\le f(x_t) - \frac{\epsilon^2}{3\beta} && \text{beta >= sqrt(rho ep)}
\end{aligned}$$

In either case, function value decreases $\ge \frac{\epsilon^2}{3\beta}$. Then

$$T \le \frac{f(x_0) - f(x^*)}{\frac{\epsilon^2}{3\beta}} = \frac{3\beta(f(x_0) - f(x^*))}{\epsilon^2}$$

■

Finding $\epsilon-$SOSP without using Hessian information

Perturbed GD: $x_{t+1} = x_t - \eta(\nabla f(x_t) + \xi_t)$ where $\xi_t \sim \mathcal{N}(0, \frac{r^2}{d} I)$.

Theorem 9.2.2 (Perturbed GD finds SOSP). Assume f is β-smooth and ρ-Hessian-Lipschitz.

Fix $\epsilon > 0, \delta \in (0, 1)$.

If we choose $\eta = \frac{1}{\beta}$, $r = \tilde{O}(\epsilon)$ and run perturbed GD for $T = O\frac{\beta(f(x_0)-f(x^*))}{\epsilon^2}$ iterations, then w.p. $\geq 1-\delta$, at least one of the iterates is an $(\epsilon, \sqrt{\rho\epsilon})$ SOSP.

Remark. Intuition: a random perturbation will have a component that aligns with the bottom eigenvector of $\nabla^2 f(x_t)$.

What if we don't want to modify GD at all?

9.3 Finding SOSP with vanilla GD

Theorem 9.3.1. Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function and $\tilde{x}$ be a stationary point of f but not a SOSP ($\nabla f(\tilde{x}) = 0, \lambda_{\min}\left(\nabla^2 f(\tilde{x})\right) < 0$). Consider GD: $x_{t+1} = x_t - \eta \nabla f(x_t)$ with $\eta < \frac{1}{\beta}$ where x_0 is initialized **randomly** whose probability distribution γ's support has non-zero Lebesgue measure (*e.g.*, $\mathcal{N}(0, \sigma^2)$), then

$$\Pr\left[\lim_{t\to 0} x_t = \tilde{x}\right] = 0$$

The above theorem is proved using Stable Manifold Theorem.

Example (Non-Convex quadratic function). $f(x) = \frac{1}{2}x^T A x, A = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1, \ldots, \lambda_k > 0 > \lambda_{k+1}, \ldots, \lambda_d$.

We have $\nabla f(x) = Ax$ and $\nabla^2 f(x) = A$. Then $\nabla f(x) = 0 \Rightarrow x = 0$.

It's not a SOSP because $\lambda_{\min}(A) < 0$.

*why is it impossible to converge to $\tilde{x} = 0$?

A: GD: $x_0 \in \mathbb{R}^d$, $x_{t+1} = x_t - \eta A x_t = (I - \eta A)x_t$, then $x_t = (I - \eta A)^T x_0$. We have

$$\begin{bmatrix} x_{t,1} \\ \vdots \\ x_{t,d} \end{bmatrix} = \mathrm{diag}(1 - \eta\lambda_i)^T \begin{bmatrix} x_{0,1} \\ \vdots \\ x_{0,d} \end{bmatrix}$$

To generalize, $x_{t,i} = (1 - \eta\lambda_i)^T x_{0,i}, i = 1, \ldots, d$.

We have $\eta < \frac{1}{\beta} = \frac{1}{\max_{i\in[d]}|\lambda_i|}$.

If $\lambda_i > 0$, then $0 < 1 - \eta\lambda_i < 1$, implying that $x_{t,i} \to 0$ as $t \to \infty$.

If $\lambda_i < 0$, then $1 - \eta\lambda_i > 1$, implying that $x_{t,i} \to 0$ only if $x_{0,i} = 0$.

This means that the only points that can converge to $\tilde{x} = 0$ is $[*, \ldots, *, 0, \ldots, 0]^T$ for k *'s and d-k 0's, which has measure 0.

9.4 Landscape Analysis

Hopefully, in a concrete problem all SOSPs are global minima.

- all local minima are global
- all saddle points are strict

Example (Top Eigenvector Problem). (1-PCA) Setting: $M \in \mathbb{R}^{d\times d}$, $M \succ 0$. Eigenvalues: $\lambda_1 > \ldots > \lambda_d > 0$ with corresponding Eigenvectors $v_1, \ldots, v_d$ with $\|v_i\|_2 = 1$.

Consider the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{2}\left\|xx^T - M\right\|_F^2$$

Global min: $x \pm \sqrt{v_1}v_1$.

In HW2, we will show all SPs are $\pm\sqrt{\lambda_i v_i}$, and only SOSPs are $\pm\sqrt{\lambda_i v_i}$.

Remark. We still need to check smoothness, but f is not globally smooth over $\mathbb{R}^d$.

If $f(x)$ is bounded, then x is bounded, then $\left\|\nabla^2 f(x)\right\|_2$ is bounded.

9.5 Trajectory Analysis

Even if the Landscape is bad, it might not be the end of the world!

Trajectory analysis is about understanding what happens along the Trajectory.

We will study one setting: deep linear networks.

$\mathcal{X} = \mathbb{R}^{d_\mathcal{X}}, \mathcal{Y} = \mathbb{R}^{d_\mathcal{Y}}$, and $\mathcal{H} = \{x \mapsto W_L W_{L-1}\cdots W_2 W_1 x \colon W_j \in \mathbb{R}^{d_j \times d_{j-1}}, j = 1,\ldots,L\}$, where $x \in \mathbb{R}^{d_0}, y \in \mathbb{R}^{d_L}$.

We have the **empirical risk**

$$\min_{W_j} \frac{1}{n}\sum_{i=1}^{n} \ell(W_L,\ldots,W_1 x_i, y_i) = f(W_1,\ldots,W_L)$$

We have the **observations**:

If $\min\{d_1,\ldots,d_{L-1}\} \geq \min\{d_0, d_L\}$, then $\mathcal{H}$ is the same as $\{x \mapsto Wx \colon W \in \mathbb{R}^{d_L \times d_0}\}$.

If $\min\{d_1,\ldots,d_{L-1}\} = k < \min\{d_0, d_L\}$, then $\mathcal{H}$ is the same as $\{x \mapsto Wx \colon W \in \mathbb{R}^{d_L \times d_0}, \text{rank}(W) \leq k\}$.

But it changes the optimization problem (convex $\to$ nonconvex).

We will consider an even simpler setting: $d_0 = d_1 = \ldots = d_L = 1$, where all matrices are 1x1 scalars, and there's just one example $(x_1 = 1, y_1 = 1)$.

$$f(w)_{w\in\mathbb{R}^L} = f(w_1,\ldots,w_L) = \frac{1}{2}\left(w_L\cdots w_1 - 1\right)$$

We will consider GD: $w(\circ) \in \mathbb{R}^L$, $w(t+1) = w(t) - \eta\nabla f(w(t))$.

*Attempt at Landscape analysis:

$$\frac{\partial f(w)}{\partial w_i} = (w_L,\ldots,w_1 - 1)\cdot\prod_{k\neq i} w_k$$

$$\nabla f(x) = 0 \Leftrightarrow \begin{cases} w_L\cdots w_1 \colon \text{global min, good} \\ \prod_{k\neq i} w_k = 0 (\forall i) \colon \text{at least } 2 w_i\text{'s are 0.} \end{cases}$$

We also can compute the hessian, $\frac{\partial^2 f(w)}{\partial w_i^2} = \left(\prod_{k\neq i} w_k\right)^2$.

$$i \neq j : \frac{\partial^2 f(w)}{\partial w_i \partial w_j} = 2\left(\prod_{k\neq i} w_k\right)\left(\prod_{k\neq j} w_k\right) - \left(\prod k \neq i, k \neq j w_k\right)$$

observation 1: If at least 3 w_i's are 0, then $\nabla^2 f(w) = 0$. (non-strict saddle points, bad!)

observation 2: f is not β-smooth for any finite β, even if $f(w)$ is bounded.

$$w = \left(\epsilon^{l-1}, \frac{1}{\epsilon}, \frac{1}{\epsilon}, \cdots, \frac{1}{\epsilon}\right), f(w) = 0, \text{global min}$$

But: if $\epsilon \to 0$, $\frac{\partial^2 f(w)}{\partial w_1^2} = \left(\frac{1}{\epsilon^{l-1}}\right)^2 \to \infty$, which gives that $\left\|\nabla^2 f(w)\right\|_2 \to \infty$, implying that it cannot satisfy the smoothness condition.

Main Result: GD starting from the set

$$\{w \in \mathbb{R}^d : w > 0, \prod_{i=1}^{L} w_i \leq 1\}$$

will linearly converge to a global minimum of f.

To prove this, we will show that along the GD trajectory, f is β-smooth and α-PL for some α, β.

Then we can apply the convergence proof for smooth & PL functions.

Definition 9.5.1. For $w \in \mathbb{R}^L$, $w > 0$, we say that w is **C-balanced** ($c \geq 1$) if $x_i \leq c x_j$ for all $i, j \in [L]$.

The initial point $w(0)$ is c-balanced for $c = \max_{i,j} \frac{w_i(0)}{w_j(0)}$.

Lemma 9.5.1 (GD preserves c-balancedness)**.** Let $w > 0$ be c-balanced and $\prod_{i=1}^{L} w_i \leq 1$. Then $w' = w - \eta \nabla f(w)$ is also c-balanced and $w' \geq w$.

Proof.

$$\frac{\partial f(w)}{\partial w_i} = (w_1 \cdots w_L - 1) \prod_{k \neq i} w_k \leq 0 \Rightarrow w_i' \geq w_i$$

which proves the second point.

For the first one, let $a = -(w_1 \cdots w_L - 1)$ ■

Lemma 9.5.2 (Consequence of c-balancedness)**.** Suppose $w > 0$ is c-balanced. Then $\forall I \subseteq [L]$,

$$\prod_{i \notin I} w_i \leq C^{|I|} \left(\prod_{i=1}^{L} w_i\right)^{1-|I|/L}$$

Proof. ... ■

Lemma 9.5.3 (smoothness)**.** Let $w > 0$ be c-balanced and $\prod_{i=1}^{L} w_i \leq 1$. Then $\left\|\nabla^2 f(w)\right\|_2 \leq 3Lc^2$.

Proof. ... ■

Lemma 9.5.4 (GD with small stepsize doesn't overshoot)**.** Let $g : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Let $x \in \mathbb{R}^d$ be a point with $\nabla g(x) \neq 0$. Suppose g is β-smooth over the line segment between x and $x' = x - \eta \nabla g(x)$, for $\eta < \frac{1}{\beta}$. Then $\nabla g(x') \neq 0$.

Proof. ... ■

Lemma 9.5.5. Let $w > 0$ be c-balanced and $\prod_{i=1}^{L} w_i \leq 1$. Let $\beta = 3Lc^2$ and $\eta \leq \frac{1}{\beta}$. Let

$w' = w - \eta \nabla f(w)$. Then:

$$\prod_{i=1}^{L} w_i' \leq 1$$

Lemma 9.5.6 (α-PL). Let $w > 0$ be c-balanced and $\delta \leq \prod_{i=1}^{L} w_i \leq 1$. Then $\|\nabla f(w)\|_2^2 \geq \frac{2L\delta^{2-\frac{2}{L}}}{c^2} f(w)$

Proof. ... ■

Part III

Deep Learning

Chapter 10

Introduction to Implicit Regularization

10.1 Background

Recall that

$$L(h_\theta) = L_S(h_\theta) + (L(h_\theta) - L_S(h_\theta))$$

where $L(h_\theta)$ is referred to as **Population/test risk**, $L_S(h_\theta)$ is referred to as **Empirical risk**, and their difference $(L(h_\theta) - L_S(h_\theta))$ is referred to as **Generalization Gap**.

In the past chapters, we aim at answering the following to questions:

1. **Optimization**: How to make the empirical risk small?
   - Classical approach: Make sure $L_S(h_\theta)$ is **convex** or has **good landscape properties** (e.g. all SOSPs are global min).
2. **Generalization**: How to make the generalization gap small?
   - Classical approach: Use **uniform convergence** to bound the generalization gap based on certain **complexity measures** of the function class

$$L(\hat{h}) - L_S(\hat{h}_{\mathrm{ERM}}) \le \sup_{h \in \mathcal{H}} (L(h) - L_S(h)) \le \sqrt{\frac{\text{complexity}(\mathcal{H})}{\text{number of samples}}}$$

Challenges in DL Optimization Theory

Objective functions in DL always don't have globally nice landscape:

- non-convex, non-smooth, bad local minima
- The generic approach is insufficient
- Overparametrized network makes many generalization bounds vacuous

Key Perspective: algorithm matters

We can no longer analyze optimization and generalization separately in DL (unlike classical ML).

Optimization: Although the overall landscape is complex, trajectory of a specific algorithm (e.g. gradient descent) might only encounter a special part of the landscape where the optimization problem becomes easy.

Generalization: Specific algorithms used in practice like (stochastic) gradient descent may be implicitly biased towards special solutions that generalize well.

10.2 Motivating Examples of Implicit Regularization

Optimization algorithms could implicitly prefer certain special solutions.

We are going to first see two simple examples where implicit regularization can be characterized exactly.

(S)GD: $L(\theta) = L_S(h_\theta) = \frac{1}{n}\sum_{i=1}^n \ell\left(h_\theta(x_i), y_i\right)$.

For GD: $\theta(0)$ initialization, update $\theta(t+1) = \theta(t) - \eta\nabla L\left(\theta(t)\right), t = 0, 1, \ldots$.

For SGD: $\theta(t+1) = \theta(t) - \eta\nabla L_{B(t)}\left(\theta(t)\right), t = 0, 1, \ldots$ where $B(t) \subseteq [n], L_{B(t)}(\theta) = \frac{1}{|B|}\sum_{i\in B} \ell\left(h_\theta(x_i), y_i\right)$.

Remark: if B is sampled randomly, $\mathbb{E}_B\left[L_B(\theta)\right] = L(\theta)$. We do random shuffle in practice to approximately achieve this goal.

Setting 1: Overparametrized linear regression

We are given $\mathcal{H} = \{x \mapsto \beta^T x \colon \beta \in \mathbb{R}^d\}$, $S = \{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. We have the empirical risk defined as $L(\beta) = \frac{1}{2n}\sum_{i=1}^n \left(\beta^T x_i - y_i\right)^2 = \frac{1}{2n}\|X\beta - y\|_2^2$.

We define $X = \begin{bmatrix} \ldots, x_1^T, \ldots \\ \ldots, x_2^T, \ldots \\ \vdots \\ \ldots, x_n^T, \ldots \end{bmatrix} \in \mathbb{R}^{n\times d}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$. We assume $d > n$ and $\operatorname{rank}(X) = n$.

Fact: $\exists$ infinitely many $\beta \in \mathbb{R}^d$ s.t. $L(\beta) = 0$ or $X\beta = y$ since we have n linearly independent equations and d variables.

Theorem 10.2.1 (Implicit Regularization for linear regression)**.** Suppose we run SGD starting from zero initialization:

$$\begin{aligned} \beta(0) &= \vec{0} \\ \beta(t+1) &= \beta(t) - \eta \cdot \nabla L_{B(t)}\left(\beta(t)\right), \quad B(t) \in [n] \end{aligned}$$

Suppose it converges to zero loss: $L\left(\beta(t)\right) \to 0 \text{as} t \to \infty$.

Then:

$$\lim_{t\to\infty} \beta(t) = \operatorname{argmin}\|\beta\|_2 \text{ s.t. } L(\beta) = 0$$

Remark. We call such solutions as min-ℓ_2-norm solution.

Proof. We will prove 3 statements and combining them together yield the final result.

① Any iterate $\beta(t)$ during SGD is in $\operatorname{span}\{x_1, \ldots, x_n\} = \{\sum_{i=1}^n \lambda_i x_i \colon \lambda_1, \ldots, \lambda_n \in \mathbb{R}\} = \{X^T\lambda \colon \lambda \in \mathbb{R}^n\}$.

② The min-ℓ_2-norm solution is in $\operatorname{span}\{x_1, \ldots, x_n\}$.

③ There is only one β

First Statement on subspace

We have $L_B(\beta) = \frac{1}{2|B|}\sum_{i\in B}\left(\beta^T x_i - y_i\right)^2$ and

$\nabla L_B(\beta) = \frac{1}{|B|}\sum_{i\in B}\left(\beta^T x_i - y_i\right)x_i = \sum_{i\in B}\lambda_i x_i \in \operatorname{span}$ by taking $\lambda_i = \frac{1}{|B|}\left(\beta^T x_i - y_i\right)$.

This implies that $\beta(t) \in \operatorname{span}\{x_1, \ldots, x_n\}$.

Second Statement on min-norm

Let β be a min-ℓ_2-norm solution.

$\beta = \beta_1 + \beta_1^\perp$, where $\beta_1 \in \operatorname{span}\{x_1, \ldots, x_n\}$ and $\beta_1^\perp \in (\operatorname{span}\{x_1, \ldots, x_n\})^\perp$.

First, β_1 is still an optimal solution. We have $L(\beta_1) = 0$ and $y = X\beta = X(\beta_1 + \beta_1^\perp) = X\beta_1$.

Next, $\|\beta_1\|_2 \leq \|\beta\|_2$.

To see this, we have $\|\beta\|_2^2 = \left\|\beta_1 + \beta_1^\perp\right\|_2^2 = \|\beta_1\|_2^2 + 2\langle\beta_1, \beta_1^\perp\rangle + \left\|\beta_1^\perp\right\|_2^2 \geq \|\beta_1\|_2^2$.

This implies that $\|\beta\|_2 = \|\beta_1\|_2$ as we know $\|\beta\|_2 \leq \|\beta_1\|_2$ by definition and $\beta = \beta_1 \in$ span.

Third Statement on uniqueness

If β is an optimal solution in the span, then we have

$$\begin{cases} X\beta = y \\ \beta = X^T\lambda, \lambda \in \mathbb{R}^n \end{cases}$$

■

Comparison to explicit regularization

$$\hat{\beta}_\lambda = \text{argmin}_\beta\{L(\beta) + \lambda\|\beta\|_2^2\}$$

As $\lambda \to 0$, $\hat{\beta}_\lambda \to$ argmin$\|\beta\|_2$ s.t. $L(\beta) = 0$.

Setting 2: A two-layer model for overparametrized linear regression

We are given $\mathcal{H} = \{x \mapsto (w \circ w)^T x \colon w \in \mathbb{R}^d\}$ where $x \in \mathbb{R}^d$.

After re-parametrization, we still have a linear model $x \mapsto \beta^T x$ for $\beta = w \circ w$. We have an additional constraint $\beta \geq 0$ (entry-wise).

We assume that $\exists\beta \geq 0$ s.t. $X\beta = y$.

New loss function: $\tilde{L}(w) = \frac{1}{2n}\|X(w \circ w) - y\|_2^2$

GD: $w(t+1) = w(t) - \eta\nabla\tilde{L}(w(t))$

Definition 10.2.1 (Gradient Flow)**.** Gradient Flow is the limit of GD as $\eta \to 0$.

First, rescale time: $w(t+\eta) = w(t) - \eta\nabla\tilde{L}(w(t))$.
Second, take $\eta \to 0$: $w(t+dt) = w(t) - dt \cdot \tilde{L}(w(t))$.

$$\dot{w}(t) = \frac{dw(t)}{dt} = \frac{w(t+dt) - w(t)}{dt} = -\tilde{L}(w(t))$$

where $\dot{w}(t)$ is a differential equation defined as the gradient flow.

Theorem 10.2.2 (Implicit regularization for 2-layer model for linear regression)**.** We consider initialization

$$w(0) = \alpha\vec{1} = \begin{pmatrix} \alpha \\ \vdots \\ \alpha \end{pmatrix}$$

where $\alpha > 0$ is small and GF: $\dot{w}(t) = -\tilde{L}(w(t))$.

We define $\beta(t) = w(t) \circ w(t)$. Suppose $\tilde{L}(w(t)) \to 0$ as $t \to \infty$.

Define $\beta_\alpha = \lim_{t\to\infty}(w(t)\circ w(t))$, then

$$\lim_{\alpha\to\beta_\alpha} = \operatorname{argmin}\|\beta\|_1 \quad \text{s.t. } X\beta = y,\ \beta \geq 0$$

Proof. We do the calculations first: $\tilde{L}(w(t)) = \frac{1}{2n}\|X(w\circ w) - y\|_2^2$ and $\nabla\tilde{L}(w(t)) = \frac{2}{n}X^T(X(w\circ w) - y)\circ w$, $\dot{w}(t) = -\nabla\tilde{L}(w(t))$.

Look at the dynamics of $\beta(t) = w(t)\circ w(t)$. (next do some analysis on time differential of $\beta = w(t)\circ w(t)$)

$$\begin{aligned}\dot{\beta}(t) &= w(t)\circ\dot{w}(t) + \dot{w}(t)\circ w(t)\\ &= 2w(t)\circ\dot{w}(t)\\ &= -\frac{4}{n}w(t)X^T(X(w\circ w) - y)\circ w(t)\\ &= -\frac{4}{\beta}\beta(t)\circ X^T(X\beta(t) - y)\end{aligned}$$

Here we got of $w(t)$ and we can compare the usual linear regression setting: $L(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2$ and here $\dot{\beta}(t) = \frac{1}{n}X^T(X\beta(t) - y)$

The additional $\beta(t)$ in setting 2 will change the implicit regularization from ℓ_2 to ℓ_1.

We will do the following to complete the proof:

1. Introduce mirror descent (MD), which is a generalization of GD
2. Show a general theorem to characterize the implicit regularization of MD
3. apply the MD theorem to the update differential equation of $\dot{\beta}(t)$

■

10.2.1 Mirror Descent

$$\min_{z\in\mathbb{R}^d} f(z), f \text{ is convex}$$

Recall that GD: $z(t+1) = z(t) - \eta\nabla f(z(t))$.

This is equivalent to $z(t+1) = \operatorname{argmin}_{z\in\mathbb{R}^d}\{(f(z(t)) + \langle\nabla f(z(t)), z - z(t)\rangle) + \left(\frac{1}{2\eta}\|z - z(t)\|_2^2\right)\}$

where the first term means 1st-order Taylor expansion at $z(t)$, and the second term implies that we don't want to go too far from the point.

GD uses ℓ_2-norm to measure distance. More generally, we can use other "distances".

Definition 10.2.2 (Bregman divergence). Let ϕ be a twice-differentiable and strictly convex function. Its Bragman divergence is

$$D_\phi(x, z) = \phi(x) - \phi(z) - \langle\nabla\phi(z), x - z\rangle$$

Remark. This definition measures the difference between the real point and its Taylor expansion.

Properties:

① $D_\phi(x, z) \geq 0$ and $D_\phi(x, z) = 0$ if and only if $x = z$.

② Fix z, $D_\phi(x,z)$ is strictly convex in x.

Definition 10.2.3 (Mirror Descent Algorithm). $z(t+1) = \text{argmin}_z\{f(z(t)) + \langle \nabla f(z(t)), z - z(t)\rangle + \frac{1}{\eta} D_\phi(z, z(t))\}$

Remark. $\frac{1}{\eta} D_\phi(z, z(t))$ is strictly convex in z.

Compared with standard GD, we substitute the ℓ_2 norm with a generalized version of norm.

Examples:

$\phi(z)$	$D_\phi(x,z)$
$\frac{1}{2}\|z\|_2^2$	$\frac{1}{2}\|x-z\|_2^2$
$\sum_j z_j \log z_j - \sum_j z_j (z>0)$	$\sum_j x_j \log \frac{x_j}{z_j} - \sum_j x_j + \sum_j z_j$

The bottom left refers to the unormalized negative entropy: $H(p) = \sum p_j \log \frac{1}{p_j}$ and the bottom right refers to the unormalized KL-divergence: $D(p||q) = \sum_j p_j \log \frac{p_j}{q_j}$.

Given $L(\beta) = \frac{1}{2n}\|X\beta - y\|_2^2$

Theorem 10.2.3 (Implicit regularization of MD for linear regression). Consider MD on $L(\beta)$ with ϕ, initialized at $\beta(0)$:

$$\beta(t+1) = \text{argmin}_\beta\{f(\beta(t)) + \langle \nabla f(\beta(t)), \beta - \beta(t)\rangle + \frac{1}{\eta} D_\phi(\beta, \beta(t))\}$$

Suppose $L(\beta(t)) \to 0$ as $t \to \infty$. Then:

$$\lim_{t\to\infty} \beta(t) = \text{argmin} D_\phi(\beta, \beta(0))$$
$$\text{s.t.} X\beta = y$$

Remark. Set $\phi(z) = \frac{1}{2}\|z\|_2^2$, $D_\phi(\beta, \beta(0)) = \frac{1}{2}\|\beta - \beta(0)\|_2^2$, which recovers the first GD theorem we talked about before.

Proof Sketch Only. We will show 3 statements

① $\nabla\phi(\beta(t)) - \nabla\phi(\beta(0)) \in \text{span}\{x_1, \ldots, x_n\}$

② The optimal solution that minimizes $D_\phi(\beta, \beta(0))$ has to satisfy $\nabla\phi(\beta) - \nabla\phi(\beta(0)) \in \text{span}\{x_1, \ldots, x_n\}$

③ There is only one optimal solution that satisfies

$$\nabla\phi(\beta) - \nabla\phi(\beta(0)) \in \text{span}\{x_1, \ldots, x_n\}$$

■

Now we have the sufficient techniques to apply the latest theorem to the original setting 2: $\tilde{L}(w) = \frac{1}{2}\|X(w \circ w) - y\|_2^2$.

Recall that $\dot{\beta} = -\frac{4}{\eta}\beta(t) \circ X^T(X\beta(t) - y)$.

Derive the continuous version of MD:

$$z(t+1) = \text{argmin}_z\{f(z(t)) + \langle \nabla f(z(t)), z - z(t)\rangle + \frac{1}{\eta} D_\phi(z, z(t))\}$$

Take $\eta \to 0$:

Do a 2nd-order Taylor expansion of $D_\phi(z, z(t))$:

$$D_\phi\left(z, z(t)\right) = \frac{1}{2}\left(z - z(t)\right)^T \nabla^2\phi\left(z(t)\right)\left(z - z(t)\right) + o\left(\left\|z - z(t)\right\|_2^2\right)$$

We can treat the last term $o\left(\left\|z - z(t)\right\|_2^2\right) \to 0$ and plug the rest terms back to the MD update equation.

Solving the quadratic minimization, we get:

$$z(t+1) = z(t) - \eta\left(\nabla^2\phi\left(z(t)\right)\right)^{-1}\nabla f\left(z(t)\right)$$

Take $\eta \to 0$: $\dot{z}(t) = -\left(\nabla^2\phi\left(z(t)\right)\right)^{-1}\nabla f\left(z(t)\right)$.

This is continuous-time MD. Finally

$$\begin{aligned}\dot{\beta}(t) &= -\frac{4}{n}\mathrm{Diag}\left(\beta(t)\right)\cdot X^T\left(X\beta(t) - y\right)\\ &= -4\mathrm{Diag}\left(\beta(t)\right)\cdot\nabla L\left(\beta(t)\right)\end{aligned}$$

Appropriate $\phi(\beta) = \sum_j \beta_j \log\beta_j - \sum_j \beta_j$, and we have

$$\left[\nabla\phi(\beta)\right]_j = \log\beta_j, \quad \left[\nabla^2\phi(\beta)\right]_{ij} = \begin{cases}\frac{1}{\beta_j}, i = j\\ 0, i \neq j\end{cases}$$

According to the last Theorem in this section, MD has to converge to

$$\begin{aligned}\mathrm{argmin} D_\phi\left(\beta, \beta(0)\right) \text{ s.t. } & X\beta = y\\ & \beta \geq 0\end{aligned}$$

$$D_\phi\left(\beta, \beta(0)\right) = \sum_j \log\frac{\beta_j}{\beta(0)_j} - \sum_j \beta_j + \sum_j \beta(0)_j \qquad = \sum_j \log\beta_j + \left(\log\frac{1}{\alpha^2} - 1\right)\left\|\beta\right\|_1 + \alpha^2 d$$

As $\alpha \to 0$, the first term is a constant, the second term goes to ∞ and the last term goes to 0.

This is why we get $\mathrm{argmin}\left\|\beta\right\|_1$ and proves the previous theorem we left with.

Chapter 11

Implicit Regularization of GD

11.1 Implicit regularization in classification

Opening Remark: We will look at linear model and (2) deep homogeneous network

Example (Linear Logisitc Regression). We have the following setup:

- $S = \{(x_i, y_i)\}_{i=1}^n \sim_{\text{i.i.d}} P, x \in \mathbb{R}^d, y \in \{1, -1\}$
- $\mathcal{H} = \{x \mapsto \beta^T x \colon x \in \mathbb{R}^d\}$
- For
- loistic loss $\ell(y', y) = \log\left(1 + e^{-yy'}\right)$ is convex and smooth
- Empirical risk $L_S(\beta) = \frac{1}{n}\sum_{i=1}^n \ell\left(\beta^T x_i, y_i\right)$
- 0-1 Classification error: $\text{Err}_S(\beta) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}[y_i \beta^T x_i] \leq 0$
- Population logistic risk and 0-1 error: $L(\beta), \text{Err}(\beta)$.

Assumption: The training data are linearly separable: $\exists \beta \in \mathbb{R}^d$ s.t. $y_i \beta^T x_i > 0, \forall i \in [n]$.

Observations:

- Only the direction of β matters for classification error:

$$\text{sign}\left(c\beta^T x\right) = \text{sign}\left(\beta^T x\right), \forall c > 0 \Rightarrow \text{Err}_S(\beta) = \text{Err}_S(c\beta), \forall c > 0$$

- Scaling up a separator β will drive empirical logistic loss to 0.

  If $y_i \beta^T x_i > 0, \forall i$, then

$$L_S(c\beta) \to 0 \text{ as } c \to +\infty$$

  (for any separator β, $\infty \cdot \beta$ is a "global min" of L_S).

- $\exists$ an infinite number of directions that are separators.

Implicit Regularization Question: which direction is found by GD?

A: max-margin/SVM solution

Definition 11.1.1. (unormalized) margin: $\gamma(\beta) \coloneqq \min_{i \in [n]} y_i \beta^T x_i$
normalized margin: $\overline{\gamma}(\beta) = \frac{\gamma(\beta)}{\|\beta\|_2}$

Remark. $\overline{\gamma}$ is a scale-invariant

Definition 11.1.2 (max-margin/SVM solution). ...

Remark. Why is max margin good?

Recall margin-based generalization bound:

$$\text{Err}(\beta) \leq \frac{4}{\gamma(\beta)} \cdot \frac{\|\beta\|_2 R}{\sqrt{n}} + (\text{small terms}) \qquad = \frac{4}{\overline{\gamma}(\beta)} \cdot \frac{R}{\sqrt{n}} + (\text{small terms})$$

Theorem 11.1.1 (Implicit Regularization for linear logistic regression). GF on $L_S(\beta)$ $\left(\dot{\beta}(t) = -\nabla L_S\left(\beta(t)\right)\right)$ starting from an arbitrary initialization $\beta(0) \in \mathbb{R}^d$ converges to the max-margin solution:

$$\overline{\gamma}\left(\beta(t)\right) \to \max_{\beta \in \mathbb{R}^d} \overline{\gamma}(\beta) \text{ as } t \to \infty$$

proof intuition (not real proof). First, $L_S\left(\beta(t)\right) \to 0$ as $t \to \infty$ by a standard smooth convex optimization argument.
Think of $_S\left(\beta t\right) \approx 0$ for large t

Second, $\gamma\left(\beta(t)\right) \to \infty$. ■

Part IV

Appendix

Chapter 12

Calculus

12.1 Taylor Expansion

In 1-dimension, for a function $f\colon \mathbb{R} \to \mathbb{R}$, 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 $f\colon \mathbb{R}^d \to \mathbb{R}$.

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:

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

12.2 Linear Algebra

Definition 12.2.1. 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 $A - B \succeq 0$, $\lambda_{\min}(A) \cdot I \preceq A\lambda_{\max}(A) \cdot I$