Contents

Color code: highlight, key concepts and definitions, to be added.

Chapter 1

On Deep Leaning

1.1 Chapter 4 Numerical Optimization

Problems of under- and over-flow.

Find x that minimizes the criterion function $f(x)$.

- Gradient descent method as an example of first-order optimization.
- Using Hessian matrix, we have second order optimization, that's related to the eigenvalue-eigenvector of the Hessian matrix.
- KKT condition.

1.2 Chapter 5 Basics of Machine Learning

5.7, 5.8 on supervised and unsupervised learning.

1.3 Chapter 6 DNN

Deep feedforward networks has no feedback. We use the composition of a network of functions $f(x)$ to approxiamte $f^*(x)$, our training set is *training points* and *label* $y \sim f^*(x)$. The final layer is *output layer* other layers are *hidden layers*.

We can generalize the linear models by applying liear model to transformed input $\phi(x)$, where ϕ is a set of features describing x or a new representation of x.

Deep learning tries to learn ϕ: $y = \phi(x;\theta)'w$.

Example 1.3.1. Learn the XOR function.

1.3.1 6.2 How to choose cost function and output layer

Given training set (x, y) where x is a vector, loss function from likelihood analysis is

$$J(\theta) = -E_{\text{sample}} \log P_{\text{model}}(y \mid x)$$

What we learn from the training set is not $\hat{y}$ that approximates y, but the conditional distribution. Regression is about the mean of the conditional distribution. If we use L^2 loss, then we can learn the mean, if we use L^1 loss we can learn the median.

The *output layer design* usually uses the following functions:

1. suppose y is normal, then a *linear* output layer is used, commonly for regression problem.
2. For Bernoulli distribution, we can use sigmoid:

$$\frac{1}{1+e^{-x}}$$

   when we use log likelihood, we can avoid the problem of having a derivative that's too close to 0 and achieve better numerical performance.. The overparametrization problem won't matter in practice.

3. For classification, most commonly used output layer is *softmax*. The *overparametrization* problem won't matter in practice.

Sometimes we consider Gaussian Mixture model, but then we need to estimate the covariance matrices and the results aree usually unstable because we need matrix division?

1.3.2 6.3 How to Choose Hidden Layers

So far there isn't a theory guidance on that. Commonly used hidden layer include the *ReLU*. Other more complicated choices don't seem to perform better in empirical studies.

Chapter 2

From Reading Group on Durrett

Reference: PTE, Folland and Tao.

2.1 Measure and Integration

2.1.1 Measure and Probability Spaces

First of all, why should we discuss measure theory. It's for a better integration theory than the usual Riemann integral. Because we need strong conditions for arguments like:

$$\lim_n \int f_n = \int \lim_n f_n$$

to hold for Riemann integral. TBA: Relationship with Riemann Integral We'll talk about

1. Measure space, construction of measure, Lebesgue measures;
2. Integration theory
3. Convergence concepts
4. Differentiation

Let's start with an abstract measure space. Let Ω be a space. A σ-algebra $\mathcal{F}$ is a collection of subsets called *measurable sets* of Ω such that it's closed under complements and countable unions. Also the $\varnothing \in \mathcal{F}$. We can assign *measure* μ to members in $\mathcal{F}$. The

reason we need to restrict attention from the power set to $\mathcal{F}$ is that there are examples that violate some properties we want a measure to have. μ satisfies:

$$\mu(\varnothing) = 0 \quad \text{and} \quad \mu\left(\bigsqcup_i^{\infty} A_i\right) = \sum_i \mu(A_i) \quad \text{for} \quad (A_i) \subset \mathcal{F}$$

The triplet $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*.

A *measurable function* is $X : (\Omega, \mathcal{F}, \mu) \to (S, \mathcal{H})$ such that for all $H \in \mathcal{H}$, $X^{-1}(H) \in \mathcal{F}$. Meaning that we can assign a measure to the subsets of $\mathcal{H}$ based on the measure μ and X. Given two measurable spaces, we can check whether a function X is measurable.

2.1.1 Lemma. *Suppose $\mathcal{H} = \sigma(\mathcal{G})$, if for all $B \in \mathcal{G}$, $X^{-1}(B) \in \mathcal{F}$, then X is measurable.*

Proof. Let $\mathcal{H}' = \left\{B \subset S \mid X^{-1}(B) \in \mathcal{F}\right\}$, then we can show $\mathcal{H}'$ is a σ-algebra because X^{-1} preserves complements and union. Then we have $\mathcal{G} \subset \sigma(\mathcal{G}) = \mathcal{H} \subset \mathcal{H}'$, meaning that all $H \in \mathcal{H}$ satisfies the condition. □

A measure space is a *probability space* if $\mu = 1$, we write it as $(\Omega, \mathcal{F}, P)$. A random variable/vector is a measurable function on probability space $X : \Omega \to \mathbb{R}/\mathbb{R}^d$ with the Borel σ-algebra.

Distribution and Density

Distribution is the probability induced by a random variable X, such that $\mu(A) = P(X \in A)$. *Density* is defined via some methods(To be added). Its use is:

$$\int g(X)\, \mathrm{d}P = \int g(x) f(x)\, \mathrm{d}x$$

Check for Measurability of Random Variables

For random variables X_n, Y_n, $X_n + Y_n$ and $\sup X_n$ are measurable.

2.1.2 Construction of Measures

Motivation

Now we want to show that we can actually find measure on $\mathbb{R}$ and $\mathbb{R}^d$ that are the most useful spaces. We start from abstract construction method inspired by the following observation of measure on $\mathbb{R}$.

In order to have a natural measure λ on $\mathbb{R}$, we want the following properties:

1. $\lambda_0(a, b] = b - a$;
2. Measure of union of intervals should be the sum;(we can't have uncountable sum, hence restricted to countable union)
3. Apply to all intervals(we can't apply to the power set, hence restrict to the measurable sets).

Now consider the set $\mathcal{S}_{\mathbb{R}} = \{(a, b] : a \leq b \in \mathbb{R}\}$, it's too small, we want to extend the natural measure funtion λ_0 to a larger set(we can extend to $\mathcal{B}_{\mathbb{R}}$, even to $\mathcal{L}_{\mathbb{R}}$).

Semi-algebra $\mathcal{S}$ is collection of subsets of a space Ω that satisfies the following conditions: closed under intersection and the complements are union of finite disjoint sets in $\mathcal{S}$. Let μ_0 be a what I call *semi-pre-measure* to (*algebra*, *premeasure*) to (outer-measurable sets, outer measure) to (sigma-algebra, measure) where the *semi-pre-measure* μ_0 satisfies if both $S_N = \bigsqcup^N S_j, S_\infty = \bigsqcup^\infty S_j \in \mathcal{S}$

$$\mu_0(\varnothing) = 0; \quad \mu_0(S_N) = \sum^N \mu_0 S_j; \quad \mu_0(S_\infty) \leq \sum^\infty \mu_0 S_j$$

We follow the steps of extension

$$(\mathcal{S}, \mu_0) \to (\mathcal{A}, \mu_1) \to (\mathcal{M}, \mu^*) \to (\sigma(\mathcal{S}), \mu)$$

that is semi-algebra and *pre-measure* μ_1 satisfies $\mu_1 \bigsqcup_i^\infty A_i = \sum_i^\infty \mu_1 A_i$ as long as the union is also in $\mathcal{A}$. And we will show that $\mathcal{B}_{\mathbb{R}} \subset \sigma(\mathcal{S})$.

Construction of Abstract Measure

2.1.2 $(\mathcal{S}, \mu_0) \to (\mathcal{A}, \mu_1)$. *We have that* $\mathcal{A} = \left\{\bigsqcup_j^N S_j : S_j \in \mathcal{S}\right\}$ *is an algebra. And* $\mu_1(A) = \sum_j^N \mu_0 S_j$ *is a premeasure.*

Proof. $\mu_1(\varnothing) = 0$. Suppose $A = \bigsqcup^{\infty} A_i \in \mathcal{A}$, then there exists $T_l \in \mathcal{S} : l = 1, \dots N_T$, such that $A = \bigsqcup_l T_l$, also for each $A_j = \bigsqcup_k^{N_j} S_{j,k}$. □

Let an *outer measure* be a function over the power set such that:

$$\mu^*(\varnothing) = 0 \quad \text{and} \quad \mu^*\left(\bigsqcup^{\infty} A_j\right) \leq \sum_j \mu^* A_j \quad \text{and} \quad \mu^*(A) \leq \mu^*(B) \quad \text{for} \quad A \subset B$$

Also let $\mathcal{M}$ be the *outer-measurable sets*, that is the collections of A such that

$$\mu^*(E) = \mu^*(E \bigcap A) + \mu^*(E \bigcap A^c) \quad \text{for all} \quad E \subset X$$

2.1.3 $(\mathcal{A}, \mu_1) \to (\mathcal{M}, \mu^*)$ **Caratheodory's Theorem applied to premeasures.** *We can find the outer measure induced by* μ_1*:*

$$\mu_1^*(B) = \inf\left\{\sum_j^{\infty} \mu_1 A_j : B \subset \bigcup_j A_j\right\}$$

we show μ_1^* *is indeed an outer measure, Caratheodory states that for outer measures,* $\mathcal{M}$ *of* μ_1^* *is a* σ*-algebra and* $\mu^*|_{\mathcal{M}}$ *is complete.*

Proof.

1. Monotone property is easy, for countable subadditivity, we use ε-room method, find a cover C_{jk} of A_j such that :
$$\mu_1^*(A_j) > \sum_k \mu_1^* C_{jk} - \varepsilon$$
2. To show that $\mathcal{M}$ is a σ-algebra we only need to show that for $E_j \in \mathcal{M}$ and any $B \subset \Omega$, we have:
$$\mu^*(B) \geq \mu^*\left(B \bigcap E\right) + \mu^*\left(B \bigcap E\right)$$
which can be seen by considering finite sum, taking out one E_j at a time and then take limit.
3. Show that the restriction is complete.

□

2.1.4 $(\mathcal{M}, \mu^*) \to (\sigma(\mathcal{S}), \mu)$. *We can restrict* μ^* *to define* $\mu = \mu^*|_{\sigma(\mathcal{S})}$*, and the resulting measure is complete.*

Proof. We show that $\sigma(\mathcal{S}) \subset \mathcal{M}$ and the restriction is complete. □

2.1.3 Application: Borel and Lebesgue-Stietjes Measure on $\mathbb{R}$

We verify that for any nondecreasing right-continuous function $F : \mathbb{R} \to \mathbb{R}$, the semi-algebra $\mathcal{S}_{\mathbb{R}}$ together with the semi-pre-measure defined by

$$\lambda_0(a, b] = F(b) - F(a)$$

satisfies the conditions set in the previous section. $\mathcal{A}_{\mathbb{R}}$ be the algebra generated by $\mathcal{S}_{\mathbb{R}}$ in turn will generate $\mathcal{B}_{\mathbb{R}}$. We can extend it to a something larger, a complete measure whose domain contains $\mathcal{B}_{\mathbb{R}}$. It will be called the *Lebesgue-Stietjes measure* and *Lebesgue measurable sets* $\mathcal{M}_\lambda$. The restriction on $\mathcal{B}$ is called *Borel measure*. It's just the completion.

2.1.5 Regularity of $\mathcal{M}_\lambda$. *Lebesgue measurable sets are of simple form if you allow for a small error.*

1. $\lambda(A) = \inf\{\lambda U : A \subset U, U \text{ is open}\} = \sup\{\lambda K : K \subset A, K \text{ is compact}\}$
2. $A = G_\delta \setminus N_1 = F_\sigma \bigcup N_2$ *where* N *are null sets.*
3. *Littlewood's First Principle: Suppose* $\lambda(A) < \infty$ *then for all* $\varepsilon > 0$*, there exists* B *that is a finite union of open intervals such that* $A \triangle B < \varepsilon$.

Proof. □

2.1.4 Integration

Integration and measure are the two sides of same coin: integration is the generalization of measure from the space of *indicator functions* $\mathbf{1}_A$ of measurable sets to a larger set of functions.

For *simple functions* $f_s = \sum_j^N \mathbf{1}_{A_j}$, we can define $\mu f_s = \sum_j^N \mu(A_j)$

Then for *nonnegative measurable* functions f, we can define

$$\mu f = \sup\{\mu f_s : f_s \leq f \quad \text{pointwise}\}$$

If $f = f^+ - f^-$ and the integrals of two nonnegative parts are not ∞, then we say f is *integrable* or in L^1 and $\mu f = \mu f^+ - \mu f^-$.

2.1.6 Littlewood's Second Principle. *Measurable and integrable functions can be approximated by simple well-behaved functions.*

2.1.7 Properties of Integral.

1. *Monotone $\mu f \le \mu g$ if $f \le g$ pointwise.*
2. *Linearity for $a > 0$, $\mu(af) = a\mu f$, $\mu(f+g) = \mu f + \mu g$.*
3. *Monoton Convergence: if $0 \le f_n \uparrow f$ a.e., then $\mu f_n \uparrow \mu f$.*

Proof. Folland proves by choosing simple function and a scaling factor α. Tao's proof is essentially the same. PTE proves via Fatou's lemma.

Suppose $f_n \uparrow f$ a.e., if μf exists, we know μf_n must converge because it's a nondecreasing bounded sequence of numbers. Also $\mu f_n \le \mu f$ by monotonicity.

To show that $\mu f \le \lim_n \mu f_n$, where $\mu f = \sup\{\mu g : g \text{ is simple}, g \le f\}$, we only need to show that for all simple $g \le f$, we have $\mu g \le \mu f_n$. We can use ε-room method, let $E_n = \{\omega : f_n \ge (1-\varepsilon)g\}$, $E_n \uparrow \Omega$, and we have $\int f_n \ge \int_{E_n} (1-\varepsilon)g$, taking lim, we have $\lim \int f_n \ge (1-\varepsilon)\int g$ for all $\varepsilon > 0$. □

Remark. These three properties fully characterise the integral. That is given a functional that satisfies the three properties we can find a measure for which the function is the integral.

2.1.8 Inequalities and Controls. *Useful inequalities and controls*

1. *Jensen's inequality: suppose we have a convex function $\Phi(x)$, then $\mu(\Phi(f)) \ge \Phi\mu f$*
2. *Holder's inequality. For $p, q : \frac{1}{p} + \frac{1}{q} = 1$, we have $\|fg\|_1 \le \|f\|_p \|g\|_q$*
3. *Minkowski's inequality: $\|f+g\|_p \le \|f\|_p + \|g\|_p$.*
4. *Chebyshev's Inequality:*
5. *For nonnegative measurable functions, $\mu f < \infty \implies f < \infty$ a.e., $\mu f = 0$ implies $f = 0$ a.e.*
6. *Moment and tail behaviour, if μ is finite,*

$$\mu f^r < \infty \implies \sum x^{r-1} \mu(f > x) < \infty$$

2.1.5 Countable Product Space and Lebesgue Measure on $\mathbb{R}^d$

Product measurable and *product measure*.

2.1.9 Fubini-Tonelli.

2.1.6 Convergence Concepts

2.1.10 Monotone, Fatou and Dominated Convergence.

2.1.11 Uniform Integrability.

2.1.7 Differentiation