MATHEMATICAL FOUNDATIONS OF MACHINE LEARNING

20ANMAG469P1, FALL TERM 2020-2021)

HÔNG VÂN LÊ *

Contents

1. Basic mathematical problems in machine learning 4
1.1. Learning, inductive learning and machine learning 4
1.2. A brief history of machine learning 6
1.3. Main tasks of current machine learning 7
1.4. Main types of machine learning 10
1.5. Basic mathematical problems in machine learning 11
1.6. Conclusion 12
2. Mathematical models for supervised learning 12
2.1. Discriminative model of supervised learning 13
2.2. Generative model of supervised learning 17
2.3. Empirical Risk Minimization and overfitting 21
2.4. Conclusion 23
3. Mathematical models for unsupervised learning 23
3.1. Mathematical models for density estimation 23
3.2. Mathematical models for clustering 29
3.3. Mathematical models for dimension reduction and manifold learning 31
3.4. Conclusion 32
4. Mathematical model for reinforcement learning 33
4.1. A setting of a reinforcement learning 34
4.2. Markov decision process 35
4.3. Existence and uniqueness of the optimal policy 35
4.4. Conclusion 36
5. Distances on statistical models, the Fisher metric and maximum likelihood estimator 36
5.1. The space of all probability measures and total variation norm 36
5.2. The Fisher metric on a statistical model 39
5.3. MSE and Cramér-Rao inequality 41
5.4. Efficient estimators and MLE 44
5.5. Consistency of MLE 44

Date: November 28, 2020.

* Institute of Mathematics of ASCR, Zitna 25, 11567 Praha 1, email: hvle@math.cas.cz.

5.6. Conclusion 45
6. Consistency of a learning algorithm 45
6.1. Consistent learning algorithm and its sample complexity 45
6.2. Uniformly consistent learning and the VC-dimension 49
6.3. Fundamental theorem of binary classification 51
6.4. Conclusions 53
7. Generalization ability of a learning machine 53
7.1. Covering number and sample complexity 54
7.2. Rademacher complexities and sample complexity 57
7.3. Model selection 59
7.4. Undecidability of learnability 61
7.5. Conclusion 61
8. Support vector machines 62
8.1. Linear classifier and hard SVM 62
8.2. Soft SVM 65
8.3. Sample complexities of SVM 67
8.4. Conclusion 68
9. Kernel based SVMs 69
9.1. Kernel trick 69
9.2. PSD kernels and reproducing kernel Hilbert spaces 71
9.3. Kernel based SVMs and their generalization ability 74
9.4. Conclusion 75
10. Neural networks 75
10.1. Neural networks as a model of computation 76
10.2. The expressive power of neural networks 79
10.3. Sample complexities of neural networks 81
10.4. Conclusion 82
11. Training neural networks 82
11.1. Gradient and subgradient descend 83
11.2. Stochastic gradient descend (SGD) 85
11.3. Online gradient descend and online learnability 87
11.4. Conclusion 87
12. Bayesian machine learning 88
12.1. Bayesian concept of learning 88
12.2. Applications of Bayesian machine learning 89
12.3. Consistency 90
12.4. Conclusion 90
Appendix A. The Radon-Nikodym theorem and regular conditional probability 90
A.1. Dominating measures and the Radon-Nikodym theorem 90
A.2. Conditional expectation and regular conditional measure 91
A.3. Joint distribution, regular conditional probability and Bayes' theorem 95
A.4. Disintegration and regular conditional probability 96

Appendix B. Law of large numbers and concentration-of-measure inequalities 97
B.1. Laws of large numbers 97
B.2. Markov's inequality 98
B.3. Chebyshev's inequality 98
B.4. Hoeffding's inequality 98
B.5. Bernstein's inequality 99
B.6. McDiarmid's inequality 99
Appendix C. The Kolmogorov theorem 99
Appendix D. Probabilistic morphisms and the category of statistical models 99
References 101

It is not knowledge, but the act of learning ... which grants the greatest enjoyment.
Carl Friedrich Gauss

Machine learning is an interdisciplinary field in the intersection of mathematical statistics and computer sciences. Machine learning studies statistical models and algorithms for deriving predictors, or meaningful patterns from empirical data. Machine learning techniques are applied in search engine, natural language processing, image detection, data mining [1], robotics etc. In our course we address the following questions: What is the mathematical model of learning? How to quantify the difficulty/hardness/complexity of a learning problem? How to choose a learning model and learning algorithm? How to measure success of machine learning?

The syllabus of our course:

1. Supervised learning, unsupervised learning, reinforcement learning.
2. Generalization ability of machine learning.
3. Support vector machine, Kernel machine.
4. Neural networks and deep learning.

Recommended Literature.

1. S. Shalev-Shwart, and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

2. L. Deveroye, L. Györfi and G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer 1996.

3. M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, MIT Press, 2012.

[1]The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.

1. Basic mathematical problems in machine learning

Machine learning is the foundation of countless important applications including speech recognition, image detection, self-driving car and many others in the emerging field of artificial intelligence. Machine learning techniques are developed using many mathematical theories. In my lecture course I shall explain the mathematical model of machine learning and how do we design a machine which shall learn successfully.

In my today lecture I shall discuss the following questions,

1. What are learning, inductive learning and machine learning.
2. History of machine learning and artificial intelligence.
3. Current tasks and main types of machine learning.
4. Basic mathematical problems in machine learning.

1.1. **Learning, inductive learning and machine learning.**

To start our discussion on machine learning let us begin first with the notion of learning. Every one from us knows what is learning from our experiences at the very early age.

(a) Small children learn their language is very simple and often erroneous. Gradually they speak freely with less and less mistakes. Their way of learning is *inductive learning*: from examples of words and phrases they learn the rules of combinations of these words and phrases into meaningful sentences.

(b) In school we learn mathematics, physics, biology, chemistry by following the instructions of our teachers and those in textbooks. We learn general rules and apply them to particular cases. This type of learning is *deductive learning*. Of course we also learn inductively in school by searching similar patterns in new problems and then apply appropriate methods possibly with modifications for solving these problems.

(c) Experimental physicists design experiments and observe the outcomes of experiments to validate/support or dispute/refute a statement/conjecture on the nature of the observables. In other words, experimental physicists learn about the dependence of certain features of the observables from empirical data which are outcomes of the experiments. This type of learning is *inductive learning*.

In mathematical theory of machine learning, or more general, in mathematical theory of learning we consider only *inductive learning*. Deductive learning is not very interesting; essentially it is equivalent to performing a set of computations using a finite set of rules and a knowledge database. Classical computer programs learn or gain some new information by deductive learning.

Let me suggest a definition of learning, that will be updated later to be more and more precise.

Definition 1.1. *A learning* is a process of gaining new knowledge, more precisely, new correlations of features of observable by examination of empirical data of the observable. Furthermore, a learning is successful if the correlations can be tested in examination of new data and will be more precise with the increase of data.

The above definition is an expansion of Vapnik's mathematical postulation: "Learning is a problem of function estimation on the basis of empirical data", see Definition 2.1.

Example 1.2. A classical example of learning is that of learning a physical law by curve fitting to data. In mathematical terms, a physical law is expressed by a function f, and data are the value y_i of f at observable points x_i. The goal of learning in this case is to estimate the unknown f from a set of pairs $(x_1, y_1), \cdots, (x_m, y_m)$. Usually we assume that f belongs to a finite dimensional family of functions of observable points x. For instance, f is assumed to be a polynomial of degree d over $\mathbb{R}$, i.e., we can write

$$f = f_w(x) := \sum_{j=0}^{d} w_j x^j \text{ where } w = (w_0, \cdots, w_d) \in \mathbb{R}^{d+1}.$$

In this case, to estimate $f = f_w$ is the same as to estimate the parameter w of f_w, observing the data (x_i, y_i). The most popular method of curve fitting is *the least square method* which quantifies *the error* $R(w)$ *of the estimation* of the parameter w in terms of the value

$$R(w) := \sum_{i=1}^{m} (f_w(x_i) - y_i)^2 \tag{1.1}$$

which the desired function f should minimize. If the measurement generating the data (x_i, y_i) were exact, then $f(x_i)$ would be equal to y_i and the learning problem is an interpolation problem. But in general one expects the values y_i to be affected by noise.

The least square technique, going back to Gauss and Legendre [2], which is computational efficient and relies on numerical linear algebra, solves this minimization problem.

In the case of measurement noise, which is the reality according to quantum physics, we need to use the language of probability theory to model the noise and therefore to use tools of mathematical statistics in learning theory. That is why statistical learning theory is important part of machine learning theory.

[2]The least-squares method is usually credited to Carl Friedrich Gauss (1809), but it was first published by Adrien-Marie Legendre (1805).

1.2. **A brief history of machine learning.** Machine learning was born as a domain of artificial intelligence and it was reorganized as a separated field only in the 1990s. Below I recall several important events when the concept of machine learning has been discussed by famous mathematicians and computer scientists.

• In 1948 John von Neumann suggested that machine can do any thing that peoples are able to do. [3]

• In 1950 Alan Turing asked "Can machines think?" in "Computing Machine and Intelligence" and proposed the famous Turing test. The Turing test is carried out as an imitation game. On one side of a computer screen sits a human judge, whose job is to chat to an unknown gamer on the other side. Most of those gamers will be humans; one will be a chatbot with the purpose of tricking the judge into thinking that it is the real human.

• In 1956 John McCarthy coined the term "artificial intelligence".

• In 1959, Arthur Samuel, the American pioneer in the field of computer gaming and artificial intelligence, defined machine learning as a field of study that gives computers the ability to learn without being explicitly programmed. The Samuel Checkers-playing Program appears to be the world's first self-learning program, and as such a very early demonstration of the fundamental concept of artificial intelligence (AI).

In the early days of AI, statistical and probabilistic methods were employed. Perceptrons which are simple models used in statistics were used for classification problems in machine learning. Perceptrons were later developed into more complicated neural networks. Because of many theoretical problems and because of small capacity of hardware memory and slow speed of computers, statistical methods were out of favour. By 1980, expert systems, which were based on knowledge database, and inductive logic programming had come to dominate AI. Neural networks returned back to machine learning with success in the mid-1980s with the reinvention of a new algorithm, called back-propagation, and thanks to increasing speed of computers and increasing hardware memory.

Machine learning, reorganized as a separate field, started to flourish in the 1990s. The current trend is benefited from Internet.

In the book by Russel and Norvig "Artificial Intelligence a modern Approach" (2010) AI encompass the following domains:
- natural language processing,
- knowledge representation,
- automated reasoning to use the stored information to answer questions and to draw new conclusions;

[3]According to Jaynes [Jaynes2003, p.8], who attended the talk by J. von Neumann on computers given in Princeton in 1948, J. von Neumann replied to the question from the audience "But of course, a mere machine can't really think, can it?" as follows "You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that".

- machine learning to adapt to new circumstances and to detect and extrapolate patterns,
- computer vision to perceive objects,
- robotics.

All the listed above domains of artificial intelligence except robotics are now considered domains of machine learning.

• Representation learning and feature engineering are part of machine learning replacing the field of knowledge representation.

• Pattern detection and recognition become a part of machine learning.

• Robotics becomes a combination of machine learning and mechatronics.

Why did such a move from artificial intelligence to machine learning happen?

The answer is that we are able to formalize most concepts and model problems of artificial intelligence using mathematical language and represent as well as unify them in such a way that we can apply mathematical methods to solve many problems in terms of algorithms that machine are able to perform.

As a final remark on the history of machine learning I would like to note that data science, much hyped in 2018, has the same goal as machine learning: Data science seeks actionable and consistent pattern for predictive uses. [4].

Now I shall briefly describe main tasks of current machine learning and methods to perform these tasks.

1.3. **Main tasks of current machine learning.**

• *Classification task* is a construction of a mapping from a set of items we are interested in to a smaller *countable* set of other items, which are regarded as possible "features" of the items in the domain set, using given data of the values of the desired functions at given items. For example, document classification may assign items with features, such as politics, email spam, sports, or weather, while image classification may assign items with categories such as landscape, portrait, or animal. Items in the target set are also called categories in machine learning community and we shall use this terminology from now on. The number of categories in such tasks can be unbounded as in OCR, text classification, or speech recognition.

As we have remarked in the classical example of learning (Example 1.2), usually we have ambiguous/incorrect measurement and we call uncertainty of the exactness of the measurement a noise to our measurement. If every thing would be exact, the classification task is the classical interpolation function problem in mathematics: the features are the values of an unknown function on the set of items which we need to construct knowing its values at given items using given data. (Don't give your data free on internet!)

[4]according to Dhar, V. (2013), Data science and prediction, see also `https://en.wikipedia.org/wiki/Data_science`

• *Regression task* is a construction of a real-valued function on the set of items, using given data of the values of the desired function at given items. Examples of regression tasks include learning physical law by curve fitting to data (Example 1.2) with application to predictions of stock values or variations of economic variables. In a regression task, the error of the prediction, which is also called the error of the estimation in Example 1.2, depends on the magnitude of the *distance between the true and predicted values*, in contrast with the classification problem, where there is typically no notion of closeness between various categories. As in the classification task, in regression problems we also need to take into account a noise to a measurement [5].

• *Density estimation task* is a construction (or estimation) of the distribution of observed items. In other words, given a sample space $(\mathcal{X}, \Sigma_{\mathcal{X}})$, where $\Sigma_{\mathcal{X}}$ denotes the σ-algebra of a measurable space $\mathcal{X}$, and a finite number of observed items $x_1, \cdots, x_l \in \mathcal{X}$, usually assumed to be i.i.d. for the sake of simplicity, such that x_i are subjected to unknown probability measures $\mathbf{p}_u$ on $(\mathcal{X}, \Sigma_{\mathcal{X}})$, we need to estimate $\mathbf{p}_u$. This is one of basic problems in statistics. Over one hundred year ago Karl Pearson (1980-1962), the founder of the modern statistics [6], proposed that all observations come from some probability distribution and the purpose of sciences is to estimate the parameter of these distributions, assuming that they belong to a known finite dimensional family S of probability measures on $(\mathcal{X}, \Sigma_{\mathcal{X}})$. Density estimation problem has been proposed by Ronald Fisher (1980-1962), the father of modern statistics and experiment designs [7], as a key element of his simplification of statistical theory, namely he assumed the existence of a density function $p(\xi)$ that governs the randomness of a problem of interest.

Digression. A measure ν on $(\mathcal{X}, \Sigma_{\mathcal{X}})$ is called *dominated by a measure μ on $(\mathcal{X}, \Sigma_{\mathcal{X}})$* (or *absolutely continuous with respect to μ*), if $\nu(A) = 0$ for every

[5]The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean of population). For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context: movement toward the mean of a statistical population. Galton's method of investigation is non-standard at that time: first he collected the data, then he guessed the relationship model of the events.

[6] He founded the world's first university statistics department at University College London in 1911, the Biometrical Society and Biometrika, the first journal of mathematical statistics and biometry.

[7]Fisher introduced the main models of statistical inference in the unified framework of parametric statistics. He described different problems of estimating functions from given data (the problems of discriminant analysis, regression analysis, and density estimation) as the problems of parameter estimation of specific (parametric) models and suggested the maximum likelihood method for estimating the unknown parameters in all these models.

set $A \in \Sigma_{\mathcal{X}}$ with $\mu(A) = 0$. Notation: $\nu \ll \mu$. By the Radon-Nykodym theorem, see Appendix, Subsection A.1, we can write

$$\nu = f \cdot \mu$$

where $f \in L^1(\mathcal{X}, \mu)$ is *the density function of* ν *w.r.t.* μ.

For example, the multivariate Gaussian (family of) probability distributions, also denoted by MVN (multivariate normal), is a muti-dimensional generalization of the 2-dimensional family of normal distributions. It is a $\frac{n(n+3)}{2}$-dimensional parameterized statistical model of dominated measures on $\mathbb{R}^n$ whose density function $\mathcal{N}$ is given as follows

$$\mathcal{N}(x|\,\mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^2} \exp\big(-\frac{1}{2}\Sigma^{-1}(x-\mu, x-\mu)\big) \tag{1.2}$$

Here the couple (μ, Σ) is the parameter of the family, where $x, \mu \in \mathbb{R}^n$ and Σ is a positive definite quadratic form (symmetric bilinear form) on $\mathbb{R}^n$. Using the canonical Euclidean metric on $\mathbb{R}^n$, we identify $\mathbb{R}^n$ with its dual space $(\mathbb{R}^n)^*$. Thus the inverse Σ^{-1} is regarded as a quadratic form on $\mathbb{R}^n$ and the norm $|\Sigma|$ is also well-defined.

The classical problem of density estimation is the problem of estimation of the unknown probability measure $\mathbf{p}_u \in S$, knowing observed data as formulated above, assuming that S consists of measures dominated by a measure μ. As in the classical example of learning by curve fitting (Example 1.2), usually we assume that the dominated measure family S is finite dimensional, e.g., S is a MVN, whose density depends on the parameter (μ, Σ). In this case we need to estimate (μ, Σ) based on observables $(x_1, \cdots, x_m) \in \mathcal{X}^m$.

• *Ranking task* orders observed items according to some criterion. Web search, e.g. returning web pages relevant to a search query, is a typical ranking example. This task may look similar to the classification problem, but not quite the same, since in the ranking task we have to find a map from a subset of instances to the set of natural numbers. Thus ranking task is a problem to find ordered structure $(S_n, \leq)$ on the subset S_n of a set of instances.

• *Clustering task* partitions items into (homogeneous) regions. Clustering is often performed to analyze very large data sets. Clustering is one of the most widely used techniques for exploratory data analysis. In all disciplines, from social sciences to biology to computer science, people try to get a first intuition about their data by identifying meaningful groups among the data points. For example, computational biologists cluster genes on the basis of similarities in their expression in different experiments; retailers cluster customers, on the basis of their customer profiles, for the purpose of targeted marketing; and astronomers cluster stars on the basis of their spacial proximity.

• *Dimensionality reduction or manifold learning* transforms an initial representation of items in high dimensional space into a representation of the items in a space of lower dimension while preserving some properties of the initial representation. A common example involves pre-processing digital images in computer vision tasks. Many of dimensional reduction techniques are linear. When the technique is non-linear we speak about manifold learning technique. We can regard clustering as dimension reduction too.

1.4. **Main types of machine learning.** The type of a machine learning task is defined by the type of *interaction* between *the learner* and *the environment*. In particular, we consider *types of training data*, i.e., the data available to the learner before making decision and prediction, and the type of the outcomes.

Main types of machine learning are supervised, unsupervised and reinforcement.

• In *supervised learning* training data are *labelled*, i.e., each item in the training data set consists of the pair of a known instance and its feature, also called label. Examples of labeled data are emails that are labeled "spam" or "no spam" and medical histories that are labeled with the occurrence or absence of a certain disease. In these cases the learners output would be a spam filter and a diagnostic program, respectively. Most of classification and regression problems of machine learning belong to supervised learning. We also interpret a learning machine in supervised learning as a student who gives his supervisor a known instance and the supervisor answers with the known feature.

• In *unsupervised learning* there is *no additional label* attached to the training data and *the task is to describe the structure* of data. Density estimation, clustering and dimensionality reduction are examples of unsupervised learning problems. Regarding a learning machine in unsupervised learning as a student, then the student has to learn by himself without teacher. This learning is harder but happens more often in life. At the current time, unsupervised learning is most exciting part of machine learning where many new deep mathematical methods, e.g., persistent homology and other topological methods, are invented to described the structure of data. Furthermore, many tasks in supervised learning can be reduced to the density estimation problem, which is an unsupervised learning.

Remark 1.3. A machine learning task can be performed using supervised learning or unsupervised learning or some intermediate type between supervised learning and unsupervised learning. For instance, let us consider the problem of anomaly detection, i.e., the task of finding samples that are *inconsistent with the rest of the data*. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for

instances that seem to fit least to the remainder of the data set. Unsupervised anomaly can be regarded as a clustering task. Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier .

• *Reinforcement learning* is the type of machine learning where a learner actively interacts with the environment to achieve a certain goal[8]. For examples, reinforcement learning is used in self-driving car and chess play. In reinforcement learning the learner collects information through a course of actions by interacting with the environment. This active interaction justifies the terminology of *an agent* used to refer to the learner. The achievement of the agent's goal is typically measured by the reward he receives from the environment and which he seeks to maximize. In reinforcement learning the training data are not labeled, i.e., the supervisor does not directly give answers to the students questions. Instead, the supervisor evaluates the students behavior and gives feedback about it in terms of rewards.

1.5. **Basic mathematical problems in machine learning.** Let me recall that a learning is a process of gaining knowledge on a feature of observable by examination of partially available data. The learning is successful if we can make a prediction on unseen data, which improves when we have more data. For that purpose we need to quantify the notion of success of learning prediction. The classification task is a typical task in supervised learning where we can explain how and why a learning machine works and how and why machine learns successfully.

Question 1.4. *How to construct a mathematical model of learning?*

In general, this question is as large as mathematics and sciences. In many concrete cases, we have a family of candidate models and a search for the most suitable model is called a *model selection*.

Question 1.5. *How to quantify the difficulty/complexity of a learning problem?*

We quantify the difficulty of a problem in terms of its time complexity, which is the minimum time needed for performing computer program to solve a problem, and in terms of its resource complexity which measures the capacity of data storage and energy resource needed to solve the problem. If the complexity of a problem is very large then we cannot not learn it. So Question 1.5 contains the sub-question "*why can we learn a problem?*"

Question 1.6. *How to choose a learning algorithm?*

[8]The term "reinforcement in the context of animal learning came into use well after Thorndike's expression of the Law of Effect, first appearing in this context (to the best of our knowledge) in the 1927 English translation of Pavlov's monograph on conditioned reflexes. Pavlov described reinforcement as the strengthening of a pattern of behavior due to an animal receiving a stimulus (a reinforcer) in an appropriate temporal relationship with another stimulus or with a response. [SB2018, p.16]

Clearly we want to have a best learning algorithm, once we know a model of a machine learning which specifies the set of possible predictors (decisions) and the associated error/reward function. We want to minimize error and maximize the reward, so a best learning algorithm is a solution of the corresponding optimization problem. By Definition 1.1, a learning process is successful, if its prediction/estimation improves with the increase of data. Thus the notion of success of learning process requires a mathematical treatment of the asymptotic rate of the error/reward of the predictor in the presence of complexity of the problem.

1.6. **Conclusion.** Machine learning is learning with computer performed algorithms, whose performance improves with increasing volume of empirical data. In machine learning we use probability theory and mathematical statistics to model incomplete information and the random nature of the observed data, in other words, we build mathematical/statistical models for problems whose solutions are learning algorithms. To build mathematical models for problems in machine learning we need to specify the type of the problem we want to solve and the type of available data, which result in the type of machine learning model, which we also call the type of machine learning.

One of the main problems of mathematical foundations of machine learning concerns the quantification of the difficulty/complexity of a learning problem. This problem is ultimately related to the question of learnability of a learning problem, i.e., if a learning problem admits a successful learning algorithm in the sense of Definition 1.1. We shall discuss in our course several characterizations of learnability. Recently S. Ben-David, P. Hrubes, S Moran, A Shpilka and A. Yehudayoff proved that the question of learnability of a learning problem is not always decidable [BHMSY2019]. Other important problems are the problem of formalizing the "meaning" or "information content" of a scientific data, and the challenge of mathematical optimization in the presence of large input sizes.

2. Mathematical models for supervised learning

What I cannot create, I do not understand.
Richard Feynman

Last week we discussed the concept of learning and examined several examples. Today I shall specify the concept of learning by presenting basic mathematical models of supervised learning. A model for machine learning must be able to make predictions and improves their ability to make predictions in light of new data.

The model of supervised learning I present today is based on Vapnik's statistical learning theory, which starts from the following concise concept of learning.

Definition 2.1. ([Vapnik2000, p. 17]) Learning is a problem of function estimation on the basis of empirical data.

There are two main model types for machine learning: *discriminative model* and *generative model*. They are distinguished by the *type of functions we want to estimate* for understanding the feature of observable.

2.1. **Discriminative model of supervised learning.** Let us consider the following toy example of a binary classification task, which like regression tasks (Example 1.2), is a typical example of tasks in supervised learning.

Example 2.2 (Toy example)**.** We are working for a dermatological clinic and we want to develop a program for prediction of a certain type of skin diseases by examination of image scanning of the skin condition of patients. First we need to define the set of essential parameters of the skin image. For example, the area of the spot in trouble, which has value in an interval $I_1 \subset \mathbb{R}$, the thickness of the skin spot in trouble, which has value in an interval $I_2 \subset \mathbb{R}$, the color of the skin spot in trouble, which can be quantified by a real number in an interval I_3, the age of the patient, which can be divided in five age groups A_1, A_2, A_3, A_4, A_5. We want to construct a function $f : \mathcal{X} := \cup_{i=1}^{5} I_1 \times I_2 \times I_3 \times \{A_i\} \to \mathcal{Y} := \{\pm 1\}$, whose value predicts "Yes" or "No" possibility of the type of disease. Such a function f can be identified with the subset $f^{-1}(1) \subset \mathcal{X}$. Our function f shall depend on data at time n i.e., depend on the set S_n consisting of n parameters $x_1, \cdots, x_n \in \mathcal{X}$ with given "label" $f(x_1), \cdots, f(x_n)$. Thus our function should be written as f_{S_n}, and our construction of a family $\{f_{S_n},\, n \in \mathbb{N}^+\}$ is successful, if as n grows the value $f_{S_n}(y)$ on new data $y \in \mathcal{X}$ is closer to the true value $f(y)$.

- Since we shall use probability theory to model uncertainty, all spaces in learning theory are measurable spaces and all mappings are measurable mappings, unless otherwise specified.
- We shall denote the space of all measurable mappings between measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ by $\mathrm{Meas}(\mathcal{X}, \mathcal{Y})$.

A discriminative model of supervised learning consists of the following components.

• *A domain space* $\mathcal{X}$ (also called *an input space*) consists of elements, whose features we wish to learn. In general case elements of x is distributed by an unknown probability measure $\mu_{\mathcal{X}}$. In other words, the probability that x belongs to a subset $A \subset \mathcal{X}$ is $\mu_{\mathcal{X}}(A)$. Usually we don't know the distribution $\mu_{\mathcal{X}}$.

In Example 2.2 of a classification task, the domain set is also denoted by $\mathcal{X}$. In Example 1.2 of learning a physical law, a regression talk, the domain set $\mathcal{X}$ is the domain $\mathbb{R}$.

• *An output space* $\mathcal{Y}$, also called *a label set*, consists of possible features (also called labels) y of inputs $x \in \mathcal{X}$. We are interested in finding a predictor/mapping $h : \mathcal{X} \to \mathcal{Y}$ such that the feature of x is $h(x)$. If such a mapping h exists and is measurable, the feature $h(x)$ is distributed by the measure $h_*(\mu_{\mathcal{X}})$. In general, such a function does not exist, and we model the stochastic relation between x and y via a probability measure $\mu_{\mathcal{X}\times\mathcal{Y}}$ on the space $(\mathcal{X} \times \mathcal{Y})$, i.e., the probability of $(x, y) \in A \subset \mathcal{X} \times \mathcal{Y}$ being a labeled pair is equal to $\mu_{\mathcal{X}\times\mathcal{Y}}(A)$. In general we don't know $\mu_{\mathcal{X}\times\mathcal{Y}}$.

In Example 2.2 the label set is also denoted by $\mathcal{Y}$. In Example 1.2 of learning a physical law the label set is the set $\mathbb{R}$ of all possible value of $f(x)$.

• *A training data* is a sequence $S = \{(x_1, y_1), \cdots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ of observed labeled pairs, which are usually assumed to be i.i.d. (independently identically distributed). In this case S is distributed by the product measure $\mu^n_{\mathcal{X}\times\mathcal{Y}}$ on $(\mathcal{X} \times \mathcal{Y})^n$. The number n is called *the size of* S. The training data S is thought as given by a "supervisor".

• A *hypothesis space* $\mathcal{H} \subset Meas(\mathcal{X}, \mathcal{Y})$ of possible predictors $h : \mathcal{X} \to \mathcal{Y}$. Since we are working in the category of mesurable spaces, we assume that h is a measurable map.

For instance, in Example 2.2 we may wish to choose $\mathcal{H}$ consisting of subsets $D_i \subset \mathcal{X}_i := I_1 \times I_2 \times I_3 \times \{A_i\}, i \in \{1, 2, 3, 4, 5\}$, and D_i is a cuboid in $I_1 \times I_2 \times I_3$.

In Example 1.2 we already let
$\mathcal{H} := \{h : \mathbb{R} \to \mathbb{R} |\ h \text{ is a polynomial of degree at most } d\ \} \cong \mathbb{R}^{d+1}$.

• *The aim of a learner* is to find *a best prediction rule* A that assigns a training data S to a prediction $h_S \in \mathcal{H}$. In other words the learner needs to find a rule, more precisely, *an algorithm*

$$A : \bigcup_{n\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H},\ S \to h_S \tag{2.1}$$

such that $h_S(x)$ predicts the label of (unseen) instance x with the less error.

• *An error function*, also called *a risk function*, measures the discrepancy between a hypothesis $h \in \mathcal{H}$ and an ideal predictor. An error function is a central notion in learning theory. After Wald's fundamental work [Wald1950], an error function, denoted by $R : \mathcal{H} \to \mathbb{R}$, is usually defined as the averaged discrepancy of $h(x)$ and y, where (x, y) runs over $\mathcal{X} \times \mathcal{Y}$, see also (3.14) for an alternative definition. The averaging is calculated using the probability measure $\mu := \mu_{\mathcal{X}\times\mathcal{Y}}$ that governs the distribution of labeled pairs (x, y). Thus a risk/error function R must depend on μ, so we denote

it by R_μ. It is accepted that the risk function $R_\mu : \mathcal{H} \to \mathbb{R}$ of a supervised learning model with a hypothesis space $\mathcal{H}$ is defined as follows.

$$R^L_\mu(h) := \int_{\mathcal{X}\times\mathcal{Y}} L(x, y, h)\, d\mu \tag{2.2}$$

where $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ is *an instantaneous loss function* that measures the discrepancy between the value of a prediction/hypothesis h at x and the possible feature y:

$$L(x, y, h) := d(y, h(x)). \tag{2.3}$$

Here $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a non-negative function that vanishes at the diagonal $\{(y, y)|\, y \in \mathcal{Y}\}$ of $\mathcal{Y} \times \mathcal{Y}$. Such a function d will be called *a quasi-distance function*. For example $d(y, y') = |y - y'|^2$. By taking averaging over $(\mathcal{X} \times \mathcal{Y})$ using μ, we effectively count only the points (x, y) which are correlated as labeled pairs.

Note the expected risk function R^L_μ is well defined on $\mathcal{H}$ only if $L(x, y, h) \in L^1(\mathcal{X} \times \mathcal{Y}, \mu)$ for all $h \in \mathcal{H}$.

• The main question of supervised learning theory is to find necessary and sufficient conditions for the existence of a prediction rule A in (2.1) such that, as the size of S goes to infinity, the error of h_S converges to the error of an ideal predictor, or more precisely, to the infimum of the error of $h \in \mathcal{H}$, and then to construct such a prediction rule A.

We summarize our discussion in the following

Definition 2.3. *A discriminative model of supervised learning* consists of a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, P_{\mathcal{X}\times\mathcal{Y}})$ where $\mathcal{X}$ is a measurable input space, $\mathcal{Y}$ is a measurable output space, $\mathcal{H} \subset Meas(\mathcal{X}, \mathcal{Y})$ is a hypothesis space, $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ is an instantaneous loss function and *a statistical model* $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$[9]. An optimal predictor $h_{S_n} \in \mathcal{H}$ should minimize the risk function R^L_μ based on empirical labeled data $S_n \in (\mathcal{X} \times \mathcal{Y})^n$ that have been distributed by un unknown measure μ^n where $\mu \in P_{\mathcal{X}\times\mathcal{Y}}$.

Remark 2.4. (1) In our discriminative model of supervised learning we model the random nature of training data $S \in (\mathcal{X} \times \mathcal{Y})^n$ via a probability measure μ^n on $(\mathcal{X} \times \mathcal{Y})^n$, where $\mu = \mu_{\mathcal{X}\times\mathcal{Y}} \in P_{\mathcal{X}\times\mathcal{Y}}$. Thus we assume implicitly that the training data are i.i.d.. In general, the distribution $\mu_{\mathcal{X}\times\mathcal{Y}}$ of labeled pairs (x, y) does not correlate with the distribution $\mu_{\mathcal{X}}$ of $x \in \mathcal{X}$. The main difficulty in search for the best prediction rule A is that we don't know μ, we know only training data S distributed by μ^n.

(2) Note that the expected risk function R^L_μ is well-defined on $\mathcal{H}$ only if $L(x, y, h) \in L^1(\mathcal{X} \times \mathcal{Y}, \mu)$ for all $h \in \mathcal{H}$. We don't know μ exactly, but based

[9]A statistical model $\mathcal{P}_{\mathcal{X}}$ is a subset of the set $\mathcal{P}(\mathcal{X})$ of all probability measures on a measurable space $\mathcal{X}$, see [Le2020] for a more in depth discussion on statistical models.

on our *prior knowledge*, we know that $\mu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$. Thus we should assume that $L \in L^1(\mathcal{X}\times\mathcal{Y}, \nu)$ for any $\nu \in \mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$.

(3) The quasi-distance function $d : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ induces a quasi-distance function $d^n : \mathcal{Y}^n \times \mathcal{Y}^n \to \mathbb{R}$ as follows

$$d^n([y_1, \cdots, y_n], [y'_1, \cdots, y'_n]) = \sum_{i=1}^n d(y_i, y'_i), \tag{2.4}$$

and therefore it induces the expected loss function $R_{\mu^n}^{L(d^n)} : \mathcal{H} \to \mathbb{R}$ as follows

$$\begin{aligned} R_{\mu^n}^{L(d^n)}(h) &= \int_{(\mathcal{X}\times\mathcal{Y})^n} d^n([y_1, \cdots, y_n], [h(x_1), \cdots, h(x_n)])d\mu^n \\ &= n\int_{\mathcal{X}\times\mathcal{Y}} L(x, y, h)d\mu. \end{aligned} \tag{2.5}$$

Thus it suffices to consider only the loss function $R_\mu^{L(d)}(h)$, if S is a sequence of i.i.d. observables.

(4) Now we shall show that the classical case of learning a physical law by curve fitting to data, see Example 1.2, is a "classical limit" of our discriminative model of supervised learning. In the classical learning problem, since we know the exact position $S := \{(x_1, y_1), \cdots, (x_n, y_n)\} \in (\mathcal{X}\times\mathcal{Y})^n$, we assign the Dirac probability measure $\mu_S^\delta := \delta_{x_1,y_1} \times \cdots \times \delta_{x_n,y_n}$ to the space $(\mathcal{X}\times\mathcal{Y})^n$. [10] Now let $d(y, y') = |y - y'|^2$. Then we have

$$R_{\mu_S^\delta}^{L(d^n)}(h) = \sum_{i=1}^n |h(x_i) - y_i|^2. \tag{2.6}$$

Note that the RHS of (2.6) coincides with the error of estimation in (1.1), i.e., our *empirical risk function* $R_{\mu_S^\delta}^{L(d^n)}$ defined on $\mathcal{H} = \mathbb{R}^{d+1}$ coincides with the error function R in (1.1).

Example 2.5 (0-1 loss)**.** Let $\mathcal{H} = Meas(\mathcal{X}, \mathcal{Y})$. The *0-1 instantaneous loss function* $L^{(0-1)} : \mathcal{X}\times\mathcal{Y}\times\mathcal{H} \to \{0, 1\}$ is defined as follows:

$$L^{(0-1)}(x, y, h) := d(y, h(x)) = 1 - \delta^y_{h(x)}. \tag{2.7}$$

The corresponding expected 0-1 loss determines the probability of the answer $h(x)$ that does not correlate with x:

$$R^{(0-1)}_{\mu_{\mathcal{X}\times\mathcal{Y}}}(h) = \mu_{\mathcal{X}\times\mathcal{Y}}\{(x, y) \in \mathcal{X}\times\mathcal{Y} |\, h(x) \neq y\} = 1 - \mu_{\mathcal{X}\times\mathcal{Y}}(\{x, h(x)\}). \tag{2.8}$$

Example 2.6. Assume that $x \in \mathcal{X}$ is distributed by a probability measure $\mu_\mathcal{X}$ and its feature y is defined by $y = h(x)$ where $h : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping. Denote by $\Gamma_h : \mathcal{X} \to \mathcal{X}\times\mathcal{Y}$, $x \mapsto (x, h(x))$, the graph of h. Then (x, y) is distributed by the push-forward measure $\mu_h := (\Gamma_h)_*(\mu_\mathcal{X})$, where

$$(\Gamma_h)_*\mu_\mathcal{X}(A) = \mu_\mathcal{X}\big(\Gamma_h^{-1}(A)\big) = \mu_\mathcal{X}\Big(\Gamma_h^{-1}(A \cap \Gamma_h(\mathcal{X}))\Big). \tag{2.9}$$

[10] If $(x_1, y_1), \cdots, (x_n, y_n)$ are i.i.d. then the measure $\mu_{S_n} := \frac{1}{n}\sum_{i=1}^n \delta_{i_i,y_i} \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is called *the empirical measure defined by the data* S_n.

Let us compute the expected 0-1 loss function for a mapping $f \in \mathcal{H} := Meas(\mathcal{X}, \mathcal{Y})$ w.r.t. the measure μ_h. By (2.8) and by (2.9), we have

$$R^{(0-1)}_{\mu_h}(f) = 1 - \mu_{\mathcal{X}}(x|f(x) = h(x)). \tag{2.10}$$

Hence $R^{(0-1)}_{\mu_h}(f) = 0$ iff $f = h$ $\mu_{\mathcal{X}}$-a.e..

2.2. **Generative model of supervised learning.** In many cases, a discriminative model of supervised learning may not yield a successful learning algorithm because the hypothesis space $\mathcal{H}$ is too small and cannot approximate a desired prediction for a feature $\in \mathcal{Y}$ of instance $x \in \mathcal{X}$ with a satisfying accuracy, i.e., *the optimal performance error of the class* $\mathcal{H}$

$$R^L_{\mu,\mathcal{H}} := \inf_{h\in\mathcal{H}} R^L_\mu(h) \tag{2.11}$$

that represents the optimal performance of a learner using $\mathcal{H}$ is quite large.

One of possible reasons of this failure is that the stochastic relation between $y \in \mathcal{Y}$ and $x \in \mathcal{X}$, cannot be accurately expressed by any function $h \in Meas(\mathcal{X}, \mathcal{Y})$. Recall that a stochastic correlation between x and its feature y is expressed in terms of a measure $\mu_{\mathcal{X}\times\mathcal{Y}}$ on $(\mathcal{X} \times \mathcal{Y}, \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}})$. This measure cannot always be represented in terms of $(\Gamma_h)_*\mu_{\mathcal{X}}$ for some measurable mapping $h : \mathcal{X} \to \mathcal{Y}$ and a probability measure $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$, see the exercise below.

Exercise 2.7. Let $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$ be Euclidean spaces. For a topological space $\mathcal{Z}$ we shall consider the natural Borel σ-algebra $\mathcal{B}(\mathcal{Z})$. The observed feature $y \in \mathcal{Y}$ of an input $x \in \mathcal{X}$ satisfies the following relation: $|y|^2+|x|^2 = 1$, where $|.|$ denotes the Euclidean norm on $\mathcal{Y}$ and $\mathcal{X}$ respectively. Denote by S^{n+m-1} the unit sphere in the product space $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^{n+m}$. Then y is a feature of x if and only if $(x, y) \in S$. Since $S \in \Sigma_{\mathcal{X}\times\mathcal{Y}} = \mathcal{B}(\mathbb{R}^{n+m})$, this correlation is expressed via a measure $\mu_S \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ with support in S, i.e., for any $D \in \Sigma_{\mathcal{X}\times\mathcal{Y}}$ we have $\mu_S(D) = \mu_S(A \cap S)$. Prove that μ_S cannot represented as a graph $(\Gamma_h)_*\mu_{\mathcal{X}}$ for any $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$ and any measurable mapping $h : \mathcal{X} \to \mathcal{Y}$.

Hint. Let $sppt(\mu_{\mathcal{X}})$ denotes the support the measure $\mu_{\mathcal{X}}$, i.e.,

$$sppt(\mathcal{X}) : \{x \in \mathcal{X}| \, \forall x \in U_x \in \mathcal{B}(\mathcal{X}) \text{ we have } \mu_{\mathcal{X}}(U_x) > 0\}.$$

Show that $sppt((\Gamma_h)_*\mu_{\mathcal{X}}) = \Gamma_h(sppt(\mu_{\mathcal{X}}))$ and S cannot be represented as $\Gamma_h(sppt(\mu_{\mathcal{X}}))$.

Now we wish to estimate the stochastic relation between x and y, more precisely, to *estimate the probability that an element $y \in \mathcal{Y}$ is a feature of x*. This probability is expressed as *the conditional probability* $P(y \in B|x)$ - the probability that a feature y of $x \in \mathcal{X}$ belongs to the subset $B \subset \Sigma_{\mathcal{Y}}$.

Digression. *Conditional probability* is one of most basic concepts in probability theory. In general, we always have a prior information before taking

a decision, e.g. before estimating the probability of a future event. Conditional probability $P(A|B)$ formalizes the probability of an event A given the knowledge that event B happens. Here we assume that A, B are elements of the sigma-algebra $\Sigma_{\mathcal{X}}$ of a probability space $(\mathcal{X}, \Sigma_{\mathcal{X}}, P)$. If $\mathcal{X}$ is countable, the concept of conditional probability can be defined straightforward:

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \tag{2.12}$$

It is not hard to see that, given B, the conditional probability $P(\cdot|B)$ defined in (2.12) is a probability measure on $\mathcal{X}$, $A \mapsto P(A|B)$, which is called *the conditional probability measure given* B. In its turn, by taking integration over $\mathcal{X}$ using the conditional probability $P(\cdot|B)$, we obtain the notion of *conditional expectation*, given B, which shall be denoted by $\mathbb{E}_{P(\cdot|B)}$, see Subsection A.2.2.

In general case when $\mathcal{X}$ is not countable, the definition of conditional probability is more subtle, especially when we have to define $P(A|B)$, where $P(B) = 0$. A typical situation is the case $B = h^{-1}(z_0)$, where $h : \mathcal{X} \to \mathcal{Z}$ is a measurable mapping and $h_*(P)(z_0) = 0$. To treat this important case we need to define first the notion of conditional expectation, see Subsection A.2 in Appendix.

In generative models of supervised learning, we expresses the stochastic correlation between input $x \in \mathcal{X}$ and label $y \in \mathcal{Y}$ via a probability measure $\mu_{\mathcal{X}\times\mathcal{Y}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, assuming usually $\Sigma_{\mathcal{X}\times\mathcal{Y}} = \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}}$. Denote by $\Pi_{\mathcal{X}}$ (resp. $\Pi_{\mathcal{Y}}$) the natural projection from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{X}$ (resp. to $\mathcal{Y}$). One is then interested in the conditional probability

$$\mu_{\mathcal{Y}|\mathcal{X}}(B|x) := \frac{d(\Pi_{\mathcal{Y}})_*(1_{B\times\mathcal{X}}\mu_{\mathcal{X}\times\mathcal{Y}})}{d(\Pi_{\mathcal{Y}})_*\mu_{\mathcal{X}\times\mathcal{Y}}}(x) \tag{2.13}$$

for $B \in \Sigma_{\mathcal{Y}}$ and $x \in \mathcal{X}$. Here the equality (2.13) should be understood as an equivalence class of functions in $L^1(\mathcal{X}, (\Pi_{\mathcal{X}})_*\mu_{\mathcal{X}\times\mathcal{Y}})$. For simplification and practical purposes, we assume that the conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is regular, i.e. $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is a probability measure on $\mathcal{Y}$ for all $x \in \mathcal{X}$, and the function $x \mapsto \mu_{\mathcal{Y}|\mathcal{X}}(B|x)$ is measurable for every $B \in \Sigma_{\mathcal{Y}}$, see also Appendix A.3. Thus under this regularity condition the problem of supervised learning is equivalent to the problem of finding a Markov kernel, also called a probabilistic morphism [JLT2020], $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ describing the correlation between labeled data y and input data x.

A further simplification is to assume that a family of regular conditional measures $\mathbf{p}(x) := \mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ is dominated by a measure $\mu_0 \in \mathcal{P}(\mathcal{Y})$, and hence we need to estimate the density function $p(y|x) := d\mathbf{p}(x)/d\mu_0$ understood as a representative of the equivalence class of the function $d\mathbf{p}(x)/d\mu_0 \in L^1(\mathcal{Y}, \mu_0)$.

Definition 2.8. *A generative model of supervised learning* is a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})})$, where $\mathcal{X}$ and $\mathcal{Y}$ are measurable spaces, $\mathcal{H}$ is a family

of measurable mappings $T : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ and $L : \mathcal{X} \times \mathcal{P}(\mathcal{Y}) \times \mathcal{H} \to \mathbb{R}$ is an instantaneous loss function and $\mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})} \subset \mathcal{P}(\mathcal{X} \times \mathcal{P}(\mathcal{Y}))$.

A classical generative model of supervised learning is a special case of a generative model $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})})$, where $\mathcal{H}$ is a family of density functions $p(y|x)$.

Remark 2.9. (1) By Formula (A.15) the knowledge of a joint distribution $\mu_{\mathcal{X}\times\mathcal{Y}}$ implies the knowledge of the conditional probability measure $\mu_{\mathcal{Y}|\mathcal{X}}(B|x)$; and conversely, by (A.13) the joint distribution $\mu_{\mathcal{X}\times\mathcal{Y}}$ can be recovered from the conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$, $x \in \mathcal{X}$, given the marginal probability measure $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$.

(2) The knowledge of a joint distribution $\mu_{\mathcal{X}\times\mathcal{Y}}$ leads to a solution of a discriminative models $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mu_{\mathcal{X}\times\mathcal{Y}})$ as follows. First, knowing $\mu := \mu_{\mathcal{X}\times\mathcal{Y}}$ we compute the expected risk R^L_μ for an instantaneous loss function L. Then we determine a minimizing sequence $\{h_i \in \mathcal{H}\}$ of R^L_μ, i.e.,

$$\lim_{n\to\infty} R^L_\mu(h_i) = R^L_{\mu,\mathcal{H}}.$$

Hence generative models give more information than discriminative models. This is a version of the Bayes principle stating that if the probability distribution μ of labeled pairs is known, then the optimal predictor h_μ can be expressed explicitly.

Example 2.10 (The Bayes Optimal Predictor)**.** Let us consider a discriminative model of supervised learning $(\mathcal{X}, \mathcal{Y} = \{0, 1\}, \mathcal{H} = \mathcal{Y}^{\mathcal{X}}, L^{(0-1)}, P_{\mathcal{X}\times\mathcal{Y}} = \{\mu\})$, where $\mathcal{X}$ be an arbitrary sample space. By Remark A.14 there exists regular conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ [11]. Then we define the Bayes Optimal Predictor $f_\mu \in \mathcal{H}$ by setting for $x \in \mathcal{X}$, see

$$f_\mu(x) = \begin{cases} 1 & \text{if } \mu_{\mathcal{Y}|\mathcal{X}}(\{y = 1\}|x) \geq 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

We claim that the Bayes Optimal Predictor f_μ is optimal. In other words for every classifier $h \in \mathcal{Y}^{\mathcal{X}}$ we have

$$R^{(0-1)}_\mu(f_\mu) \leq R^{(0-1)}_\mu(h). \tag{2.14}$$

To prove our claim, note that for any $h \in \mathcal{Y}^{\mathcal{X}}$ we have

$$R^{(0-1)}_\mu(h) \overset{(2.8)}{=} \mathbb{E}_\mu(|h(x) - y|)\, d\mu \overset{(A.16)}{=} \int_{\mathcal{X}}\int_{\mathcal{Y}} |h(x) - y| d\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)\, d\mu_{\mathcal{X}}. \tag{2.15}$$

By (2.15), it suffices to prove that for all $x \in \mathcal{X}$ we have

$$\mathbb{E}_\mu(h(x)) := \int_{\mathcal{Y}} |h(x) - y| d\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x) \geq \int_{\mathcal{Y}} |f_\mu(x) - y| d\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x). \tag{2.16}$$

To prove (2.16), we rewrite

[11] the function $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is defined up to a $\mu_{\mathcal{X}}$-zero set uniquely, where $\mu_{\mathcal{X}} = (\Pi_{\mathcal{X}})_*(\mu)$ is the marginal measure on $\mathcal{X}$.

$$\mathbb{E}_\mu(h(x)) = \begin{cases} 1 - \mu_{\mathcal{Y}|\mathcal{X}}(\{y=1\}|x) & h(x)=1 \\ \mu_{\mathcal{Y}|\mathcal{X}}(\{y=1\}|x) & h(x)=0. \end{cases}$$

Given $x \in \mathcal{X}$, our aim is to minimize $\mathbb{E}_\mu(h(x))$. If $1 - \mu_{\mathcal{Y}|\mathcal{X}}(\{y=1\}|x) \leq \mu_{\mathcal{Y}|\mathcal{X}}(\{y=1\}|x)$ we set $h(x) := 1$, and $h(x) := 0$ otherwise. By rearranging the inequality into the form of $\mu_{\mathcal{Y}|\mathcal{X}}(\{y=1\}|x) \geq 1/2$ we get the original definition of Bayes optimal predictor $f_\mu(x)$.

Exercise 2.11 (Regression optimal Bayesian estimator)**.** In regression problem the output space $\mathcal{Y}$ is $\mathbb{R}$. Let us define the following embedding

$$i_1 : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{Y}} : \, [i_1(f)](x,y) := f(x),$$

$$i_2 : \mathbb{R}^{\mathcal{Y}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{Y}} : \, [i_2(f)](x,y) := f(y).$$

(These embeddings are adjoint to the projections: $X : \mathcal{X}\times\mathbb{R} \overset{\Pi_{\mathcal{X}}}{\to} \mathcal{X}$ and $\mathcal{X}\times\mathbb{R} \overset{\Pi_{\mathbb{R}}}{\to} \mathbb{R}$.) For a given probability measure μ on $\mathcal{X}\times\mathbb{R}$ we set

$$L^2(\mathcal{X}, (\Pi_{\mathcal{X}})_*\mu) = \{f \in \mathbb{R}^{\mathcal{X}} |\, i_1(f) \in L^2(\mathcal{X}\times\mathbb{R}, \mu)\},$$

$$L^2(\mathbb{R}, (\Pi_{\mathbb{R}})_*\mu) = \{f \in \mathbb{R}^{\mathbb{R}} |\, i_2(f) \in L^2(\mathcal{X}\times\mathbb{R}, \mu)\}.$$

Now we let $\mathcal{F} := L^2(\mathcal{X}, (\Pi_{\mathcal{X}})_*\mu) \subset \mathcal{Y}^{\mathcal{X}}$. Let Y denote the identity function on $\mathbb{R}$ i.e., $Y(y) = y$. Assume that $Y \in L^2(\mathbb{R}, (\Pi_{\mathbb{R}})_*\mu)$ and define the quadratic loss function $L : \mathcal{X}\times\mathcal{Y}\times\mathcal{F} \to \mathbb{R}$ (cf. (2.7))

$$L(x,y,h) = d^2(y, h(x)) := |y - h(x)|^2, \tag{2.17}$$

$$R^L_\mu(h) = \mathbb{E}_\mu\big(|Y(y) - h(x)|^2\big) = |i_2(Y) - i_1(h)|^2_{L^2(\mathcal{X}\times\mathbb{R},\mu)}. \tag{2.18}$$

The expected risk R^L_μ is called the *L_2-risk*, also known as *mean squared error* (MSE). Show that the *regression function* $r(x) := \mathbb{E}_\mu\big(i_2(Y)|X = x\big)$ belongs to $\mathcal{F}$ and minimizes the $L_2(\mu)$-risk.

Hint. Establish the following formula for the L_2-risk

$$R_\mu(h) = \mathbb{E}_\mu(|i_2(Y) - i_1(r)|^2) + \mathbb{E}_\mu(|i_1(h) - i_1(r)|^2). \tag{2.19}$$

Example 2.12. Let us revisit Example 2.2 of prediction of a skin disease. A generative model in this case consists of quadruple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L)$ where $\mathcal{X} = \cup_{i=1}^5 I_1 \times I_2 \times I_2 \times \{A_i\}$, $\mathcal{Y} = \{0,1\}$ and $\mathcal{H} = \{h_\theta : \mathcal{X} \rightsquigarrow \mathcal{Y} | \theta \in \Theta\}$, where Θ is a parameter set, and $L : \mathcal{X}\times\mathcal{Y}\times\mathcal{H} \to \mathbb{R}$ is an instantaneous loss. In the classical setting, $h_\theta(x) = p(y|x)\mu_0$, where $\mu_0 \in \mathcal{P}(\mathcal{Y})$ is the counting measure. This generative model gives us clearer picture of disease, but in most cases we need to take binary solution, which is given in a disriminative model.

2.3. **Empirical Risk Minimization and overfitting.** In a discriminative model of supervised learning our aim is to construct a prediction rule A that assigns a predictor h_S to each sequence

$$S = \{(x_1, y_1), \cdots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

of i.i.d. labeled data such that the expected error $R^L_\mu(h_S)$ tends to the optimal performance error $R^L_{\mu,\mathcal{H}}$ of the class $\mathcal{H}$. One of most popular ways to find a prediction rule A is to use the Empirical Risk Minimization.

For a loss function

$$L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R},$$

and a training data $S \in (\mathcal{X} \times \mathcal{Y})^n$ we define *the empirical risk* of a predictor h as follows

$$\hat{R}^L_S(h) := \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, h). \tag{2.20}$$

If L is fixed, then we also omit the superscript L.

The empirical risk is a function of two variables: the "empirical data" S and the predictor h. Given S a learner can compute $\hat{R}_S(h)$ for any function $h : \mathcal{X} \to \mathcal{Y}$. A minimizer of the empirical risk should also "approximately" minimize the expected risk. This is the *empirical risk minimization principle*, abbreviated as ERM.

Remark 2.13. We note that

$$\hat{R}^L_S(h) = R^L_{\mu_S}(h) \tag{2.21}$$

where μ_S is the empirical measure on $(\mathcal{X} \times \mathcal{Y})$ associated to S, cf. (2.6). If h is fixed, by the weak law of large numbers, see e.g. Proposition B.2 in the Appendix, the RHS of (2.21) converges in probability to the expected risk $R^L_\mu(h)$, so we could hope to find a condition under which the RHS of (2.21) for a sequence of h_S, instead of h, converges to $R^L_{\mu,\mathcal{H}}$.

Example 2.14. In this example we shall show the failure of ERM in certain cases. We shall consider a disriminative model $(\mathcal{X}, \mathcal{Y}, \mathcal{H} = \mathcal{Y}^{\mathcal{X}}, L, \mathcal{P}(\mathcal{X} \times \mathcal{Y}))$ where $\mathcal{Y} = \{0, 1\}$ and L is the 0-1 loss function, see Example 2.5. Then the empirical 0-1 risk $\hat{R}^{0-1}_S$ is defined as follows:

$$\hat{R}^{0-1}_S(h) := \frac{|i \in [n] : \, h(x_i) \neq y_i|}{n} \tag{2.22}$$

for a training data $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ and a function $h : \mathcal{X} \to \mathcal{Y}$. We also often call $R^{0-1}_S(h)$ - *the training error* or *the empirical error*.

Now we assume that labeled data (x, y) is generated by a map $f : \mathcal{X} \to \mathcal{Y}$, i.e., $y = f(x)$, and further more, x is distributed by a measure $\mu_{\mathcal{X}}$ on $\mathcal{X}$. Then $(x, f(x))$ is distributed by the measure $\mu_f = (\Gamma_f)_*(\mu_{\mathcal{X}})$, see Example 2.6. Since $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$, we have $f \in \mathcal{H}$, and clearly $R^{0-1}_{\mu_f}(f) = 0$. Next, for any given $\varepsilon > 0$ and any $n \in \mathbb{N}^+$, we shall find a map $f \in \mathcal{Y}^{\mathcal{X}}$, a measure $\mu_{\mathcal{X}}$,

and a predictor h_{S_n} such that $\hat{R}^{0-1}_{S_n}(h_{S_n}) = 0$ and $R^{0-1}_{\mu_f}(h_{S_n}) = \varepsilon$, which shall imply that the ERM is invalid in this case.

Set

$$h_{S_n}(x) = \begin{cases} f(x_i) & \text{if there exists } i \in [n] \text{ s.t. } x_i = x \\ 0 & \text{otherwise.} \end{cases} \tag{2.23}$$

Clearly $\hat{R}^{0-1}_{S_n}(h_{S_n}) = 0$. We also note that $h_{S_n}(x) = 0$ except a finite (at most n) number of points x in $\mathcal{X}$.

Let $\mathcal{X}$ be any Borel subset in $\mathbb{R}^k$, $k \geq 1$. Let μ_0 be the Lebesgue measure on I^k.We decompose $\mathcal{X}$ into a disjoint union of two measurable subsets A_1 and A_2 such that $\mu_{\mathcal{X}}(A_1) = \varepsilon$. Let $f : \mathcal{X} \to \mathbb{Z}_2$ be equal 1_{A_1} - the indicator function of A_1. By (2.8) we have

$$R_{\mu_f}(h_{S_n}) = \mu_{\mathcal{X}}(\{x \in \mathcal{X} |\, h_{S_n}(x) \neq 1_{A_1}(x)\}). \tag{2.24}$$

Since $h_{S_n}(x) = 0$ a.e. on $\mathcal{X}$ it follows from (2.24) that

$$R_{\mu_f}(h_{S_n}) = \mu_{\mathcal{X}}(A_1) = \varepsilon.$$

Such a predictor h_{S_n} is said to be *overfitting*, i.e., it fits well to training data but not real life.

The idea for the construction of the counter-example of ERM principle is simple: the minimizer h_n of the empirical risk R^{0-1}_S, constructed using the values $f(x_n)$ of a training data $S = (x_1, x_2 \cdots, x_n)$, has empirical risk 0. On the other hand, the set $[S]$ of a training data $S = \{x_1, \cdots, x_n\}$ is countable and hence is a null set in the space $\mathcal{X} = I^k$. Since h_n vanishes outside the set $[S]$, its expected 0-1 risk is equal to the measure of the subset $A_2 = f^{-1}(1)$ of measure ε.

Exercise 2.15 (Empirical risk minimization)**.** Let $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ and $\mathcal{F} := \{h : \mathcal{X} \to \mathcal{Y} |\, \exists v \in \mathbb{R}^d : h(x) = \langle v, x \rangle\}$ be the class of linear functions in $\mathcal{Y}^{\mathcal{X}}$. For $S = \big((x_i, y_i)\big)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ and the quadratic loss L (defined in (2.17)), find the hypothesis $h_S \in \mathcal{F}$ that minimizes the empirical risk $\hat{R}^L_S$.

The phenomenon of overfitting suggests the following questions concerning ERM principle [Vapnik2000, p.21].

1) Can we learn in discriminative model of supervised learning using the ERM principle?

2) If we can learn, we would like to know the rate of convergence of the learning process as well as construction method of learning algorithms.

We shall address these questions later in our course and recommend the books by Vapnik [Vapnik1998, Vapnik2000, Vapnik2006] on statistical learning theory for further reading.

Exercise 2.16. $^{(\star)}$ Describe a ranking task as a problem in supervised learning.

Hint: see [SSBD2014, §17.4, p. 238].

Remark 2.17. (1) We omit the ERM principle for generative models of supervised learning and refer the reader to [Vapnik1998, §1.9.2, p. 36-37].

(2) ERM principle was a starting point of departure of classical parametric statistics to statistical learning theory founded Vapnik [Vapnik1998, p. 7].

2.4. **Conclusion.** A discriminative model for supervised learning consists of a quintuple $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, P_{\mathcal{X}\times\mathcal{Y}})$ where $\mathcal{H} \subset Meas(\mathcal{X}, \mathcal{Y})$ and $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}} \subset \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. In this model the aim of a learner is to find a prediction rule $A : S \mapsto h_S \in \mathcal{H}$ for any sequence of i.i.d. labeled training data $S = S_n \in (\mathcal{X} \times \mathcal{Y})^n$ of size n such that $(x, h_S(x))$ approximates the training data best, i.e. $R^L_\mu(h_S)$ is smallest as possible. The ERM principle suggests that we could choose h_S to be the minimizer of the empirical risk $\hat{R}^L_S$ instead of the unknown function R^L_μ, and we hope that as the size of S increases the expected error $R^L_\mu(h_S)$ converges to the optimal performance error $R^L_{\mu,\mathcal{H}}$. Without further condition on $\mathcal{H}$ and L the ERM principle does not work.

A generative model of supervised learning is a discriminative model $(\mathcal{X}, \mathcal{P}(\mathcal{Y}), \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})})$ whose training data are of the form $\{(x_1, \delta_{y_1}), \cdots, (x_n, \delta_{y_n})\}$. Discriminative models can be regarded as "limit" cases of generative models.

Finally I would like to remark that there is another popular type of generative models, called Bayesian models, which we shall learn at the end of our course.

3. Mathematical models for unsupervised learning

Last week we discussed models $(\mathcal{X}, \mathcal{Y}, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})})$ of supervised learning where $\mathcal{H} \subset Prob(\mathcal{X}, \mathcal{Y}) := Meas(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ is a subset of probabilistic mappings from $\mathcal{X}$ to $\mathcal{Y}$. The error function $L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \times \mathcal{P}_{\mathcal{X}\times\mathcal{P}(\mathcal{Y})} \to \mathbb{R}$ is a central notion of learning theory that specifies the idea of "best approximation", "best predictor". [12] In the supervised learning, the risk function R^L can be derived from an instantaneous loss function L that measures discrepancy between y and $h(x)$.

Today we shall study models of machine learning for several important tasks in unsupervised learning: density estimation, clustering, dimension reduction and manifold learning. We shall focus on the choice of error functions that measures the accuracy of an estimator or the fitness of a decision that should provides us minimizing sequences that converges fast to the true unknown probability measure. Since the training data are unlabelled, we have to figure out what error an instantaneous loss function L measure.

3.1. **Mathematical models for density estimation.** Recall that for a measurable space $\mathcal{X}$ we denote by $\mathcal{P}(\mathcal{X})$ the space of all probability measures on $\mathcal{X}$. In classical (frequentist) density estimation problem we are given a sequence of observables $S_n = (x_1, \cdots, x_n) \in \mathcal{X}^n$, which are i.i.d. by

[12] The notion of a loss function has been introduced by A. Wald in his statistical decision theory [Wald1950].

un unknown probability measure μ_u. We have to estimate the measure $\mu_u \in \mathcal{P}(\mathcal{X})$, using S_n. Usually, having a prior knowledge, we assume that μ_u belongs to a statistical model $P_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$. Simplifying further, we assume that $P_\mathcal{X}$ consists of probability measures that are dominated by a measure $\mu_0 \in \mathcal{P}(\mathcal{X})$. Thus we regard $P_\mathcal{X}$ as a family of density functions on $\mathcal{X}$. Usually, we assume that $P_\mathcal{X}$ is parameterized by a "nice" parameter set Θ: $\mathcal{P}_\mathcal{X} = \mathbf{p}(\Theta)$, e.g. Θ is an open subset in $\mathbb{R}^n$, where $\mathbf{p}$ is a parameterization.

Remark 3.1. The initial object of statistical investigations is a sample

$$Z_n = (z_1, \cdots, z_n) \in \mathcal{Z}^n$$

where $\mathcal{Z}$ is a sample space. We wish to find the stochastic relation between z and its feature $y \in \mathcal{Y}$, knowing Z_n [13]. Since the stochastic relation between $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ can be expressed in terms of a probability measure on $\mathcal{Z} \times \mathcal{Y}$, finding a stochastic relation between z and y could be reduced to a density estimation problem. In a general feature estimation problem, we regard a feature y of z as a function of the distribution $P_u \in \mathcal{P}_\mathcal{X}$ on a generative model $\mathcal{X}$ that is attached to our initial statistical investigation of stochastic relation between a sample $z \in \mathcal{Z}$ and its feature $y \in \mathcal{Y}$. Then our initial problem can be reformulated as a φ-density estimation problem: to find a map $\varphi \circ \hat{\sigma}_n : \mathcal{X}^n \stackrel{\hat{\sigma}_n}{\to} \mathcal{P}_\mathcal{X} \stackrel{\varphi}{\to} \mathcal{Y}$, see [Le2020, Definition 8] (Definition **??** below), where $\hat{\sigma}_n : \mathcal{X}^n \to \mathcal{P}_\mathcal{X}$ is a (nonparametric) estimator and $\mathcal{Y}$ is assumed to be a topological vector space.

Example 3.2. (1) In supervised learning with an input space $\mathcal{X}$ and a label space $\mathcal{Y}$ we are interested in the stochastic relation between $x \in \mathcal{X}$ and its label $y \in \mathcal{Y}$, which is expressed via a measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ that governs the distribution of labelled pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Finding μ is a density estimation problem, assuming that we are given a sequence of i.i.d. labelled pairs $\{(x_1, y_1), \cdots, (x_n, y_n)\}$. In practice, we are interested only in knowing the conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$, which is regular under very general assumptions [Faden1985]. Then finding the conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is equivalent to finding a measurable mapping $\overline{T} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, or equivalently, a probabilistic morphism $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Usually $\mathcal{Y}$ is represented as a subset in $\mathbb{R}^n$ and the knowledge of $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is often not required, it is sufficient to determine one of its characteristics, for example the regression function

$$r(x) = \int_\mathcal{Y} y d\mu_{\mathcal{Y}|\mathcal{X}}(y|x).$$

In this case, the map $\varphi : \mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to Map(\mathcal{X}, \mathbb{R})$ is defined as the composition of the mappings defined above

$$\mathcal{P}(\mathcal{X} \times \mathcal{Y}) \to \mathrm{Probm}(\mathcal{X}, \mathcal{Y}) \to Map(\mathcal{X}, \mathbb{R}),$$

[13] This setting also works for the density estimation problem, since $\mathcal{P}(\mathcal{X})$ has the natural σ-algebra Σ_w, see Appendix D

where $\mathrm{Probm}(\mathcal{X}, \mathcal{Y})$ denotes the space of probabilistic morphisms from $\mathcal{X}$ to $\mathcal{Y}$.

(2) A classical example of a φ-map is the moment of a probability measure in a 1-dimensional statistical model $\mathbf{p}(\Theta)$, where Θ is an interval or the real line [Borovkov1998, p. 55]. Given a real function $g(x)$ we define

$$\varphi(\mathbf{p}(\theta)) := \int g(x) d\mathbf{p}(\theta).$$

Under a certain condition this map is 1-1.

3.1.1. *Classical parametric density estimation problems.* We assume further that the image $\mathbf{p}(\Theta) = \mathcal{P}_{\mathcal{X}}$ is a dominated measure family, i.e., there exists a measure $\mu_0 \in \mathcal{P}(\mathcal{X})$ such that $\mathbf{p}(\theta) = p_\theta \mu_0$ for all $\theta \in \Theta$, where $p_\theta \in L^1(\mathcal{X}, \mu_0)$. Thus the probability measure $\mathbf{p}(\theta)$ is specified by its density p_θ.

• In this lecture we shall assume that Θ is an open subset of $\mathbb{R}^n$ and $\mathbf{p}$ is a continuous 1-1 map. This condition always holds in classical theory of density estimations. Thus we shall identify $P_{\mathcal{X}}$ with Θ and the parametric density estimation in this case is equivalent to estimation of the parameter $\theta \in \Theta$, assuming knowledge of S_n.

As in mathematical models for supervised learning, we define an expected risk function $R_\mu : \Theta \to \mathbb{R}$ by averaging an *instantaneous loss function* [14]

$$L : \mathcal{X} \times \Theta \to \mathbb{R}.$$

Usually the instantaneous loss function L is given by the minus log-likelihood function [15]

$$L(x, \theta) = -\log p_\theta(x). \tag{3.1}$$

Given the instantaneous loss log-likelihood function L in (3.1), the *expected risk function*

$$R^L_\mu : \Theta \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$$

is *the expected log-likelihood function*, cf. (2.2)

$$R^L(\theta, \mu_u) := R^L_{\mu_u}(\theta) = -\int_{\mathcal{X}} \log p_\theta(x) p_u(x) d\mu_0 \tag{3.2}$$

where $\mu_u = p_u \cdot \mu_0$, which we have to estimate. Given a data $S_n = (x_1, \cdots, x_n) \in \mathcal{X}^n$, by (3.1), the corresponding empirical risk function is

$$\hat{R}^L_{S_n}(\theta) := R^L_{\mu_{S_n}}(\theta) = -\sum_{i=1}^n \log p_\theta(x_i) = -\log[p^n_\theta(S_n)], \tag{3.3}$$

[14] note that in this example the instantaneous loss function depends only on two variables $x \in \mathcal{X}$, $\theta \in \Theta$.

[15] Note that the log-likelihood function $\log p_\theta(x)$ does not depend on the choice of a dominating measure μ_0, i.e., if we replace μ_0 by μ_0' that also dominates $\mathbf{p}(\theta)$ then $\log \frac{d\mathbf{p}(\theta)}{d\mu_0} = \log \frac{d\mathbf{p}(\theta)}{d\mu_0'}$.

where $p^n_\theta(S_n)$ is the density of the probability measure μ^n_θ on $\mathcal{X}^n$. It follows that the minimizer θ of the empirical risk $\hat{R}^L_{S_n} = R^L_{\mu_{S_n}}$ is the maximizer of the log-likelihood function $\log[p^n_\theta(S_n)]$, cf. (2.5).

According to ERM principle, which is inspired by the weak law of large numbers, the minimizer θ of $\hat{R}^L_{S_n}$ should provide an "approximation" of the density p_u of the unknown probability measure μ_u that governs the distribution of i.i.d $x_i \in \mathcal{X}$.

Remark 3.3. (i) The instantaneous loss function $L(x, y, h)$ in a supervised learning measures the discrepancy between the value of a prediction $h(x)$ and a possible feature y, see (2.3).

(ii) In density estimation problem it is harder to figure out a correct choice of an instantaneous loss function $L : \mathcal{X} \times \Theta \to \mathbb{R}$, or more generally, an instantaneous function $L : \mathcal{X} \times \Theta \times \mathbf{p}(\Theta)$ as in the next example of non-parametric estimation, see (3.12). (In Theorem 5.17 we shall give a justification of the instantaneous loss function given by the minus log-likelihood function (3.1) using the notion of efficient estimator and the Cramér-Rao inequality).

For $\mathcal{X} = \mathbb{R}$ and $\mu_0 = dx$ being the Lebesgue measure, we have another justification of the choice of the risk function being the expected log-likelihood function [Vapnik1998, p. 30]. Namely this risk function has the following nice properties.

(1) The minimum of the risk function $R^L_{\mu_u} : \Theta \to \mathbb{R}$ (if exists) is attained at θ^* such that the density function $p(\theta^*)$ may differ from p_u at a μ_0-zero set.
(2) The risk function $R^L_{\mu_u}$ satisfies the Bretagnolle-Huber inequality

$$\int_{\mathcal{X}} |p(x,\theta) - p_u(x)|dx \leq 2\sqrt{1 - \exp(R^L_{\mu_u}(\theta) - R^L_{\mu_u}(\theta_u))}.$$

(iii) Note that minimizing the expected log-likelihood function $R_\mu(\theta)$ is the same as minimizing the following modified risk function [Vapnik2000, p.32]

$$R^*_{\mu_u}(\theta) := R^L_{\mu_u}(\theta) + \int_{\mathcal{X}} \log p_u(x) p_u(x) d\mu_0 = -\int_{\mathcal{X}} \log \frac{p_\theta(x)}{p_u(x)} p_u(x) d\mu_0. \tag{3.4}$$

The expression on the RHS of (3.4) is the Kullback-Leibler divergence $KL(p_\theta\mu_0|\mu_u)$ that is used in statistics for measuring the divergence between $p_\theta\mu_0$ and $\mu_u = p_u\mu_0$. The Kullback-Leibler divergence $KL(\mu|\mu')$ is defined for probability measures $(\mu, \mu') \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $\mu \ll \mu'$, see also Remark 5.9 below. It is a *divergence*, i.e., it satisfies the following properties:

$$KL(\mu|\mu') \geq 0 \text{ and } KL(\mu|\mu') = 0 \text{ iff } \mu = \mu'. \tag{3.5}$$

By (3.4), a maximizer of the expected risk function $R^L_{\mu_u}(\mu_\theta)$ on Θ minimizes the KL-divergence $KL(\mu_\theta|\mu_u)$ regarded as a function on Θ. The relations in (3.5) justify the choice of the expected risk function $R^L_{\mu_u}$.

(iv) It is important to find natural distances as well as quasi-distance functions on $\mathcal{P}(\mathcal{X})$ that satisfying certain natural statistical requirement. This problem has been considered in information geometry, see [Amari2016, AJLS2017] for further reading.

(v) If the parameter set Θ is large, e.g., $\Theta = \mathcal{P}_{\mathcal{X}} = \mathcal{P}(\mathcal{X})$ where $\dim(\mathcal{X}) \geq 1$, then $\mathcal{P}_{\mathcal{X}}$ cannot be dominated by a probability measure $\mu_0 \in \mathcal{P}(\mathcal{X})$. In this case we have to consider *nonparametric* techniques.

3.1.2. *Nonparametric density estimation technique KDE.* For estimating density functions on $\mathcal{X} = \mathbb{R}^m$ using empirical data $S_n \in \mathcal{X}^n$ we use a popular technique called *the kernel density estimation* (KDE) cf. [Tsybakov2009, p. 2]. For understanding the idea of KDE we shall consider only the case $\mathcal{X} = \mathbb{R}$ and $\mu_0 = dx$. We need to estimate the unknown measure $\mu_u := p_u dx \in \mathcal{P}(\mathbb{R})$. Let

$$F_{\mu_u}(t) = \int_{-\infty}^{t} p_u(x)dx = \mu_u((-\infty, t))$$

be the corresponding *cumulative distribution function*. Then we have

$$p_u(t) = \frac{F_{\mu_u}(t+h) - F_{\mu_u}(t-h)}{2h} + O(h). \tag{3.6}$$

Since μ_u governs the distribution of $S_n = (x_1, \cdots, x_n) \in \mathcal{X}^n$, we replace μ_u in the RHS of (3.6) by μ_{S_n} and obtain *the Rosenblatt estimator*

$$\hat{p}^R_{S_n}(h;t) := \frac{\hat{F}_{S_n}(t+h) - \hat{F}_{S_n}(t-h)}{2h}, \tag{3.7}$$

where $\hat{F}_{S_n}$ is *the empirical distribution function* and h is the *bandwidth* of the estimator.

$$\hat{F}_{S_n}(t) = F_{\mu_{S_n}}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(x_i \leq t)}.$$

Furthermore, the LHS of (3.7) can be rewritten in the following form

$$\hat{p}^R_{S_n}(h;t) = \frac{1}{2nh}\sum_{i=1}^{n} 1_{(t-h \leq x_i \leq t+h)} = \frac{1}{nh}\sum_{i=1}^{n} K_0(\frac{x_i - t}{h}) \tag{3.8}$$

where $K_0(u) := \frac{1}{2}1_{(-1 \leq u \leq 1)}$. A simple generalization of the Rosenblatt estimator is given by

$$\hat{p}^{PR}_{S_n}(h;t) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x_i - t}{h}) \tag{3.9}$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)du = 1$. Such a function K is called *kernel* and the parameter h is called *the bandwidth* of the *kernel density estimator* (3.9), also called *the Parzen-Rosenblatt estimator*. Popular kernels are the rectangular kernel $K(u) = \frac{1}{2}I_{(|u| \leq 1)}$, the Gaussian kernel $K(u) = (2\pi)^{-1/2}\exp(u^2/2)$ etc., see e.g., [Tsybakov2009, p. 3].

To measure the accuracy of the estimator $\hat{p}^{PR}_{S_n}(h;\cdot)$ we first need to define an expected loss function R^L on the hypothesis class $\mathcal{H}$ of possible densities of probability measures we are interested in. In the given case we choose $\mathcal{H}$ to be defined as follows

$$\mathcal{H} = \mathcal{P}(\beta, L) := \{p \in \mathbb{R}^{\mathbb{R}}_{\geq 0} \cap \Sigma_{\mathbb{R}}(\beta, L) | \int_{\mathbb{R}} p(x)\, dx = 1\} \tag{3.10}$$

where $\Sigma_{\mathbb{R}}(\beta, L)$ is the Hölder class on $\mathbb{R}$ of the set of $l = \lfloor \beta \rfloor$ [16] times differentiable functions $f : \mathbb{R} \to \mathbb{R}$ whose derivative $f^{(l)}$ satisfies

$$|f^{(l)}(x) - f^{(l)}(x')| \leq L|x - x'|^{\beta - l},\ \forall\, x, x' \in \mathbb{R}. \tag{3.11}$$

Let $\mathcal{P}_{\mathbb{R}} \subset \mathcal{P}(\mathbb{R})$ be a statistical model of possible distributions $\mu' = p'dx$ that contains our unknown distribution $\mu_u = p_u dx$. By definition $\mathcal{P}_{\mathbb{R}} = \mathcal{H}$. A natural candidate for an *instantaneous loss function*

$$L : \mathbb{R} \times \mathcal{H} \times \mathcal{P}_{\mathbb{R}} \to \mathbb{R} \tag{3.12}$$

is the square distance function (cf. (2.17))

$$L(x, p, p') := |p(x) - p'(x)|^2. \tag{3.13}$$

This induces the *expected loss at $x_0 \in \mathbb{R}$ function* as follows

$$R^L := R^L_{|x_0} : \mathcal{H} \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}, (p, p') \mapsto |p(x_0) - p'(x_0)|^2. \tag{3.14}$$

Let

$$\mathcal{A}_n \subset \mathcal{H}^{\mathbb{R}^n}$$

be the set of possible estimators $\hat{p}_n : \mathbb{R}^n \to \mathcal{H}$, $S_n \mapsto \hat{p}_{S_n} \in \mathcal{H}$. Then we define *the expected accuracy at x_0 function*

$$MR^L_{|x_0} : \mathcal{A}_n \to \mathbb{R}$$

by

$$MR^L_{|x_0}(\hat{p}_n) := MSE_{|x_0}(\hat{p}_n) := \mathbb{E}_{\mu^n_u}(R^L_{|x_0}(\hat{p}_n(S_n), p_u)), \tag{3.15}$$

where $\mu_u = p_u dx$ and S_n is distributed by μ^n_u.

For instance, for the Parzen-Rosenblatt estimator $\hat{p}^{PR}_n(h;), S_n \mapsto \hat{\rho}^{PR}_{S_n}(h;)$, we have

$$MSE_{x_0}(\hat{p}^{PR}_n(h;)) = \mathbb{E}_{\mu^n_u}[(\hat{p}^{PR}_{S_n}(h;)(x_0) - p_u(x_0))^2]. \tag{3.16}$$

Note that the RHS of (3.16) measures the expected accuracy at x_0 of the estimator $\hat{p}^{PR}_{S_n}(x_0)$ *averaging over* $S_n \in \mathbb{R}^n$. This is an important concept of the accuracy of an estimator in the presence of uncertainty.

If the kernel function K used in the kernel density estimator $\hat{p}^{PR}_{S_n}(h :)$ is square integrable, and assuming the infinite dimensional statistical model $P = \mathcal{H}$ of density functions is given in (3.11), the mean expected risk $MSE(\hat{p}^{PR}_{S_n}(h;)_{x_0}$, regarded as a function of x_0, converges to zero uniformly

[16] $\lfloor \beta \rfloor$ denotes the greatest integer strictly less than $\beta \in \mathbb{R}$

on $\mathbb{R}$ as n goes to infinity and h goes to zero. Namely, letting $h = n^{-\frac{1}{2\beta+1}}$, we have[Tsybakov2009, Theorem 1.1, p. 9]

$$\sup_{x_0\in\mathbb{R}}\sup_{p\in\mathcal{H}}\mathbb{E}_{(pdx)^n}[\hat{p}^{PR}_{S_n}(x_0) - p(x_0)^2] \leq C(K,\mathcal{H}) \leq n^{-\frac{2\beta}{2\beta+1}}, \tag{3.17}$$

where the expectation is taken over the space $\mathbb{R}^n \ni S_n$ and $C(K,\mathcal{H})$ is a constant depending only on K, L and β. Thus our estimator converges in probability to the true density function point-wise.

Remark 3.4. We have considered two models for density estimation using different instantaneous functions to measure the accuracy of our estimators in terms of an *expected loss function*. In the first model of density estimation $(\mathcal{X}, \mathcal{H} = \Theta, L : \mathcal{X}\times\mathcal{H} \to \mathbb{R}, \mathcal{P}_{\mathcal{X}} = \mathcal{H})$ we let $L(\theta, x) := \log p_\theta(x)$. Similarly, as in the mathematical model of supervised learning, the ERM says that a minimizer of the empirical risk function $R^L_{\mu_{S_n}}(\theta) = -\log[p^n_\theta(S_n)]$ should be the *best estimator* i.e., it should minimize the expected loss function $R^L_{\mu_u} : \Theta \to \mathbb{R}$. Note that a minimizer of $R^L_{\mu_{S_n}}$ is the MLE estimator: θ must maximize the value $\log p_{\theta^n}(S_n)$ given a data $S_n \in \mathcal{X}^n$.

In the second model for density estimation $(\mathcal{X} = \mathbb{R}^m, \mathcal{H}, L, \mathcal{P}_{\mathcal{X}})$ an instantaneous function $L : \mathcal{X}\times\mathcal{H}\times\mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ depends also on the unknown probability measure $\mu_u \in \mathcal{P}_{\mathcal{X}}$, namely as a distance between the value of the estimator and the true density, see (3.12). Using L we define a new concept of *the expected accuracy function* $MR^L_{|x_0}(\hat{\rho}_n)$ at a point x_0 of an estimator $\hat{\rho}_n : \mathcal{X}^n \to \mathcal{H}$ that measures the accuracy of $\hat{\rho}_n$ at a given point x_0, averaging over observable data $S_n \in \mathcal{X}^n$ with the help of the unknown distribution generating our sample data on $\mathcal{X}$ and hence on $\mathcal{X}^n$. In the second model we don't discuss how to find a "best" estimator $\hat{\rho}_n : \mathcal{X}^n \to \mathcal{H}$, we find the Parzen-Rosenblatt estimator by heuristic argument. The notion of a mean expected loss function is an important concept in theory of generalization ability of machine learning that gives an answer to Question 1.5: how to quantify the difficulty/complexity of a learning problem.

3.2. **Mathematical models for clustering.** Given an input measurable space $\mathcal{X}$, clustering is the process of grouping similar objects $x \in \mathcal{X}$ together. There are two possible types of grouping in machine learning: *a partitional clustering*, where we partition the objects into disjoint sets; and *a hierarchical clustering*, where we create a nested tree of partitions. We propose the following unified notion of a clustering.

Let Ω_k denote a finite sample space of k elements $\omega_1, \cdots, \omega_k$.

Definition 3.5. A *clustering of* $\mathcal{X}$ is a measurable mapping $\chi_k : \mathcal{X} \to \Omega_k$. A clustering $\chi_k : \mathcal{X} \to \Omega_k$ is called *hierarchical*, if there exists a number $l \leq k-1$ such that $\chi_k = \chi_{lk} \circ \chi_l : \mathcal{X} \overset{\chi_l}{\to} \Omega_l \overset{\chi_{lk}}{\to} \Omega_k$.

A *probabilistic clustering of* $\mathcal{X}$ is a probabilistic morphism $\chi_k : \mathcal{X} \rightsquigarrow \Omega_k$, i.e. a measurable mapping $\overline{\chi_k} : \mathcal{X} \to \mathcal{P}(\Omega_k)$. A probabilistic clustering

$\chi_k : \mathcal{X} \rightsquigarrow \Omega_k$ is called *hierarchical*, if there exists a number $l \leq k-1$ such that $\chi_k = \chi_{lk} \circ \chi_l : \mathcal{X} \overset{\chi_l}{\rightsquigarrow} \Omega_l \overset{\chi_{lk}}{\rightsquigarrow} \Omega_k$.

Now we can define a model of clustering similarly as in the case of density estimations. To formalize the notion of similarity we need a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, i.e. d is nonnegative, symmetric, satisfying triangle inequality, and $d(x,y) = 0$ iff $x = y$. Then we pick up a set $\mathcal{C}$ of possible clusterings, i.e., $\mathcal{C}$ is a subset of all measurable/probabilistic morphisms $\cup_{i=1}^{\infty} \chi_k : \mathcal{X} \rightsquigarrow \Omega_k$ and the goal of a clustering algorithm is to find a clustering $\chi_k \in \mathcal{C}$ of minimal cost

$$G : (\mathcal{X}, d) \times \mathcal{C} \to \mathbb{R}. \tag{3.18}$$

The cost function G is also called *the objective function.*

Example 3.6. The *k-means objective function* is one of the most popular clustering objectives. In k-means, given a clustering $\chi_k : \mathcal{X} \to \Omega_k = \{\omega_1, \cdots, \omega_k\}$, the data is partitioned into disjoint sets $C_1 = \chi_k^{-1}(\omega_1), \cdots, C_k = \chi_k^{-1}(\omega_k)$ and each C_i is represented by its centroid $c_i := c(C_i) \in (\mathcal{X}', d')$, where $\mathcal{X} \subset \mathcal{X}'$ and $d = d'_{|\mathcal{X}}$. The centroid c_i is defined as the point of $\mathcal{X}'$ that minimizes the averaging distance to points in $\mathcal{C}_i$ using a measure μ, i.e.:

$$c_i := \arg\min_{x' \in \mathcal{X}'} \int_{C_i} d(x, x') d\mu(x).$$

Note that we assume that μ is known and it should govern the distribution of x in $\mathcal{X}$.

Then we define the k-means objective function G_k as follows

$$G_k((\mathcal{X}, d), \chi_k) := \sum_{i=1}^{k} \int_{C_i} d(x, c_i) d\mu. \tag{3.19}$$

The k-means objective function G is used in digital communication tasks, where the members of $\mathcal{X}$ may be viewed as a collection of signals that have to be transmitted. For further reading on k-means algorithms, see [SSBD2014, p. 313], [HTF2008, §14.3.6, p. 505].

Remark 3.7. (1) The above formulation of clustering is deterministic. For probabilistic clustering, where the output is a measurable mapping $\chi_k : \mathcal{X} \to \mathcal{P}(\Omega_k)$ i.e., $\chi_k(x) = (p_1(x), \cdots, p_k(x)) \in \mathcal{P}(\Omega_k)$, we modify the cost function G_k as follows

$$G_k((\mathcal{X}, d), \chi_k) = \sum_{i=1}^{k} \int_{C_i} d(x, c_i) p_i(x) d\mu \tag{3.20}$$

where μ is a probability measure on $\mathcal{X}$.

(2) The main difference between a clustering problem and a density estimation problem (resp. an unsupervised learning problem) is that in the first one the distribution of elements on $\mathcal{X}$ is known. Thus clustering algorithms are essential classical optimization algorithms. As it turns out, most of the

resulting optimization problems are NP-hard, and some are even NP-hard to approximate.

(3) The hardest part in clustering is to find a suitable function d, and it should depend on concrete class of clustering problems. There is no "universal" rule for clustering algorithm [CM2010]. At the present, clustering problem has not formulated as a problem in statistical learning theory, where we can define the notion of learning success.

3.3. **Mathematical models for dimension reduction and manifold learning.** In machine learning we represent an input space $\mathcal{X}$ as a subset in a Euclidean space $\mathbb{R}^d$. We call d *the dimension of the input space* $\mathcal{X}$, or also as accepted in the ML community, *the dimension of input* x. As the dimension of input x grows, any learning problem significantly gets harder and harder. Handling high-dimensional data is cumbersome in practice, which is often referred to as *the curse of dimensionality*. Hence various methods of dimensionality reduction are introduced. Dimension reduction is the process of taking data in a high dimensional space and mapping it into a new space whose dimension is much smaller.

• *Classical (linear) dimension reduction methods.* Given original data $S_m := \{x_i \in \mathbb{R}^d | \, i \in [1, m]\}$ we want to embed it into $\mathbb{R}^n$, $n < d$. A simple way to do it is to find a linear transformation $W \in Hom(\mathbb{R}^d, \mathbb{R}^n)$ such that

$$W(S_m) := \{W(x_i) | \, x_i \in S_m\} \subset \mathbb{R}^n.$$

To find the "best" transformation $W = W_{(S_m)}$, we define an error function on the space $Hom(\mathbb{R}^d, \mathbb{R}^n)$ and solve the associated optimization problem.

Example 3.8. A popular linear method for dimension reduction is called *Principal Component Analysis* (PCA). Given $S_m \subset \mathbb{R}^d$, we use a linear transformation $W \in Hom(\mathbb{R}^d, \mathbb{R}^n)$, where $n < d$, to embed S_m into $\mathbb{R}^n$. Then, a second linear transformation $U \in Hom(\mathbb{R}^n, \mathbb{R}^d)$ can be used to (approximately) recover S_m from its compression $W(S_m)$. In PCA, we search for W and U to be a minimizer of the following *reconstruction error* function $R_{S_m} : Hom(\mathbb{R}^d, \mathbb{R}^n) \times Hom(\mathbb{R}^n, \mathbb{R}^d) \to \mathbb{R}$

$$\hat{R}_{S_m}(W, U) = \sum_{i=1}^{m} ||x_i - UW(x_i)||^2 \tag{3.21}$$

where $||, ||$ denotes the quadratic norm on the Euclidean space $\mathbb{R}^d$.

Exercise 3.9. ([SSBD2014, Lemma 23.1, p.324]) Let (W, U) be a minimizer of $\hat{R}_{S_m}$ defined in (3.21). Show that U can be chosen as an orthogonal embedding and $W \circ U = Id_{|\mathbb{R}^n}$.

Hint. First we show that if a solution (W, U) of (3.21) exists, then there is a solution (W', U') of (3.21) such that $\dim \ker(U'W') = d - n$.

Let $\mathrm{Proj}_g(\mathbb{R}^d, \mathbb{R}^n)$ denotes the set of all orthogonal projections from $\mathbb{R}^d$ to $\mathbb{R}^n$ and $\mathrm{Emb}_g(\mathbb{R}^n, \mathbb{R}^d)$ the set of all isometric embeddings from $\mathbb{R}^m$ to $\mathbb{R}^d$.

Let $\mathcal{F} \subset \mathrm{Emb}_g(\mathbb{R}^d, \mathbb{R}^n) \times \mathrm{Proj}_g(\mathbb{R}^n, \mathbb{R}^d)$ be the subset of all pairs (W, U) of transformations such that $W \circ U = Id_{|\mathbb{R}^n}$. Exercise 3.9 implies that any minimizer (W, U) of $\hat{R}_{S_m}$ is an element of $\mathcal{F}$.

Exercise 3.10. ([SSBD2014, Theorem 3.23, p. 325]) Let $C(S_m) \in \mathrm{End}(\mathbb{R}^d)$ be defined as follows

$$C(S_m)(v) := \sum_{i=1}^{m} \langle x_i, v \rangle x_i.$$

Assume that $\xi_1, \cdots, \xi_d \in \mathbb{R}^d$ are eigenvectors of $C(S_m)$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Show that any $(W, U) \in \mathcal{F}$ with $W(x_j) = 0$ for all $j \geq m + 1$ is a solution of (3.21).

Thus a PCA problem can be solved using linear algebra method.

• *Manifold learning and autoencoder.* In real life, data are not concentrated on a linear subspace of $\mathbb{R}^d$, which implies that we could use PCA to reduce the dimension of the data, but around a submanifold $M \subset \mathbb{R}^d$, where the PCA method does not apply. The current challenge in ML community is that to reduce representation of data in $\mathbb{R}^d$ not using all the data in $\mathbb{R}^d$ but only use only data concentrated around M. For that purpose we use autoencoder, which is a non-linear analogue of PCA.

In an auto-encoder we learn a pair of functions: an *encoder function* $\psi : \mathbb{R}^d \to \mathbb{R}^n$, and a *decoder function* $\varphi : \mathbb{R}^n \to \mathbb{R}^d$. The goal of the learning process is to P find a pair of functions (ψ, φ) such that *the reconstruction error*

$$R_{S_m}(\psi, \varphi) := \sum_{i=1}^{m} ||x_i - \varphi(\psi(x_i))||^2$$

is small. We therefore must restrict ψ and φ in some way. In PCA, we constrain $k < d$ and further restrict ψ and φ to be linear functions.

Remark 3.11. (1) Modern autoencoders have generalized the idea of an encoder and a decoder beyond deterministic functions to *probabilistic morphisms* $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$. For further reading I recommend [SSBD2014], [GBC2016] and [Bishop2006, §12.2, p.570].

(2) At the current state, dimension reduction as well as manifold learning are deterministic optimization problems with a large number of variables and these problems are not yet formulated as inductive learning problems.

3.4. **Conclusion.** A mathematical model for supervised learning and unsupervised (inductive) learning (density estimation problem) consists of a quadruple $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ where

• $\mathcal{Z}$ is a measurable space of inputs z whose features we want to learn. For example $\mathcal{Z} = \mathcal{X} \times \mathcal{P}(\mathcal{Y})$ in classification and regression problems,

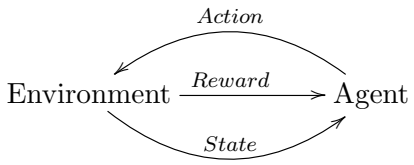
• $\mathcal{H}$ is a decision space containing all possible decisions we have to find based on a sequence of observable $(z_1, \cdots, z_n) \in \mathcal{Z}^n$. For example, $\mathcal{H}$ is a hypothesis space in the space $Meas(\mathcal{X}, \mathcal{Y})$ in the discriminative model for

supervised learning.
• $\mathcal{P}_{\mathcal{Z}} \subset \mathcal{P}(\mathcal{Z})$ is a *statistical model* that containing all possible distribution μ_u that governs the distribution of i.i.d. sequence $z_1, \cdots, z_n$.
• An *instantaneous loss function* $L : \mathcal{Z} \times \mathcal{H} \times \mathcal{P}_{\mathcal{Z}} \to \mathbb{R}_{\geq 0}$ which generates the *expected loss function* $R^L : \mathcal{H} \times \mathcal{P}_{\mathcal{Z}} \to \mathbb{R}_{\geq 0}$ by setting $R^L(h, \mu) = \mathbb{R}^L_\mu(h) := \mathbb{E}_\mu L(z, h, \mu)$, where μ is the unknown probability measure that governs the distribution of observable z_i, or the loss function evaluated at z_0: $R^L_{|z_0}(h, \mu) := \mathbb{E}_{\delta_{z_0}} L(z, h, \mu) = L(z_0, h, \mu)$. Let us denote by μ_0 for μ or δ_{z_0} used in the definition of the expected loss function in both the cases.
• A *learning rule* $\hat{\rho}_n : \mathcal{Z}^n \to \mathcal{H}$, which is usually chosen to be a minimizer of the empirical loss function $R^L_{\mu_{S_n}}$.
• An μ_0-*expected accuracy function* $L_{\mu_0}(\hat{\rho}_n) := \mathbb{E}_{\mu_0^n} R^L_{\mu_0} L(\hat{\rho}_n(S_n))$, where μ_0 is the unknown probability measure that governs the distribution of observables z_i and μ_0 is defined above.
• Supervised learning is distinguished from unsupervised learning in their aim of learning: supervised learning is interested only in a concrete answer: given an input $x \in \mathcal{X}$ what is probability that $y \in \mathcal{Y}$ is a feature of x, so the decision space $\mathcal{H}$ is a subspace of the space $Meas(\mathcal{X}, \mathcal{P}(\mathcal{Y}))$ of all probabilistic mappings from $\mathcal{X}$ to $\mathcal{Y}$. In unsupervised learning we are interested in a structure of $\mathcal{Z}$, e.g., the probability measure $\mu_{\mathcal{Z}}$ that governs the distribution of $z_i \in \mathcal{Z}$, or the homogeneity of $\mathcal{Z}$ in clustering problem, etc.
• To define a decision rule/ algorithm $A_n : \mathcal{Z}^n \to \mathcal{H}$ we usually use the ERM principle that minimize the expected loss $R^L_{\mu_{S_n}} : \mathcal{H} \to \mathbb{R}$ for each given data $S_n \in \mathcal{Z}^n$, or some heuristic method as in Parzen-Rosenblatt nonparametric estimation. To estimate the success of learning rule A_n is to estimate the accuracy of the algorithm A_n as n grows to infinity. As we are working with uncertainty, the accuracy of A_n should be computed by averaging over the $\mathcal{X}^n$ using the unknown probability measure μ_u.

4. Mathematical model for reinforcement learning

Today we shall discuss a mathematical model for reinforcement learning, in which a learner, also called *an agent*, has to learn to *interact with environment* optimally by *correlating the immediate reward*, given to the agent by the environment at every action of the agent, *with his past action* to maximize the *cumulative rewards*.

A typical example of a reinforcement learning is a car-driver, a chess player and all sequence decision maker who has to take a decision in an environment with uncertainty and whose decision at every step changes his environment. Furthermore, his goal is to maximize his cumulative rewards. This goal and the presence of uncertainty in his environment are the features of a (statistical, inductive) learning in Definition 1.1, if we shall interpret his cumulative rewards as a feature of knowledge.



4.1. **A setting of a reinforcement learning.** The knowledge the learner acquires during the course of reinforcement learning shall be measured by an accumulative reward, which the learner has to maximize by choosing an optimal decision, also is called an optimal *policy*.

To formalize the notion that the environment changes after the agent takes action, we define the notion of the *state* of an environment, and the change of the environment after taking action is formalized as a function of the action of the agent and the state before taking the action. Furthermore, actions are usually assumed in discrete time steps but can be continuously taken in time. For today lecture we shall assume that agent's actions are taken in discrete time steps, called *decision epochs*, in the set $E := \{0, \cdots, T\}$ of all decision epochs. This model can be straightforwardly generalized to a continuous-times where action are taken at arbitrary points in time.

(a) At each time $t \in E$, the agent observes a state s_t in the set S of possible states of environment, and selects an action a_t in the set A of all possible actions.

(b) The environment moves to a new state $s_{t+1} \in S$ with conditional probability $Pr(s_{t+1}|a_t, s_t)$ and an *immediate reward* r_t is given to the agent. [17]

(c) The *discount cumulative reward* along the trajectory $h(t) = [s_1, a_1, \cdots, s_t, a_t, s_{t+1}]$ [18] of states and actions of the agent is the discount sum of *immediate rewards* $r_t = r(s_t, a_t, s_{t+1})$.

(d) A *policy* of an agent at a state s is his choice of a course of the actions a based on state s and given trajectory of his past state/action.

(e) The *value* $V_\pi(s)$ *of a policy* π at a state s is the expected discount cumulative rewards averaging over all random paths starting at s and following π.

(f) The goal of the agent in a reinforcement learning is to choose the optimal policy with maximal value *at any state* s.

Remark 4.1. Mathematical model of reinforcement learning is a *Markov decision process* (MDP) since we assume the following *Markov property* in modeling reinforcement learning: the conditional probability distribution $Pr(s_{t+1}|a_t, s_t)$ for the next state depends only on the current state and action of the system; the reward $r_t(s_t, a_t, s_{t+1})$ depends only on the current state and the action and the next state of the system.

[17] so the reward function is deterministic.

[18] h for history

The main problem for an agent in an MDP environment is to determine the action to take at each state, i.e., to define a optimal policy that maximizes $V_\pi(s)$ for any s.

4.2. **Markov decision process.** A MDP consists of the following components:

- (S, μ) is a countable probability space of possible *states*,
- A is a countable set of possible *actions*,
- E the set of decision epochs $\{0, \cdots, T\}$. If E is finite, the MDP is said to have a *finite horizon*.
- A *policy* is a probabilistic morphism $\pi : S \rightsquigarrow A$,
- $\mathcal{R} \subset \mathbb{R}_{\geq 0}$ is a bounded measurable set of possible *rewards*,
- By (b), there is a function $\mathrm{s} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{S}$ such that $Pr(s_{t+1}|a_t, s_t) = \mathrm{s}(s_{t+1}, a_t, a_t)$. Similarly there is a function $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{R}$ such that $r_t = r(s_t, a_t, s_{t+1})$ is the immediate reward given to the agent at the state a_t that selects action a_t and moves to state s_{t+1}.
- A *path* is a sequence $h = [s_1, a_1, \cdots, s_T, a_T, s_{T+1}]$.
- A *policy* $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$ is defined uniquely by its value $\pi(a_t|s_1)$.
- Given policy π and a path h of states

$$h = [s_1(h), a_1(h), \cdots, s_T(h)],$$

the return of π along the path h is defined as

$$R(\pi, h) := \sum_{\tau=0}^{T-1} \gamma^\tau r(s_{1+\tau}(h), \pi(s_{1+\tau}(h))) \tag{4.1}$$

where $\gamma \in [0, 1)$ is called *discount factor* for future rewards.

- The agent's objective is to find a policy $\pi : S \rightsquigarrow A$ that maximizes his expected *discounted reward/return* $V_\pi(s)$ at state s, which is defined as follows

$$V_\pi(s) := \mathbb{E}_{\mu_s}[R(\pi, h)|s_1(h) = s],$$

where the expectation are over the countable space of all paths $h = [s_1, a_1, \cdots, s_T]$ beginning at the state $s_1 = s$ and following π. Namely

$$\mu_s(h) = \mu(s_1)\Pi_{t=1}^{T}\mathrm{s}(s_{t+1}|s_t, a_t)\pi(a_t|s_t). \tag{4.2}$$

4.3. **Existence and uniqueness of the optimal policy.** We assume that the MDP is finite, i.e., both S and A are finite sets, then there exists unique an optimal policy π such that the value $V_\pi(s)$ is maximal for all $s \in S$. This is a consequence of the following Bellman equation for $V_\pi(s)$.

$$V_\pi(s) = E[r(s, \pi(s))] + \gamma \sum_{s'} \mathrm{s}(s')V_\pi(s'). \tag{4.3}$$

Since $\gamma < 1$, Bellmann's equation has a unique solution [MRT2012, Theorem 14.1, p. 318].

(1) If the environment is known, i.e., if the probabilistic mapping π and function r are known, then finding an optimal policy is an *planning problem*, for which there are known several algorithms [MRT2012, §14.4., p. 319].

(2) If the environment is unknown, then there exists algorithm for finding the unique solution, using stochastic approximation, see [MRT2012, §14.5].

4.4. **Conclusion.** In reinforcement learning we have to consider the discount value of policy. Then the Bellmann equation implies the uniqueness of the solution, which can be solved effectively in an known environment and in an unknown environment. A good book on Reinforcement learning is the book by M Sugiyama, Introduction to Statistical Machine Learning, 2016 [Sugiyama2016].

5. Distances on statistical models, the Fisher metric and maximum likelihood estimator

For many problems in supervised learning and unsupervised learning that involve statistical models $\mathcal{P}_\mathcal{X}$, especially in density estimation problem, it is important to have a distance function on $\mathcal{P}_\mathcal{X}$, e.g. for a construction of a loss function in a nonparametric density estimation, or to endow $\mathcal{P}_\mathcal{X}$ with a metric structure.

In today lecture we shall survey some natural metrics on statistical models $\mathcal{P}_\mathcal{X}$ by regarding them as subsets in the Banach space $\mathcal{S}(\mathcal{X})$ with the total variation norm. In many cases $\mathcal{P}_\mathcal{X}$ can be endowed with Riemannian like metric, which we call the diffeological Fisher metric. We shall use the Fisher metric to justify the use of the popular maximum likelihood estimator (MLE), which is obtained from the ERM for the expected log-likelihood function. The popularity of MLE stems from its asymptotic accuracy, also called consistency, which holds under mild conditions. This type of asymptotic accuracy also holds for Parzen-Rosenblatt's estimator that we have learned.

We shall also clarify the relation between the Fisher metric and another choice of the error function for parametric density estimation - the Kullback-Leibler divergence, considered in Remark 3.3.

5.1. **The space of all probability measures and total variation norm.** We begin today lecture with our investigation of a natural geometry and topology of $\mathcal{P}(\mathcal{X})$ for an arbitrary measurable space $(\mathcal{X}, \Sigma)$.

Let us fix some notations. Recall that a signed finite measure μ on $\mathcal{X}$ is a function $\mu : \Sigma \to \mathbb{R}$ which satisfies all axioms of a measure except that μ needs not take non-negative value. Now we set

$$\mathcal{M}(\mathcal{X}) := \{\mu \ : \ \mu \text{ a finite measure on } \mathcal{X}\},$$

$$\mathcal{S}(\mathcal{X}) := \{\mu \ : \ \mu \text{ a signed finite measure on } \mathcal{X}\}.$$

It is known that $\mathcal{S}(\mathcal{X})$ is a Banach space whose norm is given by the total variation of a signed measure, defined as

$$\|\mu\|_{TV} \;:=\; \sup \sum_{i=1}^{n} |\mu(A_i)|$$

where the supremum is taken over all finite partitions $\mathcal{X} = A_1 \dot\cup \ldots \dot\cup A_n$ with disjoint sets $A_i \in \Sigma(\mathcal{X})$ (see e.g. [Halmos1950]). Here, the symbol $\dot\cup$ stands for the disjoint union of sets.

Let me describe the total variation norm using the Jordan decomposition theorem for signed measures, which is an analogue of the decomposition theorem for a measurable function. For a measurable function $\phi : \mathcal{X} \to [-\infty, \infty]$ we define $\phi_+ := \max(\phi, 0)$ and $\phi_- := \max(-\phi, 0)$, so that $\phi_\pm \geq 0$ are measurable with disjoint support, and

$$\phi = \phi_+ - \phi_- \qquad\qquad |\phi| = \phi_+ + \phi_-. \tag{5.1}$$

Similarly, by the *Jordan decomposition theorem*, each measure $\mu \in \mathcal{S}(\mathcal{X})$ can be decomposed uniquely as

$$\mu = \mu_+ - \mu_- \qquad \text{with } \mu_\pm \in \mathcal{M}(\mathcal{X}),\ \mu_+ \perp \mu_-. \tag{5.2}$$

That is, there is a *Hahn decomposition* $\mathcal{X} = \mathcal{X}_+ \dot\cup \mathcal{X}_-$ with $\mu_+(\mathcal{X}_-) = \mu_-(\mathcal{X}_+) = 0$ (in this case the measures μ_+ and μ_- are called *mutually singular*). Thus, if we define

$$|\mu| := \mu_+ + \mu_- \in \mathcal{M}(\mathcal{X}),$$

then (5.2) implies

$$|\mu(A)| \leq |\mu|(A) \qquad \text{for all } \mu \in \mathcal{S}(\mathcal{X}) \text{ and } A \in \Sigma(\mathcal{X}), \tag{5.3}$$

so that

$$\|\mu\|_{TV} = \| \, |\mu| \, \|_{TV} = |\mu|(\mathcal{X}).$$

In particular,

$$\mathcal{P}(\mathcal{X}) = \{\mu \in \mathcal{M}(\mathcal{X}) \;:\; \|\mu\|_{TV} = 1\}.$$

Given a measure $\mu_0 \in \mathcal{M}(\mathcal{X})$, we let

$$\mathcal{S}(\mathcal{X}, \mu_0) := \{\mu \in \mathcal{S}(\mathcal{X}) \;:\; \mu \text{ is dominated by } \mu_0\}.$$

By the Radon-Nikodym theorem, we may canonically identify $\mathcal{S}(\mathcal{X}, \mu_0)$ with $L^1(\mathcal{X}, \mu_0)$ by the correspondence

$$\imath_{can} : L^1(\mathcal{X}, \mu_0) \longrightarrow \mathcal{S}(\mathcal{X}, \mu_0), \qquad \phi \longmapsto \phi \, \mu_0. \tag{5.4}$$

Observe that $\imath_{can}$ is an isomorphism of Banach spaces, since evidently

$$\|\phi\|_{L^1(\mathcal{X}, \mu_0)} = \int_{\mathcal{X}} |\phi| \, d\mu_0 = \|\phi \, \mu_0\|_{TV}.$$

Example 5.1. Let $\mathcal{X}_n := \{\omega_1, \cdots, \omega_n\}$ be a finite set of n elementary events. Let δ_{ω_i} denote the Dirac measure concentrated at ω_i. Then

$$\mathcal{S}(\mathcal{X}_n) = \{\mu = \sum_{i=1}^n x_i\delta_{\omega_i} |\, x_i \in \mathbb{R}\} = \mathbb{R}^n(x_1, \cdots, x_n)$$

and

$$\mathcal{M}(\mathcal{X}_n) = \{\sum_{i=1}^n x_i\delta_{\omega_i} |\, x_i \in \mathbb{R}_{\geq 0}\} = \mathbb{R}^n_{\geq 0}.$$

For $\mu \in \mathcal{M}(\mathcal{X}_n)$ of the form

$$\mu = \sum_{i=1}^k c_i\delta_i,\ c_i > 0$$

we have $||\mu||_{TV} = \sum c_i$. Thus the space $L^1(\mathcal{X}_n, \mu)$ with the total variation norm is isomorphic to $\mathbb{R}^k$ with the l^1-norm. The space $\mathcal{P}(\mathcal{X}_n)$ with the induced total variation topology is homeomorphic to a $(n-1)$-dimensional simplex $\{(c_1, \cdots, c_n) \in \mathbb{R}^n_+ |\ \sum_i c_i = 1\}$.

Exercise 5.2. ([JLS2017, Lemma 3.1, p. 146], cf. [Neveu1970, Ex. IV.1.3]) For any countable family of signed measures $\{\mu_n \in \mathcal{S}(\mathcal{X})\}$ show that there exists a measure $\mu \in \mathcal{M}(\mathcal{X})$ dominating all measures μ_n.

Since $\mathcal{S}(\mathcal{X})$ is a Banach space, it is a metric space with the total variation distance

$$d_V(\mu_1, \mu_2) = \int_{\mathcal{X}} d|\mu_1 - \mu_2|(x) \tag{5.5}$$

and the associated topology on $\mathcal{S}(\mathcal{X})$ is called *strong topology*. Since $\mathcal{M}(\mathcal{X})$ and $\mathcal{S}(\mathcal{X})$ are subsets in $\mathcal{S}(\mathcal{X})$ they inherit the total variation metric and the strong topology.

The popular *Hellinger distance* d_H and *Bhattacharyya distance* d_B on $\mathcal{P}(\mathcal{X})$ are defined as follows. Let μ_1, μ_2 be dominated by a measure λ. Then

$$d^2_H(\mu_1, \mu_2) = \frac{1}{2}\int_{\mathcal{X}} \Big(\sqrt{\frac{d\mu_1}{d\lambda}} - \sqrt{\frac{d\mu_2}{d\lambda}}\Big)^2 d\lambda, \tag{5.6}$$

$$d_B(\mu_1, \mu_2) = \ln \int_{\mathcal{X}} \sqrt{\frac{d\mu_1}{d\lambda} \cdot \frac{d\mu_2}{d\lambda}} d\lambda. \tag{5.7}$$

Sin $d^2_H(\mu_1, \mu_2) \leq d_V(\mu_1, \mu_2) \leq \sqrt{2}d_H(\mu_1, \mu_2)$ the topology generated by d_H is the strong topology. Since $d^2_H(\mu_1, \mu_2) = 1 - e^{-d_B(\mu_1, \mu_2)}$, the topology generated by d_B is also the strong topology.

Remark 5.3. (cf. [AJLS2017, Appendix C]) On infinite dimensional Banach spaces there exists two notions of topology (convergences): strong topology and weak topology. This leads to two notions of differentiable mappings between Banach spaces.

Let V and W be Banach spaces and $U \subset V$ an open subset. Denote by $Lin(V,W)$ the space of all continuous linear map from V to W. A map $\phi : U \to W$ is called *(Fréchet differentiable at $x \in U$*, if there exists $d_x\phi \in Lin(V,W)$ such that

$$\lim_{h\to 0} \frac{\|\phi(x+h) - \phi(x) - d_x\phi(h)\|_W}{\|h\|_V} = 0. \tag{5.8}$$

In this case, $d_x\phi$ is called the *differential of ϕ at x*. Moreover, ϕ is called *continuously differentiable* or shortly a *C^1-map*, if it is differentiable at every $x \in U$, and the map $d\phi : U \to Lin(V,W)$, $x \mapsto d_x\phi$, is continuous. Furthermore, a differentiable map $c : (-\varepsilon, \varepsilon) \to W$ is called a *curve in W*.

Similarly, a map $\varphi : U \to W$ is called *weakly differerentiable at $x \in U$*, if there exists $d_x\varphi \in Lin(V,W)$ such that

$$\frac{\varphi(x+v) - \varphi(v) - d\varphi_x(v)}{\|v\|_V} \overset{weakly}{\to} 0. \tag{5.9}$$

In this case, $d_x\phi$ is called the *weak differential of ϕ at x*. Moreover, ϕ is called *weakly continuously differentiable* or shortly a *weakly C^1-map*, if it is differentiable at every $x \in U$, and the map $d\phi : U \to Lin(V,W)$, $x \mapsto d_x\phi$, is weakly continuous.

A map ϕ from an open subset Θ of a Banach space V to a subset $\mathcal{X}$ of a Banach space W is called *differentiable*, (resp. *weakly differentiable*) if the composition $i \circ \phi : \Theta \to W$ is differentiable (resp. weakly differentiable), where $i : \mathcal{X} \to W$ is the inclusion. The notion of differentiable mappings (resp. weakly differentiable mappings) can be further generalized to the general setting of diffeology [IZ2013], [Le2020], (resp. [LT2020]).

5.2. **The Fisher metric on a statistical model.** Given a statistical model $\mathcal{P}_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$ we shall endow $\mathcal{P}_\mathcal{X}$ with a nice geometric structure induced from the Banach space $(\mathcal{S}(\mathcal{X}), ||, ||_{TV})$.

Definition 5.4. (cf [AJLS2017, Definition 3.2, p. 141]) (1) Let $(V, \|\cdot\|)$ be a Banach space, $\mathcal{X} \overset{i}{\hookrightarrow} V$ an arbitrary subset, where i denotes the inclusion, and $x_0 \in \mathcal{X}$. Then $v \in V$ is called *a tangent vector of $\mathcal{X}$ at x_0*, if there is a C^1-map $c : \mathbb{R} \to \mathcal{X}$, i.e., the composition $+ \ i \circ c : \mathbb{R} \to V$ is a C^1-map, such that $c(0) = x_0$ and $\dot{c}(0) = v$.

(2) *The tangent (double) cone $C_x\mathcal{X}$* at a point $x \in \mathcal{X}$ is defined as the subset of the tangent space $T_xV = V$ that consists of of tangent vectors of $\mathcal{X}$ at x. The *tangent space $T_x\mathcal{X}$* is the linear hull of the tangent cone $C_x\mathcal{X}$.

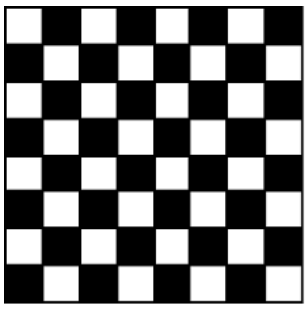
(3) *The tangent cone fibration $C\mathcal{X}$* (resp. *the tangent fibration $T\mathcal{X}$*) is the union $\cup_{x\in\mathcal{X}} C_x\mathcal{X}$ (resp. $\cup_{x\in\mathcal{X}} T_x\mathcal{X}$), which is a subset of $V \times V$ and therefore it is endowed with the induced topology from $V \times V$.

Example 5.5. Let us consider a mixture family $\mathcal{P}_\mathcal{X}$ of probability measures $p_\eta\mu_0$ on $\mathcal{X}$ that are dominated by $\mu_0 \in \mathcal{P}(\mathcal{X})$, where the density functions

p_η are of the following form

$$p_\eta(x) := g^1(x)\eta_1 + g^2(x)\eta_1 + g^3(x)(1-\eta_1-\eta_2) \text{ for } x \in \mathcal{X}. \tag{5.10}$$

Here g^i, for $i = 1, 2, 3$, are nonnegative functions on $\mathcal{X}$ such that $\mathbb{E}_{\mu_0}(g^i) = 1$ and $\eta = (\eta_1, \eta_2) \in D_b \subset \mathbb{R}^2$ is a parameter, which will be specified as follows. Let us divide the square $D = [0,1] \times [0,1] \subset \mathbb{R}^2$ in smaller squares and color them in black and white like a chessboard. Let D_b be the closure of the subset of D colored in black. If η is an interior point of D_b then $C_{p_\eta}\mathcal{P}_\mathcal{X} = \mathbb{R}^2$. If η is a boundary point of D_b then $C_{p_\eta}\mathcal{P}_\mathcal{X} = \mathbb{R}$. If η is a corner point of D_b, then $C_{p_\eta}\mathcal{P}_\mathcal{X}$ consists of two intersecting lines.



Exercise 5.6. ([Bogachev2010, Corollary 3.3.2, p.77], cf. [AJLS2017, Theorem 3.1, p. 142]) Let $\mathcal{P}_\mathcal{X}$ be a statistical model. Show that any $v \in C_\xi\mathcal{P}_\mathcal{X}$ is dominated by ξ. Hence *the logarithmic representation of* v

$$\log v := dv/d\xi$$

is an element of $L^1(\mathcal{X}, \xi)$.

Next we want to put a Riemannian metric on $\mathcal{P}_\mathcal{X}$ i.e., to put a positive quadratic form $\mathfrak{g}$ on each tangent space $T_\xi\mathcal{P}_\mathcal{X}$. By Exercise 5.6, the logarithmic representation $\log(C_\xi\mathcal{P}_\mathcal{X})$ of $C_\xi\mathcal{P}_\mathcal{X}$ is a subspace in $L^1(\mathcal{X}, \xi)$. The space $L^1(\mathcal{X}, \xi)$ does not have a natural metric but its subspace $L^2(\mathcal{X}, \xi)$ is a Hilbert space.

Definition 5.7. (1) A statistical model $\mathcal{P}_\mathcal{X}$ that satisfies

$$\log(C_\xi\mathcal{P}_\mathcal{X}) \subset L^2(\mathcal{X}, \xi) \tag{5.11}$$

for all $\xi \in P$ is called *almost 2-integrable*.

(2) Assume that $\mathcal{P}_\mathcal{X}$ is an almost 2-integrable statistical model. For each $v, w \in C_\xi P$ *the Fisher metric on* $\mathcal{P}_\mathcal{X}$ is defined as follows

$$\mathfrak{g}(v, w) := \langle \log v, \log w \rangle_{L^2(\mathcal{X},\xi)} = \int_\mathcal{X} \log v \cdot \log w \, d\xi. \tag{5.12}$$

(3) An almost 2-integrable statistical model $\mathcal{P}_\mathcal{X}$ is called *2-integrable*, if the function $v \mapsto |v|_\mathfrak{g}$ is continuous on $C\mathcal{P}_\mathcal{X}$.

Since $T_\xi\mathcal{P}_\mathcal{X}$ is a linear hull of $C_\xi\mathcal{P}_\mathcal{X}$, the formula (5.12) extends uniquely to a positive quadratic form on $T_\xi\mathcal{P}_\mathcal{X}$, which is also called *the Fisher metric*.

Example 5.8. Let $\mathcal{P}_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$ be a 2-integrable statistical model that is parameterized by a differentiable map $\mathbf{p} : \Theta \to \mathcal{P}_\mathcal{X}$, $\theta \mapsto p_\theta \mu_0$, where Θ is an open subset in $\mathbb{R}^n$. It follows from (5.12) that the Fisher metric on $\mathcal{P}_\mathcal{X}$ has the following form

$$\mathfrak{g}_{|\mathbf{p}(\theta)}(d\mathbf{p}(v), d\mathbf{p}(w)) = \int_\mathcal{X} \frac{\partial_v p_\theta}{p_\theta} \cdot \frac{\partial_w p_\theta}{p_\theta} p_\theta d\mu_0, \tag{5.13}$$

for any $v, w \in T_\theta \Theta$.

Remark 5.9. (1) The Fisher metric has been defined by Fisher in 1925 to characterize "information" of a statistical model. One of most notable applications of the Fisher metric is the Cramér-Rao inequality which measures our ability to have a good density estimator in terms of geometry of the underlying statistical model, see Theorem 5.14 below.

(2) The Fisher metric $\mathfrak{g}_{\mathbf{p}(\theta)}(d\mathbf{p}(v), d\mathbf{p}(v))$ of a parameterized statistical model $\mathcal{P}_\mathcal{X}$ of dominated measures in Example 5.8 can be obtained from the Taylor expansion of the Kullback-Leibler divergence $I(\mathbf{p}(\theta), \mathbf{p}(\theta + \varepsilon v))$, assuming that $\log p_\theta$ is continuously differentiable in all partial derivative in θ up to order 3. Indeed we have

$$I(\mathbf{p}(\theta), \mathbf{p}(\theta + \varepsilon v)) = \int_\mathcal{X} p_\theta(x) \log \frac{p_\theta(x)}{p_{\theta+\varepsilon v}(x)} d\mu_0$$

$$= -\varepsilon \int_\mathcal{X} p_\theta(x) \partial_v \log p_\theta(x) d\mu_0 \tag{5.14}$$

$$- \varepsilon^2 \int_\mathcal{X} p_\theta(x) (\partial_v)^2 \log p_\theta(x) d\mu_0 + O(\varepsilon^3). \tag{5.15}$$

Since $\log_\theta(x)$ is continuously differentiable in θ up to order 3, we can apply differentiation under the integral sign, see e.g. [Jost2005, Theorem 16.11, p. 213] to (5.14), which then must vanish, and integration by part to (5.15). Hence we obtain

$$I(\mathbf{p}(\theta), \mathbf{p}(\theta + \varepsilon v)) = \varepsilon^2 \mathfrak{g}_{\mathbf{p}(\theta)}(d\mathbf{p}(v), d\mathbf{p}(v)) + O(\varepsilon^3)$$

what is required to prove.

5.3. **MSE and Cramér-Rao inequality.** Given a statistical model $\mathcal{P}_\mathcal{X} \subset \mathcal{P}(\mathcal{X})$, we wish to measure the accuracy (also called the efficiency) of a nonparametric *estimator* $\hat{\sigma} : \mathcal{X} \to P$ via MSE. As we have remarked in Remark 5.19 usually we need only to estimate a φ-coordinate of $\mathcal{P}_\mathcal{X}$ where V is a real Hilbert vector space with a scalar product $\langle \cdot, \cdot \rangle$ and the associated norm $\| \cdot \|$ and $\varphi : \mathcal{P}_\mathcal{X} \to V$ is a "coordinate" map.

Example 5.10. (1) Let us reconsider the example of nonparametric density estimation where $\mathcal{X} = \mathbb{R}^k$ and $\mathcal{H} \subset C^\infty(\mathbb{R}^k)$, see Subsubsection 3.1.2. Let $x_0 \in \mathbb{R}^k$ and $S_n \in (\mathbb{R}^k)^n$. Then we have considered estimators $\hat{\rho}_{S_n} \in \mathcal{H}$ and its associated φ-estimators $\hat{\rho}_{S_n}(x_0) \in \mathbb{R}^k$, where $\varphi : \mathcal{H} \to \mathbb{R}^k$, $f \mapsto f(x_0)$, is the evaluation mapping.

(2) Assume that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and positive definite kernel function and V be the associated RKHS. Then we define the *kernel mean embedding* $\varphi : \mathcal{P}(\mathcal{X}) \to V$ as follows [MFSS2017]

$$\varphi(\xi) := \int_{\mathcal{X}} k(x, \cdot) d\xi(x).$$

Denote by $Map(\mathcal{P}_{\mathcal{X}}, V)$ the space of all mappings $\varphi : \mathcal{P}_{\mathcal{X}} \to V$. For $l \in V$ let $\varphi^l : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ defined by

$$\varphi^l(\xi) := \langle l, \varphi(\xi) \rangle,$$

and we set

$$L^2_{\varphi}(\mathcal{P}_{\mathcal{X}}, \mathcal{X}) := \{\hat{\sigma} : \mathcal{X} \to \mathcal{P}_{\mathcal{X}} | \, \varphi^l \circ \hat{\sigma} \in L^2(\mathcal{X}, \xi) \text{ for all } \xi \in \mathcal{P}_{\mathcal{X}} \text{ and for all } l \in V\}.$$

For $\hat{\sigma} \in L^2_{\varphi}(\mathcal{X}, \mathcal{P}_{\mathcal{X}})$ we define the φ-mean value of $\hat{\sigma}$, denoted by $\varphi_{\hat{\sigma}} : \mathcal{P}_{\mathcal{X}} \to V$, as follows

$$\varphi_{\hat{\sigma}}(\xi)(l) := \mathbb{E}_{\xi}(\varphi^l \circ \hat{\sigma}) \text{ for } \xi \in \mathcal{P}_{\mathcal{X}} \text{ and } l \in V.$$

The difference $b^{\varphi}_{\hat{\sigma}} := \varphi_{\hat{\sigma}} - \varphi \in Map(\mathcal{P}_{\mathcal{X}}, V)$ is called *the bias* of the φ-estimator $\hat{\sigma}_{\varphi}$. An estimator $\hat{\sigma}$ is called *φ-unbiased*, if $\varphi_{\hat{\sigma}} = \varphi$, equivalently, $b^{\varphi}_{\hat{\sigma}} = 0$.

Next for each $l \in V$ we define the φ-instantaneous loss functions at ξ, similar as in (3.13) (cf. (2.17))

$$L^l : \mathcal{X} \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}, \; L^l(x, \xi) = |\varphi^l \circ \hat{\sigma}(x) - \varphi^l \circ \xi|^2.$$

This leads to a function $MSE^{\varphi}[\hat{\sigma}] \in Map(\mathcal{P}_{\mathcal{X}}, Q(V))$ where $Q(V)$ is the space of quadratic forms on V, whose value at $\xi \in \mathcal{P}_{\mathcal{X}}$ is called the *mean square error of* $\varphi \circ \hat{\sigma}$ at ξ, such that for all $l \in V$ we have

$$MSE^{\varphi}_{\xi}[\hat{\sigma}](l, l) := \mathbb{E}_{\xi}[(L^l(x, \xi))^2]. \tag{5.16}$$

Similarly we define the *variance* of $\varphi \circ \hat{\sigma}$ to be an element in $Map(\mathcal{P}_{\mathcal{X}}, Q(V))$ such that such that for all $l, h \in V$ we have

$$V^{\varphi}_{\xi}[\hat{\sigma}](l, l) = \mathbb{E}_{\xi}[|(\varphi^l \circ \hat{\sigma}(x) - \mathbb{E}_{\xi}(\varphi^l \circ \hat{\sigma}))|^2].$$

Remark 5.11. For $\hat{\sigma} \in L^2_{\varphi}(\mathcal{X}, \mathcal{P}_{\mathcal{X}})$ the $\mathbb{R}$-valued mean square error $MSE^{\varphi}_{\xi}(\hat{\sigma})$ of the φ-estimator $\varphi \circ \hat{\sigma}$ at ξ is defined by

$$MSE^{\varphi}_{\xi}(\hat{\sigma}) := \mathbb{E}_{\xi}(\|\varphi \circ \hat{\sigma} - \varphi(\xi)\|^2). \tag{5.17}$$

The RHS of (5.17) is well-defined, since $\hat{\sigma} \in L^2_{\varphi}(\mathcal{X}, \mathcal{P}_{\mathcal{X}})$ and therefore

$$\langle \varphi \circ \hat{\sigma}(x), \varphi \circ \hat{\sigma}(x) \rangle \in L^1(\mathcal{X}, \xi) \text{ and } \langle \varphi \circ \hat{\sigma}(x), \varphi(\xi) \rangle \in L^2(\mathcal{X}, \xi).$$

Similarly, we define that the $\mathbb{R}$-valued variance of a φ-estimator $\varphi \circ \hat{\sigma}$ at ξ as follows

$$V^{\varphi}_{\xi}(\hat{\sigma}) := \mathbb{E}_{\xi}(\|\varphi \circ \hat{\sigma} - \mathbb{E}_{\xi}(\varphi \circ \hat{\sigma})\|^2). \tag{5.18}$$

If V has a countable basis of orthonormal vectors $v_1, \cdots, v_\infty$, then we have

$$MSE_\xi^\varphi(\hat{\sigma}) = \sum_{i=1}^{\infty} MSE_\xi^\varphi[\hat{\sigma}](v_i, v_i), \tag{5.19}$$

$$V_\xi^\varphi(\hat{\sigma}) = \sum_{i=1}^{\infty} V_\xi^\varphi[\hat{\sigma}](v_i, v_i). \tag{5.20}$$

Exercise 5.12. ([AJLS2017, (5.58), p. 279]) Prove the following formula

$$MSE_\xi^\varphi[\hat{\sigma}](l,k) = V_\xi^\varphi[\hat{\sigma}](l,k) + \langle b_{\hat{\sigma}}^\varphi(\xi), l\rangle \cdot \langle b_{\hat{\sigma}}^\varphi(\xi), k\rangle \tag{5.21}$$

for all $\xi \in \mathcal{P}_\mathcal{X}$ and all $l, k \in V$.

Now we assume that $\mathcal{P}_\mathcal{X}$ is an almost 2-integrable statistical model. For any $\xi \in \mathcal{P}_\mathcal{X}$ let $T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X}$ be the completion of $T_\xi\mathcal{P}_\mathcal{X}$ w.r.t. the Fisher metric $\mathfrak{g}$. Since $T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X}$ is a Hilbert space, the map

$$L_\mathfrak{g} : T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X} \to (T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X})',\ L_\mathfrak{g}(v)(w) := \langle v, w\rangle_\mathfrak{g},$$

is an isomorphism. Then we define the inverse $\mathfrak{g}^{-1}$ of the Fisher metric $\mathfrak{g}$ on $(T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X})'$ as follows

$$\langle L_g v, L_g w\rangle_{\mathfrak{g}^{-1}} := \langle v, w\rangle_\mathfrak{g} \tag{5.22}$$

Definition 5.13. (cf. [AJLS2017, Definition 5.18, p. 281]) Assume that $\hat{\sigma} \in L^2_\varphi(\mathcal{X}, \mathcal{P}_\mathcal{X})$. We shall call $\hat{\sigma}$ *a φ-regular estimator*, if for all $l \in V$ the function $\xi \mapsto \|\varphi^l \circ \hat{\sigma}\|_{L^2(\mathcal{X},\xi)}$ is locally bounded, i.e., for all $\xi_0 \in \mathcal{P}_\mathcal{X}$

$$\lim_{\xi \to \xi_0} \sup \|\varphi^l \circ \hat{\sigma}\|_{L^2(\mathcal{X},\xi)} < \infty.$$

In [AJLS2017, Le2020] it has been proved that, if $\hat{\sigma}$ is a φ-regular estimator, then for all $l \in V$ the function $\varphi^l_{\hat{\sigma}}$ is *differentiable*, moreover, there is differential $d\varphi^l_{\hat{\sigma}}(\xi) \in (T_\xi^\mathfrak{g}\mathcal{P}_\mathcal{X})'$ for any $\xi \in \mathcal{P}_\mathcal{X}$ such that

$$\partial_v \varphi^l_{\hat{\sigma}}(\xi) = d\varphi^l_{\hat{\sigma}}(v) \text{ for all } v \in T_\xi. \tag{5.23}$$

Here for a function f on $\mathcal{P}_\mathcal{X}$ we define $\partial_v f(\xi) := \dot{f}(c(t))$ where $c(t) \subset \mathcal{P}_\mathcal{X}$ is a curve with $c(0) = \xi$ and $\dot{c}(0) = v$.

Set

$$\nabla_\mathfrak{g}\varphi^l_{\hat{\sigma}}(\xi) := L_\mathfrak{g}^{-1}(d\varphi^l_{\hat{\sigma}}).$$

Then (5.23) is equivalent to the following equality

$$\partial_v \varphi^l_{\hat{\sigma}}(\xi) = \langle v, \nabla_\mathfrak{g}\varphi^l_{\hat{\sigma}}(\xi)\rangle_\mathfrak{g}.$$

Theorem 5.14 (The Cramér-Rao inequality)**.** *(*[Le2020, Theorem 3]*) Let $\mathcal{P}_\mathcal{X}$ be a 2-integrable statistical model, V -a Hilbert vector space, φ a V-valued function on P and $\hat{\sigma} \in L^2_\varphi(P, \mathcal{X})$ a φ-regular estimator. Then for all $l \in (\mathbb{R}^n)^*$ we have*

$$V_\xi^\varphi[\hat{\sigma}](l,l) - \|d\varphi^l_{\hat{\sigma}}\|^2_{\mathfrak{g}^{-1}}(\xi) \geq 0.$$

Example 5.15. Assume that $\hat{\sigma}$ is an unbiased estimator. Then the terms involving $b_{\hat{\sigma}} := b^{\varphi}_{\hat{\sigma}}$ vanishes. Since $\varphi_{\hat{\sigma}} = \varphi$ we have $||d\varphi^l||_{\mathfrak{g}^{-1}(\xi)} = ||d\varphi^l_{\hat{\sigma}}||^2_{\mathfrak{g}^{-1}(\xi)}$. If $\varphi : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}^n$ is a local coordinate mapping at $\xi \in \mathcal{P}_{\mathcal{X}}$ then the Cramér-Rao inequality in Theorem 5.14 becomes the well-known Cramér-Rao inequality for an unbiased estimator

$$V_\xi := V_\xi[\hat{\sigma}] \geq \mathfrak{g}^{-1}(\xi). \tag{5.24}$$

5.4. **Efficient estimators and MLE.**

Definition 5.16. Assume that $\mathcal{P}_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ is a 2-integrable statistical model and $\varphi : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}^n$ is a feature map. An estimator $\hat{\sigma} \in L^2_{\varphi}(\mathcal{P}_{\mathcal{X}}, \mathcal{X})$ is called *efficient*, if the Cramér-Rao inequality for $\hat{\sigma}$ becomes an equality, i.e., $V^{\varphi}_{\xi}[\hat{\sigma}] = ||d\varphi_{\hat{\sigma}}||^2_{\mathfrak{g}^{-1}(\xi)}$ for any $\xi \in \mathcal{P}_{\mathcal{X}}$.

Theorem 5.17. *(*[AJLS2017, Corollary 5.6, p. 291]*) Let $\mathcal{P}_{\mathcal{X}}$ be a 2-integrable statistical model parameterized by a differentiable map $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$, $\theta \mapsto p_\theta\mu_0$, where Θ is an open subset of $\mathbb{R}^n$, and $\varphi : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}^n$ a differentiable coordinate mapping, i.e., $\mathbf{p} \circ \varphi = Id$. Assume that the function $p(x, \theta) := p_\theta(x)$ has continuous partial derivatives up to order 3. If $\hat{\sigma} : \mathcal{X} \to \mathcal{P}_{\mathcal{X}}$ is an unbiased efficient estimator then $\hat{\sigma}$ is a maximum likelihood estimator (MLE), i.e.,*

$$\partial_v \log p(\theta, x)|_{\theta=\varphi\circ\hat{\sigma}(x)} = 0 \tag{5.25}$$

for all $x \in \mathcal{X}$ and all $v \in T_\theta\Theta$.

5.5. **Consistency of MLE.** Assume that $P_1 := \mathcal{P}_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ is a 2-integrable statistical model that contains an unknown probability measure μ_u governing distribution of random instance $x_i \in \mathcal{X}$. Then $P_n := \{\mu^n | \, \mu \in \mathcal{P}_{\mathcal{X}}\} \subset \mathcal{P}(\mathcal{X}^n)$ is a 2-integrable statistical model containing probability measure μ^n_u that governs the distribution of i.i.d. of random instances $(x_1, \cdots, x_n) \in \mathcal{X}^n$. Denote by $\mathfrak{g}_n$ the Fisher metric on the statistical model P_n. The map $\lambda_n : P_1 \to P_n$, $\mu \mapsto \mu^n$, is a 1-1 map, it is the restriction of the differentiable map, also denoted by λ_n

$$\lambda_n : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{X}^n),\ \lambda_n(\mu) = \mu^n.$$

It is easy to see that for any $v \in T_\mu\mathcal{P}_{\mathcal{X}}$ we have

$$\mathfrak{g}_n(d\lambda_n(v), d\lambda_n(v)) = n \cdot \mathfrak{g}_1(v, v). \tag{5.26}$$

This implies that the lower bound in the Cramér-Rao inequality (5.24) for unbiased estimators $\hat{\sigma}^*_n := \lambda_n^{-1} \circ \hat{\sigma}_n : (\mathcal{X})^n \to P$ converges to zero and there is a hope that MLE is asymptotically accurate as n goes to infinity. Now we shall give a concept of a consistent sequence φ-estimators that formalizes the notion of asymptotically accurate sequence of estimators $\hat{\sigma}^*_k : \mathcal{X}^k \to P_1$ and using it to examine MLE. [19].

[19] the notion of a consistent sequence of estimators that is asymptotically accurate has been suggested by Fisher in [Fisher1925]

Definition 5.18. (cf. [IH1981, p. 30]) Let $\mathcal{P}_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ be a statistical model, V - real Hilbert space and φ a V-valued function on P. A sequence of φ-estimators $\hat{\sigma}_k^* : (\mathcal{X})^k \to \mathcal{P}_{\mathcal{X}} \to V$ is called *a consistent sequence of φ-estimators for the value $\varphi(\mu_u)$* if for all $\delta > 0$ we have

$$\lim_{k\to\infty} \mu_u^k(\{\mathbf{x} \in \mathcal{X}^{\mathbf{k}} : |\varphi \circ \hat{\sigma}_{\mathbf{k}}^*(\mathbf{x}) - \varphi(\mu_{\mathbf{u}})| > \delta\}) = \mathbf{0}. \tag{5.27}$$

Remark 5.19. Under quite general conditions on the density functions of a statistical model $\mathcal{P}_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ of dominated measures, see e.g. [IH1981, Theorem 4.3, p. 36] the sequence of MLE's is consistent.

5.6. **Conclusion.** In this lecture we derive a natural geometry on a statistical model $\mathcal{P}_{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$, regarding it as a subset in the Banach space $\mathcal{S}(\mathcal{X})$ with the total variation norm. We define the MSE and variance of φ_k-estimator $\varphi_k \circ \hat{\sigma}_k$, which can be estimated using the Cramér-Rao inequality, assuming that $\mathcal{P}_{\mathcal{X}}$ is 2-integrable, i.e. the Fisher metric is continuous on $\mathcal{P}_{\mathcal{X}}$. It turns out that MLE is an efficient unbiased φ-estimator, when φ is the coordinate map. Furthermore, a φ-estimator is consistent if its MSE at a sample of size n converges in probability in zero as n goes to infinity.

6. Consistency of a learning algorithm

In the last lecture we learned the important concept of a consistent sequence of φ-estimators, which formalizes the notion of the asymptotic accuracy of a sequence of estimators, and applied this concept to MLE. Since the notion of a φ-estimator is a technical modification of the notion of an estimator, we shall not consider φ-estimators separately in our consideration of consistency of a learning algorithm.

In this lecture we extend the concept of a consistent sequence of estimators to the notion of consistency of a learning algorithm in a unified learning model of supervised learning and unsupervised learning. Then we shall apply this concept to examine the success of ERM algorithm in binary classification problems.

6.1. **Consistent learning algorithm and its sample complexity.** Recall that in *a unified learning model* $(\mathcal{Z}, \mathcal{H}, L, P_{\mathcal{Z}})$, see Subsection 3.4, we are given a sample space $\mathcal{Z}$, a hypothesis /decision class $\mathcal{H}$ together with a mapping $\varphi : \mathcal{H} \to V$, where V is a real Hilbert space, and instantaneous loss function $L : \mathcal{Z} \times \mathcal{H} \times P_{\mathcal{Z}} \to \mathbb{R}_+$, where $P_{\mathcal{Z}} \subset \mathcal{P}(\mathcal{Z})$ is a statistical model. We define the expected loss/risk function as follows

$$R^L : \mathcal{H} \times P_{\mathcal{Z}} \to \mathbb{R}_+,\ (h, \mu) \mapsto \mathbb{E}_\mu L(z, h, \mu),$$

We also set for $\mu \in P$

$$R_\mu^L : \mathcal{H} \to \mathbb{R},\ h \mapsto R^L(h, \mu).$$

A learning algorithm is a map

$$A : \bigcup_n \mathcal{Z}^n \to \mathcal{H},\ S \mapsto h_S$$

where S is distributed by some unknown $\mu^n \in \mathcal{P}_{\mathcal{Z}^n} = \lambda_n(\mathcal{P}_{\mathcal{Z}})$, cf. (2.1).

The expected accuracy of a learning algorithm A with data size n is defined as

$$L(A_n) := \mathbb{E}_{\mu^n} R^L_\mu(A(S_n))$$

where μ^n is the unknown probability measure that governs the distribution of observable $S_n \in \mathcal{X}^n$.

Example 6.1. In a unified learning model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ there is always a learning algorithm using ERM, if $\mathcal{H}$ is finite, or more general, assuming $\mathcal{H}$ is a topological space, $\mathcal{H}$ is compact and $R^L_{\mu_{S_n}}$ is a continuous function on $\mathcal{H}$. The ERM algorithm A_{erm} for $(\mathcal{Z}, \mathcal{H}, L, P_{\mathcal{Z}})$ is defined as follows

$$A_{erm}(S_n) := \arg_{h\in\mathcal{H}} \min \hat{R}^L_{S_n}(h), \tag{6.1}$$

where recall that $\hat{R}^L_{S_n} := R^L_{\mu_{S_n}}$. Observe that $\arg_{h\in\mathcal{H}} \min \hat{R}^L_{S_n}(h)$ may not exist if $\mathcal{H}$ is not compact. In this case we denote by $A_{erm}(S_n)$ any hypothesis that satisfies the inequality $\hat{R}^L_S(A_{erm}(S_n)) - \inf_{h\in\mathcal{H}} \hat{R}^L_{S_n}(h) = O(n^{-k})$, where $k \geq 1$ depends on the computational complexity of defining A_{erm}. In other words, A_{erm} is only asymptotically ERM.

In general we expect a close relationship between a discriminative model $\mathcal{H}$ and a statistical model $\mathcal{P}_{\mathcal{Z}}$, which generalizes the Bayes principle.

Recall that $R^L_{\mu,\mathcal{H}} = \inf_{h\in\mathcal{H}} R^L_\mu(h)$, see (2.11).

Definition 6.2. A learning algorithm A in a model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ is called *consistent*, if for any $\varepsilon \in (0, 1)$ and for every probability measure $\mu \in \mathcal{P}_{\mathcal{Z}}$,

$$\lim_{n\to\infty} \mu^n\{S \in \mathcal{Z}^n : |R^L_\mu(A(S)) - R^L_{\mu,\mathcal{H}}| \geq \varepsilon\} = 0 \tag{6.2}$$

(2) A learning algorithm A is called *uniformly consistent*, if (6.2) converges to zero uniformly on $\mathcal{P}_{\mathcal{Z}}$, i.e., for any $(\varepsilon, \delta) \in (0, 1)^2$ there exists a number $m_A(\varepsilon, \delta)$ such that for any $\mu \in \mathcal{P}_{\mathcal{Z}}$ and any $m \geq m_A(\varepsilon, \delta)$ we have

$$\mu^m\{S \in \mathcal{Z}^m : |R^L_\mu(A(S)) - R^L_{\mu,\mathcal{H}}| \leq \varepsilon\} \geq 1 - \delta. \tag{6.3}$$

If(6.3) holds we say that A predicts with *expected accuracy* ε and *expected confidence* $1 - \delta$ using m samples.

We characterize the uniform consistency of a learning algorithm A via the notion of *the sample complexity function of* A.

Definition 6.3. Let A be an algorithm on $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ and $m_A(\varepsilon, \delta)$ the minimal number $m_0 \in \mathbb{R}_+ \cup \infty$ such that (6.3) holds for any $\mu \in P$ and any $m \geq m_0$. Then the function $m_A : (\varepsilon, \delta) \mapsto m_A(\varepsilon, \delta)$ is called *the sample complexity function of algorithm* A.

Clearly a learning algorithm A is uniformly consistent if and only if m_A takes value in $\mathbb{R}_+$. Furthermore, A is consistent if and only if the sample function of A on the sub-model $(\mathcal{Z}, \mathcal{H}, L, \mu)$ takes values in $\mathbb{R}_+$ for all $\mu \in P$.

Example 6.4. Let $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ be a unified learning model. Denote by $\pi_n : \mathcal{Z}^\infty \to \mathcal{Z}^n$ the map $(z_1, \cdots, z_\infty) \mapsto (z_1, \cdots, z_n)$. A sequence $\{S_\infty \in \mathcal{Z}^\infty\}$ of i.i. instances distributed by $\mu \in \mathcal{P}_{\mathcal{Z}}$ is called *overfitting*, if there exist $\varepsilon \in (0,1)$ such that for all n we have

$$|R^L_\mu(A_{erm}(\pi_n(S_\infty)) - R^L_{\mu,\mathcal{H}}| \geq \varepsilon. \tag{6.4}$$

Thus A_{erm} is consistent, if and only if the set of all overfitting sequences $S_\infty \in \mathcal{Z}^\infty$ has μ^∞-zero measure, see Subsection C for the definition of μ^∞, equivalently, if (6.2) holds, for any $\mu \in P$. In Example 2.14 we showed the existence of a measure μ_f on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ such that any sequence $S_\infty \in \mathcal{Z}^\infty$ distributed by μ_f^∞ is overfitting. Hence the unified learning model in Example 2.14 for any $\mathcal{P}_{\mathcal{Z}}$ containing μ_f is not consistent using A_{erm}.

The following simple Lemma states that the uniform consistency of A_{erm} is a consequence of the convergence in probability of $\hat{R}^L_S(h)$ to $R^L_\mu(h)$ (by the weak law of large number) that is uniform on $\mathcal{H}$.

Lemma 6.5. *Assume that for any $(\varepsilon, \delta) \in (0,1)^2$ there exists a function $m_{\mathcal{H}}(\varepsilon, \delta)$ taking value in $\mathbb{R}_+$ such that for all $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ for all $\mu \in P$ and for all $h \in \mathcal{H}$ we have*

$$\mu^m\{S \in \mathcal{Z}^m : |\hat{R}^L_S(h) - R^L_\mu(h)| < \varepsilon\} \geq 1 - \delta \tag{6.5}$$

then A_{erm} is uniformly consistent.

Proof. For the simplicity of the exposition we assume first that $A_{erm}(S_n) = \arg\min_{h \in \mathcal{H}} R^L_{S_n}(h)$. The argument for the general case can be easily adapted from this simple case.

Given $m \geq m_{\mathcal{H}}(\varepsilon/2, \delta/2)$, $\mu \in P$ and $h_\varepsilon \in \mathcal{H}$ such that $R^L_\mu(h_\varepsilon) \leq R^L_{\mu,\mathcal{H}} + \varepsilon$ we have

$$\mu^m\{S \in \mathcal{Z}^m : |R^L_\mu(A_{erm}(S)) - R_{\mu,\mathcal{H}}| \leq 2\varepsilon\} \geq$$

$$\mu^m\{S \in \mathcal{Z}^m : R^L_\mu(A_{erm}(S)) \leq R^L_\mu(h_\varepsilon) + \varepsilon\} \geq$$

$$\mu^m\{S \in \mathcal{Z}^m : R^L_\mu(A_{erm}(S)) \leq \hat{R}^L_S(h_\varepsilon) + \frac{\varepsilon}{2} \,\&\, |\hat{R}^L_S(h_\varepsilon) - R^L_\mu(h_\varepsilon)| < \frac{\varepsilon}{2}\} \geq$$

$$\mu^n\{S \in \mathcal{Z}^m : R^L_\mu(A_{erm}(S)) \leq \hat{R}^L_S(h_\varepsilon) + \varepsilon\} - \frac{\delta}{2} \geq$$

$$\mu^n\{S \in \mathcal{Z}^n : |R^L_\mu(A_{erm}(S)) - \hat{R}^L_S(A(S))| \leq \frac{\varepsilon}{2}\} - \frac{\delta}{2} \geq 1 - \delta$$

since $\hat{R}^L_S(A(S)) < \hat{R}^L_S(h_\varepsilon)$. This completes the proof of Lemma 6.5. □

Theorem 6.6. *(cf.* [SSBD2014, Corollary 4.6, p.57]*) Let $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}(\mathcal{Z}))$ be a unified learning model. If $\mathcal{H}$ is finite and $L(\mathcal{Z} \times \mathcal{H}) \subset [0, c] \not\ni \infty$ then the ERM algorithm is uniformly consistent.*

Proof. By Lemma 6.5, it suffices to find for each $(\varepsilon, \delta) \in (0,1) \times (0,1)$ a number $m_{\mathcal{H}}(\varepsilon, \delta)$ such that for all $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ and for all $\mu \in \mathcal{P}(\mathcal{Z})$ we have

$$\mu^m\big(\bigcap_{h\in\mathcal{H}} \{S \in \mathcal{Z}^m : |\hat{R}_S^L(h) - R_\mu^L(h)| \leq \varepsilon\}\big) \geq 1 - \delta. \tag{6.6}$$

In order to prove (6.6) it suffices to establish the following inequality

$$\sum_{h\in\mathcal{H}} \mu^m\big(\{S \in \mathcal{Z}^m | \, |\hat{R}_S^L(h) - R_\mu^L(h)| > \varepsilon\}\big) < \delta. \tag{6.7}$$

Since $\#H < \infty$, it suffices to find $m_{\mathcal{H}}(\varepsilon, \delta)$ such that when $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ each summand in RHS of (6.7) is small enough. For this purpose we shall apply the well-known Hoeffding inequality, which specifies the rate of convergence in the week law of large numbers, see Subsection B.4, which shall provides us a necessary bound for the required uniform convergence in (6.3).

To apply Hoeffding's inequality to the proof of Theorem 6.6 we observe that for each $h \in \mathcal{H}$

$$\{\theta_i^h(z) := L(h, z) \in [0, c]\}$$

are i.i.d. $\mathbb{R}$-valued random variables on $\mathcal{Z}$. Furthermore we have for any $h \in \mathcal{H}$ and $S = (z_1, \cdots, z_m)$

$$\hat{R}_S^L(h) = \frac{1}{m}\sum_{i=1}^m \theta_i^h(z_i),$$

$$R_\mu^L(h) = \bar{\theta}^h.$$

Hence the Hoeffding inequality implies

$$\mu^m\{S \in \mathcal{Z}^m : |\hat{R}_S^L(h) - R_\mu^L(h)| > \varepsilon\} \leq 2\exp(-2m\varepsilon^2 c^{-2}). \tag{6.8}$$

Now pluging

$$m \geq m_{\mathcal{H}}(\varepsilon, \delta) := \frac{\log(2\#(\mathcal{H})/\delta)}{2\varepsilon^2 c^{-2}}$$

in (6.8) we obtain (6.7). This completes the proof of Theorem 6.6.

□

Definition 6.7. The function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{R}$ defined by the requirement that $m_{\mathcal{H}}(\varepsilon, \delta)$ is the least number for which (6.6) holds is called *the sample complexity of a (unified) learning model* $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$.

Remark 6.8. (1) The sample complexity $m_{\mathcal{H}}$ gives a lower bound for the rate of convergence in probability of the empirical risk to the expected risk of any $f \in \mathcal{H}$. We have proved in Lemma 6.5 that the sample complexity of the algorithm A_{erm} is upper bounded by the sample complexity $m_{\mathcal{H}}$ of $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}(\mathcal{Z}))$.

(2) The definition of the sample complexity $m_{\mathcal{H}}$ in our lecture is different from the definition of the sample complexity $m_{\mathcal{H}}$ in [SSBD2014, Definition

3.1, p. 43], where the authors consider the concept of PAC-learnability [20] of a hypothesis class $\mathcal{H}$ under a "realizability" assumption. Their $m_{\mathcal{H}}$ depends on a learning algorithm f and it is not hard to see that it is equivalent to the notion of the sample complexity m_f of the learning algorithm f in our lecture.

6.2. **Uniformly consistent learning and the VC-dimension.** In this subsection we shall examine some sufficient and necessary conditions for the existence of a uniformly consistent learning algorithm on a unified learning model with infinite hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$. First we prove a version of No-Free-Lunch theorem which asserts that there is no uniformly consistent learning algorithm on a learning model with a very large hypothesis class and a very large statistical model. Denote by $\mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y})$ the set of all probability measures $(\Gamma_f)_*(\mu_{\mathcal{X}}) \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ where $f : \mathcal{X} \to \mathcal{Y}$ is a measurable map and $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$. Thus $\mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y})$ consists of all probability distributions of labelled pairs $(x, f(x))$ where x is distributed by a probability measure $\mu_{\mathcal{X}}$ and $f \in \mathrm{Meas}(\mathcal{X}, \mathcal{Y})$.

Before stating the No-Free-Lunch theorem let us introduce the notion of a standard sample space. A measurable (sample) space $\mathcal{X}$ will be called *standard*, if its σ-algebra $\Sigma_{\mathcal{X}}$ contains every element $x \in \mathcal{X}$.

Theorem 6.9 (No-Free-Lunch-Theorem)**.** *Let $\mathcal{X}$ be an infinite standard sample space, $\mathcal{Y} = \{0, 1\}$ and $L^{(0-1)}$ the 0-1 loss function. Then there is no uniformly consistent learning algorithm on a unified learning model $(\mathcal{X}\times\mathcal{Y}, \mathcal{H} = \mathrm{Meas}(\mathcal{X}, \mathcal{Y}), L^{(0-1)}, \mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y}))$.*

Proof. To prove Theorem 6.9 it suffices to show that $m_A(\varepsilon, \delta) = \infty$ for $(\varepsilon, \delta) = (1/8, 1/8)$ and for any learning algorithm A. Assume the opposite, i.e., $m_A(1/8, 1/8) = m < \infty$. Then we shall find $\mu(m) \in \mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y})$ such that Equation (6.3) violates for m, $\mu(m)$ and $(\varepsilon, \delta) = (1/8, 1/8)$.

To describe the measure $\mu(m) = (\Gamma_f)_*\mu_{\mathcal{X}}$ for some $\mu_{\mathcal{X}} \in \mathcal{P}(\mathcal{X})$ and some $f \in \mathrm{Meas}(\mathcal{X}, \mathcal{Y})$ we need some notations. For a subset $C[k] \subset \mathcal{X}$ of k-elements let $\mu_{\mathcal{X}}^{C[k]} \in \mathcal{P}(\mathcal{X})$ be defined by

$$\mu_{\mathcal{X}}^{C[k]}(B) := \frac{\#(B \cap C[k])}{k} \text{ for any } B \subset \mathcal{X}. \tag{6.9}$$

For a map $f : \mathcal{X} \to \mathcal{Y}$ we set

$$\mu_f^{C[k]} := (\Gamma_f)_*\mu_{\mathcal{X}}^{C[k]} \in \mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y}).$$

Lemma 6.10. *Assume that $\mathcal{X}$, $\mathcal{Y}$ are finite sets and $\#\mathcal{X} \geq n + 1$. For $f \in \mathcal{Y}^{\mathcal{X}}$ set $\mu_f := (\Gamma_f)_*(\mu_{\mathcal{X}}^{\mathcal{X}}) \in \mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\mathcal{Y})$. Then for any learning algorithm $A : S \mapsto A_S$ we have*

$$\int_{\mathcal{Y}^{\mathcal{X}}} \int_{(\mathcal{X}\times\mathcal{Y})^n} R_{\mu_f}^{(0-1)}(A_S)\, d(\mu_f^n)(S) d\mu_{\mathcal{Y}^{\mathcal{X}}}^{\mathcal{Y}^{\mathcal{X}}}(f) \geq (1 - \frac{1}{\#\mathcal{Y}})(1 - \frac{n}{\#\mathcal{X}}) \tag{6.10}$$

[20] the concept of PAC-learnability has been introduced by the computer scientist L. Valiant [Valiant1984], which is equivalent to the existence of a uniformly consistent learning algorithm with account of computational complexity

Proof of Lemma 6.10. We set for $S \in (\mathcal{X} \times \mathcal{Y})^n$

$$Pr_i : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{X},\ (x_1, y_1), \cdots, (x_n, y_n) \mapsto x_i \in \mathcal{X},$$

$$\mathcal{X}_S := \bigcup_{i+1}^{n} Pr_i(S).$$

Note that S is distributed by μ_f^n means that $S = \{(x_1, f(x_1)), \cdots, (x_n, f(x_n))$ where $(x_1, \cdots, x_n)$ is distributed by the uniform (counting) probability measure $(\mu_{\mathcal{X}}^{\mathcal{X}})^n$. Let us compute and estimate the double integral in the LHS of (6.10) using (2.10) and the Fubini theorem.

$$\begin{aligned}
\mathbb{E}_{\mu_{\mathcal{Y}^{\mathcal{X}}}^{\mathcal{Y}^{\mathcal{X}}}}\Big(\mathbb{E}_{\mu_f^n}\big(R_{\mu_f}^{(0-1)}(A_S)\big)\Big) &= \frac{1}{\#\mathcal{X}}\mathbb{E}_{\mu_{\mathcal{Y}^{\mathcal{X}}}^{\mathcal{Y}^{\mathcal{X}}}}\Big(\mathbb{E}_{(\mu_{\mathcal{X}}^{\mathcal{X}})^n}\big(\sum_{x\in\mathcal{X}}(1-\delta_{f(x)}^{A_S(x)})\big)\Big) \\
&\geq \frac{1}{\#\mathcal{X}}\mathbb{E}_{\mu_{\mathcal{Y}^{\mathcal{X}}}^{\mathcal{Y}^{\mathcal{X}}}}\Big(\mathbb{E}_{(\mu_{\mathcal{X}}^{\mathcal{X}})^n}\big(\sum_{x\notin\mathcal{X}_S}(1-\delta_{f(x)}^{A_S(x)})\big)\Big) \\
&= \frac{1}{\#\mathcal{X}}\mathbb{E}_{(\mu_{\mathcal{X}}^{\mathcal{X}})^n}\Big(\sum_{x\notin\mathcal{X}_S}\mathbb{E}_{\mu_{\mathcal{Y}^{\mathcal{X}}}^{\mathcal{Y}^{\mathcal{X}}}}(1-\delta_{f(x)}^{A_S(x)})\Big) \\
&= \frac{1}{\#\mathcal{X}}\mathbb{E}_{(\mu_{\mathcal{X}}^{\mathcal{X}})^n}\Big(\#[\mathcal{X}\setminus\mathcal{X}_S]\cdot\big(1-\frac{1}{\#\mathcal{Y}}\big)\Big) \\
&\overset{\text{since } \#[\mathcal{X}\setminus\mathcal{X}_S]\geq\#\mathcal{X}-n}{\geq} (1-\frac{1}{\#\mathcal{Y}})(1-\frac{n}{\#\mathcal{X}}). \qquad (6.11)
\end{aligned}$$

This completes the proof of Lemma 6.10. □

Continuation of the proof of Theorem 6.9. Let us denote by $C[2m]$ a subset of size $2m$ in $\mathcal{X}$. Then $\mu_{\mathcal{X}}^{C[2m]} \in \mathcal{P}(\mathcal{X})$. It follows from Lemma 6.10 that there exists $f \in \mathrm{Meas}(\mathcal{X}, \mathcal{Y})$ such that, denoting $\mu := \mu_f^{C[2m]} \in \mathcal{P}_{\mathcal{X}}(\mathcal{X} \times \mathcal{Y})$, we have

$$\int_{(\mathcal{X}\times\mathcal{Y})^m} R_\mu^{(0-1)}(A_S)\, d\mu^m = \int_{(C[2m]\times\mathcal{Y})^m} R_\mu^{(0-1)}(A_S)\, d(\mu^m)(S) \geq \frac{1}{4}. \qquad (6.12)$$

Since $0 \leq R_\mu^{(0-1)} \leq 1$ we obtain from (6.12)

$$\mu^m\{S \in (\mathcal{X} \times \mathcal{Y})^m |\, R_\mu^{(0-1)}\big(A(S)\big) \geq \frac{1}{8}\} > \frac{1}{8}.$$

This implies that (6.3) does not hold for $(\varepsilon, \delta) = (1/8, 1/8)$, for any m and $\mu(m) = \mu_f^{C[m]}$. This proves Theorem 6.9. □

Remark 6.11. Given $m \in \mathbb{N}$, in the proof of Theorem 6.9 we showed that, if there is a subset $C \subset \mathcal{X}$ of size $2m$ and the restriction of $\mathcal{H}$ to C is the full set of functions in $\{0, 1\}^C$ then (6.3) does not hold for m for $(\varepsilon, \delta) = (1/8, 1/8)$ and any learning algorithm A. This implies that the existence of such a subset $C \subset \mathcal{X}$ contributes to the sample complexity of $\mathcal{H}$. This motivates the following

Definition 6.12. (1) A hypothesis class $\mathcal{H} \subset \text{Meas}(\mathcal{X}, \{0,1\})$ *shatters* a finite subset $C \subset \mathcal{X}$ if $\#\mathcal{H}|_C = 2^{\#C}$.

(2) *The VC-dimension* of a hypothesis class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$, denoted by $VC\dim(\mathcal{H})$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has *infinite VC-dimension*.

Example 6.13. (1) A hypothesis class $\mathcal{H}$ shatters a subset of one point $x_0 \in \mathcal{X}$ if and only if there are two functions $f, g \in \mathcal{H}$ such that $f(x_0) \neq g(x_0)$.

(2) Let $\mathcal{H}$ be the class of intervals in the real line, namely,

$$\mathcal{H} = \{1_{(a,b)} : a < b \in \mathbb{R}\},$$

where $1_{(a,b)} : \mathbb{R} \to \{0,1\}$ is the indicate function of the interval (a,b). Take the set $C = \{1,2\}$. Then, $\mathcal{H}$ shatters C, since all the functions in the set $\{1,2\}^{(0,1)}$ can be obtained as the restriction of some function from $\mathcal{H}$ to C. Hence $VC\dim(H) \geq 2$. Now take an arbitrary set $C = \{c_1 < c_2 < c_3\}$ and the corresponding labeling $(1,0,1)$. Clearly this labeling cannot be obtained by an interval: Any interval $h_{(a,b)}$ that contains c_1 and c_3 (and hence labels c_1 and c_3 with the value 1) must contain c_2 (and hence it labels c_2 with 0). Hence $\mathcal{H}$ does not shatter C. We therefore conclude that $VC\dim(\mathcal{H}) = 2$. Note that $\mathcal{H}$ has infinitely many elements.

Exercise 6.14 (VC-Threshold functions)**.** Consider the hypothesis class $\mathcal{F} \subset \{-1,1\}^{\mathbb{R}}$ of all threshold functions $\text{sign}^{\,b} : \mathbb{R} \to \mathbb{R}$, where $b \in \mathbb{R}$, defined by

$$\text{sign}^{\,b}(x) := \text{sign}\,(x-b)$$

Show that $VC\dim(\mathcal{F}) = 1$.

Exercise 6.15. Let us reconsider the Toy Example 2.2, where we want to predict skin diseases by examination of skin images. Recall that $\mathcal{X} = \cup_{i=1}^{5} I_1 \times I_2 \times I_3 \times \{A_i\}$ and $\mathcal{Y} = \{\pm 1\}$. Recall that a function $f \in \text{Meas}(\mathcal{X}, \mathcal{Y})$ can be identified with a measurable subset $f^{-1}(1) \subset \mathcal{X}$. Let us consider the hypothesis class $\mathcal{H} \subset \text{Meas}(\mathcal{X}, \mathcal{Y})$ consisting of cubes $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3] \times \{A_i\} \subset \mathcal{X}$, $i = 1, 5$. Prove that $VC\dim(\mathcal{H}) < \infty$.

In Remark 6.11 we observed that the finiteness of $VC\dim(\mathcal{H})$ is a necessary condition for the existence of a uniformly consistent learning algorithm on a unified learning model $(\mathcal{X}, \mathcal{H}, L^{(0-1)}, \mathcal{P}_{\mathcal{X}}(\mathcal{X} \times \mathcal{Y}))$. In the next section we shall show that the finiteness of $VC\dim(\mathcal{H})$ is also a sufficient condition for the uniform consistency of A_{erm} on $(\mathcal{X}, \mathcal{H}, L^{(0-1)}, \mathcal{P}_{\mathcal{X}}(\mathcal{X} \times \mathcal{Y}))$.

6.3. **Fundamental theorem of binary classification.**

Theorem 6.16 (Fundamental theorem of binary classification)**.** *A learning model* $(\mathcal{X}, \mathcal{H} \subset \text{Meas}(\mathcal{X}, \{0,1\}), L^{(0-1)}, \mathcal{P}(\mathcal{X} \times \{0,1\}))$ *has a uniformly consistent learning algorithm, if and only if* $VC\dim(\mathcal{H}) < \infty$.

Outline of the proof. Note that the "only if" assertion of Theorem 6.16 follows from Remark 6.11. Thus we need only to prove the "if" assertion. By Lemma 6.5 it suffices to show that if $VC\dim(\mathcal{H}) = k < \infty$ then $m_{\mathcal{H}}(\varepsilon, \delta) < \infty$ for all $(\varepsilon, \delta) \in (0,1)^2$. In other words we need to find a lower bound for the LHS of (6.5) in terms of the VC-dimension, which is an upper bound of the RHS of (6.5), when $\varepsilon \in (0,1)$ and m is sufficiently large. This shall be done in three steps.

In step 1, setting $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and omitting superscript L of the risk function R, we use the Markov inequality to obtain the following lower bound for the rate of the convergence in probability of the empirical risk of $h \in \mathcal{H}$ to the expected risk of h:

$$\mu^m\{S \in \mathcal{Z}^m : |R_\mu(h) - \hat{R}_S(h)| < a\} \geq 1 - \frac{\mathbb{E}_{\mu^m}|R_\mu(h) - R_S(h)|}{a} \tag{6.13}$$

for any $a > 0$ and any $h \in \mathcal{H}$.

In step 2 we define *the growth function* $\Gamma_{\mathcal{H}}(m) : \mathbb{N} \to \mathbb{N}$ and use it to upper bound the RHS of (6.13) in Lemma 6.20.

Definition 6.17 (Growth function)**.** Let $\mathcal{F} \subset \mathrm{Meas}(\mathcal{X}, \mathcal{Y})$ be a class of functions with finite target space $\mathcal{Y}$. *The growth function* $\Gamma_{\mathcal{F}}$ assigned to $\mathcal{F}$ is then defined for all $n \in \mathbb{N}$ as

$$\Gamma_{\mathcal{F}}(n) := \max_{\Sigma \subset \mathcal{X} | \, \#\Sigma = n} \#\mathcal{F}|_\Sigma.$$

We also set $\Gamma(0) = 1$.

Example 6.18. Consider the set $\mathcal{F} : \{sign^b | \, b \in \mathbb{R}\}$ of all threshold functions. Given a set of distinct points $\{x_1, \cdots, x_n\} = \Sigma \subset \mathbb{R}$, there are $n+1$ functions in $\mathcal{F}|_\Sigma$ corresponding to $n+1$ possible ways of placing b relative to the x_i s. Hence, in this case $\Gamma_{\mathcal{F}}(n) \geq n+1$.

Exercise 6.19. Show that $\Gamma_{\mathcal{F}}(n) = n+1$ for the set $\mathcal{F}$ of all threshold functions.

Lemma 6.20. *We have*

$$\mathbb{E}_{\mu^m}(\sup_{h \in \mathcal{H}} |R_\mu(h) - \hat{R}_S(h)|) \leq \frac{4 + \sqrt{\log\big(\Gamma_{\mathcal{H}}(2m)\big)}}{\sqrt{2m}} \tag{6.14}$$

for every $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

Proof. Using the following identity

$$R_\mu(h) = \int_{\mathcal{X}^m} \hat{R}_{S'}(h) d\mu^m(S') \tag{6.15}$$

we obtain

$$\mathbb{E}_{\mu^m}(\sup_{h \in \mathcal{H}} |R_\mu(h) - \hat{R}_S(h)|) = \mathbb{E}_{\mu^n}[\sup_{h \in \mathcal{H}} | \int_{\mathcal{X}^m} |\hat{R}_S(h) - \hat{R}_{S'}(h) d\mu^m(S')|] \leq$$

$$\mathbb{E}_{\mu^n\times\mu^n}[\sup_{h\in\mathcal{H}}|\hat{R}_S(h)-\hat{R}_{S'}(h)]\leq$$
$$\leq\frac{4+\sqrt{\log\big(\Gamma_{\mathcal{H}}(2m)\big)}}{\sqrt{2m}}.$$

(For the last inequality see, [SSBD2014, p. 76-77]).

□

In step 3 we use the following Vapnik-Chervonenski-Lemma, also known as Sauer's Lemma, whose proof can be found in [SSBD2014, p. 74-75].

Lemma 6.21. *Let $\mathcal{H}\subset\mathrm{Meas}(\mathcal{X},\{0,1\})$ be a hypothesis class with $VC\dim(H)=d<\infty$. Then, for all $n\in\mathbb{N}$ we have*

$$\Gamma_{\mathcal{H}}(n)\leq\sum_{i=0}^{d}\binom{n}{i}.$$

In particular, if $n>d+1$ then $\Gamma_{\mathcal{H}}(n)\leq(en/d)^d$.

It follows from Lemma 6.21 that for any $(\varepsilon,\delta)\in(0,1)^2$ there exists m such that

$$\frac{4+\sqrt{\log\big(\Gamma_{\mathcal{H}}(2m)\big)}}{\delta\sqrt{2m}}<\varepsilon$$

and therefore by (6.14) for any $(\varepsilon,\delta)\in(0,1)^2$ the value $m_{\mathcal{H}}(\varepsilon,\delta)$ is finite. This completes the proof of Theorem 6.16.

6.4. **Conclusions.** In this lecture we define the notion of the (uniform) consistency of a learning algorithm A on a unified learning model $(\mathcal{Z},\mathcal{H},L,P)$ and characterize this notion via the sample complexity of A. We relate the consistency of the ERM algorithm with the uniform convergence of the law of large numbers over the hypothesis space $\mathcal{H}$ and use it to prove the uniform consistency of ERM in the binary classification problem $(\mathcal{X},\mathcal{H}\subset\mathrm{Meas}(\mathcal{X},\{0,1\}),L^{(0-1)},\mathcal{P}(\mathcal{X}\times\{0,1\})$. We show that the finiteness of the VC-dimension of $\mathcal{H}$ is a necessary and sufficient condition for the existence of a uniform consistent learning algorithm on a binary classification problem $(\mathcal{X}\times\{0,1\},\mathcal{H}\subset\mathrm{Meas}(\mathcal{X},\{0,1\}),L^{(0-1)},\mathcal{P}_{\mathcal{X}}(\mathcal{X}\times\{0,1\})$ (Theorem 6.16).

7. Generalization ability of a learning machine

In the last lecture we measured the uniform consistency of a learning algorithm A in a unified learning model $(\mathcal{Z},\mathcal{H},L,\mathcal{P}_{\mathcal{Z}})$ via the sample complexity function $m_A:(0,1)^2\to\mathbb{R}_+\cup\infty$. The sample complexity $m_A(\varepsilon,\delta)$ is the less number of samples that A requires in order to make a prediction with ε accuracy and $(1-\delta)$ confidence. In 1984 Valiant suggested a PAC-model of learning, which corresponds to the notion of uniform (w.r.t. $\mathcal{P}_{\mathcal{Z}}$) consistency of A, which has moreover to be efficiently computable, i.e. the function $m_A(\varepsilon,\delta)$ must be polynomial in ε^{-1} and δ^{-1} [Valiant1984]. Furthermore Valiant also requires that A is efficiently computable, which can

be expressed in terms of the computational complexity of A, see [SSBD2014, Chapter 8] for discussion on running time of A. In Valiant's PAC theory the uniform consistency requirement is a natural requirement, which distinguishes generalizability of a leaning machine from the consistency of an estimator in classical mathematical statistics.

Thus, given a learning machine $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}}, A)$, where A is a learning algorithm, the generalization ability of $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}}, A)$ is measured in terms of the sample complexity of A. In Lemma 6.5 of the previous lecture we gave upper bounds for the sample complexity $m_{A_{erm}}(\varepsilon, \delta)$ in terms of the sample complexity $m_{\mathcal{H}}(\varepsilon/2, \delta/2)$ of the learning model. Then we showed that in a binary classification problem the sample complexity $m_{\mathcal{H}}$ takes finite values if and only if the VC-dimension of $\mathcal{H}$ is finite (Theorem 6.16).

Today we shall discuss two further methods of upper bounds for the sample complexities $m_{\mathcal{H}}$ and $m_{A_{erm}}$ of some important learning models. Then we discuss the problem of learning model selection.

7.1. **Covering number and sample complexity.** In the binary classification problem of supervised learning the VC-dimension is a combinatorial characterization of the hypothesis class $\mathcal{H}$, which carries no topology, since the domain $\mathcal{X}$ and the target space $\mathcal{Y}$ are discrete. The expected zero-one loss function is therefore the preferred choice of a risk function. In [CS2001] Cucker-Smale estimated the sample complexity $m_{\mathcal{H}}$ of a discriminative model $(\mathcal{X}, \mathcal{Y} = \mathbb{R}^n, \mathcal{H}, L_2, \mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n))$ for a regression problem, i.e., $\mathcal{Y} = \mathbb{R}^n$, where $\mathcal{X}$ is a topological space and $\mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n)$ denotes the set of all Borel probability measures on $\mathcal{X} \times \mathbb{R}^n$.

Before stating Cucker-Smale's results we introduce necessary notations.

• Let $C_n(\mathcal{X})$ be the Banach space of continuous bounded $\mathbb{R}^n$-valued functions on $\mathcal{X}$ with the C^0-norm [21]

$$||f||_{C^0} = \sup_{x \in \mathcal{X}} ||f(x)||.$$

• For $f \in C_n(\mathcal{X})$ let $L_f \in C_n(\mathcal{X} \times \mathbb{R}^n)$ be the quadratic instantaneous loss function valued at f (cf. (2.17))

$$L_f(x, y) = L_2(x, y, f) := \|f(x) - y\|^2. \tag{7.1}$$

Then for $\rho \in \mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n)$ we have $MSE_\rho(f) = \mathbb{E}_\rho(L_f)$ is the expected loss function, cf. (5.17).

• For a function $g \in C_n(\mathcal{X} \times \mathbb{R}^n)$ denote by $V_\rho(g)$ its variance, see (5.18),

$$V_\rho(g) = \mathbb{E}_\rho(||g - \mathbb{E}_\rho(g)||^2) = \mathbb{E}_\rho(||g||^2) - ||\mathbb{E}_\rho g||^2.$$

• For a compact hypothesis class $\mathcal{H} \subset C_n(\mathcal{X})$ define the following quantities for $f \in \mathcal{H}$

$$MSE_{\rho,\mathcal{H}}(f) := MSE_\rho(f) - \min_{f \in \mathcal{H}} MSE_\rho(f), \tag{7.2}$$

[21] In [CS2001, p.8] the authors used the L_∞-norm, but they considered only the subspace of continuous functions

which is called *the estimation error of f*, or *the sample error of f* [CS2001], and we set

$$V_\rho(\mathcal{H}) := \sup_{f\in\mathcal{H}} V_\rho(L_f).$$

• For $s \in \mathbb{R}_+$ we define *the covering number*

$\mathcal{N}(\mathcal{H}, s) := \min\{l \in \mathbb{N}|$ there exists l disks in $\mathcal{H}$ with radius s covering $\mathcal{H}\}$.

• For $S = (z_1, \cdots, z_m) \in \mathcal{Z}^m := (\mathcal{X} \times \mathcal{Y})^m$, where $z_i = (x_i, y_i)$, denote by f_S the minimizer of the empirical risk function $MSE_S : \mathcal{H} \to \mathbb{R}$

$$MSE_S(f) := \frac{1}{m}\sum_{i=1}^{m} ||f(x_i) - y_i||^2 = MSE_{\mu_S}(f).$$

The existence of f_S follows from the compactness of $\mathcal{H}$ and the continuity of the functional MSE_S on $\mathcal{H}$.

Theorem 7.1. *(*[CS2001, Theorem C]*) Let $\mathcal{H}$ be a compact subset of $C_1(\mathcal{X})$. Assume that for all $f \in \mathcal{H}$ we have $|f(x) - y| \le M$ ρ-almost everywhere, where ρ is a probability measure on $\mathcal{X} \times \mathbb{R}$. Then for all $\varepsilon > 0$*

$$\rho^m\{S \in \mathcal{Z}^m |\, MSE_{\rho,\mathcal{H}}(f_S) \le \varepsilon\} \ge 1 - \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{16M}) 2e^{-\frac{m\varepsilon^2}{8(4V_\rho(\mathcal{H}) + \frac{1}{3}M^2\varepsilon)}}. \tag{7.3}$$

Theorem 7.1 implies that for any $n < \infty$ the ERM algorithm is consistent on the unified learning model $(\mathcal{X}, \mathcal{H} \subset C_n(\mathcal{X}), L_2, \mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n))$, where L_2 is the quadratic loss function, if $\mathcal{H}$ is compact, since we can increase the confidence of our estimate, given the accuracy ε, by increasing the number m of samples.

The proof of Theorem 7.1 is based on Lemma 6.5. By Remark 6.8, to prove that the sample complexity of A_{erm} is bounded it suffices to prove that the sample complexity $m_{\mathcal{H}}$ of the learning model $(\mathcal{X}, \mathcal{H} \subset C(\mathcal{X}), L_2, \mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n))$ is bounded.

Outline of the proof of the Cucker-Smale theorem. The strategy of the proof of Theorem 7.1 is similar to that of Fundamental Theorem of binary classification (Theorem 6.16).

In the first step we prove the following Lemma, which gives a lower bound on the rate of the convergence in probability of empirical risk $MSE_S(f)$ to the expected risk $MSE_\rho(f)$ for a given $f \in \mathcal{H}$, i.e. to give a lower bound for the sample complexity $m_{\mathcal{H}}$.

Lemma 7.2. *(*[CS2001, Theorem A, p.8]*) Let $M > 0$ and $f \in C_1(\mathcal{X})$ such that $|f(x) - y| \le M$ ρ-a.e.. Then for all $\varepsilon > 0$ we have*

$$\rho^m\{S \in \mathcal{Z}^m :\, |MSE_\rho(f) - MSE_S(f)| \le \varepsilon\} \ge 1 - 2e^{(-\frac{m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M^2\varepsilon)})}$$

where $\sigma^2 = V_\rho(L_f)$.

Lemma 7.2 is a version of the inequality (6.8), for which we used the Hoeffding inequality. The Hoeffding inequality does not involve the variance and Cucker-Smale used the Bernstein inequality instead of the Hoeffding inequality, see Appendix B.

In the second step, letting $l = \mathcal{N}(\mathcal{H}, \frac{\varepsilon}{4M})$ we cover $\mathcal{H}$ by l disks $D_1, \cdots, D_l$. Then we reduce the problem of estimating upper bound for the sample complexity $m_{\mathcal{H}}$ to the problem of estimating upper bound for the sample complexities m_{D_j}. Namely we have the following easy inequality [CS2001, Lemma 1, p. 13]

$$\rho^m\{S \in \mathcal{Z}^m | \sup_{f\in\mathcal{H}} |MSE_\rho(f) - MSE_S(f)| \geq \varepsilon\} \leq$$

$$\sum_{j=1}^{l} \rho^m\{S \in \mathcal{Z}^m | \sup_{f\in D_j} |MSE_\rho(f) - MSE_S(f)| \geq \varepsilon\}. \tag{7.4}$$

In the last third step we proof the following

Lemma 7.3. *(*[CS2001, Proposition 3, p. 12]*) Let $f_1, f_2 \in C_1(\mathcal{X})$. If $|f_j(x) - y| \leq M$ on a set $U \subset \mathcal{Z}$ of full measure for $j = 1, 2$ then for any $S \in U^m$ we have*

$$|MSE_S(f_1) - MSE_S(f_2)|| \leq 4M||f_1 - f_2||_{C_0}.$$

Lemma 7.3 implies that for all $S \in U^m$

$$\sup_{f\in D_j} |MSE_\rho(f) - MSE_S(f)| \geq 2\varepsilon \Longrightarrow |MSE_\rho(f_j) - MSE_S(f_j)| \geq \varepsilon.$$

Combining the last relation with (7.4), we derive the following desired upper estimate for the sample complexity $m_{\mathcal{H}}$.

Proposition 7.4. *Assume that for all $f \in \mathcal{H}$ we have $|f(x)-y| \leq M$ ρ-a.e.. Then for all $\varepsilon > 0$ we have*

$$\rho^m\{S \in \mathcal{Z}^m : \sup_{f\in\mathcal{H}} |MSE_\rho(f)-MSE_S(f)| \leq \varepsilon\} \geq 1-\mathcal{N}(\mathcal{H}, \frac{\varepsilon}{8M})2e^{(-\frac{m\varepsilon^2}{4(2\sigma^2+\frac{1}{3}M^2\varepsilon)}}$$

where $\sigma^2 = \sup_{f\in\mathcal{H}} V_\rho(L_f)$.

This completes the proof of Theorem 7.1.

Exercise 7.5. Derive from Theorem 7.1 an upper bound for the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta)$ of the learning model $(\mathcal{X}, \mathcal{H} \subset C_n(\mathcal{X}), L_2, \mathcal{P}_B(\mathcal{X}\times\mathbb{R}^n))$, where $\mathcal{H}$ is compact and $\mathcal{P}_B(\mathcal{X} \times \mathbb{R}^n)$ is the space of Borel measures on the topological space $\mathcal{X} \times \mathbb{R}^n$.

Remark 7.6. If the hypothesis class $\mathcal{H}$ in Theorem 7.1 is a convex subset in $\mathcal{H}$ then Cucker-Smale got an improved estimation of the sample complexity $m_{A_{erm}}$ [CS2001, Theorem C*].

7.2. **Rademacher complexities and sample complexity.** Rademacher complexities of a learning model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ are refined versions of the sample complexity $m_{\mathcal{H}}$ by adding a parameter $S \in \mathcal{Z}^n$ for the empirical Rademacher complexity, and by adding a parameter $\mu \in P$ for the expected version of the (expected) Rademacher complexity. Rademacher complexities are designed for *data-depending estimation of upper bounds* of the sample complexity $m_{A_{erm}}$ of the ERM algorithm.

For a learning model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ we set (cf. (7.1))

$$\mathcal{L}_{\mathcal{H}} := \{L_h(z) := L(z,h) |\, h \in \mathcal{H}\}. \tag{7.5}$$

Definition 7.7 (Rademacher complexity)**.** Let $S \in \mathcal{Z}^n$. *The empirical Rademacher complexity* of a family $\mathcal{G} \subset Meas(\mathcal{Z}, \mathbb{R})$ w.r.t. a sample S is defined as follows

$$\hat{\mathcal{R}}_S(\mathcal{G}) := \mathbb{E}_{(\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2})^n}[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)],$$

where $\{\sigma_i \in \mathbb{Z}_2 = \{\pm 1\} |\, i \in [1,n]\}$ and $\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2}$ is the counting measure on $\mathbb{Z}_2$, see (6.9).

If S is distributed according to a probability measure μ^n on $\mathcal{Z}^n$, where $\mu \in \mathcal{P}_{\mathcal{Z}}$, then *the Rademacher n-complexity of a family* $\mathcal{G} \subset Meas(\mathcal{Z}, \mathbb{R})$ *w.r.t.* μ is defined by averaging the empirical Rademacher complexity

$$\mathcal{R}_{n,\mu}(\mathcal{G}) := \mathbb{E}_{\mu^n}[\hat{\mathcal{R}}_S(\mathcal{G})].$$

Let $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ be a learning model and $\mu \in \mathcal{P}_{\mathcal{Z}}$. *The Rademacher n-complexity* $\mathcal{R}_n(\mathcal{Z}, \mathcal{H}, L, \mu)$ (resp. *the Rademacher empirical n-complexity* $\hat{\mathcal{R}}_S(\mathcal{Z}, \mathcal{H}, L)$) is defined to be the complexity $\mathcal{R}_{n,\mu}(\mathcal{L}_{\mathcal{H}})$ (resp. the empirical complexity $\hat{\mathcal{R}}_S(\mathcal{L}_{\mathcal{H}})$), where $\mathcal{L}_{\mathcal{H}}$ is the family associated to the model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ by (7.5).

Remark 7.8. The supremum

$$\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)],$$

measures the correlation between $\mathcal{G}$ and the random noise $\sigma := (\sigma_1, \cdots, \sigma_n) \in \mathbb{Z}_2^n$ over the sample $S = (z_1, \cdots, z_n) \in \mathcal{Z}^n$. This describes the richness of the family $\mathcal{G}$.

Example 7.9. Let us consider a learning model $(\mathcal{X}\times\mathbb{Z}_2, \mathcal{H} \subset (\mathbb{Z}_2)^{\mathcal{X}}, L^{(0-1)}, \mu)$. For a sample $S = \{(x_1, y_1), \cdots, (x_m, y_m)\} \in (\mathcal{X} \times \mathbb{Z}_2)^m$ denote by $Pr(S)$ the sample $(x_1, \cdots, x_m) \in \mathcal{X}^m$. We shall show that

$$\hat{R}_S(\mathcal{L}_{\mathcal{H}}^{(0-1)}) = \frac{1}{2}\hat{R}_{Pr(S)}(\mathcal{H}). \tag{7.6}$$

Using the identity

$$L^{(0-1)}(x, y, h) = 1 - \delta_y^{h(x)} = \frac{1}{2}(1 - y_i h(x_i))$$

we compute

$$\hat{R}_S(\mathcal{L}_{\mathcal{H}}^{(0-1)}) = \mathbb{E}_{(\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2})^m}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i(1-\delta_{y_i}^{h(x_i)})]$$

$$= \mathbb{E}_{(\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2})^m}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \frac{1-y_ih(x_i)}{2}]$$

$$= \frac{1}{2}\mathbb{E}_{(\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2})^m}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m -\sigma_i y_i h(x_i)]$$

$$= \frac{1}{2}\mathbb{E}_{(\mu_{\mathbb{Z}_2}^{\mathbb{Z}_2})^m}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)] = \frac{1}{2}\hat{R}_{Pr(S)}(\mathcal{H})$$

which is required to prove.

We have the following relation between the empirical Rademacher complexity and the Rademacher complexity, using the McDiarmid concentration inequality, see (B.7) and [MRT2012, (3.14), p.36]

$$\mu^n\{S\in\mathcal{Z}^n | \mathcal{R}_{n,\mu}(\mathcal{L}_{\mathcal{H}}) \leq \mathcal{R}_S(\mathcal{L}_{\mathcal{H}}) + \sqrt{\frac{\ln(2/\delta)}{2m}}\} \geq 1-\delta/2. \tag{7.7}$$

In Theorem 7.10 below we express lower bounds for the rate of convergence of the empirical risk to the expected risk of a hypothesis $h\in\mathcal{H}$ in probability in terms of Rademacher (empirical) complexities. As a consequence we have lower bounds for the sample complexity of A_{erm}.

Theorem 7.10. *(see e.g.* [SSBD2014, Theorems 26.3, 26.5, p. 377- 378]*) Assume that $(\mathcal{Z},\mathcal{H},L,\mu)$ is a learning model with $|L(z,h)|<c$ for all $z\in\mathcal{Z}$ and all $h\in\mathcal{H}$. Then for any $\delta>0$ and any $h\in\mathcal{H}$ we have*

$$\mu^n\{S\in\mathcal{Z}^n |\, R_\mu^L(h) - R_S^L(h) \leq \mathcal{R}_{n,\mu}(\mathcal{L}_{\mathcal{H}}) + c\sqrt{\frac{2\ln(2/\delta)}{n}}\} \geq 1-\delta, \tag{7.8}$$

$$\mu^n\{S\in\mathcal{Z}^n |\, R_\mu^L(h) - R_S^L(h) \leq \mathcal{R}_S(\mathcal{L}_{\mathcal{H}}) + 4c\sqrt{\frac{2\ln(4/\delta)}{n}}\} \geq 1-\delta. \tag{7.9}$$

(7.10)

$$\mu^n\{S\in\mathcal{Z}^n | R_\mu^L(A_{erm}(S)) - R_\mu^L(h) \leq 2Rr_S(\mathcal{L}_{\mathcal{H}}) + 5c\sqrt{\frac{2\ln(8/\delta)}{\delta}}\} \geq 1-\delta.$$

$$\mathbb{E}_{\mu^n}(R_\mu^L(A_{erm}(S)) - R_{\mu,\mathcal{H}}^L) \leq 2Rr_{n,\mu}(\mathcal{L}_{\mathcal{H}}). \tag{7.11}$$

It follows from (7.11), using the Markov inequality, the following bound for the sample complexity $m_{A_{erm}}$ in terms of Rademacher complexity

$$\mu^n\{S\in\mathcal{Z}^n |\, R_\mu^L(A_{erm}) - R_{\mu,\mathcal{H}}^L \leq \frac{2\mathcal{R}_{n,\mu}^L(\mathcal{L}_{\mathcal{H}})}{\delta}\} \geq 1-\delta. \tag{7.12}$$

Remark 7.11. (1) The first two assertions of Theorem 7.10 give an upper bound of a "half" of the sample complexity $m_{\mathcal{H}}$ of a unified learning model $(\mathcal{Z}, \mathcal{H}, L, \mu)$ by the (empirical) Rademacher complexity $\mathcal{R}_n(\mathcal{L}_{\mathcal{H}})$ of the associated family $\mathcal{L}_{\mathcal{H}}$. The last assertion of Theorem 7.10 is derived from the second assertion and the Hoeffding inequality.

(2) For the binary classification problem $(\mathcal{X}\times\{0,1\}, \mathcal{H} \subset \{0,1\}^{\mathcal{X}}, L^{(0-1)}, \mathcal{P}(\mathcal{X}\times\{0,1\})$ there exists a close relationship between the Rademacher complexity and the growth function $\Gamma_{\mathcal{H}}(m)$, see [MRT2012, Lemma 3.1, Theorem 3.2, p. 37] for detailed discussion.

7.3. **Model selection.** Up to now we assume that a learning model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$ is given and we want to know when a learning algorithm A has a generalization ability, i.e., it is consistent uniformly w.r.t. the statistical model $\mathcal{P}_{\mathcal{Z}}$. The generalization ability of A is measured in sample complexity of A. If A is ERM then the sample complexity of A is upper bounded by by sample complexity of $\mathcal{H}$.

In practice, we are given a learning task and have to approach it with a choice of a hypothesis class $\mathcal{H}$, and assuming the loss function is chosen, therefore we have the learning model $(\mathcal{Z}, \mathcal{H}, L, \mathcal{P}_{\mathcal{Z}})$, where $\mathcal{P}_{\mathcal{Z}}$ should be in correlation with the choice of $\mathcal{H}$.

There are many possible choices of $\mathcal{H}$. Usually we have a preferred sets of algorithms whose output -predictors form different classes $\mathcal{H}_\alpha, \alpha \in A$ of hypothesis classes with parameter $\alpha \in A$. The choice of a right $\alpha \in A$ is called *a model selection*.

To select a right model, we need to understand dependence of the error of a learning algorithm A on the choice of hyothesis class $\mathcal{H}$.

7.3.1. *Error decomposition.* To understand the performance of a learning algorithm A we decompose the error of the predictor $h = A(S), S \in \mathcal{Z}^n$, as a sum of its estimation error (or the sample error of A [22] and the approximation error of the underlying hypothesis class $\mathcal{H}$.

We assume that the maximum domain of the expected loss function R^L_μ is a subspace $\mathcal{H}_{L,\mu} \supset \mathcal{H}$, given L and a probability measure $\mu \in \mathcal{P}$.

We define *the Bayes risk* of the learning problem R^L_μ on the maximal domain $\mathcal{H}_{L,\mu}$ as follows

$$R^L_{b,\mu} := \inf_{h\in\mathcal{H}_{L,\mu}} R^L_\mu(h)$$

Recall that $R^L_{\mu,\mathcal{H}} := \inf_{h\in\mathcal{H}} R^L_\mu(h)$ quantify the optimal performance of a learner in $\mathcal{H}$. Then we decompose the difference between the expected risk of a predictor $h \in \mathcal{H}$ and the Bayes risk as follows:

$$R^L_\mu(h) - R^L_{b,\mu} = (R^L_\mu(h) - R^L_{\mu,\mathcal{H}}) + (R^L_{\mu,\mathcal{H}} - R^L_{b,\mu}). \tag{7.13}$$

[22] If $h = A_{erm}(S)$ is a minimizer of the empirical risk $\hat{R}^L_S$, then the estimation error of $A_{erm}(S)$ is also called *the sample error* [CS2001, p. 9], cf. (7.2.

The first term in the RHS of (7.13) is called *the estimation error* of h or the sample error of the predictor $h = A(S)$, and the second term is called *the approximation error*.

The approximation error quantifies how well the hypothesis class $\mathcal{H}$ is suited for the problem under consideration. The estimation error measures how well the hypothesis h performs relative to best hypotheses in $\mathcal{H}$. Typically, the approximation error will decrease when enlarging $\mathcal{H}$ but the sample error will increase as demonstrated in No-Free-Lunch Theorem 6.9, because P should be enlarged as $\mathcal{H}$ will be enlarged.

Example 7.12. (cf. Exercise 2.11). Let us compute the error decomposition of a discriminative model $(\mathcal{X} \times \mathbb{R}, \mathcal{H} \subset \mathbb{R}^{\mathcal{X}}, L_2, \mathcal{P}_B(\mathcal{X} \times \mathbb{R})$ for regression problem. Let $\pi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$ denote the natural projection. Then for any $\rho \in \mathcal{P}_B(\mathcal{X} \times \mathbb{R})$, the push-forward measure $\pi_*(\rho) \in \mathcal{P}(\mathcal{X})$ and the conditional probability measure $\rho(y|x)$ on each fiber $\pi^{-1}(x) = \mathbb{R}$ are related via the Disintegration Theorem (Theorem A.12)

$$\int_{\mathcal{X}\times\mathbb{R}} \varphi(x,y)d\rho = \int_{\mathcal{X}} (\int_{\mathbb{R}} \varphi(x,y)d\rho(y|x))d\pi_*(\rho)$$

for $\varphi \in L^1(\rho)$.

Let us compute the Bayes risk $R^L_{b,\rho}$ for $L = L_2$. The maximal subspace $\mathcal{H}_{L_2,\rho}$ where the expected loss R^L_ρ is well defined is the space $L^2(\mathcal{X}, \pi_*(\rho))$. We claim that the regression function of ρ

$$r_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x)$$

minimizes the MSE_ρ defined on the hypothesis space $\mathcal{H}_{L_2,\rho} = L^2(\mathcal{X} \times \mathbb{R}, \pi_(\rho))$. Indeed, for any $f \in L^2(\mathcal{X}, \pi_*(\rho))$ by the Disintegration Theorem we have

$$\begin{aligned} MSE_\rho(f) = \int_{\mathcal{X}\times\mathbb{R}} \|f(x) - y\|^2 d\rho = \int_{\mathcal{X}} \|f(x) - r_\rho(x)\|^2 d\pi_*(\rho) \\ + \int_{\mathcal{X}} \int_{\mathbb{R}} |y - r_\rho(x)|^2 d\rho(y|x) d\pi_*(\rho). \end{aligned} \tag{7.14}$$

The last term in (7.14) is $MSE_\rho(r_\rho)$. Hence the Bayes risk $R^L_{b,\rho}$ of the class $\mathcal{H}_{L_2,\rho} = L^2(\mathcal{X} \times \mathbb{R}, \pi_*(\rho))$ is $MSE_\rho(r_\rho)$.

Now assume that $\mathcal{H} \subset \mathcal{H}_{L,\rho}$ is a compact andf_{min} is a minimzer of MSE_ρ in $\mathcal{H}$. Then we have the following decomposition of the error for $h \in \mathcal{H}$

$$MSE_\rho(h) - MSE_\rho(r_\rho) = \|h - f_{min}\|_{L_2(\mathcal{X},\pi_*(\rho))} + \|f_{min} - r_\rho\|_{L^2(\mathcal{X},\pi_*(\rho))}.$$

7.3.2. *Validation and cross-validation.* An important empirical approach in model selection is validation and its refinement - (k-fold) cross-validation. Validation is used for model selection as follows. We first train different algorithms (or the same algorithm with different parameters) on the given training set S. Let $\mathcal{H} := \{h_1, \cdots, h_r\}$ be the set of all output predictors of the different algorithms. Now, to choose a single predictor from $\mathcal{H}$ we sample

a fresh validation set $S' \in \mathcal{Z}^m$ and choose the predictor h_i that minimizes the error over the validation set.

The basic idea of *cross-validation* is to partition the training set $S = (z_1, \cdots, z_n) \in \mathcal{Z}^n$ into two sets $S = S_1 \cup S_2$ where $S_1 \in \mathcal{Z}^k$ and $S_2 \in \mathcal{Z}^{n-k}$. The set S_1 is used for training each of the candidate models, and the second set S_2 is used for deciding which of them yields the best results.

The *n-cross validation* is a refinement of cross-validation by partition of the training set into n-subsets and use one of them for testing the and repeat the procedure $(n-1)$-time for other testing subsets.

7.4. **Undecidability of learnability.** Kurt Gödel proved in 1940 that the negation of the continuum hypothesis, i.e., the existence of a set with intermediate cardinality, could not be proved in standard set theory. The second half of the independence of the continuum hypothesis i.e., unprovability of the nonexistence of an intermediate-sized set was proved in 1963 by Paul Cohen. Thus Gödel and Cohen showed that not every thing is provable. Recently, using similar ideas, S. Ben-David, P. Hrubes, S. Moran, A. Shpilka and A. Yehudayoff showed that learnability can be undecidable [BHMSY2019]. Specifically, they consider a learning problem called *estimating the maximum (EMX)* for a triple $(\mathcal{X}, \mathcal{F}, \mathcal{P}_\mathcal{X})$ is formulated as follows. Given a family $\mathcal{F} \subset \Sigma_\mathcal{X}$ of measurable subsets of some sample space $\mathcal{X}$, find an element in $\mathcal{F}$ whose measure w.r.t. an unknown probability distribution $\mu \in \mathcal{P}_\mathcal{X}$ is close to maximal. This should be done based on a finite sample generated i.i.d. from μ. (A solution of such an EMX problem can be used for advertisements firms who want to pose ad to the largest cluster of a population divided by a common interest.) Now let $\mathcal{F}$ be the family of all finite subsets in the interval $[0,1] = \mathcal{X}$ and $\mathcal{P}_\mathcal{X}$ consists of all probability measures with finite support in $\mathcal{X}$. They show that the standard axioms of mathematics cannot be used to prove that we can solve the EMX problem for given $(\mathcal{X}, \mathcal{F}, \mathcal{P}_\mathcal{X})$ nor they can be used to prove that we cannot solve this problem. In other words, they show that the learnability statement is independent of the ZFC axioms (ZermeloFraenkel set theory with the axiom of choice).

7.5. **Conclusion.** In this lecture we use several complexities of a learning model $(\mathcal{Z}, \mathcal{H}, L, P)$ for obtaining upper bounds of the sample complexity of the ERM algorithm A_{erm}. Among them Rademacher complexities are the most sophisticated ones that measure the capacity of a hypothesis class on a specific sample, which can be used to bound the difference between empirical and expected error, and thus the excess generalization error of empirical risk minimization. To find an ideal hypothesis class $\mathcal{H}$ for a learning problem we have to take into account the error decomposition of a learning model and the resulting bias-variance trade-off and use empirical cross validation methods. Finally the learnability of a learning problem can be undecidable.

8. Support vector machines

In this lecture we shall consider a class of simple supervised learning machines for binary classification problems and apply results in the previous lectures on consistent learning algorithms. Our learning machines are $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}(V), L, \mathcal{P}(V \times \mathbb{Z}_2), A)$ where V is a real Hilbert space, $\mathcal{H}_{lin}(V)$ consists of linear classifiers on V, defined below, L is the $(0-1)$ loss function (resp. a regularized $(0-1)$ loss function) and A is a hard SVM algorithm (resp. a soft SVM algorithm), which we shall learn in today lecture. The original SVM algorithm is the hard SVM algorithm, which was invented by Vapnik and Chervonenkis in 1963. The current standard incarnation (soft margin) was proposed by Cortes and Vapnik in 1993 and published in 1995.

8.1. **Linear classifier and hard SVM.** For $(w, b) \in V \times \mathbb{R}$ we set

$$f_{(w,b)}(x) := \langle w, x\rangle + b. \tag{8.1}$$

Definition 8.1. *A linear classifier* is a function $\operatorname{sign} f_{(w,b)} : V \to \mathbb{Z}_2$, $x \mapsto \operatorname{sign} f_{(w,b)}(x) \in \{-1, 1\} = \mathbb{Z}_2$.

We identify each linear classifier with the half space $H^+_{(w,b)} = (\operatorname{sign} f_{(w,b)})^{-1}(1) = f^{-1}_{(w,b)}(\mathbb{R}_{\geq 0}) \subset V$ and set $H_{(w,b)} := f^{-1}_{(w,b)}(0) \subset V$. Note that each hyperplane $H_{(w,b)} \subset V$ defines $H^+_{(w,b)}$ up to a reflection of V around $H_{(w,b)}$ and therefore defines the affine function $f_{(w,b)}$ up to a multiplicative factor $\lambda \in \mathbb{R}^*$. Let

$$\mathcal{H}_A(V) := \{H_{(w,b)} \subset V |\, (w, b) \in V \times \mathbb{R}\}$$

be the set of of all hyperplanes in the affine space V. Then $\mathcal{H}_{lin}(V)$ is a double cover of $\mathcal{H}_A(V)$ with the natural projection $\pi : \mathcal{H}_{lin}(V) \to \mathcal{H}_A(V)$ defined above.

Definition 8.2. A training sample $S = (x_1, y_1), \cdots, (x_m, y_m) \in (V \times \{\pm 1\})^m$ is called *separable*, if there is a half space $H^+_{(w,b)} \subset V$ that correctly classifies S, i.e. for all $i \in [1, m]$ we have $x_i \in H^+_{(w,b)}$ iff $y_i = 1$.

Remark 8.3. (1) A half space $H^+_{(w,b)}$ correctly classifies S if and only if the empirical risk function $\hat{R}^{(0-1)}_S(\operatorname{sign} f_{(w,b)}) = 0$.

(2) Write $S = S_+ \cup S_-$ where

$$S_\pm := \{(x, y) \in S |\, y = \pm 1\}.$$

Let $Pr : (V \times \{\pm 1\})^m \to V^m$ denote the canonical projection. Then S is separable if and only if there exists a hyper-plane $H_{(w,b)}$ that separates $[\mathrm{Pr}(S_+)]$ and $[Pr(S_-)]$, where recall that $[(x_1, \cdots, x_m)] = \cup_{i=1}^m \{x_i\} \subset V$. In this case we say that $H_{(w,b)}$ *correctly separates* S.

(3) If a training sample S is separable then the separating hyperplane is not unique, and hence there are many minimizers of the empirical risk function $\hat{R}^{(0-1)}_S$. Thus, given S, we need to find a strategy for selecting one of these ERM's, or equivalently for selecting a separating hyperplane $H_{(w,b)}$,

since the associated half-space $H^+_{(w,b)}$ is defined by $H_{(w,b)}$ and any training value (x_i, y_i). The standard approach in the SVM framework is to choose $H_{(w,b)}$ that maximizes the distance to the closest points $x_i \in [Pr(S)]$. This approach is called *the hard SVM rule*. To formulate the hard SVM rule we need a formula for the distance of a point to a hyperplane $H_{(w,b)}$.

Lemma 8.4 (Distance to a hyperplane)**.** *Let V be a real Hilbert space and $H_{(w,b)} := \{z \in V | \langle z, w\rangle + b = 0\}$. The distance of a point $x \in V$ to $H_{(w,b)}$ is given by*

$$\rho(x, H_{(w,b)}) := \inf_{z \in H_{(w,b)}} ||x - z|| = \frac{|\langle x, w\rangle + b|}{||w||}. \tag{8.2}$$

Proof. Since $H_{(w,b)} = H_{(w,b)/\lambda}$ for all $\lambda > 0$, it suffices to prove (8.2) for the case $||w|| = 1$ and hence we can assume that $w = e_1$. Now formula (8.2) follows immediately, noting that $H_{(e_1,b)} = H_{(e_1,0)} - be_1$. □

Let $H_{(w,b)}$ separate $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ correctly. Then we have

$$y_i = sign(\langle x_i, w\rangle + b),$$
$$\Longrightarrow |\langle x_i, w\rangle + b| = y_i(\langle x, w\rangle + b).$$

Hence, by Lemma 8.4, the distance between $H_{(w,b)}$ and S is

$$\rho(S, H_{(w,b)}) := \min_i \rho(x_i, H_{(w,b)}) = \frac{\min_i y_i(\langle x_i, w\rangle + b)}{||w||}. \tag{8.3}$$

The distance $\rho(S, H_{(w,b)})$ is also called *the margin of a hyperplane $H_{(w,b)}$ w.r.t. S*. The hyperplanes, that are parallel to the separating hyperplane and passing through the closest points on the negative or positive sides are called *marginal*.

Set

$$\mathcal{H}_S := \{H_{(w,b)} \in \pi(\mathcal{H}_{lin}) = \mathcal{H}_A(V) | \, H_{(w,b)} \text{ separates } S \text{ correctly } \},$$

$$A^*_{hs}(S) := arg \max_{H_{(w,b)'} \in \mathcal{H}_S} \rho(S, H_{(w,b)'}), \tag{8.4}$$

$$A_{hs} : \cup_m (V \times \mathbb{Z}_2)^m \to \mathcal{H}_{lin},$$
$$A_{hs}(S) \in \pi^{-1}(A^*_{hs}(S)) \text{ and } A_{hs}(S) \cap \{(x, -1) | \, x \in V\} = \emptyset.$$

Definition 8.5. *A hard SVM* is a learning machine $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}(V), L^{(0-1)}, \mathcal{P}(V \times \mathbb{Z}_2), A_{hs})$.

The domain of the optimization problem in (8.4) is $\mathcal{H}_S$, which is not easy to determine. So we replace this problem by another optimization problem over a larger convex domain as follows.

Lemma 8.6. *For $S = \{(x_1, y_1), \cdots, (x_m, y_m)\}$ we have*

$$A_{hs}(S) = H^+_{(w,b)} \text{ where } (w, b) = arg \max_{(w,b):||w|| \leq 1} \min_i y_i(\langle w, x_i\rangle + b). \tag{8.5}$$

Proof. If $H_{(w,b)}$ separates S then $\rho(S, H_{(w,b)}) = \min_i y_i(\langle w, x_i\rangle + b)$. Since the constraint $||w|| \leq 1$ does not effect on $H_{(w,b)}$, which is invariant under a positive rescaling, (8.3) implies that

$$\max_{(w,b):||w||\leq 1} \min_i y_i(\langle w, x_i\rangle + b) \geq \max_{H_{(w,b)'}\in\mathcal{H}_S} \rho(S, H_{(w,b)'}). \tag{8.6}$$

Next we observe that if $H_{(w,b)} \notin \mathcal{H}_S$ then

$$\min_i y_i(\langle w, x_i\rangle + b) < 0.$$

Combining this with (8.6) we obtain

$$\max_{(w,b):||w||\leq 1} \min_i y_i(\langle w, x_i\rangle + b) = \max_{(w,b):H_{(w,b)}\in\mathcal{H}_S} \min_i y_i(\langle w, x_i\rangle + b).$$

This completes the proof of Lemma 8.6. □

A solution $A_{hs}(S)$ of the equation (8.5) maximizes the enumerator of the far RHS of (8.3) under the constraint $||w|| \leq 1$. In the Proposition below we shall show that $A_{hs}(S)$ can be found as a solution to the dual optimization problem of minimizing the dominator of the RHS (8.3) under the constraint that the enumerator of the far RHS of (8.3) has to be fixed.

Proposition 8.7. *A solution to the following optimization problem, which is called Hard-SVM,*

$$(w_0, b_0) = arg\min_{w,b}\{||w||^2 : \ y_i(\langle w, x_i\rangle + b) \geq 1 \text{ for all } i\} \tag{8.7}$$

produces a solution $(w, b) := (w_0/||w_0||, b_0/||w_0||)$ *of the optimization problem (8.5).*

Proof. Let (w_0, b_0) be a solution of (8.7). We shall show that $(w_0/||w_0||, b_0/||w_0||)$ is a solution of (8.5). It suffices to show that the margin of the hyperplane $H_{(w_0,b_0)}$ is greater than or equal to the margin of the hyperplane associated to a (and hence any) solution of (8.5).

Let (w^*, b^*) be a solution of Equation (8.5). Set

$$\gamma^* := \min_i y_i(\langle w^*, x_i\rangle + b^*)$$

which is the margin of the hyperplane $H_{(w^*,b^*)}$ by (8.3). Therefore for all i we have

$$y_i(\langle w^*, x_i\rangle + b^*) \geq \gamma^*$$

or equivalently

$$y_i^*(\langle \frac{w^*}{\gamma^*}, x_i\rangle + \frac{b^*}{\gamma^*}) \geq 1.$$

Hence the pair $(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*})$ satisfies the condition of the quadratic optimization problem in (8.7). It follows that

$$||w_0|| \leq ||\frac{w^*}{\gamma^*}|| = \frac{1}{\gamma^*}.$$

Hence for all i we have

$$y_i(\frac{w_0}{||w_0||} + \frac{b_0}{||w_0||}) = \frac{y_i(\langle w_0, x_i\rangle + b_0)}{||w_0||} \geq \frac{1}{||w_0||} \geq \gamma^*.$$

This implies that the margin of $H_{(w_0,b_0)}$ satisfies the required condition. This completes the proof of Proposition 8.7. □

Remark 8.8. (1) The optimization problem of (8.4) is a specific instance of quadratic programming (QP), a family of problems extensively studied in optimization. A variety of commercial and open-source solvers are available for solving convex QP problems. It is well-known that there is a unique solution of (8.4).

(2) In practice, when we have a sample set S of large size, then S is not separable, thus the application of hard SVM is limited.

Exercise 8.9. (1) Show that the vector w_0 of the solution (w_0, b_0) in (8.7) of the SVM problem is a linear combination of the training set vectors $x_1, \cdots, x_m$.

(2) Show that x_i lies on the marginal hyperplanes $\langle w_0, x\rangle + b_0 = \pm 1$.

A vector x_i appears in the linear expansion of the weight vector w_0 in Exercise 8.9 is called *a support vector*.

8.2. **Soft SVM.** Now we consider the case when the sample set S is not separable. There are at least two possibilities to overcome this difficulty. The first one is to find a nonlinear embedding of patterns into a high-dimensional space. To realize this approach we use a kernel trick that embeds the patterns in an infinite dimensional Hilbert space space, which we shall learn in the next lecture. The second way is to seek a predictor $sign\, f_{(w,b)}$ such that $H_{(w,b)} = f^{-1}_{(w,b)}(0)$ still has maximal margin in some sense. More precisely, we shall relax the hard SVM rule (8.7) by replacing the constraint

$$y_i(\langle w, x_i\rangle + b) \geq 1 \tag{8.8}$$

by the relaxed constraint

$$y_i(\langle w, x_i\rangle) + b \geq 1 - \xi_i \tag{8.9}$$

where $\xi_i \geq 0$ are called *the slack variables*. The slack variables are commonly used in optimization to define relaxed versions of some constraints. In our case a slack variable ξ_i measures the distance by which vector x_i violates the original inequality in the LHS of (8.8).

The relaxed hard SVM rule is called *the soft SVM rule*.

Definition 8.10. *The soft SVM algorithm* $A_{ss} : (\mathbb{R}^d \times \mathbb{Z}_2)^m \to H_{lin}(\mathbb{R}^d)$ *with slack variables* $\{\xi \in \mathbb{R}^m_{\geq 0}\}$ is defined as follows

$$A_{ss}(S) = sign\, f_{(w_0,b_0)}(S)$$

where (w_0, b_0) satisfies the following equation with $\xi = (\xi_1, \cdots, \xi_m)$

$$(w_0, b_0, \xi) = arg\min_{w,b,\xi}(\lambda||w||^2 + \frac{1}{m}||\xi||_{l_1}) \tag{8.10}$$

$$\text{s. t. } \forall i,\ y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0. \tag{8.11}$$

In what follows we shall show that the soft SVM algorithm A_{ss} is a solution of a regularized loss minimization rule, which is a refinement of the ERM rule.

Digression *Regularized Loss Minimization* (RLM) is a learning algorithm on a learning model $(\mathcal{Z}, \mathcal{H}, L, P)$ in which we jointly minimize the empirical risk $\hat{R}_S^L$ and a regularization function. Formally, a regularization function is a mapping $R : \mathcal{H} \to \mathbb{R}$ and the regularized loss minimization rule is a map $A_{rlm} : \mathcal{Z}^n \to \mathcal{H}$ such that $A_{rlm}(S)$ is a minimizer of the empirical regularized loss function $\tilde{R}_S^L := \hat{R}_S^L + R : \mathcal{H} \to \mathbb{R}$.

As the ERM algorithm works under certain condition, the RLM algorithm also works under certain conditions, see e.g. [SSBD2014, Chapter 13] for detailed discussion.

The loss function for the soft SVM learning machine is the hinge loss function $L^{hinge} : \mathcal{H}_{lin}(V) \times (V \times \{\pm 1\}) \to \mathbb{R}$ defined as follows

$$L^{hinge}(h_{(w,b)}, (x, y)) := \max\{0, 1 - y(\langle w, x\rangle + b)\}. \tag{8.12}$$

Hence the empirical hinge risk function is defined as follows for $S = \{(x_1, y_1) \cdots, (x_m, y_m)\}$

$$R_S^{hinge}(h_{(w,b)}) = \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i(\langle w, x_i\rangle + b\}.$$

Lemma 8.11. *The Equation (8.10) with constraint (8.11) for A_{ss} is equivalent to the following regularized risk minimization problem, which does not depend on the slack variables ξ:*

$$A_{ss}(S) = arg\min_{f_{(w,b)}} \big(\lambda||w||^2 + R_S^{hinge}(f_{(w,b)})\big) \in \mathcal{H}_{lin}. \tag{8.13}$$

Proof. Let us fix (w_0, b_0) and minimize the RHS of (8.10) under the constraint (8.11). It is straightforward to see that $\xi_i = L^{hinge}\big((w, b), (x_i, y_i)\big)$. Using this and comparing (8.10) with (8.13), we complete the proof of Lemma 8.11. □

From Lemma 8.11 we obtain immediately the following

Corollary 8.12 (Definition)**.** *A soft SVM is a learning machine* $(V \times \mathbb{Z}_2, \mathcal{H}_{lin},$
$L^{hinge}, \mathcal{P}(V \times \mathbb{Z}_2), A_{rlm})$.

Remark 8.13. The hinge loss function L^{hinge} enjoys several good properties that justify the preference of L^{hinger} as a loss function over the zero-one loss function $L^{(0-1)}$, see [SSBD2014, Subsection 12.3, p. 167] for discussion.

8.3. **Sample complexities of SVM.**

Exercise 8.14. Prove that $VC\dim H_{lin}(V) = \dim V + 1$.

Hint. To show that any $d+1$ points S_{d+1} in $\mathbb{R}^d$ can be shattered by $\mathcal{H}_{lin}(V)$, it suffices to show that there is a haft space separating any subset in S_{d+1}. Next we show that there is a configuration S_{d+2} of $d+2$ points in $\mathbb{R}^d$ that cannot be shattered by $\mathcal{H}_{lin}(V)$, if S_{d+2} consisting of $(d+1)$ vertices of a convex set containing an interior point.

From the Fundamental Theorem of binary classification 6.16 and Exercise (8.14) we obtain immediately the following

Proposition 8.15. *The binary classification problem* $(V\times\mathbb{Z}_2, \mathcal{H}_{lin}, L^{(0-1)}, \mathcal{P}(V\times\mathbb{Z}_2))$ *ha a uniformly consistent learning algorithm if and only if* V *is finite dimensional.*

In what follow we shall show the uniform consistent of hard SVM and soft SVM replacing the statistical model $\mathcal{P}(\mathbb{R}^d \times \mathbb{Z}_2)$ by a smaller class $\mathcal{P}_{(\gamma,\rho)}(V \times \mathbb{Z}_2)$ to be defined below.

Definition 8.16. ([SSBD2014, Definition 15.3]) Let $\mu \in \mathcal{P}(V \times \mathbb{Z}_2)$. We say that μ *is separable with a* (γ, ρ)*-margin* if there exists $(w^*, b^*) \in V \times \mathbb{R}$ such that $\|w\| = 1$ and such that

$$\mu\{(x, y) \in V \times \mathbb{Z}_2 |\, y(\langle w^*, x\rangle + b^*) \geq \gamma \text{ and } ||x|| \leq \rho\} = 1.$$

Similarly, we say that μ *is separable with a* (γ, ρ)*-margin using a homogeneous half-space* if the preceding holds with a half-space defined by a vector $(w^*, 0)$.

Definition 8.16 means that the set of labeled pairs $(x, y) \in V \times \mathbb{Z}_2$ that satisfy the condition

$$y(\langle w^*, x\rangle + b^*) \geq \gamma \text{ and } ||x|| \leq \rho$$

has a full μ-measure, where $\mu \in \mathcal{P}(V \times \mathbb{Z}_2)$ is a separable measure with (γ, ρ)-margin.

Remark 8.17. (1) We regard an affine function $f_{(w,b)} : V \to \mathbb{R}$ as a linear function $f_{w'} : V' \to \mathbb{R}$ where $V' = \langle e_1\rangle_{\otimes\mathbb{R}} \oplus V$, i.e., we incorporate the bias term b of $f_{(w,b)}$ in (8.1) into the term w as an extra coordinate. More precisely we set

$$w' := be_1 + v \text{ and } x' := e_1 + x.$$

Then

$$f_{(w,b)}(x) = f_{w'}(x').$$

Note that the natural projection of the zero set $f_{w'}^{-1}(0) \subset V'$ to V is the zero set $H_{(w,b)}$ of $f_{(w,b)}$.

(2) By Remark 8.17 (1), we can always assume that a separable measure with (γ, ρ)-margin is a one that uses a homogeneous half-space by enlarging the instance space V.

Denote by $\mathcal{P}_{(\gamma,\rho)}(V \times \mathbb{Z}_2)$ the subset of $\mathcal{P}(V \times \mathbb{Z}_2)$ that consists of separable measures with a (γ, ρ)-margin using a homogeneous half-space. Using the Rademacher complexity, see [SSBD2014, Theorem 26.13, p. 384], we have the following estimate of the sample complexity of the learning machine $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}, L^{(0-1)}, P_{(\gamma,\rho)}(V \times \mathbb{Z}_2), A_{hs})$.

Theorem 8.18. *(*[SSBD2014, Theorem 15.4, p. 206]*) Let $\mu \in \mathcal{P}_{(\gamma,\rho)}(V \times \mathbb{Z}_2)$. Then we have*

$$\mu^m\{S \in (V \times \mathbb{Z}_2)^m |\, R_\mu^{(0-1)}(A_{hs}(S)) \leq \sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}\} \geq 1 - \delta.$$

Denote by $\mathcal{P}_\rho(V \times \mathbb{Z}_2)$ the set of of probability measures on $V \times \mathbb{Z}_2$ whose support lies in $B(0, \rho) \times \mathbb{Z}_2$ where $B(0, \rho)$ is the ball of radius ρ centered at the origin of V. Now we shall examine the sample complexity of the soft SVM learning machine $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}, L^{hinge}, P_\rho(V \times \mathbb{Z}_2), A_{ss})$. The following theorem is a a consequence of Lemma 8.11 and a general result on the sample complexity of RLM under certain conditions, see [SSBD2014, Corollary 13.8, p. 179].

Theorem 8.19. *(*[SSBD2014, Corollary 15.7, p. 208]*) Let $\mu \in \mathcal{P}_\rho(V \times \mathbb{Z}_2)$. Then for every $r > 0$ we have*

$$\begin{aligned} \mathbb{E}_{\mu^m}\Big(R_\mu^{(0-1)}\big(A_{ss}(S)\big)\Big) &\leq \mathbb{E}_{\mu^m}\Big(R_\mu^{hinge}\big(A_{ss}(S)\big)\Big) \\ &\leq \min_{w \in B(0,r)} R_\mu^{hinge}(h_w) + \sqrt{\frac{8\rho^2 r^2}{m}}. \end{aligned} \tag{8.14}$$

Exercise 8.20. Using the Markov inequality, derive from Theorem 8.19 an upper bound for the sample complexity of the soft SVM.

Theorem 8.19 and Exercise 8.20 imply that we can control the sample complexity of a soft SVM algorithm as a function of the norm of the underlying Hilbert space V, independently of the dimension of V. This becomes highly significant when we learn classifiers $h : V \to \mathbb{Z}_2$ via embeddings into high dimensional feature spaces.

8.4. **Conclusion.** In this section we consider two classes of learning machines for binary classification. The first class consists of learning models $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}, L^{(0-1)}, \mathcal{P}(V \times \mathbb{Z}_2))$ with finite VC-dimension iff and only if V is finite dimensional. If $\mu \in \mathcal{P}(V \times \mathbb{Z}_2)$ is separable with (γ, ρ)-margin then we can upper bound the sample complexity of the hard SVM algorithm A_{hs} for $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}, L^{(0-1)}, \mu)$ using the ration ρ/γ and Rademacher's complexity.

The second class consists of learning machines $(V \times \mathbb{Z}_2, \mathcal{H}_{lin}, L^{hinge}, \mathcal{P}_\rho(V \times \mathbb{Z}_2), A_{ss})$. The soft SVM algorithm A_{ss} is a solution of a regularized ERM and therefore we can apply here general results on sample complexity of regularized ERM algorithms.

9. Kernel based SVMs

In the previous lecture we considered the hypothesis class $\mathcal{H}_{lin}$ of linear classifiers. A linear classifier $\operatorname{sign} f_{(w,b)}$ correctly classifies a training sample $S \subset (V \times \{\pm 1\})^m$ iff the zero set $H_{(w,b)}$ of $f_{(w,b)}$ separates the subsets $[Pr(S_-)]$ and $[Pr(S_+)]$. By Radon's theorem, any set of distinct $(d+2)$ points in $\mathbb{R}^d$ can be partitioned into two subsets that cannot be separated by a hyperplane in $\mathbb{R}^d$. Thus it is reasonable to enlarge the hypothesis class $\mathcal{H}_{lin}$ by adding polynomial classifiers, or to enlarge the dimension d of the Euclidean space $\mathbb{R}^d$. However, the computational complexity of SVM with learning by polynomial embedding $\{(x, f(x)) | \, x \in \mathbb{R}^d\}$ may be computationally expensive. In this lecture we shall learn the kernel based SVMs, which are implementations of the idea to enlarge the space $\mathbb{R}^d$ containing sample points $x_1, \cdots, x_m$. The term "kernels" is used in this context to describe inner products in the feature, enlarged space. Namely we are interested in classifiers of the form

$$\operatorname{sign} \tilde{h} : \mathcal{X} \to \mathbb{Z}_2, \, \tilde{h}(x) := \langle h, \psi(x) \rangle,$$

where h is an element in a Hilbert space W, $\psi : \mathcal{X} \to W$ is a "feature map" and the kernel function $K_\varphi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined

$$K_\psi(x, y) := \langle \psi(x), \psi(y) \rangle. \tag{9.1}$$

We shall see that to solve the hard SVM optimization problem (8.7) for $h \in W$ it suffices to know K. This kernel trick requires less computational complexity than the one for learning $\psi : \mathcal{X} \to W$.

9.1. **Kernel trick.** It is known that a solution of the hard SVM equation can be expressed as a linear combination of support vectors (Exercise 8.9). If the number of the support vectors is less than the dimension of the instance space, then this expressibility as a linear combination of support vectors simplifies the search for a solution of the hard SVM equation. Below we shall show that this expressibility is a consequence of the Representer Theorem concerning solutions of a special optimization problem. The optimization problem we are interested in is of the following form:

$$w_0 = arg \min_{w \in W} \Big(f\big(\langle w, \psi(x_1) \rangle, \cdots, \langle w, \psi(x_m) \rangle\big) + R(\|w\|) \Big) \tag{9.2}$$

where $x_i \in \mathcal{X}$, w and $\psi(x_i)$ are elements of a Hilbert space W, $f : \mathbb{R}^m \to \mathbb{R}$ is an arbitrary function and $R : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically non-decreasing function. The map $\psi : \mathcal{X} \to W$ is often called *the feature map*, and W is called *the feature space*.

The following examples show that the optimization problem for the hard (resp. soft) SVM algorithm is an instance of the optimization problem (9.2).

Example 9.1. Let $S = \{(x_1, y_1), \cdots, (x_m, y_m) \in (V \times \mathbb{Z}_2)^m$.

(1) Plugging in Equation (9.2)

$$R(a) := a^2,$$

$$f(a_1, \cdots, a_m) := \begin{cases} 0 & \text{if } y_i(a_i) \geq 1 \text{ for all } i \\ \infty & \text{otherwise} \end{cases}$$

we obtain Equation (8.7) of a hard SVM for homogeneous vectors $(w, 0)$, replacing a_i by $\langle w, x_i\rangle$. The general case of non-homogeneous solutions (w, b) is reduced to the homogeneous case, using Remark 8.17.

(2) Plugging in Equation (9.2)

$$R(a) := \lambda a^2,$$

$$f(a_1, \cdots, a_m) := \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i a_i\}$$

we obtain Equation (8.13) of a soft SVM for a homogeneous solution $A_{ss}(S)$, identifying $A_{ss}(S)$ with its parameter $(w, 0)$, S with $\{(x_1, y_1) \cdots, (x_m, y_m)\}$ and replacing a_i with $\langle w, x_i\rangle$.

Theorem 9.2 (Representer Theorem)**.** *Let $\psi : \mathcal{X} \to W$ be a feature mapping from an instance space $\mathcal{X}$ to a Hilbert space W and w_0 a solution of (9.2). Then the projection of w_0 to the subspace $\langle \psi(x_1), \cdots, \psi(x_m)\rangle_{\otimes \mathbb{R}}$ in W is also a solution of (9.2).*

Proof. Assume that w_0 is a solution of (9.2). Then we can write

$$w_0 = \sum_{i=1}^{m} \alpha_i \psi(x_i) + u$$

where $\langle u, \psi(x_i)\rangle = 0$ for all i. Set $\bar{w}_0 := w_0 - u$. Then

$$\|\bar{w}_0\| \leq \|w_0\| \tag{9.3}$$

and since $\langle \bar{w}_0, \psi(x_i)\rangle = \langle w_0, \psi(x_i)\rangle$ we have

$$f\big(\langle \bar{w}_0, \psi(x_1)\rangle, \cdots, \langle \bar{w}_0, \psi(x_m)\rangle\big) = f\big(\langle w_0, \psi(x_1)\rangle, \cdots, \langle w_0, \psi(x_m)\rangle\big). \tag{9.4}$$

From (9.3), (9.4) and taking into account the monotonicity of R, we conclude that $\bar{w}_0$ is also a solution of (9.2). This completes the proof of Theorem 9.2. □

The Representer Theorem suggests that we could restrict ourselves to a search for a solution of Equation (9.2) in a finite dimensional subspace $W_1 \subset W$ generated by vectors $\psi(x_i)_{i=1}^m$. In what follows we shall describe a method to solve the minimization problem of (9.2) on W_1, which is called *the kernel trick*.

Let

• $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $K(x, x') := \langle \psi(x), \psi(x')\rangle$ be *a kernel function*,

• $G = (G_{ij}) := K(x_i, x_j)$ - *a Gram matrix*,
• $w_0 = \sum_{i=1}^m \alpha_j \psi(x_j) \in W_1$ - a solution of Equation (9.2).

Then $\alpha = (\alpha_1, \cdots, \alpha_m)$ is a solution of the following minimization problem

$$arg \min_{\alpha \in \mathbb{R}^m} f\big(\sum_{j=1}^m \alpha_j G_{j1}, \cdots, \sum_{j=1}^m \alpha_j G_{jm}\big) + R\Big(\sqrt{\sum_{i,j=1}^m \alpha_i \alpha_j G_{ji}}\Big). \tag{9.5}$$

Recall that the solution $w_0 = \sum_{i=1}^m \alpha_j \psi(x_j)$ of the hard (resp. soft) SVM optimization problem, where $(\alpha_1, \cdots, \alpha_m)$ is a solution of (9.5), produces a "nonlinear" classifier $\hat{w}_0 : \mathcal{X} \to \mathbb{Z}_2$ associated to as follows

$$\hat{w}_0(x) := sign\, w_0(x)$$

where

$$w_0(x) := \langle w_0, \psi(x) \rangle = \sum_{i=1}^m \alpha_j \langle \psi(x_j), \psi(x) \rangle = \sum_{j=1}^m \alpha_j K(x_j, x). \tag{9.6}$$

To compute (9.6) we need to know only the kernel function K and not the mapping ψ, nor the inner product $\langle , \rangle$ on the Hilbert space W.

This motivates the following question.

Problem 9.3. *Find a sufficient and necessary condition for a kernel function, also called a kernel, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that K can be written as $K(x, x') = \langle \psi(x), \psi(x') \rangle$ for a feature mapping $\psi : \mathcal{X} \to W$, where W is a real Hilbert space.*

Definition 9.4. If K satisfies the condition in Problem 9.3 we shall say that K *is generated by a (feature) mapping* ψ. The target Hilbert space is also called *a feature space*.

9.2. **PSD kernels and reproducing kernel Hilbert spaces.**

9.2.1. *Positive semi-definite kernel.*

Definition 9.5. Let $\mathcal{X}$ be an arbitrary set. A map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *positive semi-definite kernel* (PSD kernel) iff for all $x_1, \cdots, x_m$ the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive semi-definite.

Theorem 9.6. *A kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is induced by a feature map to a Hilbert space if and only if it is positive semi-definite.*

Proof. 1) First let us prove the "only if" assertion of Theorem 9.6. Assume that $K(x, x') = \langle \psi(x), \psi(x') \rangle$ for a mapping $\psi : \mathcal{X} \to W$, where W is a Hilbert space. Given m points $x_1, \cdots, x_m \in \mathcal{X}$ we consider the subspace $W_m \subset W$ generated by $\psi(x_1), \cdots, \psi(x_m)$. Using the positive definite of the inner product on W_m, we conclude that the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive semi-definite. This proves the "only if" part of Theorem 9.6

2) Now let us prove the "if" part. Assume that $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semi-definite. For each $x \in \mathcal{X}$ let $K_x \in \mathbb{R}^{\mathcal{X}}$ be the function defined by

$$K_x(y) := K(x, y).$$

Denote by

$$W := \{f \in \mathbb{R}^{\mathcal{X}} \mid f = \sum_{i=1}^{N(f)} a_i K_{x_i},\ a_i \in \mathbb{R} \text{ and } N(f) < \infty\}.$$

Then W is equipped with the following inner-product

$$\langle \sum_i \alpha_i K_{x_i}, \sum_j \beta_j K_{y_j} \rangle := \sum_{i,j} \alpha_i \beta_j K(x_i, y_j).$$

The PSD property of K implies that the inner product is positive semi-definite, i.e.

$$\langle \sum_i \alpha_i K_{x_i}, \sum_j \alpha_i K_{x_i} \rangle \geq 0.$$

Since the inner product is positive semi-definite, the Cauchy-Schwarz inequality implies for $f \in W$ and $x \in \mathcal{X}$

$$\langle f, K_x \rangle^2 \leq \langle f, f \rangle \langle K_x, K_x \rangle. \tag{9.7}$$

Since for all x, y we have $K_y(x) = K(y, x) = \langle K_y, K_x \rangle$, it follows that for all $f \in W$ we have

$$f(x) = \langle f, K_x \rangle. \tag{9.8}$$

Using (9.8), we obtain from (9.7) for all $x \in \mathcal{X}$

$$|f(x)|^2 \leq \langle f, f \rangle K(x, x).$$

This proves that the inner product on W is positive definite. Hence W is a pre-Hilbert space. Let $\mathcal{H}$ be the completion of W. The map $x \mapsto K_x$ is the desired mapping from $\mathcal{X}$ to $\mathcal{H}$. This completes the proof of Theorem 9.6. □

Example 9.7. (1) (Polynomial kernels). Assume that P is a polynomial in one variable with non-negative coefficients. Then the polynomial kernel of the form $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $(x, y) \mapsto P(\langle x, y \rangle)$ is a PSD kernel. This follows from the observations that if $K_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $i = 1, 2$, are PDS kernel then $(K_1 + K_2)(x, y) := K_1(x, y) + K_2(x, y)$ is a PSD kernel, and $(K_1 \cdot K_2)(x, y) := K_1(x, y) \cdot K_2(x, y)$ is a PSD kernel. In particular, $K(x, y) := (1 + \langle x, y \rangle)^2$ is a PSD kernel.

(2) (Exponential kernel). For any $\gamma > 0$ the kernel $K(x, y) := \exp\big(\gamma \cdot \langle x, y \rangle\big)$ is a PSD kernel, since it is the limit of a polynomials in $\langle x, y \rangle$ with non-negative coefficients.

Exercise 9.8. (1) Show that the Gaussian kernel $K(x, y) := \exp\big(-\frac{\gamma}{2}||x - y||^2\big)$ is a PSD kernel.

(2) Let $\mathcal{X} = B(0, 1)$ - the open ball of radius 1 centered at the origin $0 \in \mathbb{R}^d$. Show that $K(x, y) := (1 - \langle x, y \rangle)^{-p}$ is a PSD kernel for any $p \in \mathbb{N}^+$.

9.2.2. *Reproducing kernel Hilbert space.* Given a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exist many feature maps $\varphi : \mathcal{X} \to W$ such that K is generated by a feature map $\varphi : \mathcal{X} \to W$. Indeed, if $K(x,x) = \langle \varphi(x), \varphi(x) \rangle$ then $K(x,x) = \langle e \circ \varphi(x), e \circ \varphi(x) \rangle$ for any isometric embedding $e : W \to W'$. However, there is a canonical choice for the feature space W, a so-called reproducing kernel Hilbert space.

Definition 9.9 (Reproducing kernel Hilbert space)**.** Let $\mathcal{X}$ be an instance set and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ a real Hilbert space of functions on $\mathcal{X}$ with the unique vector space structure such that for $x \in \mathcal{X}$ the evaluation map

$$ev_x : \mathcal{H} \to \mathbb{R}, ev_x(f) := f(x)$$

is a linear map. [23] Then $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS) on $\mathcal{X}$ if for all $x \in \mathcal{X}$ the linear map ev_x is bounded i.e.,

$$\sup_{f \in B(0,1) \subset \mathcal{H}} ev_x(f) < \infty.$$

Remark 9.10. Let $\mathcal{H}$ be a RKHS on $\mathcal{X}$ and $x \in \mathcal{X}$. Since ev_x is bounded, by the Riesz representation theorem, there is a function $k_x \in \mathcal{H}$ so that $f(x) = \langle f, k_x \rangle$ for all $f \in \mathcal{H}$. Then the kernel

$$K(x,y) := \langle k_x, k_y \rangle$$

is a PSD kernel. K is called *the reproducing kernel of* $\mathcal{H}$.

Thus every RKHS $\mathcal{H}$ on $\mathcal{X}$ produces a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conversely, Theorem 9.11 below asserts that every PSD kernel reproduces a RKHS $\mathcal{H}$.

Theorem 9.11. *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PSD kernel. There there exists a unique RKHS $\mathcal{H}(K)$ such that K is the reproducing kernel of $\mathcal{H}(K)$.*

Proof. Given a PSD kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, for any $x \in \mathcal{X}$ let us define a function K_x on $\mathcal{X}$ by

$$K_x(y) := K(x,y) \text{ for all } y \in \mathcal{X}.$$

Let us denote by $\mathcal{H}_0(K)$ the pre-Hilbert space whose elements are finite linear combination of functions K_x and the inner product on $\mathcal{H}_0(K)$ is defined on the generating elements K_x as follows

$$\forall x, y \in \mathcal{X} \text{ we have } \langle K_x, K_y \rangle := K(x,y). \tag{9.9}$$

From (9.9) we conclude that the evaluation map $ev_x : \mathcal{H}_0(K) \to \mathbb{R}, ev_x(K_y) = K(y,x)$ is a linear bounded map for all x, y, since

$$||ev_x|| = \max_{||K_y||=1} ev_x(K_y) = \max_{||K_y||=1} K(y,x) \leq \sqrt{K(x,x)}.$$

[23] In other words, the vector structure on $\mathcal{H}$ is induced from the vector structure on $\mathbb{R}$ via the evaluation map.

Hence the evaluation map ev_x extends to a bounded linear map on the completion $\mathcal{H}(K)$ of $\mathcal{H}_0(K)$. It follows that $\mathcal{H}(K)$ is a RKHS.

To show the uniqueness of a RKHS $\mathcal{H}$ such that K is the reproducing kernel of $\mathcal{H}$, we assume that there exists another RKHS $\mathcal{H}'$ such that for all $x, y \in \mathcal{X}$ there exist $k_x, k_y \in \mathcal{H}'$ with the following properties

$$K(x,y) = \langle k_x, k_y\rangle \text{ and } f(x) = \langle f, k_x\rangle \text{ for all } f \in \mathcal{H}.$$

We define a map $g : \mathcal{H}(K) \to \mathcal{H}'$ by setting $g(K_x) = k_x$. It is not hard to see that g is an isometric embedding. To show that g extends to an isometry it suffices to show that the set k_x is dense in $\mathcal{H}'$. Assume the opposite, i.e. there exists $f \in \mathcal{H}'$ such that $\langle f, k_x\rangle = 0$ for all x. But this implies that $f(x) = 0$ for all x and hence $f = 0$. This completes the proof of Theorem 9.11. □

9.3. **Kernel based SVMs and their generalization ability.**

9.3.1. *Kernel based SVMs.* Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PSD kernel. Denote by $\mathcal{H}(K)$ the RHKS of functions on $\mathcal{X}$ that produces K. Each function $h \in \mathcal{H}(K)$ defines a binary classifier

$$\operatorname{sign} h : \mathcal{X} \to \mathbb{Z}_2.$$

Denote by K_{lin} the set of all binary classifiers $\operatorname{sign} h$ where $h \in \mathcal{H}(K)$. Using the Representer Theorem 9.2 and Example 9.1 (1), we replace the algorithm A_{hs} of a hard SVM by a kernel based algorithm.

Definition 9.12. *A kernel based hard SVM* is a learning machine $(\mathcal{X}\times\mathbb{Z}_2, K_{lin}, L^{(0-1)}, \mathcal{P}(\mathcal{X}\times\mathbb{Z}_2), A_{hk})$, where for $S = \{(x_1, y_1), \cdots, (x_m, y_m)\} \in (\mathcal{X} \times \mathbb{Z}_2)^m$ and $x \in \mathcal{X}$ we have

$$A_{hk}(S)(x) = \operatorname{sign} \sum_{i=1}^{m} \alpha_j K(x_j, x) \in \mathbb{Z}_2.$$

Here $\alpha := (\alpha_1, \cdots, \alpha_m)$ is the solution of the following optimization problem (9.10)

$$\alpha = arg\min\Big(f\big(\sum_{j=1}^{m} \alpha_j K(x_j, x_1), \cdots, \sum_{j=1}^{m} \alpha_j K(x_j, x_m)\big) + R\big(\sqrt{\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j)}\big)\Big),$$

where R and f are defined in Example 9.1(1).

Similarly, using the Representer Theorem 9.2 and Example 9.1(2)s, we replace the algorithm A_{ss} of a soft SVM by a kernel based algorithm.

Definition 9.13. *A kernel based soft SVM* is a learning machine $(\mathcal{X}\times\mathbb{Z}_2, K_{lin}, L^{hinge}, \mathcal{P}(\mathcal{X}\times\mathbb{Z}_2), A_{sk})$, where for $S = \{(x_1, y_1), \cdots, (x_m, y_m)\} \in (\mathcal{X} \times \mathbb{Z}_2)^m$ and $x \in \mathcal{X}$ we have

$$A_{sk}(S)(x) = \operatorname{sign} \sum_{i=1}^{m} \alpha_j K(x_j, x) \in \mathbb{Z}_2.$$

Here $\alpha := (\alpha_1, \cdots, \alpha_m)$ is the solution of the following optimization problem (9.11)

$$\alpha = arg\min\Big(f(\sum_{j=1}^{m}\alpha_j K(x_j, x_1), \cdots, \sum_{j=1}^{m}\alpha_j K(x_j, x_m)) + R(\sqrt{\sum_{i,j=1}^{m}\alpha_i\alpha_j K(x_i, x_j)})\Big),$$

where R and f are defined in Example 9.1(2).

9.3.2. *Generalization ability of kernel based SVMs.*

• The advantage of working with kernels rather than directly optimizing in the feature space is that in some situations the dimension of the feature space is extremely large while implementing the kernel function is very simple. Moreover, in many cases, the computational time complexity of solving (9.10) is a polynomial on the variable of the size of x_i, $i \in [1, m]$, see [SSBD2014, p. 221-223].

• The upper bound for the sample complexity of a s hard SVM in Theorem 8.18 is also valid for the sample complexity of the kernel based hard SVM [SSBD2014, Theorem 26.3, p. 384] after translating the condition of separability with (γ, ρ)-margin of a measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathbb{Z}_2)$ in terms of the kernel function.

• The upper bound for the sample complexity of soft SVM in Theorem 8.19 is also valid for the sample complexity of the kernel based soft SVM, after translating the support condition a measure $\mu \in \mathcal{P}(\mathcal{X} \times \mathbb{Z}_2)$ in terms of the kernel function.

9.4. **Conclusion.** In this section we learn the kernel trick, which simplifies the algorithm of solving hard SVM and soft SVM optimization problem, using embedding of patterns into a Hilbert space. The kernel trick is based on the theory of RKHS and has many applications, e.g., for defining a feature map $\varphi : \mathcal{P}(\mathcal{X}) \to V$, where V is a RHKS, see e.g. [MFSS2017]. The main difficulty of the kernel method is that we still have no general method of selecting a suitable kernel for a concrete problem. Another open problem is to improve the upper bound for sample complexity of SVM algorithm, i.e., to find new conditions on $\mu \in P$ such that the sample complexity of A_{hk}, A_{sk} which is computed w.r.t. μ is bounded.

10. Neural networks

In the last lecture we examined kernel based SVMs which are generalizations of linear classifiers sign $f_{(w,b)}$. Today we shall study further generalizations of linear classifiers which are artificial neural networks, shortened as neural networks. Neural networks are inspired by the structure of neural networks in the brain. The idea behind neural networks is that many neurons can be joined together by communication links to carry out complex computations.

A neural network is a model of computation. Therefore, a neural network also denotes a hypothesis class of a learning model consisting of functions realizable by a neural network.

In today lecture we shall investigate several types of neural networks, their expressive power, i.e., the class of functions that can be realized as elements in a hypothesis class of a neural network. In the next lecture we shall discuss the current learning algorithm for learning machines whose hypothesis class is a neural network.

10.1. **Neural networks as a model of computation.** A neural network has a graphical representation for multivariate functions of multi-variables $h_{V,E,\sigma,w} : \mathbb{R}^n \to \mathbb{R}^m$ (in the case of a feedforward neural network), or a sequence of multivariate functions of multi-variables $\{h^i_{V,E,\sigma,w} : \mathbb{R}^n \to \mathbb{R}^m | \, i \in \mathbb{N}\}$ (in the case of a recurrent neural network).

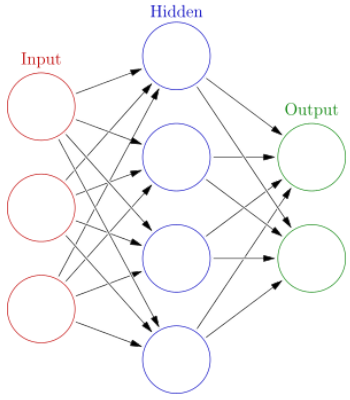
• The quadruple (V, E, σ, w) consists of

+ *the network graph* (V, E), also called *the underlying graph of the network*, where V is the set of nodes $\mathfrak{n}$, also called *neurons*, and E is the set of directed edges connecting nodes V.

+ a family of functions $\sigma_\mathfrak{n} : \mathbb{R} \to \mathbb{R}$, also called *the activation function of neuron* $\mathfrak{n}$. Usually $\sigma_\mathfrak{n} = \sigma$ is independent of $\mathfrak{n}$. Most common activation functions are:
- the sign function, $\sigma(x) = \text{sign}\,(x)$,
- the threshold function, $\sigma(x) = 1_{\mathbb{R}^+}(x)$,
- the sigmoid function $\sigma(x) := \frac{1}{1+e^{-x}}$, which is a smooth approximation to the threshold function;

+ $w : E \to \mathbb{R}$ - *the weight function of the network.*

• *The networks architecture* of a neural network is the triple $G = (V, E, \sigma)$.



• A weight $w : E \to \mathbb{R}$ endows each neuron $\mathfrak{n}$ with a computing instruction of type " input-output". *The input* $I(\mathfrak{n})$ *of a neuron* $\mathfrak{n}$ is equal to the weighted sum of the outputs of all the neurons connected to it: $I(\mathfrak{n}) = \sum w(\mathfrak{n}'\mathfrak{n})O(\mathfrak{n}')$, where $\mathfrak{n}'\mathfrak{n} \in E$ is a directed edge and $O(\mathfrak{n}')$ is the output of the neuron $\mathfrak{n}'$ in the network.

• *The output* $O(\mathfrak{n})$ *of a neuron* $\mathfrak{n}$ is obtained from the input $I(\mathfrak{n})$ as follows: $O(\mathfrak{n}) = \sigma(I(\mathfrak{n}))$.

• *The i-th input nodes* give the output x_i. If the input space is $\mathbb{R}^n$ then we have $n + 1$ input-nodes, one of them is the "constant" neuron, whose output is 1.

• There is a neuron in the hidden layer that has no incoming edges. This neuron will output the constant $\sigma(0)$.

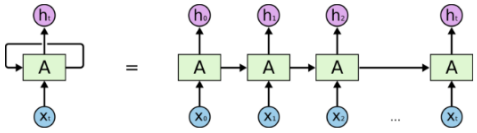
• *A feedforward neural network* (FNN) has underlying acyclic directed graph. Each FNN (E, V, w, σ) represents a multivariate multivariable function $h_{V,E,\sigma,w} : \mathbb{R}^n \to \mathbb{R}^m$ which is obtained by composing the computing instruction of each neuron on directed paths from input neurons to output neurons. For each architecture (V, E, σ) of a FNN we denote by

$$\mathcal{H}_{V,E,\sigma} = \{h_{V,E,\sigma,w} : \ w \in \mathbb{R}^E\}$$

the underlying hypothesis class of functions from the input space to the output space of the network.

• *A recurrent neural network* (RNN) has underlying directed graph with a cycle. By unrolling cycles in a RNN in discrete time $n \in \mathbb{N}$, a RNN defines a map $r : \mathbb{N}^+ :\to \{\text{FNN}\}$ such that $[r(n)] \subset [r(n + 1)]$, where

$[r(n)]$ is the underlying graph of $r(n)$, see [GBC2016, §10.2, p. 368] and [Graves2012, §3.2, p. 22]. Thus a RNN can be regarded as a sequence of multivariate multivariable functions which serves as in a discriminative model for supervised sequence labelling.



An unrolled recurrent neural network.

Digression The goal of supervised sequence labelling is to assign sequences of labels, drawn from a label space, to sequences of input data. Denoting the set of labels by $\mathcal{Y}$, and the set of input data by $\mathcal{X}$ then $\mathcal{Y}^\infty$ is the space of all sequences of labels and $\mathcal{X}^\infty$ is the space of all sequences of input data. For example, one might wish to transcribe a sequence of acoustic features with spoken words (speech recognition), or a sequence of video frames with hand gestures (gesture recognition). If the sequences are assumed to be independent and identically distributed, we recover the basic framework of pattern classification. In practice, this assumption may not be the case.

Example 10.1. *A multilayer perceptron (MLP)* is a special FNN that has vertices arranged in a disjoint union of layers $V = \cup_{l=0}^{n} V_i$ such that every edge in E connects nodes in neighboring layers V_i, V_{i+1}. The *depth* of the MLP is m. V_0 is called *the input layer*, V_n is called *the output layer*, the other layer is called *hidden*.

(1) A MLP

$$\phi(x) = \begin{cases} 0 & \text{if } \psi(x) \leq 1/2, \\ 1 & \text{otherwise} \end{cases} \tag{10.1}$$

for

$$\psi(x) = c_0 + \sum_{i=1}^{d} c_i x^i,$$

$c_i \in \mathbb{R}$, is a perceptron (linear classifier) with the activation function sign $(\psi(x) - 1/2)$. This neural network has no hidden layer and c_i are weights of the neural network.

(2) In a feedforward neural network with one hidden layer one takes $\phi(x)$ of the form (10.1) and

$$\psi(x) = c_0 + \sum_{i=1}^{k} c_i \sigma(\psi_i(x))$$

where c_i are the weights of the network and σ is a activation function.

Remark 10.2. Neural networks are abstraction of biological neural networks, the connection weights w represent the strength of the synapses between the neurons and the activation function σ is usually an abstraction representing the rate of action potential firing in the cell. In its simplest form, this function is binary, that is, either the neuron is firing or not. We can consider activation function as a filter of relevant information.

In the remaining part of today lecture we consider only FNNs. In particular under NN's we mean FNNs.

10.2. **The expressive power of neural networks.** In this section we want to address the following

Question 10.3. *What type of functions can be implemented using neural networks.*

First we consider representations of Boolian functions by neural networks.

Proposition 10.4 (Representation of Boolean functions)**.** *(* [SSBD2014, Claim 20.1, p. 271]*) Every Boolean function $f : \mathbb{Z}_2^d \to \mathbb{Z}_2$ can be represented exactly by a feedforward neural network $\mathcal{H}_{V,E,\text{sign}}$ with a single hidden layer containing at most $2^d + 1$ neurons and with the activation function $\sigma(x) =$* sign (x)*.*

Proof. Let (V, E) be a two-layer FNN with $\#V_0 = d + 1$, $\#V_1 = 2^d + 1$, $\#V_2 = 1$ and E consist of all possible edges between adjacent layers. As before $\mathbb{Z}_2 = \{\pm 1\}$. Now let $f : \mathbb{Z}_2^d \to \mathbb{Z}_2$. Let $u_i \in f^{-1}(1) \subset \mathbb{Z}_2^d$ and $k := \#f^{-1}(1)$. Set

$$g_i(x) := sign(\langle x, u_i \rangle - d + 1).$$

Then $\{g_i |\, i \in [1, k]\}$ are linear classifiers and therefore can be implemented by the neurons in V_1. Now set

$$f(x) := sign(\sum_{i=1}^{k} g_i(x) + k - 1)$$

which is also a linear classifier. This completes the proof of Proposition 10.4 □

In general case we have the following Universal Approximation Theorem, see e.g. [Haykin2008, p. 167].

Theorem 10.5. *Let φ be a nonconstant, bounded and monotone increasing continuous function. For any $m \in \mathbb{N}$, $\varepsilon > 0$ and any function $F \in C_0([0,1]^m)$ there exists an integer $m_1 \in \mathbb{N}$ and constants a_i, b_j, w_{ij} where $i \in [1, m_1]$, $j \in [1, m]$ such that*

$$f(x_1, \cdots, x_m) := \sum_{i=1}^{m_1} \alpha_i \varphi(\sum_{j=1}^{m} w_{ij} x_j + b_i)$$

for all $(x_1, \cdots, x_m) \in [0,1]^m$ we have

$$|F(x_1, \cdots, x_m) - f(x_1, \cdots, x_m)| < \varepsilon.$$

Remark 10.6. The universal approximation theorem may be viewed as a natural extension of the Weierstrass theorem stating that any continuous function over a closed interval on the real axis can be expressed in that interval as an absolutely and uniformly convergent series of polynomials. Many researchers regard the Universal Approximation Theorem as a natural generalization of Kolomogorov's theorem [Kolmogorov1957] and Lorentz's theorem [Lorentz1976] which states that every continuous function on $[0,1]^d$ can be written as

$$f(x) = \sum_{i=1}^{2d+1} F_i(\sum_{j=1}^{d} G_{ij} x^j)$$

where G_{ij} and F_i are continuous functions. In 1989 Cybenko demonstrated rigorously for the first time that a single hidden layer is sufficient to uniformly approximate any continuous function with support in a unit hypercube [Cybenko1989].

Example 10.7 (Neural networks for regression problems)**.** In neural networks with hypothesis class $\mathcal{H}_{V,E,\sigma}$ for regression problems one often chooses the activation function σ to be the sigmoid function $\sigma(a) := (1 + e^{-a})^{-1}$, and the loss function L to be L_2, i.e., $L(x, y, h_w) := \frac{1}{2}||h_w(x) - y||^2$ for $h_w \in \mathcal{H}_{V,E,\sigma}$ and $x \in \mathcal{X} = \mathbb{R}^n$, $y \in \mathcal{Y} = \mathbb{R}^m$.

Example 10.8 (Neural networks for generative models in supervised learning)**.** In generative models of supervised learning we need to estimate the conditional distribution $p(t|x)$, where $t \in \mathbb{R}$ is the target variable in a regression problem. In many regression problems p is chosen as follows [Bishop2006, (5.12), p. 232]

$$p(t|x) = \mathcal{N}(t|y(x,w), \beta^{-1}) = \frac{\beta}{\sqrt{2\pi}} \exp \frac{-\beta}{2}(t - y(x,w)), \tag{10.2}$$

where β is unknown parameter and $y(x,w)$ is the expected value of t. Thus the learning model is of the form $(\mathcal{X}, \mathcal{H}, L, P)$ where $\mathcal{H} := \{y(t,x,w)\}$ is

parameterized by a parameter w, β, and a statistical model P is a subset of $\mathcal{P}(\mathcal{X})$, since the joint distribution $\mu(x, y)$ is completely defined by the marginal distribution $\mu_{\mathcal{X}}$ and the conditional distribution $\mu(y|x)$, see Remark 2.9.

Now assume that $X = (x_1, \cdots, x_n)$ are i.i.d. by $\mu \in P$ along with labels $(t_1, \cdots, t_n)$. Then (10.2) implies

$$-\log p(t|X, w, \beta) = -\frac{n}{2}\log\beta + \frac{n}{2}\log(2\pi) + \frac{\beta}{2}\sum_{i=1}^{n}|t_n - y(x_n, w)|^2. \tag{10.3}$$

As in the density estimation problem we want to minimize the LHS of (10.2). Leaving $\beta = const$ we minimize first the β -independent component of the loss function

$$L_S(w) = \frac{1}{2}\sum_{i-1}^{n}|y(x_n, w)^2 - t_n|^2. \tag{10.4}$$

Once we know a solution w_{ML} of the equation minimizing $L_S(w)$, the value of β can be found by the following formula

$$\frac{1}{\beta_{ML}} = \frac{1}{n}\sum_{i=1}^{n}|y(x_i, w_{ML}) - t_i|^2. \tag{10.5}$$

10.3. **Sample complexities of neural networks.** A learning model $(\mathcal{X} \times \mathcal{Y}, \mathcal{H}_{V,E,\sigma}, L, P)$ is called *a neural network* if $\mathcal{H}_{V,E,\sigma}$ is a neural network and $\mathcal{H}_{V,E,\sigma} \subset \mathcal{Y}^{\mathcal{X}}$.

10.3.1. *Neural networks for binary classification problem.* In neural networks with hypothesis class $\mathcal{H}_{V,E,\sigma}$ for binary classification problems one often choose the activation function σ to be the sign function, and the loss function L to be $L^{(0-1)}$.

Proposition 10.9. *(*[SSBD2014, Theorem 20.6, p. 274]*) Let $\mathcal{H}_{V,E,sign}$ be a MLP. The VC-dimension of $\mathcal{H}_{V,E,sign}$ is $O(|E|\log|E|)$.*

Outline of the proof Assume that $\mathcal{H} := \mathcal{H}_{V,E,sign}$ consists of exactly one perceptron. In this case Proposition 10.9 is valid since by Exercise 8.14 we have $VC\dim\mathcal{H} = |E^{in}| = O(|E|\log|E|)$, where E^{in} denotes the set of directed edges coming into the perceptron.

Next, we shall reduce the proof for the general case of a neural network to the case of a single perceptron. Let $m := VC\dim(\mathcal{H})$. Using

$$\Gamma_{\mathcal{H}}(m) = 2^m, \tag{10.6}$$

to prove Proposition 10.9, it suffices to show that

$$\Gamma_{\mathcal{H}_{V,E,\sigma}}(m) \leq (em)^{|E|}, \tag{10.7}$$

since $\log_2(em) < 4\log(E)$ by (10.7) and ((10.6)).

Let $V_0, \cdots, V_T$ be the layers of (E, V). For $t \in [1, T]$ denote by $\mathcal{H}_t$ the neural network $\mathcal{H}_{W_t, E_t, sign}$ where W_t consists of inputs neurons in V_{t-1} and output neurons in V_t and E_t consists of edges of $\mathcal{H}$ that connect V_{t-1} with V_t. Now we decompose

$$\mathcal{H} = \mathcal{H}_T \circ \cdots \circ \mathcal{H}_1. \tag{10.8}$$

Lemma 10.10 (Exercises)**.** *(1) (*[SSBD2014, Exercise 4, p. 282]*) Let $\mathcal{F}_1 \subset \mathcal{Z}^{\mathcal{X}}$ and $\mathcal{F}_2 \subset \mathcal{Y}^{\mathcal{Z}}$. Set $\mathcal{H} := \mathcal{F}_2 \circ \mathcal{F}_1$. Then $\Gamma_{\mathcal{H}}(n) \leq \Gamma_{\mathcal{F}_2}(n)\Gamma_{\mathcal{F}_1}(n)$.*

(2) ([SSBD2014, Exercise 3, p. 282]*) Let $\mathcal{F}_i$ be a set of function from $\mathcal{X}$ to $\mathcal{Y}_i$ for $i = 1, 2$. Then $\Gamma_{\mathcal{F}_1 \times \mathcal{F}_2}(n) \leq \Gamma_{\mathcal{F}_1}(n)\Gamma_{\mathcal{F}_2}(n)$.*

By Lemma 10.10 (1) we have

$$\Gamma_{\mathcal{H}}(m) \leq \Pi_{t=1}^{T} \Gamma_{\mathcal{H}_t}(m).$$

Next we observe that

$$\mathcal{H}_t = \mathcal{H}_{t,1} \times \cdots \times H_{t,|V_t|}. \tag{10.9}$$

Each neuron $\mathfrak{n}_i$ on V_t has $d_{t,i}$ heading edges presenting the number of the inputs for the linear classifier $\mathfrak{n}_i$. Hence $VC\dim \mathcal{H}_{t,i} = d_{t,i} < m - 1$. By Lemma 10.10 and by Vapnik-Chervonenski-Sauer-Lemma we have

$$\Gamma_{\mathcal{H}}(m) \leq \Pi_{t=1}^{T} \Pi_{i=1}^{|V_t|} (\frac{em}{d_{t,i}})^{d_{t,i}} < (em)^{|E|},$$

which completes the proof of Proposition 10.9.

It follows from Proposition 10.9 that the sample complexity of the ERM algorithm for $(V \times \mathbb{Z}_2, \mathcal{H}_{V,E,sign}, L^{(0-1)}, \mathcal{P}(V \times \mathbb{Z}_2))$ is finite. But the running time for ERM algorithm in a neural network $\mathcal{H}_{V,E,sign}$ is non-polynomial and therefore it is impractical to use it [SSBD2014, Theorem 20.7, p. 276]. The solution is to use the stochastic gradient descend, which we shall learn in the next lecture.

10.4. **Conclusion.** In this lecture we considered learning machines whose hypothesis class consisted of functions or sequence of functions that can be graphical represented by neural networks. Neural networks have good expressive power and finite VC-dimension in binary classification problems but the ERM algorithm in these networks has very high computational complexity and therefore they are unpractical.

11. Training neural networks

Training a neural network is a popular name for running a learning algorithm in a neural network learning model. We consider in this lecture only the case where the input space and the output space of a network are Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Our learning model is of the form $(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{H}_{V,E,\sigma}, L, P)$ and the learning algorithm is stochastic gradient descend (SGD), which aims to find a minimizer of the expected risk function

$R_\mu^L : \mathcal{H}_{V,E,\sigma} \to \mathbb{R}$. Since $\mathcal{H}_{V,E,\sigma}$ is parameterized by the weight functions $w \in \mathbb{R}^E \cong \mathbb{R}^{|E|}$, we regard R_μ^L as a function of variable w on $\mathbb{R}^N$, where $N = |E|$. We begin with classical (deterministic) gradient and subgradient descend of a function on $\mathbb{R}^N$ and then analyze the SGD, assuming the loss function L is convex. Under this assumption, we get an upper bound for the sample complexity of SGD. Finally we discuss SGD in general FNNs.

11.1. **Gradient and subgradient descend.** For a differentiable function f on a $\mathbb{R}^N$ denote by $\nabla_g f$ the gradient of f w.r.t. a Riemannian metric g on $\mathbb{R}^N$, i.e., for any $x \in \mathbb{R}^N$ and any $V \in \mathbb{R}^N$ we have

$$df(V) = \langle \nabla_g f, X \rangle. \tag{11.1}$$

If g is fixed, for instance g is the standard Euclidean metric on $\mathbb{R}^N$, we just write ∇f instead of $\nabla_g f$.

The negative gradient flow of f on $\mathbb{R}^N$ is a dynamic system on $\mathbb{R}^N$ defined by the following ODE with initial value $w_0 \in \mathbb{R}^N$

$$w(0) = w_0 \in \mathbb{R}^N \text{ and } \dot{w}(t) = -\nabla f(w(t)). \tag{11.2}$$

If $w(t)$ is a solution of (11.2) then $f(w(t)) < f(w(t'))$ for any $t' > t$ unless $\nabla f(w(t)) = 0$, i.e., $w(t)$ is a critical point of f.

If f is not differentiable we modify the notion of the gradient of f as follows.

Definition 11.1. Let $f : S \to \mathbb{R}$ be a function on an open convex set $S \subset \mathbb{R}^N$. A vector $v \in \mathbb{R}^N$ is called *a subgradient of f at $w \in S$* if

$$\forall u \in S,\ f(u) \geq f(w) + \langle u - w, v \rangle. \tag{11.3}$$

The set of subgradients of f at w is also called *the differential set* and denoted by $\partial f(w)$.

Exercise 11.2. (1) Show that if f is differentiable at w then $\partial f(w)$ contains a single element $\nabla f(w)$.

(2) Find a subgradient of the generalized hinge loss function $f_{a,b,c}(w) = \max\{a, 1 - b\langle w, c\rangle\}$ where $a, b \in \mathbb{R}$ and $w, c \in \mathbb{R}^N$ and $\langle \cdot, \cdot \rangle$ a scalar product.

Remark 11.3. It is known that a subgradient of a function f on a convex open domain S exists at every point $w \in S$ iff f is convex, see e.g. [SSBD2014, Lemma 14.3].

• *Gradient descend algorithm* discretizes the solution of the gradient flow equation (11.2). We begin with an arbitrary initial point $x_0 \in \mathbb{R}^N$. We set

$$w_{n+1} = w_n - \gamma_n \nabla f(w_n), \tag{11.4}$$

where $\gamma_n \in \mathbb{R}_+$ is a constant, called a "learning rate" in machine learning, to be optimized. This algorithm can be slightly modified. For example, after T iterations we set the output point $\bar{w}_T$ to be

$$\bar{w}_T := \frac{1}{T} \sum_{i=1}^{T} w_i, \tag{11.5}$$

or

$$\bar{w}_T := \arg\min_{i\in[1,T]} f(w_i). \tag{11.6}$$

If a function f on $\mathbb{R}^N$ has a critical point which is not the minimizer of f, then the gradient flow (11.2) and its discrete version (11.4) may not converge to the required minimizer of f. If f is convex, then f has only a unique critical point w_0 which is also the minimizer of f. In fact we have the following stronger assertion.

$$f(w) - f(u) \leq \langle w - u, \nabla f(w)\rangle \text{ for any } w, u \in \mathbb{R}^N. \tag{11.7}$$

It also follows from (11.7) that there exists a unique minimizer of f, and hence the gradient flow (11.2) works. Its discrete version (11.4) also works, as stated in the following.

Proposition 11.4. *(*[SSBD2014, Corollary 14.2, p. 188]*) Let f be a convex ρ-Lipschitz function on $\mathbb{R}^N$,*[24] *and let $w^* \in \arg\min_{w\in B(0,r)\subset\mathbb{R}^N} f(w)$. If we run the GD algorithm (11.4) on f for T steps with $\gamma_t = \eta = \frac{r}{\rho\sqrt{T}}$ for $t \in [1, T]$, then the output $\bar{w}_T$ defined by (11.5) satisfies*

$$f(\bar{w}_T) - f(w^*) \leq \frac{r\rho}{\sqrt{T}}.$$

Under the conditions in Proposition 11.4, for every $\varepsilon > 0$, to achieve $f(\bar{w}_T) - f(w^*) \leq \varepsilon$, it suffices to run the GD algorithm for a number of iterations that satisfies

$$T \geq \frac{r^2\rho^2}{\varepsilon^2}.$$

Lemma 11.5. *(*[SSBD2014, Lemma 14.1, p. 187]*) Let $w^*, v_1, \cdots, v_T \in \mathbb{R}^N$. Any algorithm with an initialization $w_1 = 0$ and*

$$w_{t+1} = w_t - \eta v_t \tag{11.8}$$

satisfies

$$\sum_{i=1}^{T}\langle w_t - w^*, v_t\rangle \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}||v_t||^2. \tag{11.9}$$

In particular, for every $r, \rho > 0$, if for all $t \in [1, T]$ we have $||v_t|| \leq \rho$ and if we set $\eta = (r/\rho)T^{-1/2}$ then if $||w^*|| \leq r$ we have

$$\frac{1}{T}\sum_{t=1}^{T}\langle w_t - w^*, v_t\rangle \leq \frac{r\rho}{\sqrt{T}}. \tag{11.10}$$

To apply Lemma 11.5 to Proposition 11.4 we set $v_t := \nabla f(w_t)$ and note that $||\nabla f(w_t)|| \leq \rho$ since f is ρ-convex, moreover

[24] i.e., $|f(w) - f(u)| \leq \rho\,|w - u|$

$$f(\bar{w}_T) - f(w^*) = f(\frac{1}{T}\sum_{t=1}^{T} w_t) - f(w^*)$$

$$\overset{\text{since } f \text{ is convex}}{\leq} \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^*) = \frac{1}{T}\sum_{i=1}^{T}(f(w_t) - f(w^*))$$

$$\overset{by(11.7)}{\leq} \frac{1}{T}\sum_{i=1}^{T}\langle w_t - w^*, \nabla f(w_t)\rangle.$$

• *Subgradient descend algorithm.* Comparing (11.3) with (11.7), taking into account the technical Lemma 11.5, we conclude that the gradient descend algorithm can be applied the case of non-differentiable function f that has subgradient at every point.

11.2. **Stochastic gradient descend (SGD).** Let $(\mathcal{Z}, \mathcal{H}, L, P)$ be a learning model. Given a sample $S := (z_1, \cdots, z_n) \in \mathcal{Z}^n$ consisting of observables z_i that are i.i.d. by $\mu \in P$, a SGD searches for an approximate minimizer $h_S \in \mathcal{H}$ of the function $R^L_\mu : \mathcal{H} \to \mathbb{R}$ using the following formula of "differentiation under integration", assuming L is differentiable.

$$\nabla R^L_\mu(h) = \int_{\mathcal{Z}} \nabla_h L(h, z)\, d\mu(z). \tag{11.11}$$

Here $\nabla_h L(h, z)$ is the gradient of a function L of variable h and parameter z. Thus $\nabla R^L_\mu(h)$ can be computed in two steps. First we compute $\nabla_h L(h, z_i)$ for $z_i \in S$. Then we approximate the RHS of (11.11) by the empirical gradient $\frac{1}{n}\sum_{z_i \in S} \nabla_h L(h, z_i)$ which is equal to the gradient of the empirical risk function.

$$\nabla \hat{R}^L_S(h) = \frac{1}{n}\sum_{z_i \in S} \nabla_h L(h, z_i).$$

Next we apply the algorithm for gradient flow described above to $\nabla \hat{R}^L_S(h)$. The weak law of large numbers ensures the convergence in probability of $\nabla \hat{R}^L_S(h)$ to RHS of (11.11), and heuristically the convergence of the empirical gradient descend algorithm to the gradient descend of the expected risk function R^L_μ.

There are several versions of SGD with minor modifications.

For simplicity and applications in NN we assume $\mathcal{Z} = \mathbb{R}^n \times \mathbb{R}^m$, $\mathcal{H} := \mathcal{H}_{E,V,\sigma}$ is parameterized by $w \in \mathbb{R}^N$ and L is differentiable in w. A version of SGD works as follows.

1) Choose a parameter $\eta > 0$ and $T > 0$.
2) Assume that $S = (z_1, z_2, \cdots, z_n) \in \mathcal{Z}^n$. Take arbitrary $z \in S$.
3) Set $w_1 = 0 \in \mathbb{R}^N$.

4) $w_{t+1} := w_t - \eta \nabla_w L(w_t, z)$.
5) Set the output $\bar{w}_T(z) := \frac{1}{n}\sum_{t=1}^n w_t$.

Proposition 11.6. *(*[SSBD2014, Corollary 14.12, p. 197]*) Assume L is a convex function in variable w and $\mu \in P$ governs the probability distribution of i.i.d. $z_i \in S \in (\mathbb{R}^n \times \mathbb{R}^m)$. Assume that $r, \rho \in \mathbb{R}_+$ are given with the following properties.*

1) $w^* \in \arg\min_{w \in B(0,r)} R^L_\mu(w)$.

2) The SGD is run for T iterations with $\eta = \sqrt{\frac{r^2}{\rho^2 T}}$.

3) For all $t \in [1, T]$ we have $\mathbb{E}_\mu\big(||\nabla_w L(w_t, z)||\big) \leq \rho$ (e.g., $||\nabla_w L(w_t, z)|| \leq \rho$ for all z).

4) Assume that $T \geq \frac{r^2\rho^2}{\varepsilon^2}$.

Then

$$\mathbb{E}_\mu\Big(R^L_\mu\big(\bar{w}_T(z)\big)\Big) \leq R^L_\mu\big(h(w^*)\big) + \varepsilon. \tag{11.12}$$

Exercise 11.7. Find an upper bound for the sample complexity of the SGD in Proposition 11.6.

Example 11.8. Let us consider layered FNN with $\mathcal{H} = H_{E,V,\sigma}$ where $V = V_0 \cup V_1 \cup \cdots \cup V_T$. For the loss function

$$L(x, y, w) := \frac{1}{2}||h_w(x) - y||^2$$

and a vector $v \in \mathbb{R}^N$ on $\mathbb{R}^n$ we compute the gradient of L w.r.t. the Euclidean metric on $\mathbb{R}^N$, regarding x, y as parameters:

$$\langle \nabla L(x, y, w), v\rangle = \langle h_w(x) - y, \nabla_v h_w(x)\rangle.$$

To compute $\nabla_v h_w(x) = dh(v)$ we decompose $h_w = h_T \circ \cdots \circ h_1$ as in (10.8) and using the chain rule

$$d(h_T \circ \cdots \circ h_1)(v) = dh_T \circ \cdots dh_1(v).$$

To compute dh_i we use the decomposition (10.9)

$$d(h_{t,1} \times \cdots \times h_{t,|V_t|}) = dh_{t,1} \times \cdots \times dh_{t,|V_t|}.$$

Finally for $h_{t,j} = \sigma(\sum a_j x_j)$ we have

$$dh_{t,j} = d\sigma \circ (\sum a_j dx_j).$$

The algorithm for computing the gradient ∇L w.r.t. w efficiently is called *backpropogation*. [25]

Remark 11.9. (1) In a general FNN the loss function L is not convex therefore we cannot apply Proposition 11.6. Training FNN is therefore subject to experimental tuning.

(2) Training a RNN is reduced to training of sequence of FNN given a sequence of labelled data, see [Haykin2008, §15.6, p. 806] for more details.

[25] According to [Bishop2006, p. 241] the term "backpropogation" is used in the neural computing literature to mean a variety of different things.

11.3. **Online gradient descend and online learnability.** For training neural networks one also use Online Gradient Descend (OGD), which works as an alternative method of SGD [SSBD2014, p. 300]. Let $L : \mathbb{R}^N \times \mathcal{Z} \to \mathbb{R}$ be a loss function. A version of OGD works almost like SGD
1) We choose a parameter $\eta > 0$ and $T > 0$.
2) A sample $S = (z_1, \cdots, \cdots z_T) \in \mathcal{Z}^T$ is given.
3) Set $w_1 = 0$.
4) For t$\in [1, T]$ set $v_t := \nabla_w f(w_t, z_t)$.
5) Set $w_t := w_t - \eta v_t$.

Despite on their similarity, at the moment there is no sample complexity analysis of OGD. Instead, ML community develops a concept of online learnability for understanding OGD.

11.3.1. *Setting of online-learning.* Let $(\mathcal{X}\times\mathcal{Y}, \mathcal{H}, L, P)$ be a supervised learning model. The general on-line learning setting involves T rounds. At the t-th round, $1 \leq t \leq T$, the algorithm A receives an instance $x_t \in \mathcal{X}$ and makes a prediction $A(x_t) \in \mathcal{Y}$. It then receives the true label $y_t \in \mathcal{Y}$ and computes a loss $L(A(y_t), y_t)$, where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. The goal of A is to find a predictor $A(x_t)$ that minimizes the *cumulative loss*, which is an analogue of the notion of empirical risk in our unified learning model $R_A(T) := \sum_{i=1}^{T} L(A(x_t), y_t)$ over T rounds [MRT2012, p. 148].

In the case of 0-1 loss function $L^{(0-1)}$ the value $R_A(T)$ is called the number of mistakes that A makes after T rounds.

Definition 11.10 (Mistake Bounds, Online Learnability)**.** ([SSBD2014, Definition 21.1, p. 288]) Given any sequence $S = (x_1, h^*(y_1)), \cdots, (x_T, h^*(y_T))$, where T is any integer and $h^* \in \mathcal{H}$, let $M_A(S)$ be the number of mistakes A makes on the sequence S. We denote by $M_A(\mathcal{H})$ the supremum of $M_A(S)$ over all sequences of the above form. A bound of the form $M_A(\mathcal{H}) \leq B < \infty$ is called a *mistake bound*. We say that a hypothesis class $\mathcal{H}$ is *online learnable* if there exists an algorithm A for which $M_A(\mathcal{H}) \leq B < \infty$.

Remark 11.11. 1) Similarly we also have the notion of a successful online learner in regression problems [SSBD2014, p. 300] and within this concept online gradient descent is a successful online learner whenever the loss function is convex and Lipschitz.

2) In the online learning setting the notion of certainty and therefore the notion of probability measure are absent. In particular we do not have the notion of expected risk. So there is an open question if we can justify online learning setting, using statistical learning theory.

11.4. **Conclusion.** In this section we study stochastic gradient descend as a learning algorithm which works if the loss function is convex. To apply stochastic gradient flow as a learning algorithm in FNN where the loss function is not convex one needs experimentally modify the algorithm so it does not stay in a critical point which is not the minimizer of the empirical risk function. One also trains NN with online gradient descends for which we

need a new concept of online learnability which has not yet interpreted using probability framework.

12. Bayesian machine learning

Under "Bayesian learning" one means application of Bayesian statistics to statistical learning theory. Bayesian inference is an approach to statistics in which all forms of uncertainty are expressed in terms of probability. Ultimately in Bayesian statistics we regard all unknown quantities as random variables and we consider a joint probability distribution for all of them, which contains the most complete information about the correlation between the unknown quantities. We denote by Θ the set of parameters that govern the distribution of data $x \in \mathcal{X}$ we like to estimate. The crucial observation is that the joint distribution $\mu_{\Theta\times\mathcal{X}}$ is defined by the conditional probability $P(A|x) := \mu_{\Theta|\mathcal{X}}(A|x)$ and the marginal probability measure μ_Θ on Θ by Remark 2.9 (1).

12.1. **Bayesian concept of learning.** A Bayesian approach to a problem starts with the formulation of a *Bayesian statistical model* $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ that we hope is adequate to describe the situation of interest. Here $\mu_\Theta \in \mathcal{P}(\Theta)$ and $\mathbf{p} : \Theta \to \mathcal{P}(\mathcal{X})$ is a measurable map, see Subsection A.4. We then formulated a *prior distribution* μ_Θ over the unknown parameters $\theta \in \Theta$ of the model, which is meant to capture our beliefs about the situation before seeing the data $x \in \mathcal{X}$. After observing some data, we apply Bayes' Theorem A.13, to obtain a *posterior distribution* for these unknowns, which takes account of both the prior and the data. From this posterior distribution we can compute *predictive distributions* $P(z^{n+1}|\, z^1, \cdots, z^n)$ for future observations using (12.1).

To predict the value of an unknown quantity z^{n+1}, given a sample $(z^1, \cdots, z^n)$, a prior distribution μ_Θ, one uses the following formula

$$P(z^{n+1}|\, z^1, \cdots, z^n) = \int P(z^{n+1}|\theta) P(\theta|z^1, \cdots, z^n) d\mu_\Theta \tag{12.1}$$

which is a consequence of disintegration formula (A.11).

The conditional distribution $P(z^{n+1}|\theta) := \mathbf{p}(\theta)(z^{n+1})$ is called *the sampling distribution of data* z^{n+1}, the conditional probability $P(\theta|z^1, \cdots, z^n)$ is called *posterior distribution of* θ *after observing* $(z^1, \cdots, z^n)$. The learning algorithm A gives a rule to compute $P(\theta|z^1, \cdots, z^n)$ is called *posterior distribution of* θ.

Definition 12.1. A Bayesian machine learning is a quintuple $(\Theta, \mu_\Theta, \mathcal{X}, , A)$, where $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ is a Bayesian statistical model and A is an algorithm for computing the posterior distribution $P(\theta|x_1, \cdots, x_n)$.

Example 12.2. Let us consider a simple Bayesian learning machine ($\Theta = \cup_{i=1}^\infty, \mathcal{X} = \cup_{j=1}^\infty x_j$), called *a discrete naive Bayesian learning machine.* In

naive Bayesian machine learning we make assumption that for all observable $\theta_k \in \Sigma_\Theta$ the conditional distribution $P(\theta_k|x_i)$ are mutually independent, i.e.,

$$P(x_i \mid x_{i+1}, \ldots, x_n, \theta_k) = p(x_i \mid \theta_k). \tag{12.2}$$

Then by (A.15) we have the following algorithm for computing the posterior distributions

$$P(\theta_k|x_1, \cdots, x_k) = P(\mathbf{x}) \cdot P(\theta_k)\Pi_{i=1}^n P(x_i|\theta_k), \tag{12.3}$$

where $\mathbf{x} = (x_1, \cdots, x_k)$.

Example 12.3. When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The Gaussian naive Bayesian model assume that the Bayesian statistical model $(\Theta, \mu_\Theta, \mathbf{p}, \mathcal{X})$ is Gaussian, i.e.,

$$P(x = v|\theta) = \frac{1}{\sqrt{2\pi\sigma(\theta)^2}} \exp(-\frac{(v-\theta)^2}{2\sigma(\theta)^2}).$$

Remark 12.4 (MAP estimator)**.** In many case it is difficult to compute the predictive distribution $P(x_{n+1}|x_1, \cdots, x_n)$, using (12.1). A popular solution is the *Maximum A Posterior estimator*. We take the value θ that maximize the posterior probability $P(\theta|x_1, \cdots, x_n)$ and plug it to compute $P(x_{n+1}|\theta)$.

12.2. **Applications of Bayesian machine learning.**

Example 12.5 (Bayesian neural networks)**.** ([Neal1996, §1.1.2, p. 5]) In Bayesian neural network the aim of a learner is to find a conditional probability $P(y|x^{n+1}, (x^1, y^1), \cdots, (x^n, y^n))$, where y is a label, x^{n+1} is a new input and $\{(x^i, y^i) | \, i = 1, n\}$ is training data. Let θ be a parameter of the neural network. Then we have

$$P(y|\, x^{n+1}, (x^1, y^1), \cdots, (x^n, y^n)) = \int P(y|\, x^{n+1}, \theta) P(\theta|\, (x^1, y^1), \cdots, (x^n, y^n))\, d\theta. \tag{12.4}$$

The conditional sampling probability $P(y|x^{n+1}, \theta)$ is assumed to known. Hence we can compute the LHS of (12.4), which is called predictive distribution of y.

Another application of Bayesian methods is model selection. First we enumerate all reasonable models of the data and assigning a prior belief μ_i to each of these models M_i. Then, upon observing the data x you evaluate how probable the data was under each of these models to compute $P(x|\mu_i)$. To compare two models M_i with M_j, we need to compute their relative probability given the data: $\mu_i P(x|M_i)/\mu_j P(x|M_j)$.

There are more applications of Bayesian machine learning, for example in decision theory using posterior distributions, see e.g. [Robert2007].

12.3. **Consistency.** There are many view points on Bayesian consistency, see e.g. [GV2017, Chapters 6, 7], where the authors devoted a sizable part of their book to the convergence of posterior distributions to the true value of the parameters as the amount of data increases indefinitely. On the other hand, according to [Robert2007, p. 48], in Bayesian decision theory one did not consider asymptotic theory. Firstly, the Bayesian point of view is intrinsically conditional. When conditioning on the observation $S \in \mathcal{X}^n$, there is no reason to wonder what might happen if n goes to infinity since n is fixed by the sample size. Theorizing on future values of the observations thus leads to a frequentist analysis, opposite to the imperatives of the Bayesian perspective. Secondly, even though it does not integrate asymptotic requirements, Bayesian procedures perform well in a vast majority of cases under asymptotic criteria. In a general context, Ibragimov and Hasminskii show that Bayes estimators are consistent [IH1981, chapter 1].

12.4. **Conclusion.** In our lecture we considered main ideas and some applications of Bayesian methods in machine learning. Bayesian machine learning is an emerging promising trend in machine learning that is well suitable for solving complex problems on one hand and consistent with most basic techniques of non-Bayesian machine learning. There are several problems in implementing Bayesian approach, for instance to translating our subjective prior beliefs into a mathematically formulated model and prior. There may also computational difficulties with the Bayesian approach.

Appendix A. The Radon-Nikodym theorem and regular conditional probability

The true logic of this world is the calculus of probabilities.
James Clerk Maxwell

Basis objects in probability theory (and mathematical statistics) are measurable spaces $(\mathcal{X}, \Sigma_{\mathcal{X}})$, where $\Sigma_{\mathcal{X}}$ is a σ-algebra of subsets of a space $\mathcal{X}$. We shall often write $\mathcal{X}$ if we dont need to specify $\Sigma_{\mathcal{X}}$. A *signed measure* μ on $\mathcal{X}$ is a countably additive function $\mu : \Sigma_{\mathcal{X}} \to \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$. A signed measure μ is called a (*nonnegative*) *measure*[26], if $\mu(A) \geq 0$ for all $A \in \Sigma_{\mathcal{X}}$. A measure μ is called *a probability measure* if $\mu(\mathcal{X}) = 1$. We shall denote by $\mathcal{S}(\mathcal{X})$ the set of all signed measures on $\mathcal{X}$, by $\mathcal{M}(\mathcal{X})$ the set of all finite measures μ on $\mathcal{X}$, i.e., $\mu(\mathcal{X}) < \infty$, and by $\mathcal{P}(\mathcal{X})$ the set of all probability measures on $\mathcal{X}$, i.e., $\mu(\mathcal{X}) = 1$.

A.1. **Dominating measures and the Radon-Nikodym theorem.** Let $\mu \in \mathcal{M}(\mathcal{X})$ and $\nu \in \mathcal{S}(\mathcal{X})$.
(i) The measure ν is called *absolutely continuous with respect to* μ (or *dominated by* μ) if $|\nu|(A) = 0$ for every set $A \in \Sigma_{\mathcal{X}}$ with $|\mu|(A) = 0$. In this case we write $\nu \ll \mu$.

[26] we shall omit in our lecture notes the adjective "nonnegative"

(ii) The measure ν is called *singular with respect to* μ, if there exists a set $\Omega \in \Sigma_{\mathcal{X}}$ such that

$$|\mu|(\Omega) = 0 \text{ and } |\nu|(\mathcal{X} \setminus \Omega) = 0.$$

In this case we write: $\nu \perp \mu$ and we denote by Id_ν the identity measurable mapping $(\mathcal{X}, \Sigma_{\mathcal{X}}) \to (\mathcal{X}, \nu^{-1}(\Sigma_{\mathcal{Y}}))$.

Theorem A.1. *(cf.* [Bogachev2007, Theorem 3.2.2, vol 1, p. 178]*) Let μ and ν be two finite measures on a measurable space $(\mathcal{X}, \Sigma)$. The measure ν is dominated by the measure μ precisely when there exists a function $f \in L^1(\mathcal{X}, \mu)$ such that for any $A \in \Sigma_{\mathcal{X}}$ we have*

$$\nu(A) = \int_A f \, d\mu. \tag{A.1}$$

We denote ν by $f\mu$ for μ, ν, f satisfying the equation (A.1). The function f is called the (Radon-Nikodym) density (or the Radon-Nikodym derivative) of ν w.r.t. μ. We denote f by $d\nu/d\mu$.

Remark A.2. The Radon-Nykodym derivative $d\nu/d\mu$ should be understood as the equivalence class of functions $f \in L^1(\mathcal{X}, \mu)$ that satisfy the relation (A.1) for any $A \in \Sigma_{\mathcal{X}}$. It is important in many cases to find a function on $\mathcal{X}$ which represents the equivalence class $d\nu/d\mu \in L^1(\mathcal{X}, \mu)$, see [JLT2020, Main Theorem].

A.2. **Conditional expectation and regular conditional measure.** The notion of conditioning is constantly used as a basic tool to describe and analyze systems involving randomness. Heuristic concept of conditional probability existed long before the fundamental work by Kolmogorov [Kolmogoroff1933], where the concept of conditional probability has been rigorously defined via the concept of conditional expectation.

A.2.1. *Conditional expectation.* In this section we define the notion of conditional expectation using the Radon-Nykodym theorem. We note that any sub-σ-algebra $\mathcal{B} \subset \Sigma_{\mathcal{X}}$ can be written as $\mathcal{B} = Id^{-1}(\mathcal{B})$ where $Id : (\mathcal{X}, \Sigma_{\mathcal{X}}) \to (\mathcal{X}, \mathcal{B})$ is the identity mapping, and hence a measurable mapping. In what follows, w.l.o.g. we shall assume that $\mathcal{B} := \sigma(\eta)$ where $\eta : (\mathcal{X}, \Sigma_{\mathcal{X}}) \to (\mathcal{Y}, \Sigma_{\mathcal{Y}})$ is a measurable mapping.

$$\begin{array}{ccc} (\mathcal{X}, \Sigma_{\mathcal{X}}) & \xrightarrow{\eta} & (\mathcal{Y}, \Sigma_{\mathcal{Y}}) \\ \Big\downarrow{\scriptstyle Id_\eta} & \nearrow_{\eta} & \\ (\mathcal{X}, \eta^{-1}(\Sigma_{\mathcal{Y}})) & & \end{array}$$

For a measurable mapping $\eta : (\mathcal{X}, \Sigma_{\mathcal{X}}) \to (\mathcal{Y}, \Sigma_{\mathcal{Y}})$ we denote by $\eta_* : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$ the push-forward mapping

$$\eta_*(\mu)(B) := \mu(\eta^{-1}(B)) \tag{A.2}$$

for any $B \in \Sigma_{\mathcal{Y}}$. Clearly $\eta_*(\mathcal{M}(\mathcal{X})) \subset \mathcal{M}(\mathcal{Y})$ and $\mu_*(\mathcal{P}(\mathcal{X})) \subset \mathcal{P}(\mathcal{Y})$. Denote by Id_η the identity measurable mapping $(\mathcal{X}, \Sigma_{\mathcal{X}}) \to (\mathcal{X}, \eta^{-1}(\Sigma_{\mathcal{Y}}))$.

Definition A.3. Let $\mu \in \mathcal{P}(\mathcal{X})$, $\eta : \mathcal{X} \to \mathcal{Y}$ a measurable mapping and $f \in L^1(\mathcal{X}, \mu)$. The conditional expectation $\mathbb{E}_\mu^{\sigma(\eta)} f$ is defined as follows

$$\mathbb{E}_\mu^{\sigma(\eta)} f := \frac{d(Id_\eta)_*(f\mu)}{d(Id_\eta)_*(\mu)} \in L^1(\mathcal{X}, \eta^{-1}(\Sigma_\mathcal{Y}), (Id_\eta)_*\mu). \tag{A.3}$$

Remark A.4. (1) The conditional expectation $\mathbb{E}_\mu^{\sigma(\eta)} f$ can be expressed via a function in $L^1(\mathcal{Y}, \Sigma_\mathcal{Y}, \eta_*\mu)$ as follows. Set

$$\eta_*^\mu f := \frac{d\eta_*(f\mu)}{d\eta_*(\mu)} \in L^1(\mathcal{Y}, \Sigma_\mathcal{Y}, \eta_*\mu).$$

Assume that g is a function on $\mathcal{Y}$ that represents the equivalence class $\eta_*^\mu f$. Then $g(\eta(x))$ is a representative of $\mathbb{E}_\mu^{\sigma(\eta)} f$, since for any $A \in \eta^{-1}(\Sigma_\mathcal{X})$ we have

$$f\mu(A) = \int_{\eta^{-1}(B)} f d\mu = \int_B \eta_*^\mu f d\eta_*(\mu) = \int_B g d\eta_*(\mu).$$

(2) In the probabilistic literature for $\mathcal{B} = \sigma(\eta)$ one uses the notation

$$\mathbb{E}(f|\mathcal{B}) := \mathbb{E}_\mu^\mathcal{B} f,$$

since in this case μ is assume the known probability measure and therefore we don't need to specify it in the expression $\mathbb{E}(f|\mathcal{B})$. We also use the notation $\mathbb{E}_\mu(f|\mathcal{B})$ later.

(3) There are many approaches to conditional expectations. The present approach using the Radon-Nykodym theorem is more or less the same as in in [Halmos1950]. In [JP2003, Definition 23.5, p. 200] the author lets the orthogonal projection of $f \in L^2(\mathcal{X}, \mu)$ to the subspace $L^2(\mathcal{X}, \sigma(\eta)(Id_\eta)_*\mu)$ to be the definition of conditional expectation of f and then extend this definition to the whole $L^1(\mathcal{X}, \mu)$.

A.2.2. *Conditional measure and conditional probability.* Let $\mu \in \mathcal{M}(\mathcal{X})$. *The measure μ, conditioning w.r.t. a sub-σ-algebra $\mathcal{B} \subset \Sigma_\mathcal{X}$*, is defined as follows

$$\mu(A|\mathcal{B}) := \mathbb{E}_\mu(1_A|\mathcal{B}) \in L^1(\mathcal{X}, \mathcal{B}, \mu) \tag{A.4}$$

for $A \in \Sigma_\mathcal{X}$. In probabilistic literature one omits μ in (A.4) and writes instead

$$P(A|\mathcal{B}) := \mu(A|\mathcal{B}).$$

If $\mathcal{B} = \eta^{-1}(\Sigma')$ where $\eta : (\mathcal{X}, \Sigma_\mathcal{X}) \to (\mathcal{Y}, \Sigma_\mathcal{Y})$ is a measurable map, one uses the notation

$$P(A|\eta) := \mu(A|\eta) := \mu(A|\mathcal{B}).$$

For any $A \in \Sigma$ and $B \in \mathcal{B} = \eta^{-1}(\Sigma_\mathcal{Y})$ formulas (A.3) and (A.1) imply

$$\mu(A \cap B) = \int_B 1_A d\mu = \int_B \frac{d(Id_\eta)_* 1_A \mu}{d(Id_\eta)_*\mu} d\mu = \int_B \mathbb{E}_\mu^{\sigma(\eta)}(1_A) d\mu = \int_B \mu(A|\mathcal{B}) d\mu. \tag{A.5}$$

Remark A.5. It follows from (A.5) that for any $f \in L^1(\mathcal{X}, \mu)$ and any $B \in \mathcal{B}$ we have

$$\int_B f d\mu = \int_B \mathbb{E}^{\mathcal{B}}_{\mu} f \, d\mu. \tag{A.6}$$

The RHS of (A.6) can be written as $\mathbb{E}_\mu(1_B \mathbb{E}^{\mathcal{B}}_{\mu} f)$. The equality (A.6) is also utilized to define the notion of conditional expectation, see e.g. [Bogachev2007, (10.1.2), p. 339, vol. 2]. It implies that for any bounded $\mathcal{B}$-measurable function g on $\mathcal{X}$ we have

$$\int_{\mathcal{X}} g f d\mu = \int_{\mathcal{X}} g \mathbb{E}^{\mathcal{B}}_{\mu} f d\mu. \tag{A.7}$$

If $f \in L^2(\mathcal{X}, \mu)$ then (A.7) implies the following simple formula for $\mathbb{E}^{\mathcal{B}}_{\mu}(f)$

$$\mathbb{E}^{\mathcal{B}}_{\mu}(f) = \Pi_{L^2(\mathcal{X}, \mathcal{B}, \mu)}(f) \tag{A.8}$$

where $\Pi_{L^2(\mathcal{X}, \mathcal{B}, \mu)} : L^2(\mathcal{X}, \Sigma_{\mathcal{X}}, \mu) \to L^2(\mathcal{X}, \mathcal{B}, \mu)$ is the orthogonal projection. See [JP2003, p. 200] for defining conditional expectation using (A.8).

If a measurable function $\zeta_A : \mathcal{Y} \to \mathbb{R}$ represents $\eta^{\mu}_*(1_A) \in L^1(\mathcal{Y}, \eta_* \mu)$, then one sets

$$\mu^{\mathcal{B}}(A|y) := \mu^{\mathcal{B}}(A|\eta(x) = y) := \zeta_A(y). \tag{A.9}$$

The RHS of (A.9) is called *the measure of A under conditioning $\eta = y$*. Clearly the measure of A under conditioning $\eta = y$, as a function of $y \in \mathcal{Y}$, is well-defined up to a set of zero $\eta_* \mu$-measure.

We also rewrite formula (A.9) as follows. For $E \in \Sigma_{\mathcal{Y}}$

$$\mu(A \cap \eta^{-1}(E)) = \int_E \mu^y(A) \, d\eta_*(\mu)(y) \tag{A.10}$$

where $\mu^y(A)$ is a function in $L^1(\mathcal{Y}, \Sigma_{\mathcal{Y}}, \eta_* \mu)$.

Note that it is not always the case that for $\eta_*(\mu)$-almost all $y \in \mathcal{Y}$ the set function $\mu^y(A)$ is countably addictive (any two representatives of the equivalence class of $\mu^y(A) \in L^1(\mathcal{Y}, \eta_*(\mu))$ coincide outside a set of zero $\eta_*(\mu)$-measure), see Example 10.4.9 in [Bogachev2007, p. 367, v.2]. Nevertheless this becomes possible under some additional conditions on set-theoretic or topological character, and in this case we say that $\mu^y(A)$ is a *regular conditional measure*.

A.2.3. *Regular conditional measure and Markov kernel.*

Definition A.6. [Bogachev2007, Definition 10.4.1, p. 357]) Suppose we are given a sub-σ-algebra $\mathcal{B} \subset \Sigma_{\mathcal{X}}$. A function

$$\mu^{\mathcal{B}}(\cdot, \cdot) : \Sigma_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$$

is called *a regular conditional measure* on $\Sigma_{\mathcal{X}}$ w.r.t. $\mathcal{B}$ if
(1) for every $x \in \mathcal{X}$ the function $A \mapsto \mu^{\mathcal{B}}(A, x)$ is a *measure on* $\Sigma_{\mathcal{X}}$,

(2) for every $A \in \Sigma_{\mathcal{X}}$ the function $x \mapsto \mu^{\mathcal{B}}(A, x)$ is *$\mathcal{B}$-measurable and μ-integrable*,
(3) For all $A \in \Sigma_{\mathcal{X}}$, $B \in \mathcal{B}$ the following formula holds, cf. (A.5)

$$\mu(A \cap B) = \int_B \mu^{\mathcal{B}}(A, x) d\mu(x). \tag{A.11}$$

Remark A.7. (1) Assume that $\mu^{\mathcal{B}}(A, x)$ is a regular conditional measure. Formulas (A.11) and (A.5) imply that $\mu^{\mathcal{B}}(A, x) : \mathcal{X} \to \mathbb{R}$ is a representative of the conditional measure $\mu(A|\mathcal{B}) \in L^1(\mathcal{X}, \mathcal{B}, \mu)$. Thus one also uses the notation $\mu^{\mathcal{B}}(A|x)$, defined in (A.9), instead of $\mu^{\mathcal{B}}(A, x)$.

(2) The equality (A.11) can be written in the following integral form: for every bounded $\Sigma_{\mathcal{X}}$-measurable function f and every $B \in \mathcal{B}$ one has (cf. (A.6))

$$\int_B f(x)\mu(dx) = \int_B \int_{\mathcal{X}} f(y) d\mu^{\mathcal{B}}(y, x) d\mu(x) = \int_B \mathbb{E}^{\mathcal{B}}_{\mu} f d\mu. \tag{A.12}$$

Regular conditional measures in Definition A.6 are examples of transition measures for which we shall have a generalized version of Fubini theorem (Theorem A.9)

Definition A.8. ([Bogachev2007, Definition 10.7.1, vol. 2, p. 384]) Let $(\mathcal{X}_1, \Sigma_1)$ and $(\mathcal{X}_2, \Sigma_2)$ be a pair of measurable spaces. *A transition measure for this pair* is a function $P(\cdot|\cdot) : \mathcal{X}_1 \times \Sigma_2 \to \mathbb{R}$ with the following properties:
(i) for every fixed $x \in \mathcal{X}_1$ the function $B \mapsto P(x|B)$ is a measure on Σ_2;
(ii) for every fixed $B \in \Sigma_2$ the function $x \mapsto P(x|B)$ is measurable w.r.t. $\mathcal{X}_1$.

In the case where transition measures are probabilities in the second argument, they are called *transition probabilities*. In probabilistic literature transition probability is also called *Markov kernel*, or *(probability) kernel* [Kallenberg2002, p. 20].

Theorem A.9. *(*[Bogachev2007, Theorem 10.7.2, p. 384, vol. 2]*) Let $P(\cdot|\cdot)$ be a transition probability for spaces $(\mathcal{X}_1, \Sigma_1)$ and $(\mathcal{X}_2, \Sigma_2)$ and let ν be a probability measure on Σ_1. Then there exists a unique probability measure μ on $(\mathcal{X}_1 \times \mathcal{X}_2, \Sigma_1 \otimes \Sigma_2)$ with*

$$\mu(B_1 \times B_2) = \int_{B_1} P(x|B_2) d\nu(x) \text{ for all } B_1 \in \Sigma_1,\ B_2 \in \Sigma_2. \tag{A.13}$$

In addition, given any $f \in L^1(\mu)$ for ν- a.e. $x_1 \in \mathcal{X}_1$ the function $x_2 \mapsto f(x_1, x_2)$ on $\mathcal{X}_2$ is measurable w.r.t. the completed σ-algebra $(\Sigma_2)_{P(x_1|\cdot)}$ and $P(x_1|\cdot)$-integrable, the function

$$x_1 \mapsto \int_{\mathcal{X}_2} f(x_1, x_2) dP(x_1|x_2)$$

is measurable w.r.t. $(\Sigma_1)_\nu$, and ν-integrable, and one has

$$\int_{\mathcal{X}_1 \times \mathcal{X}_2} f(x_1, x_2) d\mu(x_1, x_2) = \int_{\mathcal{X}_1} \int_{\mathcal{X}_2} f(x_1, x_2) dP(x_1|x_2) d\nu(x_1). \tag{A.14}$$

Corollary A.10. *If a parametrization $(\Theta, \Sigma_\Theta) \to \mathcal{P}(\mathcal{X})$, $\theta \mapsto \mathbf{p}_\theta$, defines a transition measure then $\mathbf{p}_\theta$ can be regarded as a regular conditional probability measure $\mu(\cdot|\theta)$ for μ defined by (A.13).*

Corollary A.11. *Assume that Θ is a topological space and Σ_Θ is a Borel σ-algebra. If the parametrization mapping $\Theta \to \mathcal{P}(\mathcal{X})$, $\theta \mapsto \mathbf{p}_\theta$, is continuous w.r.t. the strong topology, then $\mathbf{p}_\theta$ can be regarded as a regular conditional probability measure.*

Proof. Since the parametrization is continuous, for any $A \in \Sigma_\mathcal{X}$ the function $\theta \mapsto \mathbf{p}_\theta(A)$ is continuous and bounded, and hence measurable. Hence the parametrization $\Theta \to \mathcal{P}(\mathcal{X})$ defines a transition probability measure and applying Theorem A.9 we obtain Corollary A.11. □

A.3. **Joint distribution, regular conditional probability and Bayes' theorem.** Till now we define conditional probability measure $\mu(A|\mathcal{B})$ on a probability space $(\mathcal{X}, \Sigma, \mu)$ where $\mathcal{B}$ is a sub σ-algebra of Σ. We can also define conditional probability measure $\mu_{\mathcal{X}\times\mathcal{Y}}(A|\mathcal{B})$ where A is a subset of the σ-algebra $\Sigma_\mathcal{X}$ of a measurable space $\mathcal{X}$ and $\mathcal{B}$ is a sub-σ algebra of the σ-algebra $\Sigma_\mathcal{Y}$ of a measurable space $\mathcal{Y}$, if a joint probability measure $\mu_{\mathcal{X}\times\mathcal{Y}}$ on $(\mathcal{X}\times\mathcal{Y}, \Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y})$ is given.

Let $\Pi_\mathcal{X}$ and $\Pi_\mathcal{Y}$ denote the projection $(\mathcal{X}\times\mathcal{Y}, \Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}) \to (\mathcal{X}, \Sigma_\mathcal{X})$ and $(\mathcal{X}\times\mathcal{Y}, \Sigma_\mathcal{X} \otimes \Sigma_\mathcal{Y}) \to (\mathcal{Y}, \Sigma_\mathcal{Y})$ respectively. Clearly $\Pi_\mathcal{X}$ and $\Pi_\mathcal{Y}$ are measurable mappings. *The marginal probability measure* $\mu_\mathcal{X} \in \mathcal{P}(\mathcal{X})$ is defined as

$$\mu_\mathcal{X} := (\Pi_\mathcal{X})_*(\mu_{\mathcal{X}\times\mathcal{Y}}).$$

The conditional probability measure $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is defined as follows for $B \in \Sigma_\mathcal{Y}$

(A.15)

$$\mu_{\mathcal{Y}|\mathcal{X}}(B|x) := \mathbb{E}^{\sigma(\Pi_\mathcal{X})}_{\mu_{\mathcal{X}\times\mathcal{Y}}}(B\times\mathcal{X}|x) = \frac{d(\Pi_\mathcal{X})_*(1_{B\times\mathcal{X}}\mu_{\mathcal{X}\times\mathcal{Y}})}{d(\Pi_\mathcal{X})_*\mu_{\mathcal{X}\times\mathcal{Y}}}(y) \in L^1(\mathcal{X}, \mu_\mathcal{X})$$

where the equality should understood as equivalence class of functions in $L^1(\mathcal{X}, (\Pi_\mathcal{Y})_*\mu_\mathcal{X})$.

The conditional probability measure $\mu_{\mathcal{X}|\mathcal{Y}}(\cdot|y)$ is called *regular*, if for all $A \in \Sigma_\mathcal{X}$ there is a function on $\mathcal{Y}$, denoted by $\mu_{\mathcal{X}|\mathcal{Y}}(A|y)$, that represents the equivalence class $\mu_{\mathcal{X}|\mathcal{Y}}(A|y)$ in $L^1(\mathcal{Y}, (\Pi_\mathcal{Y})_*\mu_{\mathcal{X}\times\mathcal{Y}})$ such that the function $P(\cdot|\cdot) : \mathcal{Y} \times \Sigma_\mathcal{X} \to \mathbb{R}$, $(y, A) \mapsto \mu_{\mathcal{X}|\mathcal{Y}}(A|y)$ is a transition probability, i.e. a Markov kernel. A convenient way to express this condition is to use the language of probabilistic morphism, see Subsection A.4.

The following disintegration Theorem is similar to Theorem A.9

Theorem A.12 (Disintegration Theorem)**.** *Assume that $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is a regular conditional probability. Then for any $f \in L^1(\mathcal{X}\times\mathcal{Y}, \mu)$ we have*

$$\int_{\mathcal{X}\times\mathcal{Y}} f d\mu = \int_\mathcal{X}\int_\mathcal{Y} f d\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x) d\mu_\mathcal{X}. \tag{A.16}$$

Proof. It suffices to prove (A.16) for $f = 1_{B\times A}$ where $B \in \Sigma_{\mathcal{Y}}$ and $A \in \Sigma_{\mathcal{X}}$. In this case (A.16) has the following form

$$\mu(B \times A) = \int_A \mu_{\mathcal{Y}|\mathcal{X}}(B|x) d\mu_{\mathcal{X}} \tag{A.17}$$

which follows directly from (A.15). □

• The Bayes theorem stated below assumes the existence of *regular conditional measure* $\mu_{\mathcal{X}|\Theta}(A|\theta)$, [27] where μ is a joint distribution of random elements $x \in \mathcal{X}$ and $\theta \in \Theta$. Furthermore we also assume the condition that there exists a measure $\nu \in \mathcal{P}(\mathcal{X})$ such that ν dominates $\mu_\theta := \mu(\cdot|\theta)$ for all $\theta \in \Theta$.

Theorem A.13. *[Bayes' theorem]([Schervish1997, Theorem 1.31, p. 16]) Suppose that $\mathcal{X}$ has a parametric family $\{P_\theta | \, \theta \in \Theta\}$ such that $P_\theta \ll \nu$ for some $\nu \in \mathcal{P}(\mathcal{X})$ for all $\theta \in \Theta$. Let $f_{\mathcal{X}|\Theta}(x|\theta)$ denotes the conditional density of P_θ w.r.t. ν. Let μ_Θ be the prior distribution of Θ and let $\mu_{\Theta|\mathcal{X}}(\cdot|x)$ the conditional distribution of Θ given x. Then $\mu_{\Theta|\mathcal{X}} \ll \mu_\Theta$ ν- a.s. and the Radon-Nykodim derivative is*

$$\frac{d\mu_{\Theta|\mathcal{X}}}{d\mu_\Theta}(\theta|x) = \frac{f_{\mathcal{X}|\Theta}(x|\theta)}{\int_\Theta f_{\mathcal{X}|\Theta}(x|t)\, d\mu_\Theta(t)}$$

for those x s.t. the denominator is neither 0 or infinite. The prior predictive probability of the set of x values s.t. the denominator is 0 or infinite is 0, hence the posterior can be defined arbitrary for such x values.

Remark A.14. (1) For a version of Bayes' theorem without the dominance condition for a family $\{\mu_\theta | \, \theta \in \Theta\}$ see [JLT2020, Theorem 3.6].

(2) In [Faden1985, Theorem 5, p. 291] Faden proved that if $\mathcal{X}$ is a Borel subset of a Polish space then for any $\mathcal{Y}$ and any $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ there exists a regular conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x) \in L^1(\mathcal{X}, \mu_{\mathcal{X}})$. We refer the reader to Example D.2 for expressing the existence of regular conditional probability $\mu_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ in terms of probabilistic morphisms.

A.4. **Disintegration and regular conditional probability.** In previous subsections we consider two close concepts: a conditional measure, in particular, a regular conditional measure $\mu(A|\mathcal{B})$ where $A \in \Sigma_{\mathcal{X}}$ and $\mathcal{B}$ is a sub-σ-algebra of $\Sigma_{\mathcal{X}}$, see (A.4), (A.6), and a regular conditional probability measure $\mu_{\mathcal{X}|\mathcal{Y}}(A|y)$, where $A \in \Sigma_{\mathcal{X}}$ and $y \in \mathcal{Y}$, see (A.15), which is defined via the first concept using the product of spaces. The later concept is what we need in Bayesian statistics. There is a generalization of regular conditional measures defined in the both concepts that is particular convenient for understanding the concept of *regular conditional probability*. This generalization is defined via the concept of disintegration of measures.

[27] Schervish considered only parametric family of conditional distributions [Schervish1997, p.13]

Let $(\mathcal{X}, \Sigma_{\mathcal{X}}, \mu)$ be a probability space, let $\mathcal{B} \subset \Sigma_{\mathcal{X}}$ be a sub-σ-algebra, and let $\mathcal{B}_E := \mathcal{B} \cap E$ denote the restriction of $\mathcal{B}$ to $E \subset \mathcal{X}$.

Definition A.15. ([Bogachev2007, Definition 10.6.1, p. 380, vol.II]) Suppose that for each $x \in \mathcal{X}$ we are given a sub-σ-algebra $\Sigma_x \subset \Sigma_{\mathcal{X}}$ and a measure $\mu(\cdot, x)$ on Σ_x satisfying the following conditions:

(i) for every $A \in \Sigma_{\mathcal{X}}$, there exists a set $N_A \in \mathcal{B}$ such that $\mu(N_A) = 0$ and $A \in \Sigma_x$ for all $x \in N_A$, and the function $x \mapsto \mu(A, x)$ on $X \setminus N_A$ is measurable with respect to $\mathcal{B} \cap (X \setminus N_A)$ and μ-integrable.

(ii) for all $A \in \Sigma_{\mathcal{X}}$ and $B \in \mathcal{B}$ one has

$$\int_B \mu(A, x)\, d\mu(x) = \mu(A \cap B). \tag{A.18}$$

Then we shall say that the measures $\mu(\cdot, x)$ give a disintegration of the measure μ with respect to $\Sigma_{\mathcal{X}}$ and call these measures conditional measures.

Remark A.16. (1) The difference between disintegration and regular conditional probabilities defined in (A.6) and in (A.15) is that the condition of $\mathcal{B}$-measurability is weakened at the expense of admitting sets N_A of measure zero. If $\sigma_{\mathcal{X}}$ is countably generated σ-algebra, then it can be shown that these two concepts are equivalent [Bogachev2007, Proposition 10.6.2, p. 380, vol. II]. There is an example that the existence of disintegration does not imply the existence of regular conditional measures.

(2) The existence of disintegration $\mu(\cdot, x)$ with $\Sigma_x = \Sigma_0$ for all $x \in \mathcal{X}$ is equivalent to the existence of conditional measures w.r.t. $\mathcal{B}$ in sense of Doob [Doob1953, p. 26]. The existence of disintegration under dominating measures has been discussed in [CP1997].

Appendix B. Law of large numbers and concentration-of-measure inequalities

In probability theory, the *concentration of measure* is a property of a large number of variables, such as in laws of large numbers. Concentration-of-measure inequalities provide bounds on the probability that a measurable mapping $f : \mathcal{X} \to \mathbb{R}$ deviates from its mean value $\mathbb{E}_\mu(f)$, or other typical values, by a given amount. Thus concentration-of-measure inequality quantifies the rate of convergence in law of large numbers.

B.1. **Laws of large numbers.**

Proposition B.1 (Strong law of larger number)**.** *(*[Kallenberg2002, Theorem 4.23, p. 73]*) Let $\xi, \xi_1, \xi_2, \cdots, \xi_\infty \in (\mathbb{R}^\infty, \mu^\infty)$ where $\mu \in \mathcal{P}(\mathbb{R})$, and $p \in (0, 2)$. Then $n^{-1/p} \sum_{k \le n} \xi_k$ converges a.s. iff $\mathbb{E}_\mu |\xi_i|^p < \infty$ and either $p \le 1$ or $\mathbb{E}_\mu \xi_i = 0$. In that case the limit equals $\mathbb{E}_\mu \xi$ for $p = 1$ and is otherwise* 0*.*

If we replace the strong condition $\mathbb{E}_\mu |\xi_i|^p < \infty$ in the law of large numbers by weaker conditions, then we have to replace the convergence a.e. (or a.s.,

or with probability 1) in the law of large numbers by the convergence in probability in the weak law of large numbers.

Proposition B.2 (Weak law of large numbers)**.** [Kallenberg2002, Theorem 5.16, p. 95] *Let $\xi, \xi_1, \xi_2, \cdots, \xi_\infty \in (\mathbb{R}^\infty, \mu^\infty)$ where $\mu \in \mathcal{P}(\mathbb{R})$, and $p \in (0,2)$, $c \in \mathbb{R}$. Then $\tilde{\xi}_n := n^{-1/p}\sum_{k\leq n}\xi_k$ converges in probability to c, i.e., for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \mu(\tilde{\xi}_n \in \mathbb{R}|\, |\tilde{\xi}_n - c| > \varepsilon) = 0,$$

iff the following conditions holds as $r \to \infty$, depending on the value of p:
$p < 1$: $r^p\mu(\xi \in \mathbb{R}|\, |\xi| > r) \to 0$ and $c = 0$;
$p = 1$: $r\mu(\xi \in \mathbb{R}|\, |\xi| > r) \to 0$ and $\mathbb{E}_\mu(\xi|\, |\xi| \leq r) \to c$;
$p =>$: $r^p\mu(\xi \in \mathbb{R}|\, |\xi| > r) \to 0$ and $\mathbb{E}_\mu(\xi) = c = 0$.

B.2. **Markov's inequality.** For any $f \in L^1(\mathcal{X}, \mu)$ such that $f(x) \geq 0$ for all $x \in \mathcal{X}$ we have and any $t > 0$ we have

$$\mu\{x \in \mathcal{X}|\, f(x) > t\} \leq \frac{\mathbb{E}_\mu f}{t}. \tag{B.1}$$

B.3. **Chebyshev's inequality.** For any $f \in L^1(\mathcal{X}, \mu)$ and$t > 0$ we have

$$\mu(x : |f(x) - \mathbb{E}_\mu(f)| \geq t) \leq \frac{|f(x) - \mathbb{E}_\mu(x)|^2}{t^2}. \tag{B.2}$$

Remark B.3. (1) Using the following identity

$$\mathbb{E}_\mu(f) = \mathbb{E}_{f\mu}(1_\mathcal{X}) = \mathbb{E}_{f_*(f\cdot\mu)}(1_\mathbb{R}) = \int_0^\infty f_*(\mu)(t, +\infty)\, dt \tag{B.3}$$

and noting that the LHS of (B.1) is equal $f_*(\mu)(t, +\infty)$, which is a monotone function in t, we obtain the Markov inequality (B.1) immediately.

(2) For any monotone function $\varphi : (\mathbb{R}_{\geq 0}, dt) \to (\mathbb{R}_{\geq 0}, dt)$ applying (B.1) we have

$$f_*(\mu)(t, +\infty) = \varphi_* \circ f_*(\mu)(\varphi(t), +\infty) \leq \frac{E_\mu(\varphi \circ f)}{\varphi(t)}. \tag{B.4}$$

The Chebyshev equality (B.2) follows from (B.4), replacing f in (B.4) by $|g - E_\mu(g)|$ for $g \in L^1(\mathcal{X}, \mu)$, and set $\varphi(t) := t^2$.

B.4. **Hoeffding's inequality.** ([Hoeffding1963]) Let $\theta = (\theta_1, \cdots, \theta_n)$ be a sequence of i.i.d. $\mathbb{R}$-valued random variables on $\mathcal{Z}$ and $\mu \in \mathcal{P}(\mathcal{Z})$. Assume that $\mathbb{E}_\mu(\theta_i(z)) = \bar{\theta}$ for all i and $\mu\{z \in \mathcal{Z}|\, [a_i \leq \theta_i(z) \leq b]\} = 1$. Then for any $\varepsilon > 0$ we have

$$\mu^m\{\mathbf{z} \in \mathcal{Z}^m :\, \big|\frac{1}{m}\sum_{i=1}^m \theta_i(z_i) - \bar{\theta}\big| > \varepsilon\} \leq 2\exp\big(\frac{-2m\varepsilon^2}{(b-a)^2}\big), \tag{B.5}$$

where $\mathbf{z} = (z_1, \cdots, z_m)$.

B.5. **Bernstein's inequality.** Let θ be a $\mathbb{R}$-valued random variable on a probability space $(\mathcal{Z}, \mu)$ with the mean $\mathbb{E}_\mu(\theta) = \bar{\theta}$ and variance $\sigma^2 = V_\mu(\theta)$. If $|\xi - \mathbb{E}_\mu(\xi)| \leq M$ then for all $\varepsilon > 0$ we have

$$\mu^m\{\mathbf{z} \in \mathcal{Z}^m : \big|\frac{1}{m}\sum_{i=1}^m \theta_i(z_i) - \bar{\theta}\big| > \varepsilon\} \leq 2\exp\big(\frac{-m\varepsilon^2}{2(\sigma^2 + \frac{1}{3}M\varepsilon)}\big), \tag{B.6}$$

where $\mathbf{z} = (z_1, \cdots, z_m)$.

B.6. **McDiarmid's inequality.** (or Bounded Differences or Hoeffding/Azuma Inequality). Let $X_1, \cdots, X_m \in \mathcal{X}$ are i.i.d. by a probability measure μ. Assume that $f : \mathcal{X}^m \to \mathbb{R}$ satisfies the following property form some $c > 0$. For all $i \in [1, m]$ and for all $x_1, \cdots, x_m, x_i' \in \mathcal{X}$ we have

$$|f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_m)| \leq c.$$

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

$$\mu^m\{S \in \mathcal{X}^m | \, |f(S) - \mathbb{E}_{\mu^m} f(S)| \leq c\sqrt{\frac{\ln(2/\delta)}{m}}\} \geq 1 - \delta. \tag{B.7}$$

Appendix C. The Kolmogorov theorem

In theory of consistency of estimators one would like to consider infinite sequences of instances $(z_1, \cdots, z_\infty)$ and put an probability measure on such sequences. To do so ones use the Kolmogorov theorem that provides the existence of such a probability measure with useful compatibility properties.

Theorem C.1. *(*[Bogachev2007, Theorem 7.7.1, p. 95, vol 1]*) Suppose that for every finite set $\Lambda \subset T$, we are given a probability measure μ_Λ on $(\Omega_\Lambda, \mathcal{B}_\Lambda)$ such that the following consistency condition is fulfilled: if $\Lambda_1 \subset \Lambda_2$ then the image of the measure μ_{Λ_2} under the natural projection $\Omega_{\Lambda_2} \to \Omega_{\Lambda_1}$ coincides with μ_{Λ_1}. Suppose that for every $t \in T$ the measure μ_t on $\mathcal{B}_t$ possesses an approximating compact class $\mathcal{K}_t \subset \mathcal{B}_t$. Then there exists a probability measure μ on the measurable space $(\Omega := \Pi_{t\in T}\Omega_t, \mathcal{B} := \otimes_{t\in T}\mathcal{B}_t)$ such that the image of μ under the natural projection from Ω to Ω_Λ is μ_Λ for each finite set $\Lambda \subset T$.*

Appendix D. Probabilistic morphisms and the category of statistical models

In 1962 Lawere proposed a categorical approach to probability theory, where morphisms are Markov kernels, and most importantly, he supplied the space $\mathcal{P}(\mathcal{X})$ with a natural σ-algebra Σ_w, making the notion of Markov kernels and hence many constructions in probability theory and mathematical statistics functorial.

Given a measurable space $\mathcal{X}$, let $\mathcal{F}_s(\mathcal{X})$ denote the linear space of simple functions on $\mathcal{X}$. There is a natural homomorphism $I : \mathcal{F}_s(\mathcal{X}) \to \mathcal{S}^*(\mathcal{X}) := Hom(S(\mathcal{X}), \mathbb{R})$, $f \mapsto I_f$, defined by integration: $I_f(\mu) := \int_{\mathcal{X}} f d\mu$ for $f \in \mathcal{F}_s(\mathcal{X})$ and $\mu \in \mathcal{S}(\mathcal{X})$. Following Lawvere [Lawvere1962], we define Σ_w to be

the smallest σ-algebra on $\mathcal{S}(\mathcal{X})$ such that I_f is measurable for all $f \in \mathcal{F}_s(\mathcal{X})$. We also denote by Σ_w the restriction of Σ_w to $\mathcal{M}(\mathcal{X})$, $\mathcal{M}^*(\mathcal{X}) := \mathcal{M}(\mathcal{X})\backslash\{0\}$, and $\mathcal{P}(\mathcal{X})$.

• For a topological space $\mathcal{X}$ we shall consider the natural Borel σ-algebra $\mathcal{B}(\mathcal{X})$. Let $C_b(\mathcal{X})$ be the space of bounded continuous functions on a topological space $\mathcal{X}$. We denote by τ_v the smallest topology on $\mathcal{S}(\mathcal{X})$ such that for any $f \in C_b(\mathcal{X})$ the map $I_f : (\mathcal{S}(\mathcal{X}), \tau_v) \to \mathbb{R}$ is continuous. We also denote by τ_v the restriction of τ_v to $\mathcal{M}(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$. If $\mathcal{X}$ is separable and metrizable then the Borel σ-algebra on $\mathcal{P}(\mathcal{X})$ generated by τ_v coincides with Σ_w.

Definition D.1. ([JLT2020a, Definition 2.4]) *A probabilistic morphism*[28] (or *an arrow*) from a measurable space $\mathcal{X}$ to a measurable space $\mathcal{Y}$ is a measurable mapping from $\mathcal{X}$ to $(\mathcal{P}(\mathcal{Y}), \Sigma_w)$.

Example D.2. Recall that the conditional probability measure $\mu_{\mathcal{X}|\mathcal{Y}}(\cdot|y)$ is defined in (A.15). Then $\mu_{\mathcal{X}|\mathcal{Y}}(\cdot|y)$ is regular, iff there exists a measurable mapping $\overline{P} : \mathcal{Y} \to (\mathcal{P}(\mathcal{X}), \Sigma_w)$ such that for all $A \in \Sigma_{\mathcal{X}}$ the function $\overline{P}(\cdot)(A)$ represents the equivalence class of $\mu_{\mathcal{X}|\mathcal{Y}}(A|\cdot)$ in $L^1(\mathcal{Y}, (\Pi_{\mathcal{Y}})_*\mu_{\mathcal{X}\times\mathcal{Y}})$. In [JLT2020a, Proposition 3.6] we proved that this is equivalent to the existence of probabilistic morphism $\mathbf{p} : \mathcal{Y} \rightsquigarrow \mathcal{X}$ such that $\mu_{\mathcal{X}} = \mathbf{p}_*(\mu_{\mathcal{Y}})$ where $\mu_{\mathcal{X}} = \Pi_{\mathcal{X}}(\mu)$ and $\mu_{\mathcal{Y}} = \Pi_{\mathcal{Y}}(\mu)$.

We shall denote by $\overline{T} : \mathcal{X} \to (\mathcal{P}(\mathcal{Y}), \Sigma_w)$ the measurable mapping defining/generating a probabilistic mapping $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$. Similarly, for a measurable mapping $\mathbf{p} : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ we shall denote by $\underline{\mathbf{p}} : \mathcal{X} \rightsquigarrow \mathcal{Y}$ the generated probabilistic morphism. Note that a probabilistic morphism is denoted by a curved arrow and a measurable mapping by a straight arrow.

Example D.3. ([JLT2020a, Example 2.6]) Let δ_x denote the Dirac measure concentrated at x. It is known that the map $\delta : \mathcal{X} \to (\mathcal{P}(\mathcal{X}), \Sigma_w)$, $x \mapsto \delta(x) := \delta_x$, is measurable [Giry1982]. If $\mathcal{X}$ is a topological space, then the map $\delta : \mathcal{X} \to (\mathcal{P}(\mathcal{X}), \tau_v)$ is continuous, since the composition $I_f \circ \delta : \mathcal{X} \to \mathbb{R}$ is continuous for any $f \in C_b(\mathcal{X})$. Hence, if $\kappa : \mathcal{X} \to \mathcal{Y}$ is a measurable mapping between measurable spaces (resp. a continuous mapping between separable metrizable spaces), then the map $\overline{\kappa} : \mathcal{X} \stackrel{\delta\circ\kappa}{\to} \mathcal{P}(\mathcal{Y})$ is a measurable mapping (resp. a continuous mapping). We regard κ as a probabilistic morphism defined by $\delta \circ \kappa : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. In particular, the identity mapping $Id : \mathcal{X} \to \mathcal{X}$ of a measurable space $\mathcal{X}$ is a probabilistic morphism generated by $\delta : \mathcal{X} \to \mathcal{P}(\mathcal{X})$. Graphically speaking, any straight arrow (a measurable mapping) $\kappa : \mathcal{X} \to \mathcal{Y}$ between measurable spaces can be seen as a curved arrow (a probabilistic morphism).

Given a probabilistic morphism $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$, we define a linear map $S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$, called *Markov morphism*, as follows [Chentsov1972,

[28] we thank J.P. Vigneaux for suggesting us to use "probabilistic morphism" instead of "probabilistic mapping"

Lemma 5.9, p. 72]

$$S_*(T)(\mu)(B) := \int_{\mathcal{X}} \overline{T}(x)(B) d\mu(x) \tag{D.1}$$

for any $\mu \in \mathcal{S}(\mathcal{X})$ and $B \in \Sigma_{\mathcal{Y}}$.

Proposition D.4. *([JLT2020]) Assume that $T : \mathcal{X} \rightsquigarrow \mathcal{Y}$ is a probabilistic morphism.*

(1) Then T induces a linear bounded map $S_*(T) : \mathcal{S}(\mathcal{X}) \to \mathcal{S}(\mathcal{Y})$ w.r.t. the total variation norm $||\cdot||_{TV}$. The restriction $M_*(T)$ of $S_*(T)$ to $\mathcal{M}(\mathcal{X})$ (resp. $P_*(T)$ of $S_*(T)$ to $\mathcal{P}(\mathcal{X})$) maps $\mathcal{M}(\mathcal{X})$ to $\mathcal{M}(\mathcal{Y})$ (resp. $\mathcal{P}(\mathcal{X})$ to $\mathcal{P}(\mathcal{Y})$).

(2) Probabilistic morphisms are morphisms in the category of measurable spaces, i.e., for any probabilistic morphism $T_1 : \mathcal{X} \rightsquigarrow \mathcal{Y}$ and $T_2 : \mathcal{Y} \rightsquigarrow \mathcal{Z}$ we have

$$M_*(T_2 \circ T_1) = M_*(T_2) \circ M_*(T_1),\ P_*(T_2 \circ T_1) = P_*(T_2) \circ P_*(T_1). \tag{D.2}$$

(3) M_* and P_* are faithful functors.

(4) If $\nu \ll \mu \in \mathcal{M}^*(\mathcal{X})$ then $M_*(T)(\nu) \ll M_*(T)(\mu)$.

References

[Amari2016] S. Amari, Information Geometry and its applications, Springer, 2016.

[AJLS2015] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer, Information geometry and sufficient statistics, Probability Theory and related Fields, 162 (2015), 327-364, arXiv:1207.6736.

[AJLS2017] N. Ay, J. Jost, H. V. Lê and L. Schwachhöfer, Information Geometry, Springer, 2017.

[AJLS2018] N. Ay, J. Jost, H. V. Lê and L. Schwachhöfer, Parametrized measure models, Bernoulli vol. 24 Nr 3 (2018), 1692-1725, arXiv:1510.07305.

[BHMSY2019] S. Ben-David, P. Hrubes, S. Moran and A. Yehudayoff, Learnability can be undecidable, Nature Machine intelligence, 1(2019), 44-48.

[Billingsley1999] P. Billingsley, Convergence of Probability measures, 2nd edition, John Wiley ans Sohns, 1999.

[Bishop2006] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[Bogachev2007] V. I. Bogachev, Measure theory I, II, Springer, 2007.

[Bogachev2010] V. I. Bogachev, Differentiable Measures and the Malliavin Calculus, AMS, 2010.

[Bogachev2018] V. I. Bogachev, Weak convergence of measures, AMS, 2018.

[Borovkov1998] A. A. Borovkov, Mathematical statistics, Gordon and Breach Science Publishers, 1998.

[CM2010] G. Carlsson and F. Memoli, Classifying clustering schemes, arXiv:1011.527.

[CP1997] J. T. Chang and D. Pollard, Conditioning as disintegration, Statistica Neerlandica. 51(1997), 287-317.

[Chentsov1972] N. N. Chentsov, Statistical Decision Rules and Optimal Inference, Translation of mathematical monographs, AMS, Providence, Rhode Island, 1982, translation from Russian original, Nauka, Moscow, 1972.

[CS2001] F. Cucker and S. Smale, On mathematical foundations of learning, Bulletin of AMS, 39(2001), 1-49.

[Cybenko1989] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems, vol. 2(1989), pp. 303-314.

[DGL1997] L. Devroye, L. Györfi and G. Lugosi, A probabilstic theory of Pattern Recognition, Springer 1996.
[Doob1953] J. L. Doob, Stochastic processes, John Wiley & and Sons, 1953.
[Faden1985] A.M. Faden, The existence of regular conditional probabilities: necessary and sufficient conditions. Ann. Probab. 13(1985), 288 - 298.
[Fisher1925] R. A. Fisher, Theory of statistical estimation, Proceedings of the Cambridge Philosophical Society, 22(1925), 700-725.
[FHIBP2018] V. Franois-Lavet, P. Henderson, R. Islam, M.G. Bellemare, J. Pineau, An Introduction to Deep Reinforcement Learning, Foundations and Trends in Machine Learning. 11 (3-4)(2018),219-354. arXiv:1811.12560.
[Fritz2019] T. Fritz, A synthetic approach to Markov kernel, conditional independence and theorem of sufficient statistics, arXiv:1908.07021.
[GBC2016] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT, 2016.
[Ghahramani2013] Z. Ghahramani, Bayesian nonparametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A 371 (2013), 20110553.
[Ghahramani2015] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature, 521(2015), 452-459.
[Giry1982] M. Giry, A categorical approach to probability theory, In: B. Banaschewski, editor, Categorical Aspects of Topology and Analysis, Lecture Notes in Mathematics 915, 68- 85, Springer, 1982.
[GS2003] J.K. Ghosh and R.V. Ramamoorthi, Bayesian Nonparametrics, Springer, 2003.
[Graves2012] A. Graves, Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012.
[GV2017] S. Ghosal and A. van der Vaart, Fundamentals of Nonparametric Bayesian Inference, Cambridge University Press, 2017.
[JLS2017] J. Jost, H. V. Lê and L. Schwachhöfer, The Cramér-Rao inequality on singular statistical models, arXiv:1703.09403.
[Halmos1950] P.R. Halmos, Measure theory, Van Nostrand 1950.
[Haykin2008] S. Haykin, Neural Networks and Learning Machines, 2008.
[HTF2008] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer 2008.
[Hoeffding1963] W. Hoeffding, Probability inequalities for sums of bounded random variables, J. Am. Stat. Assoc., 58(301):13-30, 1963.
[IH1981] I. A. Ibragimov and R. Z. Has'minskii, Statistical Estimation: Asymptotic Theory, Springer, 1981.
[IZ2013] P. Igleasias-Zemmour, Diffeology, AMS, 2013.
[Janssen2003] A. Janssen, A nonparametric Cramér-Rao inequality, Statistics & Probability letters, 64(2003), 347-358.
[Jaynes2003] E. T. Jaynes, Probability Theory The Logic of Sciences, Cambridge University Press, 2003.
[Jost2005] J. Jost, Postmodern Analysis, Springer, 2005.
[JLT2020a] J. Jost, H. V. Lê, and T. D. Tran, Probabilistic morphisms and Bayesian nonparametrics,
[JLT2020] J. Jost, H. V. Lê and T. D. Tran, Differentiation of measures on complete Riemannian manifolds,
[JP2003] J. Jacod and P. Protter, Probability Essentials, Springer, 2. edition, 2004.
[Kallenberg2002] O. Kallenberg, Foundations of modern Probability, Springer, 2002.
[Kallenberg2017] O. Kallenberg, Random measures, Springer, 2017.
[KM1997] A. Kriegl and P. W. Michor, The Convenient Setting of Global Analysis, AMS, 1997.

[Kolmogoroff1933] A. Kolmogoroff, Grundbegriife der Wahrscheinlichkeitensrechnung. Springer, Berlin, 1933. English transl: Foundations of the Theory of Probability, Chelsea, New York, 1950.

[Kolmogorov1957] A. Kolmogorov, On the representation of continuous functions of many variables by superposition of continuous functions of one variable and additions, Dokl. Acad. Nauk USSR, 114(1957), 953-956.

[Kullback1959] S. Kullback, Information theory and statistics, John Wiley and Sons, 1959.

[Lawvere1962] W. F. Lawvere, The category of probabilistic mappings (1962). Unpublished, Available at https://ncatlab.org/nlab/files/lawvereprobability1962.pdf.

[Le2017] H.V. Lê, The uniqueness of the Fisher metric as information metric, AISM, 69 (2017), 879-896, arXiv:math/1306.1465.

[Le2020] H. V. Lê, Diffeological statistical models, the Fisher metric and probabilistic mappings, Mathematics 2020, 8, 167; doi:10.3390/math8020167, arXiv:1912.02090.

[LT2020] H. V. Lê and A. Tuzhilin, Nonparametric estimations and the diffeological Fisher metric, (2020)

[Ledoux2001] M. Ledoux, The concentration of measure phenomenon, AMS, 2001.

[Lorentz1976] G. Lorentz, The thirteen problem of Hilbert, In Proceedings of Symposia in Pure Mathematics, 28(1976), 419-430, Providence, RI.

[Lugosi2009] G. Lugosi, Concentration of measure inequalities - lecture notes, 2009, available at http://www.econ.upf.edu/∼lugosi/anu.pdf.

[LC1998] E. L. Lehmann and G. Casella, Theory of Point Estimation, Springer, 1998.

[McCullagh2002] P. McCullagh, What is a statistical model, The Annals of Statistics 2002, Vol. 30, No. 5, 1225-1310.

[MFSS2017] K. Muandet, K. Fukumizu, B. Sriperumbudur and B. Schölkopf, Kernel Mean Embedding of Distributions: A Review and Beyonds, Foundations and Trends in Machine Learning: Vol. 10: No. 1-2, pp 1-141 (2017), arXiv:1605.09522.

[MRT2012] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations of Machine Learning, MIT Press, 2012.

[Murphy2012] K. P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

[Neal1996] R. M. Neal, Bayesian Learning for Neural Networks, Springer, 1996.

[Neveu1970] J. Neveu, Bases Mathématiques du Calcul de Probabilités,deuxième edition, Masson, Paris, (1970).

[RN2010] S. J. Russell and P. Norvig, Artificial Intelligence A Modern Approach, Prentice Hall, 2010.

[Robert2007] C. P. Robert, The Bayesian Choice, From Decision-Theoretic Foundations to Computational Implementation, Springer, 2007.

[SB2018] R. S. Sutton and A. G. Barto, Reinforcement Learning, second edition, MIT, 2018.

[Shioya2016] T. Shioya, Metric Measure Geometry Gromov's Theory of Convergence and Concentration of Metrics and Measures, EMS, 2016.

[SSBD2014] S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.

[Schervish1997] M. J. Schervish, Theory of Statistics, Springer, 2d corrected printing, 1997.

[Sugiyama2015] M. Sugiyama, Statistical Reinforcement learning, CRC Press, 2015.

[Sugiyama2016] M. Sugiyama, Introduction to Statistical Machine Learning, Elsevier, 2016.

[Toetall2020] Tô T. D., Protin F., Nguyen T.T. H., Nguyen D. T., C. Piffault, Rodriguez W., Figueroa S., H. V. Lê, W. Tuschmann, N. T. Zung, Epidemic Dynamics via Wavelet Theory and Machine Learning with Applications to Covid-19, arXiv:2010.14004.

[Tsybakov2009] A. B. Tsybakov, Introduction to Nonparametric Estimation, Springer, 2009.

[Valiant1984] L. Valiant, A theory of the learnable, Communications of the ACM, 27, 1984.

[Vapnik1998] V. Vapnik, Statistical learning theory, John Willey and Sons, 1998.

[Vapnik2000] V. Vapnik, The nature of statistical learning theory, Springer, 2000.

[Vapnik2006] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, 2006.

[Wald1950] A. Wald, Statistical decision functions, Wiley, New York; Chapman & Hall, London, 1950.