553.740 Course Notes

October 20, 2023

Contents

1 Lecture 1

1.1 Administrativa

1.1.1 Git Repo

Assignments, starter code, and data will be housed in a Git repository:
https://github.com/schmidttgenstein/fa23-mli/ Please do not push anything to this repo!

1.2 Course Overview

1.2.1 Introduction

From the syllabus:

Machine Learning describes a mishmash of computational techniques for "finding patterns in data." The scope of use, analytic tools, algorithms, and results are almost too numerous to meaningfully batch all such applications under a common appellation. Still, we try. This course focuses on *supervised* machine learning (sml) which roughly deals with using historical *labeled* data to construct a predictor which will correctly label future data.

Formally, we will be operating in space $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes "input" space, and $\mathcal{Y}$ "output." While these notions are heuristic, they well frame the situation that we may easily sample data at will from $\mathcal{X}$, while sampling from $\mathcal{Y}$ may be difficult or expensive, and often we would like to decision according to how we believe $x \in \mathcal{X}$ is associated with label $y \in \mathcal{Y}$.

In this course, you will learn how to formulate the supervised learning problem in mathematical terms, how to describe a measure of performance, restrict search space for constructing models for prediction, optimize performance measure in search space, and how to check for generalization. You

will learn, also, how to implement some of these methods in code, from the ground up, as well as incorporating pre-built libraries (such as pytorch) for such tasks. Finally, you will learn how to articulate learning guarantees, and understand some of the limits of learning claims. While this course is primarily theory-centric, there will be no dearth of opportunity for employing concrete computational techniques.

1.2.2 The Standard Diagram

Describing our problem space in more detail, consider the following diagram

$$\begin{array}{ccc} \mathcal{X} \times \mathcal{Y} & \xrightarrow{\pi_{\mathcal{Y}}} & \mathcal{Y} \\ \downarrow{\scriptstyle \pi_{\mathcal{X}}} & \nearrow_{\tilde{y}} & \\ \mathcal{X} & & \end{array} \tag{1}$$

We will refer to this diagram often, and to do so give it the somewhat non-descriptive, but in our context wholly unambiguous, name 'the standard diagram.'

Traditionally, $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, defined by mapping $(x, y) \mapsto x$ (read: $\pi_{\mathcal{X}}(x, y) := x$), is taken to be "easy, efficient, or cheap" to evaluate or sample while $\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$, defined by mapping $(x, y) \mapsto y$, is computationally expensive, expensive otherwise, difficult for other reasons, or altogether infeasible. The original space $\mathcal{X} \times \mathcal{Y}$ is itself inaccessible, except for some *labeled* data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$ which provides a proxy (and incomplete!) illustration of what $\mathcal{X} \times \mathcal{Y}$ looks like. The map $\tilde{y} : \mathcal{X} \dashrightarrow \mathcal{Y}$ is a critter we'd like to construct from data S so that both $\tilde{y}(x) \approx y$ for $(x, y) \in S$ *and* for $(x, y) \in \mathcal{X} \times \mathcal{Y}$. What "$\approx$" means, how to construct $\tilde{y}$, conditions on S which are needed to make this problem feasible, etc. are all aspects of the supervised machine learning problem which we will explore in this course.

As an example, suppose $\mathcal{X} = \mathbb{R}$ denotes credit score and $\mathcal{Y} = \{0, 1\}$ loan repayment (say '1' corresponds to repayment of loan, '0' to default). Then a *point* $(x, y) \in \mathbb{R}$ represents data corresponding to a loan whose account holder has credit score x and for which the loan was either paid in full ($y = 1$) or not ($y = 0$). The reason we say $\pi_{\mathcal{X}}$ is "easy" to sample is that you may ask any person what their credit score is (more realistically: as creditor, you would see this information *at the time of application*), while loan repayment information (the "label") would not be observed until potentially many years later when the loan is finally repaid or defaults.

It is worth noting, and perhaps lingering upon the observation, that the "input-output" relation (x, y) is not necessarily functional, i.e. for two points (x, y), $(x', y') \in \mathcal{X} \times \mathcal{Y}$, $x = x'$ does not imply that $y = y'$. The stand-in for determinism is probability, i.e. we will presume that there is some joint probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$, and e.g. that $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = f(x)|x) \neq 1$.

1.2.3 Themes

Generalization Given model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$, how well does $\tilde{y}$ match the (finite) data we have $S \subsetneq \mathcal{X} \times \mathcal{Y}$—i.e. $\tilde{y}(x) \approx y$ for $(x, y) \in S$—*and* the data we don't have, $(x, y) \in \mathcal{X} \times \mathcal{Y}$?

Dimensionality Computation in high dimensions becomes harder, in part because computation is more expensive, and because there are more "corners" for data to hide in (which exacerbates the computational problem). The geometry of high dimensionality will be a recurring theme; for now, we simply observe sources of dimensionality:

1. The data "set" itself $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$. Properly speaking, this data will be presumed to be sampled $(x_i, y_i) \sim_{iid} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ which *means* that the data "set" is a *point* in the space $(\mathcal{X} \times \mathcal{Y})^m$. A reasonable question to ask, then, is: what is the measure $\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m}$? Independence tells us that it is $\prod_{j=1}^{m} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$. More on this later.

2. Size of space itself. This could include high dimensionality of $\mathcal{X}$ and/or $\mathcal{Y}$. Examples abound of high dimensional input data: numerous columned tabular data, imagery data, audio data, video data.

3. Parameter space for model $\tilde{y}$. In the case of linear regression, a model $\tilde{y}(x) = \sum_{j=0}^{n} a_j x^j$ may have arbitrarily large degree n. Or a fully connected neural network with many nodes and many layers. And so on. More generally, the space of functions $\mathcal{Y}^{\mathcal{X}} := \{\tilde{y} : \mathcal{X} \to \mathcal{Y}\}$ is even bigger.

Trade-offs Assumptions must be made and compromises allowed for in order to gain tractability in the learning problem. There is no universal solution ("no free lunch," and as you may imagine, there's a theorem for that) and formulating the setup to address one challenge may introduce other ones elsewhere (ML can sometimes feel like one giant game of whackamole).

The famed bias-variance trade-off is one example: a high complexity model may well represent the data S, which in one sense is good, but in another is bad if said model represents data *too well*, i.e. at the exclusion of modeling 'from where the data comes.'

1.3 Zeroth Assignments

Worksheet Please bring hard copy of this worksheet with you to class on Wednesday. You may not answer a question about Independence with only equality or inequality between $f_{\mathcal{X}\times\mathcal{Y}}$ and $f_{\mathcal{X}} \cdot f_{\mathcal{Y}}$: please say something to indicate how you know such equality or inequality.

Programming Assignment You may find the first programming assignment under pa0 in git repo, and starter code in the git. I suggest you follow the tutorial at Real Python real python[1] which shows you how to spin up a logistic regression model using sklearn. Scikit-Learn (also known as sklearn) is an open source ML library for python, and contains functionality for constructing numerous models. This is perhaps the only time in this course you will be asked to use this library, and if you have a preferred alternative library, you are more than welcome to use it for this assignment.

The purpose of the assignment is threefold:

1. Gain initial exposure to the *structure* of machine learning code, including object oriented programming and the typical methods included.
2. Shake off any residual rust using Python.
3. To gain deeper appreciation for the *aim* of building model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ as in diagram (6), and metrics that illustrate success.

Recall the Standard Diagram

$$\begin{array}{ccc} \mathcal{X}\times\mathcal{Y} & \xrightarrow{\pi_{\mathcal{Y}}} & \mathcal{Y} \\ \downarrow{\scriptstyle \pi_{\mathcal{X}}} & \nearrow_{\tilde{y}} & \\ \mathcal{X} & & \end{array} \tag{2}$$

This diagram provides formalism for talking about "approximating" y with model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ when $y \neq y(x)$ is not necessarily functionally determined by x.

Consider a concrete example to illustrate the problem: suppose that $\mathcal{X} = \mathbb{R}$ denotes credit score and $\mathcal{Y} = \{0, 1\}$ denotes repayment on loan, 1 denotes full repayment and 0 denotes default. A data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ corresponds to a credit score-loan repayment pair, and in real life a loan account would have these attributes associated with it, i.e. the debtor would have some credit score and their loan will (eventually) be repaid or not. (Note that "eventuality" is what, in this case, makes $\pi_{\mathcal{Y}}$ hard or expensive to evaluate.) It is possible for two different loans, belonging to two different people, to agree on credit score but disagree on outcome $y \in \mathcal{Y}$. In fact, we will likely observe both outcomes $y = 0$ and $y = 1$ associated to *any* credit score. Presumably, there should be some relation between the relative *counts* of $\#y = 1$ and credit score; in other words, one might suppose that lower credit scores correspond to accounts which in actual fact get repaid less frequently than those with high credit scores. We need mathematical language to describe and work with this phenomenon. The language is probability.

[1] You may need to sign up in order to view this content, but it is not behind a paywall.

Thus, we suppose that there is some joint probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X}\times\mathcal{Y}$. One way of contextualizing supervised machine learning is as a study of probability on the standard diagram. In this lecture, we will review probability, give intuition for probability as measure, as well as notation $\mathbb{P}_{\mathcal{X}} = \int d\mathbb{P}_{\mathcal{X}}(x)$, and define expectation $\mathbb{E}(f)$ of a random variable $f : \mathcal{X} \to \mathbb{R}$ as a *Lebesgue* integral $\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$. We kick this lecture off by emphasizing that probability has nothing to do with randomness…yet. When we return to admitting randomness into our lexicon, it will be as a *result*. For the moment, we forget any association between probability and chance, stochasticity, randomness, or any other (for now) anathema word affiliated with the notion of uncertainty.

1.4 Intuition for Measure from Calculus

We start reviewing integration in calculus to preview notation for measure. One interpretation of the integral is as "area under the curve." Another (what amounts to similar) is as measure. And one interpretation of dQ is as an infinitesimal Q element, whatever Q is. Another is as an indicator of what kind of measurement we are taking. We then express length ℓ as $\int d\ell$, area a as $\int da$, volume v as $\int dv$, measure m as $\int dm$, and so on. Note that the expression $\int dm$ is defined in terms of measure m, i.e. we suppose prior (at the very least conceptual) knowledge of the measure, irrespective of whether or not given a particular object to measure A, we are actually able to evaluate its measure $m(A)$.

Example: length We've learned that the length $\ell([a,b])$ of an interval $[a,b]$ is the difference of endpoints $b-a$. We can write this with an integral as $\ell([a,b]) = \int_a^b dx$ or more in line with notation we will use, as $\ell = \int_{[a,b]} d\ell(x)$. In this notation, the subscript is the object we are measuring, ℓ is the type of measure, and x is a dummy variable. Until we start integrating functions, we won't need it, and we could have just written $\int_{[a,b]} d\ell$. The dummy variable was kept only to clearly delineate that $d\ell$ is not the same thing as dx: we are using calculus intuition only for loose guidance.

General Formula For measure m and object to be measured A, we write $m(A) = \int_A dm$. Right now, do not put too much stock in the—what in calculus would be thought of as infinitesimal—dm term: it is *defined* as a *phrase* with the integral $\int$; neither in isolation makes sense, and the definition goes this way: $\int_A dm := m(A)$. Thus, you need to first know the measure. We'll come back to this momentarily.

Example: Area

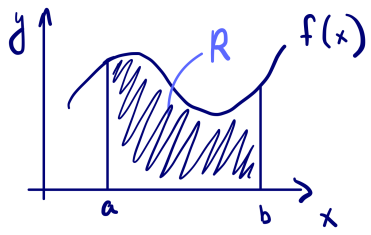

Figure 1: Area

In our notation $A(R) = \int_R dA$, where $A(R)$ is the area. If you want to import calculus intuition, when we interpret dA as a differential area element, we may express it elsewise as the product $f(x)dx$, where $f(x)$ represents height and dx the differential width. As area is equal to height times width (or length) a differential area is a length element times a differential length element ie $A(R) = \int_a^b f(x)dx$.

Example: Volume

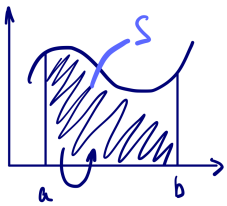

Figure 2: Volume

Using the above picture, $V(S) = \int_S dV$, where $V(S)$ is the volume. Therefore, by generalizing the above idea we can write $\mathbb{P}(A) = \int_A d\mathbb{P}$.

1.5 The Probability Measure

Now we formalize what we mean by probability measure. The first point to make is that it is a measure.

Definition 1.1. Let $\mathcal{X}$ be a set. We define a *probability measure*

$$\mathbb{P}_{\mathcal{X}} : (\text{[Some] Subsets of } \mathcal{X}) \to [0,1]$$

on $\mathcal{X}$ to be a map from (a subset of)[2] the power set of $\mathcal{X}$ to the closed interval $[0,1]$ satisfying the following two properties:

1. $\mathbb{P}_{\mathcal{X}}(\mathcal{X}) = 1$ (space is measure finite and normalized), and
2. $\mathbb{P}_{\mathcal{X}}\left(\bigsqcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}_{\mathcal{X}}(A_j)$ where $\bigsqcup_{j=1}^{\infty} A_j$ denotes disjoint union, i.e. as a set $\bigsqcup_{j=1}^{\infty} A_j = \bigcup_{j=1}^{\infty} A_j$ and $A_i \cap A_j = \varnothing$ for $i \neq j$ (countable additivity).

Recall that the power set $2^{\mathcal{X}}$ of a set $\mathcal{X}$ is defined to be the set of all subsets of $\mathcal{X}$, including $\varnothing$ and $\mathcal{X}$ itself. For example, when $\mathcal{X} = \{1,2,3\}$,

$$2^{\mathcal{X}} = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \mathcal{X}\}.$$

When $\mathcal{X} = \mathbb{R}$, the power set includes any reasonable combination (e.g. unions) of intervals (a,b) or $[a,b]$, but also many many more (see Cantor Set for a fun, but not particularly relevant, excursion into the wonders of measure theory).

The second condition is called 'countable additivity' (informally: a conservation of stuff principle) and represents the intuitive idea that if you slice and dice an object for measurement, measure each

[2] This subtle point is a technicality beyond the scope of this course, and quite frankly unnecessary for reasonably understanding measure. Measure measures subsets. But it's possible that it may not be able to measure *all* subsets. So the domain of the measure may not be *all* subsets. A lot of work goes into specification of the structure of the collection of subsets you can measure, and like topology is characterized by closure operations, e.g., if you can measure $\{A_j\}_{j\in\mathbb{N}}$ then you can measure its union $\bigcup_{j\in\mathbb{N}} A_j$. For the curious, you may look into sigma algebras for more detail.

constituent piece without double counting, and add your results, you'll end up with the same result as if you just measured the whole original unadulterated tamale.

You should check that countable additivity implies *finite* additivity $\mathbb{P}_{\mathcal{X}}\left(\sum_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \mathbb{P}_{\mathcal{X}}(A_j)$. You will need the fact that $\mathbb{P}_{\mathcal{X}}(\varnothing) = 0$, which itself is implied by conditions 1. and 2. Indeed, $\mathcal{X} = \mathcal{X} \sqcup \bigsqcup_{j=2}^{\infty} \varnothing$. And the second fact is called the Union Bound: $\mathbb{P}(\cup_{j\in\mathbb{N}} A_j) \leq \sum_{j\in\mathbb{N}} \mathbb{P}(A_j)$.

Remark 1.1. We are being fairly blasé about the *domain* "some subsets of $\mathcal{X}$" of $\mathbb{P}_{\mathcal{X}}$. Perhaps we shouldn't be. Much of the architecture constructing measure depends on the fact that some structure must be placed on the set of subsets of $\mathcal{X}$ which are suitable for measurement; in particular, that such a set comprises a so-called σ-algebra, is not necessarily (and in many cases in fact is not) the entire power set $2^{\mathcal{X}} = \{A \subset \mathcal{X}\}$, and so on. We pay lip-service to this nuance, but fret little over the possibility that we will accidentally run across both a measure $\mathbb{P}_{\mathcal{X}}$ and subset $A \subset \mathcal{X}$ for which $\mathbb{P}_{\mathcal{X}}(A)$ does not make sense (read: which $\mathbb{P}_{\mathcal{X}}$ is "incapable" of measuring). One must try very hard—you might find such a question on a measure theory qualifying exam—to come up with an example. Therefore you may reasonably suppose that any set you'd come across in real life is in fact measurable. Still, know *that* there is a potential problem: if you can construct a non-measurable set, then you can cut an apple into finitely many pieces and reassemble those finitely many pieces into *two* apples of the same size (see Banach Tarski for more information). In other words, weird things can happen with things that aren't measurable.

Definition 1.2. We define a *probability space* to be a pair $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ where $\mathcal{X}$ is a set and

$$\mathbb{P}_{\mathcal{X}} : (\text{[Some/Many] Subsets of } \mathcal{X}) \to [0, 1]$$

is a probability measure (c.f. definition 1.1).

In probabilistic terminology, measurable subsets $A \subset \mathcal{X}$ are often called 'events' and individual points $x \in \mathcal{X}$ called 'outcomes.' An outcome $x \in \mathcal{X}$ defines the *event* $\{x\} \subset \mathcal{X}$ with the single outcome $x \in \{x\}$.

Example 1 Let $\left(\mathcal{X} = [0, 1], \mathbb{P}_X([a, b]) := b - a\right)$ for $0 \leq a \leq b \leq 1$. We define notation $\int_{[a,b]} d\mathbb{P}(x) = \mathbb{P}([a, b])$.

Example 2 Let $\left(\mathcal{X} = \mathbb{R}, \mathbb{P}_X([a, b]) := \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\right)$, the so-called normal distribution with zero mean and unit variance. Observe that we use a Riemann integral to *compute* or give the rule for realizing the probability measure. The integrand $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ of the Riemann integral is called a *probability density function*. The integral $\int_{[a,b]} d\mathbb{P}_{\mathcal{X}}(x)$, by contrast, is not a Riemann integral; it is *defined* by the measure $\mathbb{P}_{\mathcal{X}}([a, b])$. (When you ask: but what is the measure?, we gave the rule for how to calculate it!)

Example 3 Let $\mathcal{Y} = [0, 1]$, $\mathbb{P}_{\mathcal{Y}}([x_1, x_2]) \times [y_1, y_2]) = (x_2 - x_1) \cdot (y_2 - y_1)$. This example is a preliminary look into independence as $\mathbb{P}_{\mathcal{X}}([x_1, x_2]) = x_2 - x_1$, $\mathbb{P}_{\mathcal{Y}}([y_1, y_2]) = y_2 - y_1$ and $\mathbb{P}_{\mathcal{X}}([x_1, x_2] \times [y_1, y_2]) = \mathbb{P}_{\mathcal{X}}([x_1, x_2])\mathbb{P}_{\mathcal{Y}}([y_1, y_2])$ We will return to this example when we discuss independence, and will want to situate as a notion which is at home in high dimension.

Independence is a high dimensional phenomenon as it is clear that we can generalize the picture in 3 to n dimensions.

Example 4 : Bernouli Random Variable (flipping a fair or unfair coin): Let $\mathcal{X} = \{0, 1\}$ where $\mathbb{P}_{\mathcal{X}}(\{1\}) = p$ and $\mathbb{P}_{\mathcal{X}}(\{0\}) = 1 - p$. This is a probability space with two outcomes that clearly satisfies the axioms stated earlier since the events or subsets of the space X, $\{0\}$ and $\{1\}$ are disjoint and the sum of their probabilities is 1:

$$\mathbb{P}_X(1) + \mathbb{P}_{\mathcal{X}}(0) = p + (1 - p) = 1 = \mathbb{P}_X(\{0\} \sqcup \{1\})$$

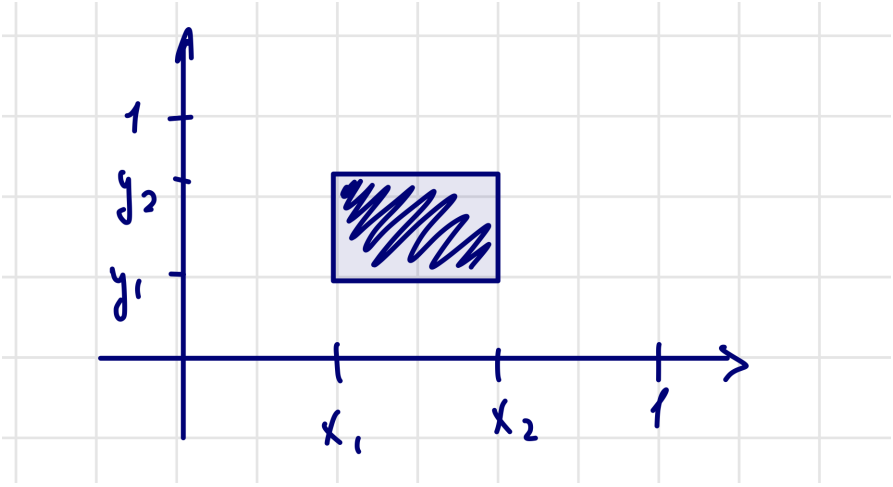

Figure 3: The area measure

2 Lecture 2

2.1 Random Variables

Recall that a probability space $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ consists of a set $\mathcal{X}$ and a measure $\mathbb{P}_{\mathcal{X}}$ (see definition 1.2). We highlighted at the beginning of the lecture that probability fundamentally operates as a theory of measure, which in essence equate it to a theory of integration, a concept that primarily revolves around numbers (values). So far, we've said nothing about the nature of the set $\mathcal{X}$, and we don't need to. We do need is a way to attribute values to outcomes $x \in \mathcal{X}$.

Definition 2.1. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a probability space. We define a *random variable* to be a *function* $f : \mathcal{X} \to \mathbb{R}$, whose codomain is $\mathbb{R}$.

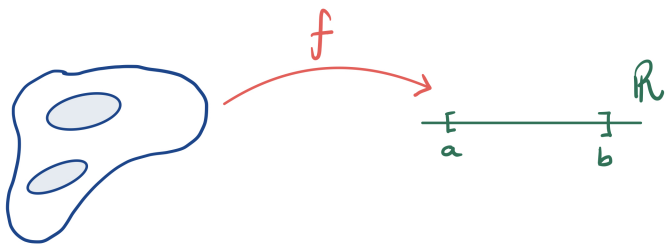

Figure 4: Random variable f

Such a function induces a measure $\mathbb{P}_{\mathbb{R}}$ on $\mathbb{R}$, defined by

$$\mathbb{P}_{\mathbb{R}}([a,b]) := \mathbb{P}_{\mathcal{X}}(f^{-1}([a,b])) = \mathbb{P}_{\mathcal{X}}(\{x \in \mathcal{X} : f(x) \in [a,b]\}). \tag{3}$$

One should check that this defines an honest probability measure on $\mathbb{R}$.

Remark 2.1. It is worth noting that certain conditions must be met by the map $f : \mathcal{X} \to \mathbb{R}$ in order to ensure that the induced measure $\mathbb{P}_{\mathbb{R}}$ is well-defined, i.e. it is a (probability) measure. In essence, we require a condition known as 'measurability,' which essentially means that f must be a *measurable* function (which really just means: f is such that the induced measure is a measure—this isn't circular!, it all comes down to saying, you cannot with impunity claim that any function whatsoever will induce a measure). Just as with the domain of $\mathbb{P}_{\mathcal{X}}$, where we suppose with little guilt that "all subsets" we encounter are measurable, we will also suppose that the functions we come across in practice are

measurable. Again, there is nuance to be appreciated, but the supposition we make will very unlikely harm any of our day to day calculations.

(For the ultra-curious, the condition of measurability stipulates that any potentially measurable set in $\mathbb{R}$ has preimage (by f^{-1}) which *is* measurable in $\mathcal{X}$.)

Remark 2.2. We noted that $f : \mathcal{X} \to \mathbb{R}$ induces a measure. In fact, this holds more generally for any (measurable) map $f : \mathcal{X} \to \mathcal{Y}$, $\mathbb{P}_{\mathcal{Y}}(B) := \mathbb{P}_{\mathcal{X}}(f^{-1}(B))$.

Definition 2.2. Let $f : \mathcal{X} \to \mathbb{R}$ be a random variable. We define *expectation of* f, denoted $\mathbb{E}(f)$, to be

$$\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x). \tag{4}$$

Equation (4) defines expectation, but we have some odd sort of critter we've not seen before on the right hand side. We must define *it*.

Definition 2.3. Let $f : \mathcal{X} \to \mathbb{R}$ be a *simple* random variable, i.e. taking *finitely* many values $a_1, \ldots, a_k$. Then we define the *Lebesgue* integral of f, denoted $\int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$, to be

$$\int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x) := \sum_{j=1}^{k} a_j \mathbb{P}_{\mathcal{X}}(f = a_j) = \sum_{j=1}^{k} a_j \mathbb{P}_{\mathcal{X}}\left(\{x \in \mathcal{X} : f(x) = a_j\}\right).$$

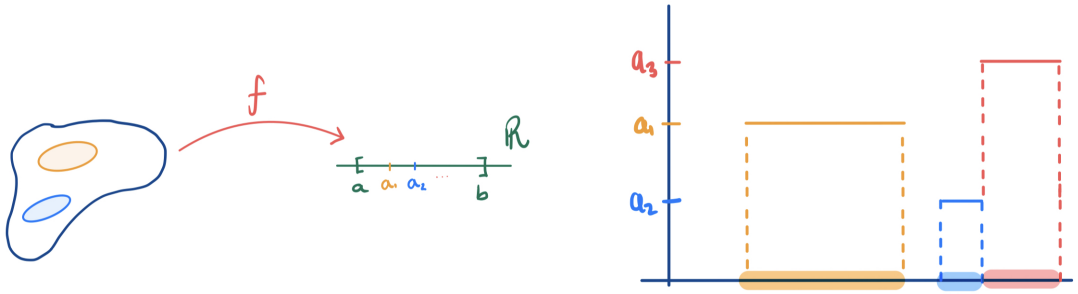

Figure 5: Simple random variable f

One may recognize this definition as the *expectation of a discrete random variable*. Indeed, that is exactly what it is. It is the Lebesgue integral! The Lebesgue integral extends to continuous random variables, but it requires some abstraction. For next time. In the meantime, note that for continuous random variable $f : \mathcal{X} \to \mathbb{R}$ (not necessarily taking finitely many values), we can *approximate* $\mathbb{E}(f)$ using the Lebesgue integral of a simple random variable (expectation of discrete r.v.), see fig. 6.

Remark 2.3. In the following, we will continue the definition of Lebesgue integral for continuous random variable. While there are theoretical reasons for insisting on the use of this abstraction, ours are more practically oriented. In fact, one may (very) often compute expectation *using* instead a Riemann integral. For example, the expectation of a mean 0 unit variance normally distributed random variable is $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} dx$ (which of course is zero). In other words, if push comes to shove and you're asked to *compute* the expectation for a continuous random variable, very often you'll have a density function in your pocket and can just multiply it by the random variable, and integrate, business as usual. We introduce Lebesgue integration for two reasons:

1. To articulate the fact that the extension from discrete to continuous random variables may *seem* confusing, and that is due in part to the nauseating head-spinning move from Lebesgue to who knows (but usually Riemann) integration without even lip-service paid to the fact that expectation of discrete r.v.s itself is a non-trivial extension of our conceptual apparatus.

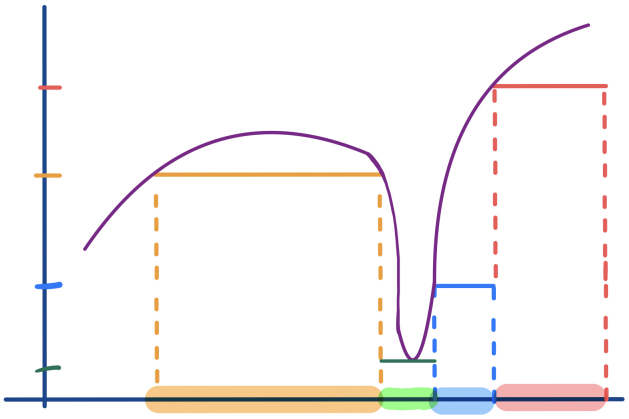

Figure 6: Approximation via simple functions

2. Ease of notation: $d\mathbb{P}_{\mathcal{X}}$ always makes sense and makes immediately obvious what our measure is. In other words, I don't always want to say: suppose that the density of a probability space exists, and anyway it might not and that doesn't matter!, for $\int_A d\mathbb{P}_{\mathcal{X}}(x)$ makes sense regardless of whether we can integrate (Riemann-wise) as $\int \varphi(x)dx$ (for density $\varphi(x)$). Related: the notation $d\mathbb{P}_{\mathcal{X}}$ collapses the distinction between continuous and discrete random variables. The distinction, in my view, is convoluted and confusing, especially when we have mixed discrete-continuous nonsense going on (e.g. in the case of binary classification $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$). Of course, successfully working with the collapse requires an comfort with the abstraction, and that may by itself be initially confusing as well.

Remark 2.4. Observe also that $\mathbb{E}(f)$ is somewhat uninformative, in a way that $\int_{\mathcal{X}} f(x)d\mathbb{P}_{\mathcal{X}}(x)$ is not. At the moment the added notational baggage of the latter may seem inconvenient, but once we start squinting at joint probability spaces $\mathcal{X} \times \mathcal{Y}$, turning them upside down, and so on, it will be imperative to be clear on how we are integrating. For example, we will see

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x, y)d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x, y) = \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}} f(x, y)d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)d\mathbb{P}_{\mathcal{X}}(x).$$

This is sortof an extension on the notational point above. I will be relying on this notation aggressively, so it's crucial to anticipate eventually becoming comfortable with standard manipulations, and recalling (any time there is lingering confusion) that $d\mathbb{P}_{\mathcal{X}}$ is defined, not in isolation, but as a package $\int_A d\mathbb{P}_{\mathcal{X}} := \mathbb{P}_{\mathcal{X}}(A)$.

2.2 Expectation

Recall definition 2.2 for which $\mathbb{E}(f) = \int_{\mathcal{X}} f(x)d\mathbb{P}_{\mathcal{X}}(x)$ for a random variable $f : \mathcal{X} \to \mathbb{R}$ on the probability space $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$. While this definition lays out the concept of expectation, it leaves us pondering what defines the integral itself. We can draw insight from discrete (or simple) random variables like in definition 2.3.

Definition 2.4. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a probability space and $f : \mathcal{X} \to \mathbb{R}^{\geq 0}$ a *nonnegative* random variable. Then we define the *Lebesgue integral* of f as

$$\int_{\mathcal{X}} f(x)d\mathbb{P}_{\mathcal{X}}(x) := \sup\left\{\mathbb{E}(\bar{f}) : 0 \leq \bar{f} \leq f \text{ is simple random variable}\right\}. \tag{5}$$

A visual illustration is provided in fig. 7.[3]

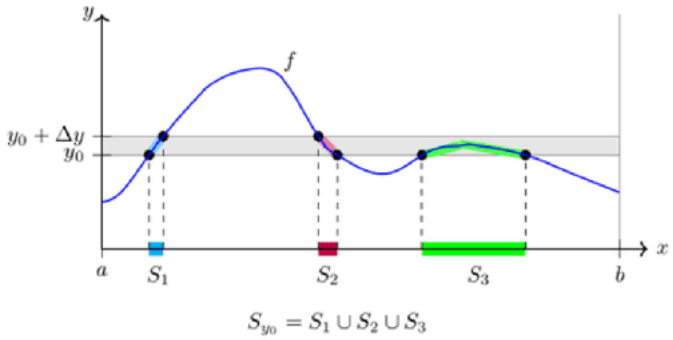

Figure 7: Visualizing Lebesgue Integration

We provide this definition because it is the definition given to us of expectation. The move to continuous random variables may seem abstract—indeed, we discretize the codomain $\mathbb{R}$ instead of domain $\mathcal{X}$ as we are used to doing with Riemann integration from calculus—but one may rest assured that in many instances expectation *may* be computed as a Riemann integral instead. In particular, when a *density* function exists, $\rho : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{P}_{\mathcal{X}}([a,b]) = \int_a^b \rho(x)dx$, we may compute expectation using the following procedure:

1. multiply the random variable f by density ρ; this will give us a function $f \cdot \rho : \mathcal{X} \to \mathbb{R}$, and
2. compute the Riemann integral $\int_{-\infty}^{\infty} f(x)\rho(x)dx$.

When the density does not exist, the definition of expectation still makes sense, and one cannot resort to this procedure for computation. Thus for the sake of understanding, one may choose to delve into the nuance of Lebesgue integration, or pretend, as we've been pretending about measurable sets and measurable functions, that "everything" can be computed as a Riemann integral.

Example 2.1. We once again point out that probability may be computed as expectation:

$$\mathbb{P}_{\mathcal{X}}(A) = 1 \cdot \mathbb{P}_{\mathcal{X}}(A) + 0 \cdot \mathbb{P}_{\mathcal{X}}(\mathcal{X} \setminus A) = \int_{\mathcal{X}} \mathbb{1}_{x \in A} \, d\mathbb{P}_{\mathcal{X}}(x) = \mathbb{E}(\mathbb{1}_{x \in A}).$$

Observe that this Lebesgue integral uses the definition from definition 2.3; we do not need to rely on any limiting procedure (as in e.g. definition 2.4).

For the sake of completeness, we define expectation for arbitrary r.v. $f : \mathcal{X} \to \mathbb{R}$.

Definition 2.5. Let $f : \mathcal{X} \to \mathbb{R}$ and suppose that $\mathbb{E}(f \cdot \mathbb{1}_{f \geq 0}) < \infty$ and $\mathbb{E}(-f \cdot \mathbb{1}_{f < 0}) < \infty$ (both expectations of non-negative random variables). Then we define

$$\mathbb{E}(f) := \mathbb{E}(f \cdot \mathbb{1}_{f \geq 0}) - \mathbb{E}(-f \cdot \mathbb{1}_{f < 0})$$

Remark 2.5. Expectation is already defined as $\mathbb{E}(f) = \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$. This definition is defining the right hand side.

Definition 2.6. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$ be probability spaces. A map $f : (\mathcal{X}, \mathbb{P}_{\mathcal{X}}) \to (\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$ is a map from $\mathcal{X}$ to $\mathcal{Y}$ that *respects the measure* i.e.

$$\mathbb{P}_{\mathcal{Y}}(B) = \mathbb{P}_{\mathcal{X}}(f^{-1}(B))$$

for all $B \subseteq \mathcal{X}$.

[3] An Introduction to the Lebesgue Integral, Ikhlas Adi, 2017

2.3 Marginalization

We now turn to induced probability measures on the standard diagram

$$\begin{array}{ccc} \mathcal{X}\times\mathcal{Y} & \xrightarrow{\pi_{\mathcal{Y}}} & \mathcal{Y} \\ \downarrow{\scriptstyle \pi_{\mathcal{X}}} & \nearrow_{\tilde{y}} & \\ \mathcal{X} & & \end{array} \tag{6}$$

Let $(\mathcal{X}\times\mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ be a joint probability measure. The projection maps $\mathcal{X}\times\mathcal{Y} \xrightarrow{\pi_{\mathcal{X}}} \mathcal{X}$ and $\mathcal{X}\times\mathcal{Y} \xrightarrow{\pi_{\mathcal{Y}}} \mathcal{Y}$ induce probability measures on $\mathcal{X}$ and $\mathcal{Y}$ defined as:

$$\mathbb{P}_{\mathcal{X}}(A) := \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\pi_{\mathcal{X}}^{-1}(A)) = \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A\times\mathcal{Y}). \tag{7}$$

The second equality follows from the fact that $\pi_{\mathcal{X}}^{-1}(A) := \{(x,y)\in\mathcal{X}\times\mathcal{Y} : \pi_{\mathcal{X}}(x,y)\in A\}$. Of course, the marginalization for $\mathbb{P}_{\mathcal{Y}}$ is defined analogously.

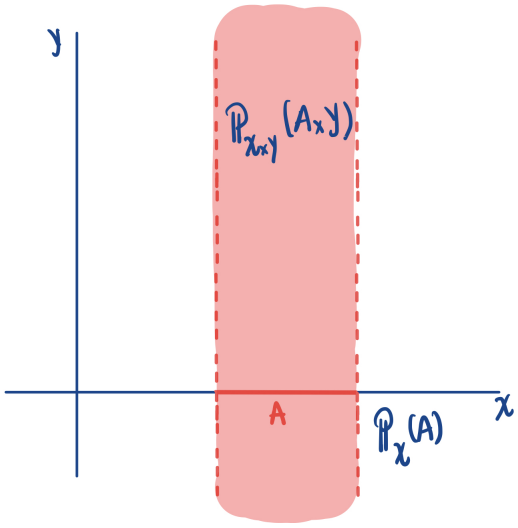

Figure 8

Quite literally, marginalization is projection: it's a mechanism for putting a probability measure on the projected space assuming the existence of probability measure upstairs.

2.4 Conditional Probability

For a probability space $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ you've likely seen the definition of conditional probability as $\mathbb{P}_{\mathcal{X}}(A|B) := \mathbb{P}_{\mathcal{X}}(A\cap B)/\mathbb{P}_{\mathcal{X}}(B)$ provided that the denominator is nonzero. It is easier to visualize this notion with joint probability. We continue with supposing that $(\mathcal{X}\times\mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ is a joint probability space. We've already defined marginal probability, so we can define the following conditional probability.

Definition 2.7. The conditional probability $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(B|A)$ is defined to be

$$\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(B|A) := \frac{\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A\times B)}{\mathbb{P}_{\mathcal{X}}(A)} \tag{8}$$

provided that the marginal $\mathbb{P}_{\mathcal{X}}(A) \neq 0$.

In general, you should be comfortable manipulating expressions with notation $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ for which we have implicit definition

$$\int_A \int_{B|A} d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)\, d\mathbb{P}_{\mathcal{X}}(x) := \int_{A\times B} d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y), \tag{9}$$

or for random variable $f : \mathcal{X} \times \mathcal{Y}$ we have

$$\int_{\mathcal{X}} \int_{\mathcal{Y}|\mathcal{X}} f(x,y) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) d\mathbb{P}_{\mathcal{X}}(x) := \int_{\mathcal{X}\times\mathcal{Y}} f(x,y) d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y). \tag{10}$$

You may recognize the original definition in eq. (8) as secretly appearing here, since $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}} d\mathbb{P}_{\mathcal{X}} = d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ may be "solved" for $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$. But remember, in addition, that this '$d\mathbb{P}$' notation itself is defined as a package $\int_A d\mathbb{P} = \mathbb{P}(A)$.

Remark 2.6. There were many questions on interpreting the critter $A|B$. In pictures I drew the rectangle $A \times B$. It is fine to think of $A|B$ pictorially as $A \times B$, but it is important to concomitantly think of the ambient space in which $A|B$ lives: $A \times B$, $A|B$, $B|A$ are all "the same" rectangle, but $A \times B \subset \mathcal{X} \times \mathcal{Y}$, $A|B \subset \mathcal{X}|B$ (which you may visualize as the rectangle $\mathcal{X} \times B$ which rectangle has probability one, *not* $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\mathcal{X} \times B)$ which may be less than one). Similarly, $B|A$ is the rectangle $A \times B$, but instead of living in $\mathcal{X} \times \mathcal{Y}$ it lives in $\mathcal{Y}|A$ (or visually the rectangle $A \times \mathcal{Y}$ with unit probability).

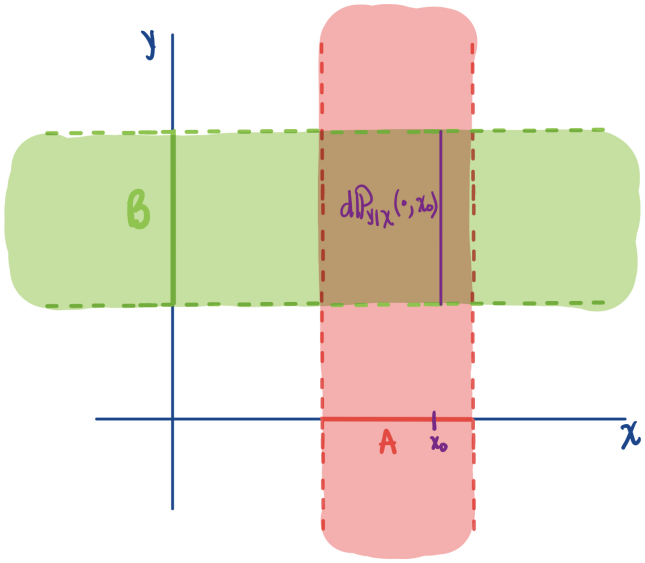

Figure 9

In math, hand-waiving can sometimes be dangerous. On the other hand, it can sometimes allow us to think reasonably about, and operationalize our intuition of, notions whose formalism is "beyond the scope of this course," all with the aim of performing computations and symbolic manipulations. If you are interested in more rigorously making sense of $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$, I recommend diving into the measure theory. The level of rigor we need is: to fluidly manipulate integral expressions involving joint probability, relying on our multivariable intuition from calculus that we may decompose high dimensional integration as sequential one-dimensional integrals. What changes with probability, is that the (measure used for the) inner one-dimensional integral may depend on the outer value.

2.5 Independence

Forget, now, our reliance on joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$, and suppose that we have two separate probability spaces $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$. From these two spaces, one may reasonably ask if we can "go the other direction" and construct a (joint) probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $(\mathcal{X} \times \mathcal{Y})$. The answer is yes.

Before doing so, let us reason about the properties we would like this measure to have. The first thing we should expect is that projection $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ *preserves measure*, i.e. that

$$\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A \times \mathcal{Y}) = \mathbb{P}_{\mathcal{X}}(\pi_{\mathcal{X}}(A \times \mathcal{Y})) = \mathbb{P}_{\mathcal{X}}(A).$$

Observe that in contrast to eq. (7), we do not define $\mathbb{P}_{\mathcal{X}}(A)$ *in terms of* $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$. Instead, the definition goes the other way:

$$\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A \times \mathcal{Y}) := \mathbb{P}_{\mathcal{X}}(A). \tag{11}$$

Of course, we would like the analogous equality to hold for $\mathcal{X} \times B$ with $\mathbb{P}_{\mathcal{Y}}(B)$.

Now this condition by itself only allows us to define sets of the form $A \times \mathcal{Y}$ or $\mathcal{X} \times B$. For general rectangle $A \times B \subset \mathcal{X} \times \mathcal{Y}$, define

$$\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A \times B) := \mathbb{P}_{\mathcal{X}}(A) \cdot \mathbb{P}_{\mathcal{Y}}(B). \tag{12}$$

Observe, in connection with our first encounter of independence as $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(b)$, we can recover this relation by observing that

$$A \times B = (A \times \mathcal{Y}) \cap (\mathcal{X} \times B), \tag{13}$$

and compute

$$\begin{aligned} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A \times B) &= \mathbb{P}_{\mathcal{X}}(A)\mathbb{P}_{\mathcal{Y}}(B) \\ &= \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A \times \mathcal{Y}) \cdot \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\mathcal{X} \times B). \end{aligned} \tag{14}$$

The intuition we extract from this construction is that independence is a most natural way to construct measure on high dimensional space using measure from its lower dimensional components. In exactly the same way that we construct area from length or volume from area and length (or volume from length).

Definition 2.8. In general, we say that $\left(\mathcal{X}^m, \mathbb{P}_{\mathcal{X}^m}\right)$ is *independent* if

$$\mathbb{P}_{\mathcal{X}^m} = (\mathbb{P}_{\mathcal{X}})^m$$

or more generally that $\left(\prod_{j=1}^{m} \mathcal{X}_j, \mathbb{P}_{\prod \mathcal{X}_j}\right)$ independent if

$$\mathbb{P}_{\prod \mathcal{X}_j} = \prod_{j=1}^{m} \mathbb{P}_{\mathcal{X}_j}.$$

Remark 2.7. One should verify that the (quasi-independence) conditions from the last problem of ws0 are a consequence of this definition (hint: marginalization).

3 Lecture 3

We continue discussing independence as a high-dimensional phenomenon, and introduce concentration.

We started with a quick review of conditional probability from last time, providing geometric interpretation for $A \times B$, $A|B$ and $B|A$ as "sets" in $\mathcal{X} \times \mathcal{Y}$ (remark 2.6).

3.1 Independence

Recall definition 2.8 which says that a joint probability measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ is independent if it decomposes as a product of marginals $\mathbb{P}_{\mathcal{X}} \cdot \mathbb{P}_{\mathcal{Y}}$. This extends to products of arbitrarily many factors.

To recognize independence, either of the following suffice:

1. conditional probability is independent of condition, or (what amounts to the same)
2. conditional probability is equal to the marginal

One may examine the independence question from worksheet for density

$$f(x,y) = cy^{-1/2}x\mathbb{1}_{y>0}\mathbb{1}_{x\geq 0}\mathbb{1}_{x^2+y^2\leq 1},$$

where the support of $f(x_0,\cdot)$—i.e. the function $f_{x_0} : \mathcal{Y} \to \mathbb{R}$ defined by $f_{x_0}(y) = f(x,y)$—is readily seen to differ according to the value of x_0, namely $\left[0, \sqrt{1-x_0^2}\right]$.

Ultimately, our purpose in going through this rigmarole is for computation. I would like you to feel comfortable doing symbolic manipulations e.g. of the form $\int_{\mathcal{X}\times\mathcal{Y}} = \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}}$. You might see this in other contexts as the "tower" property. If you are familiar with iterated and embedded expectations, then you need not refer to the geometric interpretation. Use whichever viewpoint you are most at home with when doing computations with joint probability spaces.

Example 3.1. Let us examine why we were destined to fail on the 0th programming assignment. We start with $(\mathcal{X} \times \mathcal{Y} = \mathbb{R}^k \times \{0,1\}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ independent, and suppose that we have "optimal" model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$. We bracket, for the moment, what optimality here means. Computing accuracy, we have:

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\tilde{y} = y) &= \mathbb{E}(\mathbb{1}_{\tilde{y}(x)=y}) \\
&:= \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=y}\, d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}} \mathbb{1}_{\tilde{y}(x)=y}\, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=y}\, d\mathbb{P}_{\mathcal{Y}}(y) d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=0}(1-p) + \mathbb{1}_{\tilde{y}(x)=1} p\, d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}} \max\{p, 1-p\} d\mathbb{P}_{\mathcal{X}}(x) \\
&= \max\{p,1-p\}\int_{\mathcal{X}} d\mathbb{P}_{\mathcal{X}}(x) \\
&= \max\{p,1-p\}
\end{aligned} \tag{15}$$

In the first equation, we recall that probability may be written as expectation of an indicator. This is useful!, it allows us to do this computation. The second line is the definition of expectation (definition 2.2). The third line is our implicit definition of conditional probability eq. (10), the geometry of which you should think of as iterated integration in multivariable calculus. The fourth line is independence (condition #2 above). The fifth is expectation w.r.t. $\mathbb{P}_{\mathcal{Y}}$ of $\mathbb{1}_{\tilde{y}(x)=y}$. The sixth line is optimality of $\tilde{y}$; we are trying to optimize accuracy, and as we've written it we get to a point where we can optimize *pointwise* (in x). In the seventh line, we pull out $\max\{p, 1-p\}$ since p is independent of x and in the final line we recall that $\mathbb{P}_{\mathcal{X}}(\mathcal{X}) = 1$.

If you look at the scores for your model, you should have had something pretty close to p for almost all x.

Next we interpret independence for data.

3.2 Data

Referring to the standard diagram (6), we hope and expect that most instances where supervised machine learning comes in to play, the measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X}\times\mathcal{Y}$ is not independent. Sometimes, as in the 0th programming assignment, you'll run into an evil data set, but such are not the norm. Instead, independence is a phenomenon we'd like to associate with *data*.

Definition 3.1. We say that a sequence of (labeled) points $\mathcal{S} = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \sim_{\mathtt{iid}} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ is (sampled) *independent and identically distributed* if $\mathcal{S} \in (\mathcal{X}\times\mathcal{Y})^m$ is a point in joint probability space $(\mathcal{X}\times\mathcal{Y})^m$ with independent measure $\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m} = (\mathbb{P}_{\mathcal{X}\times\mathcal{Y}})^m$.

The following picture is for the intuition in the high-dimensional space.

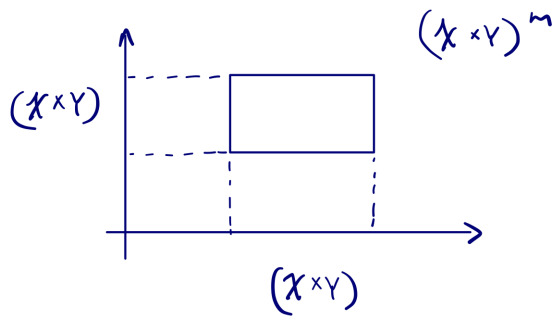

Figure 10

Remark 3.1. While seemingly pedantic, the nuance is important: a data set in ML is properly speaking a sequence. While we might not care about order, we absolutely do care about repetition.

Talk of sampling points suggests talk of randomness. You are allowed to select any point $x \in \mathcal{X}$ that you fancy. How you sample should in some sense be affiliated with your measure $\mathbb{P}_{\mathcal{X}}$. One way of handwaiving this affiliation is to say that given any set (event) $A \subset \mathcal{X}$, the probability that the point $x \in \mathcal{X}$ you picked happens to (also) be in A is $\mathbb{P}_{\mathcal{X}}(A)$. Properly speaking, we are saying in English that $\mathbb{E}(\mathbb{1}_{x\in A}) = \mathbb{P}_{\mathcal{X}}(A)$. Since we generally cannot evaluate $\mathbb{P}_{\mathcal{X}}$ directly, we would like a more data-centric way to talk about probability.

3.2.1 Law of Large Numbers (LLN)

Theorem 3.1. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a probability space and for $m \in \mathbb{N}$, $(\mathcal{X}^m, \mathbb{P}_{\mathcal{X}^m})$ independent (meaning: $\mathbb{P}_{\mathcal{X}^m} = \mathbb{P}_{\mathcal{X}}^m$), and suppose that both $\mu_{\mathcal{X}} := \mathbb{E}(x) = \int_{\mathcal{X}} x d\mathbb{P}_{\mathcal{X}}(x) < \infty$ and $\sigma_{\mathcal{X}}^2 := \mathbb{E}((x-\mu_{\mathcal{X}})^2) < \infty$. Define empirical means $s_m : \mathcal{X}^m \to \mathbb{R}$ by $(x_1, \ldots, x_m) \mapsto \frac{1}{m}(x_1 + \ldots + x_m)$. Then for $\varepsilon > 0$,

$$\lim_{m\to\infty} \mathbb{P}_{\mathcal{X}^m}\left(|s_m - \mu_{\mathcal{X}}| > \varepsilon\right) = 0.$$

Proof. WLOG suppose $\mu_x = 0$. By Chebyshev

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}^m}\left(|s_m - \mu_{\mathcal{X}}| > \varepsilon\right) &\leq \frac{\mathbb{E}((s_m-\mu_x)(s_m-\mu_x))}{\varepsilon^2} \\
&= \frac{\mathbb{E}\left(\frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m x_i x_j - \frac{2}{m}\sum_{j=1}^m x_j\mu_x + \mu_x^2\right)}{\varepsilon^2} \\
&= \frac{\frac{1}{m^2}\sum_{i,j=1}^m \mathcal{E}(x_i x_j) - \frac{2}{m}\mu_x\sum_{j=1}^m \mathcal{E}(x_j) + \mu_x^2}{\varepsilon^2} = \frac{\frac{1}{m^2}\sum_{i\neq j}\mathcal{E}(x_i)\mathcal{E}(x_j) + \frac{1}{m^2}m\mathcal{E}(x^2)}{\varepsilon^2} \xrightarrow{m\to\infty} 0
\end{aligned}$$

□

Remark 3.2. To recall, the statement of LLN measures the set of tuples $(x_1, \ldots, x_m) \in \mathcal{X}^m$ that are at least ε-far from $s_m^{-1}(\mu_{\mathcal{X}})$.

We can use the Law of Large numbers concretely with data: to say that $(x_1, \ldots, x_n) \sim_{iid} \mathbb{P}_{\mathcal{X}}$ (for the moment, we're not talking about labeled data) means, among other things, that we can expect the law of large numbers to hold for probabilities: i.e. for $A \subset \mathcal{X}$,

$$\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{x_j \in A} \xrightarrow{m\to\infty} \mathbb{E}(\mathbb{1}_{x\in A}) = \mathbb{P}_{\mathcal{X}}(A).$$

Note that convergence in this expression is "in probability," i.e. fix $\varepsilon > 0$. Then for any $\delta > 0$, one can specify $M_\delta > 0$ so that

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{x_j\in A} - \mathbb{P}_{\mathcal{X}}(A)\right| > \varepsilon\right) < \delta$$

whenever $m > M_\delta$.

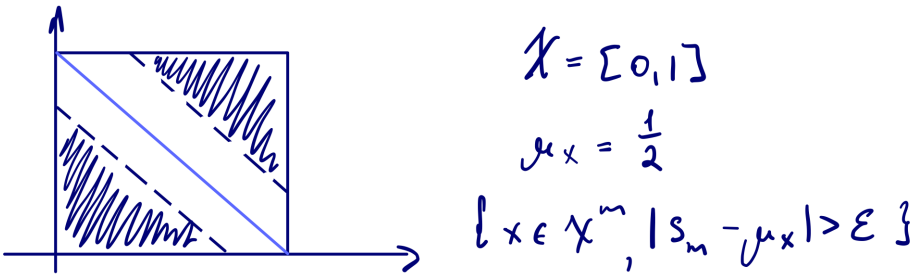

Figure 11

Remark 3.3. I didn't say this in class, but to check in simulation, generate a data sequence $(x_1, \ldots, x_m) \sim_{iid} \mathbb{P}_{\mathcal{X}}$, and check/evaluate

$$\mathbb{1}_{|s_m(x_1,\ldots,x_m) - \mathbb{P}_{\mathcal{X}}(A)| > \varepsilon}.$$

The answer will be zero or one. Most of the time it should be zero. If you do this many times, and take the empirical average of *these* results, the answer should be less than δ.

3.3 Concentration Inequalities

Concentration inequalities are sort of a heartbeat in ML. They are omnipresent and make a lot of important results work.

Proposition 3.1 (Markov)**.** $(\mathcal{X} \subset \mathbb{R}^{\geq 0}, \mathbb{P}_{\mathcal{X}})$ a non-negative probability space, then

$$\mathbb{P}_{\mathcal{X}}(x \geq t) \leq \mathbb{E}(x)/t.$$

Proof. Compute: $\mathbb{E}(x) := \int_{\mathcal{X}} x d\mathbb{P}_{\mathcal{X}}(x) = \int_{[0,t)} x d\mathbb{P}_{\mathcal{X}}(x) + \int_{[t,\infty)} x d\mathbb{P}_{\mathcal{X}}(x)$, the last equality by linearity of integration. Since $x \geq 0$, this expression is bounded above by

$$\int_{[t,\infty)} x d\mathbb{P}_{\mathcal{X}}(x) \geq \int_{[t,\infty)} t d\mathbb{P}_{\mathcal{X}}(x),$$

where the latter follows from the fact that $x \geq t$ on $x \in [t, \infty)$. Taking t out and simplifying the integral as probability proves the inequality. □

Proposition 3.2 (Chebyshev). $(\mathcal{X} \subset \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ a probability space with finite variance $\sigma^2_{\mathcal{X}}$. Then for $\varepsilon > 0$,

$$\mathbb{P}_{\mathcal{X}}(|x - \mu_{\mathcal{X}}| > \varepsilon) \leq \sigma^2_{\mathcal{X}} / \varepsilon^2.$$

Proof. Markov applied to $\{|x - \mu_x| > t\} = \{(x - \mu_x^2) > t^2\}$ □

Proposition 3.3 (Hoeffding). $(\mathcal{X} \subset [0,1], \mathbb{P}_{\mathcal{X}})$ a probability space, then

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|\left|\frac{1}{m}\sum_{j=1}^{m} x_j - \mu_x\right| > \epsilon\right)\right) \leq 2e^{-2m\epsilon^2}.$$

Hoeffding may appear a bit bizarre at first. What you should try to latch onto is that the right hand side looks like something you're already familiar with, the probability density function for normal distribution. While Chebyshev gives a quasi-distributionless way to get a handle on tail probabilities, Hoeffding is particularly nice because it gives a bound as a quadratically decreasing exponential. Moreover, it generalizes when $\mathcal{X} = [a, b]$.

Theorem 3.2 (Glivenko-Cantelli). $(\mathcal{X} = \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ a probability space, for $t \in \mathbb{R}$, define $F(t) := \mathbb{P}_{\mathcal{X}}(x \leq t)$ the cdf and $F_m(t) : \mathcal{X}^m \to \mathbb{R}$ by

$$(x_1, \ldots, x_m) \mapsto \frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{x_j \leq t}$$

the *empirical cdf*. Then

$$\mathbb{P}_{\mathcal{X}^m}\left(\sup_{t \in \mathbb{R}} \left|F_m(t) - F(t)\right| > \epsilon\right) \leq 8(m+1)e^{-\frac{m \cdot \varepsilon^2}{32}}. \tag{16}$$

The statement of this theorem is rather striking: no matter what $\mathbb{P}_{\mathcal{X}}$, you can stipulate conditions for ensuring precision specification (ε) is satisfies with arbitrarily high confidence $(1 - \delta)$ provided m is "sufficiently large." And again... totally independent of $\mathbb{P}_{\mathcal{X}}$. Sometimes, one must pause to marvel at the beauty of mathematics. The picture illustrating the idea of the theorem is below

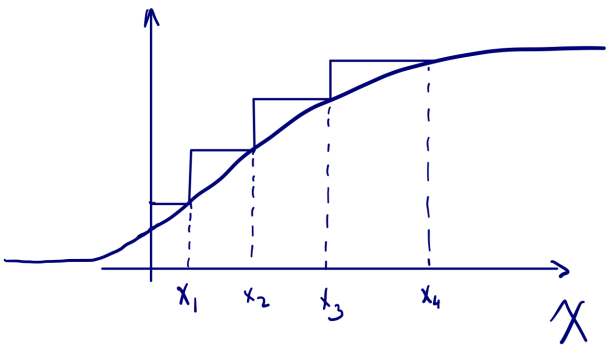

Figure 12

3.4 Intuition in High Dimension

Probability does weird things in high dimension. I made a tenuous sounding claim that a high dimensional gaussian has density concentrated at the origin (tracks our low-dimensional intuition) but probability concentrated *away* from it (what?!).

Let's say this a bit more formally. Suppose that $(\mathcal{X} = \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ is probability space with $\mathbb{P}_{\mathcal{X}}$ normal, zero mean, unit variance, and $(\mathcal{X}^m, \mathbb{P}_{\mathcal{X}^m})$ independent. For $\varepsilon > 0$,

$$\lim_{m\to\infty} \mathbb{P}_{\mathcal{X}^m}\left(\left|\|x\|^2 - m\right| > \varepsilon\right) = \lim_{m\to\infty} \mathbb{P}_{\mathcal{X}^m}\left(\left\{x \in \mathcal{X}^m : \left|\|x\|^2 - m\right| > \varepsilon\right\}\right) = 0.$$

This says that in high dimension, a gaussian concentrates around the sphere of radius $\sqrt{m}$. In particular, the *solid* sphere of radius (strictly) less than $\sqrt{m}$ is practically empty!

Here's a concrete computation which sortof illustrates the point.

Example 3.2. Consider $(\mathcal{X} = [0,1], \mathbb{P}_{\mathcal{X}})$ with uniform measure $\mathbb{P}_{\mathcal{X}}([a,b]) = (b-a)\mathbb{1}_{0\le a\le b\le 1}$. Suppose that $A \subset \mathcal{X}$ with $1 > \mathbb{P}_{\mathcal{X}}(A) \ge 1-\varepsilon$. The "hypercube" $A^m \subset \mathcal{X}^m$ in high dimension has measure

$$\mathbb{P}_{\mathcal{X}^m}(A^m) = (\mathbb{P}_{\mathcal{X}}(A))^m = (1-\varepsilon)^m,$$

the first equality by independence. For $\varepsilon > 0$, $\lim_{m\to\infty}(1-\varepsilon)^m = 0$.

4 Lecture 4

In this lecture we prove Glivenko-Cantelli inequality. Given data samples $(x_1, \ldots, x_m) \in \mathcal{X}^m$, we want to approximate the cdf $F(t) := \mathbb{P}_{\mathcal{X}}(\mathcal{X} \leq t)$ by its empirical average defined by

$$F_m(t) := \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{x_j \leq t} \tag{17}$$

Properly speaking $F_m : \mathcal{X}^m \to [0,1]^{\mathcal{X}}$ defined by mapping $\overline{x} = (x_1, \ldots, x_m) \mapsto F_m^{\overline{x}} : \mathcal{X} \to [0,1]$, the latter of which is defined by (17). Typically we drop notational dependence on $\overline{x}$.

Glivenko-Cantelli tells us that $\sup_{t \in \mathbb{R}} |F_m(t) - F(t)| \to 0$ with high probability. The rigorous statement:

Theorem 4.1 (Glivenko-Cantelli)**.** Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ be a random variable. Then

$$\mathbb{P}_{\mathcal{X}^m}\left(\sup_{t \in \mathbb{R}} |F_m(t) - F(t)| > \epsilon \right) \leq 8(m+1) e^{-\frac{m \cdot \epsilon^2}{32}} \tag{18}$$

for m sufficiently large.

The proof will be presented in a sequence of steps, which we initially enumerate and subsequently prove.

Outline of the proof

There are five main steps. For the first three, we fix $t \in \mathbb{R}$.

1. First we bound the (probability of) separation between (true) cdf $F(t)$ and empirical cdf $F_m(t)$ by a probability of separation between two *separate* empirical cdfs $F_m(t)$ and $F'_m(t)$. Notice that the probability statement in eq. (18) is w.r.t. the data sample $(x_1, \ldots, x_m) \in \mathcal{X}^m$; in this step we consider (/transport our attention to) data samples $(x_1, \ldots, x_m, x'_1, \ldots, x'_m) \in \mathcal{X}^{2m}$.This step is crucial, as we'll see in step , for turning $\sup_{t \in \mathbb{R}} = \bigcup_{t \in \mathbb{R}}$ into a finite union

2. Having introduced a second (in-distribution identical) data sample $(x'_1, \ldots, x'_m) \in \mathcal{X}^m$, we observe that separation condition $\left| \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t} \right| > \varepsilon'$ is symmetric in $x, x' \in \mathcal{X}^m$ ($\varepsilon' > 0$ is not necessarily equal to ε). We thus, without affecting the probability statement, introduce (symmetric) *Rademacher* variables $s \in \mathcal{S} := \{-1, 1\}$ for which $\mathbb{P}_{\mathcal{S}}(s = 1) = \frac{1}{2}$.
   With the Rademacher variables, we consider separation condition

$$\left| \left(\frac{1}{m} \sum_{j=1}^{m} s_j (\mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}) \right) \right| > \varepsilon'$$

   and using triangle inequality isolate each sample into its own separation condition

$$\left| \frac{1}{m} \sum_{j=1}^{m} s_j \mathbb{1}_{x_j \leq t} \right| > \varepsilon'',$$

   for some other $\varepsilon'' > 0$. We will see that this step collapses the probability calculation on $\mathcal{X}^{2m}$ to one on $\mathcal{X}^m$.

3. Next we apply law of total expectation, condition on $\mathcal{X}^m$, fixing sample $(x_1, \ldots, x_m) \in \mathcal{X}^m$, and compute $\int_{\mathcal{X}^m \times \mathcal{S}^m} = \int_{\mathcal{X}^m} \int_{\mathcal{S}^m | \mathcal{X}^m}$. The inner expectation is

$$\int_{\mathcal{S}^m | \mathcal{X}^m} \mathbb{1}_{\left| \frac{1}{m} \sum_{j=1}^{m} s_j \mathbb{1}_{x_j \leq t} \right| > \varepsilon''} d\mathbb{P}_{\mathcal{S}^m | \mathcal{X}^m}(\overline{s} | \overline{x}) = \mathbb{P}_{\mathcal{S}^m | \mathcal{X}^m}\left(\left| \frac{1}{m} \sum_{j=1}^{m} s_j \mathbb{1}_{x_j \leq t} \right| > \varepsilon'' | x_1, \ldots, x_m \right).$$

4. Supremizing over $t \in \mathbb{R}$, we observe that

$$\mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m}\left(\sup_{t\in\mathbb{R}}\left|\frac{1}{m}\sum_{j=1}^{m} s_j \mathbb{1}_{x_j\le t}\right| > \varepsilon''|\bar{x}\right) = \mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m}\left(\bigcup_{t\in\{t_0,\dots,t_m\}}\left|\frac{1}{m}\sum_{j=1}^{m} s_j \mathbb{1}_{x_j\le t}\right| > \varepsilon''|\bar{x}\right).$$

Generally, $\mathbb{P}\left(\sup_{t\in\mathbb{R}}\right) = \mathbb{P}\left(\bigcup_{t\in\mathbb{R}}\right)$; in this case, however, conditioned (i.e. *fixing*) on $x_1,\dots,x_m$, the values of $\sum_{j=1}^{m}\mathbb{1}_{x_j\le t}$ change only at $t = x_1,\dots,x_m$. Therefore, this supremum turns out to be a finite union.

5. Finally we apply Hoeffding's inequality to the inner integral to get a bound in terms of decreasing exponential and then apply the outer integral $\int_{\mathcal{X}^m}$, which recovers a probability we seek to bound.

You should go through the details of the argument at least once, get a feel for the techniques used, and make yourself at home with the outline. The first step is an obnoxiously detailed calculation, but step 2, e.g., contains a kernel of a concept we'll see again when we cover PAC learnability and complexity of hypothesis classes.

Lemma 4.1.

$$\mathbb{P}_{\mathcal{X}^m}\left(|F_m(t) - F(t)| > \epsilon\right) \le 2\mathbb{P}_{\mathcal{X}^{2m}}\left(\left|F'_m(t) - F_m(t)\right| > \frac{\epsilon}{2}\right)$$

where

$$F'_m = \frac{1}{m}\sum_{j=1}^{m}\mathbb{1}_{x'_j\le t}$$

for sample $\bar{x}' = (x'_1,\dots,x'_m) \in \mathcal{X}^m$.

Proof. Fix $t \in \mathbb{R}$, $\epsilon > 0$, and suppose that $\epsilon < |F(t) - F_m(t)|$. Inserting $0 = -F'_m(t) + F'_m(t)$ into the right-hand side and applying the triangle inequality gives

$$\epsilon < |F(t) - F_m(t)| = |F(t) - F'_m(t) + F'_m(t) - F_m(t)| \le |F(t) - F'_m(t)| + |F_m(t) - F'_m(t)|. \tag{19}$$

Supposing (19), $|F(t) - F'_m(t)| \le \frac{\epsilon}{2}$ implies that $|F_m(t) - F'_m(t)| > \frac{\epsilon}{2}$ (why?). Translating conditions into an indicator function, this implication is equivalent to inequality

$$\mathbb{1}_{|F(t)-F_m(t)|>\varepsilon}\cdot\mathbb{1}_{|F(t)-F_m(t)|\le\frac{\varepsilon}{2}} \le \mathbb{1}_{|F_m(t)-F'_m(t)|>\frac{\varepsilon}{2}}.$$

This is true for all $x_1,\dots,x_m,x'_1,\dots,x'_m$, so, taking the expectation of both sides with respect to $d\mathbb{P}_{\mathcal{X}^{2m}}(x_1,\dots,x_m,x'_1,\dots,x'_m)$ preserves the inequality:

$$\mathbb{E}\left(\mathbb{1}_{\epsilon<|F(t)-F'_m(t)|}\cdot\mathbb{1}_{\frac{\epsilon}{2}>|F(t)-F_m(t)|}\right) \le \mathbb{E}\left(\mathbb{1}_{\frac{\epsilon}{2}<|F_m(t)-F'_m(t)|}\right).$$

Recalling that expectation of an indicator is probability, we obtain

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|F(t) - F_m(t)\right| > \varepsilon\right)\mathbb{P}_{\mathcal{X}^m}\left(\left|F(t) - F'_m(t)\right| \le \frac{\varepsilon}{2}\right) \le \mathbb{P}_{\mathcal{X}^{2m}}\left(\left|F_m(t) - F'_m(t)\right| > \frac{\varepsilon}{2}\right).$$

Because Bernoulli random variable $\mathbb{1}_{x\le t}$ has expectation $F(t)$ with finite variance, we can apply Chebyshev's inequality to bound the second term in the above product

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|F(t) - F'_m(t)\right| \le \frac{\varepsilon}{2}\right) = 1 - \mathbb{P}\left(\left|F(t) - F'_m(t)\right| \ge \frac{\varepsilon}{2}\right) \ge \frac{\mathrm{Var}(\mathbb{1}_{x\le t})}{m(\epsilon/2)^2} \ge \frac{1}{2}$$

for m large enough. □

$$\mathbb{P}\left(|F(t) - F_m(t)| < \frac{\epsilon}{2}\right) \geq 1 - \frac{Var(F_m(t)}{(\epsilon/2)^2} \geq 1 - \frac{1/4m}{\epsilon^2/4} = 1 - \frac{1}{m\epsilon^2}$$

$$\mathbb{P}\left(|F(t) - F_m(t)| < \frac{\epsilon}{2}\right) \geq 1 - \frac{1}{m\epsilon^2} \geq 1 - \frac{1}{2} = \frac{1}{2}$$

One more lemma:

Lemma 4.2. (Symmetrization) Let $(\mathcal{X} \subset \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ be a random variable, and $\left(\mathcal{S} = \{-1, 1\}, \mathbb{P}_{\mathcal{S}}(1) = \frac{1}{2}\right)$ symmetric. Then random variable $p : \mathcal{X} \times \mathcal{S} \to \mathcal{X}$ defined by mapping $(x, s) \mapsto s \cdot x$ is symmetric mean zero, i.e.

$$\mathbb{P}_{\mathcal{X}}(sx > t) = \mathbb{P}_{\mathcal{X}\times\mathcal{S}}(sx < -t) \tag{20}$$

When $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ is symmetric, both of these tail probabilities equal the original tail probability $\mathbb{P}_{\mathcal{X}}(x > t)$.

Proof. Let's consider the embeddings $\iota_+ : \mathcal{X} \hookrightarrow \mathcal{X} \times S$, $\iota_- : \mathcal{X} \hookrightarrow \mathcal{X} \times S$ defined by

$$\iota_+(x) := (x, 1) \text{ and } \iota_-(x) := (x, -1)$$

We want to induce via $\mathcal{X}$ a measure on $\mathcal{X} \times S$. If we naively try to define $\mathbb{P}_{\mathcal{X}\times S}([a, b] \times \{s\}) := \mathbb{P}_{\mathcal{X}}(\iota_+^{-1}([a, b] \times \{s\}))$, then for $s = -1$ we will have

$$\mathbb{P}_{\mathcal{X}\times S}([a, b] \times \{-1\}) = \mathbb{P}_{\mathcal{X}}(\iota_+^{-1}([a, b] \times \{-1\})) = \mathbb{P}_{\mathcal{X}}(\emptyset) = 0,$$

Instead, ι_+, ι_- will induce a conditional measure $\mathbb{P}_{\mathcal{X}|S}$ on $\mathcal{X}|s$, and then we extend to $\mathcal{X} \times S$ by the law of total probability. Specifically,

$$\mathbb{P}_{\mathcal{X}\times S}([a, b] \times \{s\}) := \int_S \mathbb{P}_{\mathcal{X}|S}([a, b]|s) d\mathbb{P}_S(s) = \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}([a, b]|1) + \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}([a, b]| - 1)$$

In particular,

$$\mathbb{P}_{\mathcal{X}|S}(x > t|s = 1) := \mathbb{P}_{\mathcal{X}}(x > t) =: \mathbb{P}_{\mathcal{X}|S}(x > t|s = -1) \tag{21}$$

then

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}}(p > t) &:= \mathbb{P}_{\mathcal{X}\times S}(x \cdot s > t) \\
&= \sum_{s\in S} \mathbb{P}_S(s)\mathbb{P}_{\mathcal{X}|S}(x > t|s) && (22) \\
&= \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}(x > t|s = 1) + \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}(-x > t|s = -1) \\
&= \frac{1}{2}\mathbb{P}_{\mathcal{X}}(x > t) + \frac{1}{2}\mathbb{P}_{\mathcal{X}}(x < -t) && (23) \\
&= \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}(x > t|s = -1) + \frac{1}{2}\mathbb{P}_{\mathcal{X}|S}(-x > t|s = 1) \\
&= \mathbb{P}_{\mathcal{X}}(p < -t)
\end{aligned}$$

In the second step (22), we apply the law of total probability, and as a result of (21), we arrive at (23). □

We may apply this result to symmetric random variable $\mathbb{1}_{x\leq t} - \mathbb{1}_{x'\leq t}$ on $\mathcal{X}^2$ with relevant modification:

$$\mathbb{P}_{\mathcal{X}^2}\left(\left|\mathbb{1}_{x\leq t} - \mathbb{1}_{x'\leq t}\right| > \varepsilon\right) = \mathbb{P}_{\mathcal{X}^2\times\mathcal{S}}\left(\left|s(\mathbb{1}_{x\leq t} - \mathbb{1}_{x'\leq t})\right| > \varepsilon\right), \tag{24}$$

and extend application *mutatis mutandis* to $\mathcal{X}^{2m} \times \mathcal{S}^m$ for separation of empirical cdfs.

Proof of Glivenko-Cantelli. In lemma 4.1, we showed that

$$\mathbb{P}_{\mathcal{X}^m}\left(|F(t) - F_m(t)| > \varepsilon\right) \leq 2\mathbb{P}_{\mathcal{X}^{2m}}\left(|F_m(t) - F'_m(t)| > \frac{\varepsilon}{2}\right).$$

Rewriting the right hand side, we wish to bound the probability of event

$$\left|\frac{1}{m}\sum_{j=1}^{m} \mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}\right| > \frac{\varepsilon}{2}.$$

Applying (the appropriate generalization of) eq. (24), we observe that

$$\mathbb{P}_{\mathcal{X}^{2m}}\left(|F_m(t) - F'_m(t)| > \frac{\varepsilon}{2}\right) = \mathbb{P}_{\mathcal{X}^{2m}\times\mathcal{S}^m}\left(\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}\right)\right| > \frac{\varepsilon}{2}\right).$$

Applying the triangle inequality to the term inside absolute value, we bound

$$\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}\right)\right| \leq \left|\frac{1}{m}\sum_{j=1}^{m} s_j(\mathbb{1}_{x_j \leq t})\right| + \left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x'_j \leq t}\right)\right|.$$

Therefore, a bound of

$$\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}\right)\right| > \frac{\varepsilon}{2}$$

on the left hand side implies a bound of

$$\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \frac{\varepsilon}{4} \text{ or } \left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x'_j \leq t}\right)\right| > \frac{\varepsilon}{4}$$

on the right hand side.

Implication of conditions translates to inclusion of sets (events) which translates to a bound on probability:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{X}^{2m}\times\mathcal{S}^m}\left(\left|\tfrac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t} - \mathbb{1}_{x'_j \leq t}\right)\right| > \tfrac{\varepsilon}{2}\right) \\
&\leq \mathbb{P}_{\mathcal{X}^{2m}\times\mathcal{S}^m}\left(\left\{\left|\tfrac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \tfrac{\varepsilon}{4}\right\} \cup \left\{\left|\tfrac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x'_j \leq t}\right)\right| > \tfrac{\varepsilon}{4}\right\}\right) \\
&\leq 2\mathbb{P}_{\mathcal{X}^{2m}\times\mathcal{S}^m}\left(\left|\tfrac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \tfrac{\varepsilon}{4}\right) \\
&= 2\mathbb{P}_{\mathcal{X}^{m}\times\mathcal{S}^m}\left(\left|\tfrac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \tfrac{\varepsilon}{4}\right).
\end{aligned} \tag{25}$$

The third line follows from the union bound, together with the condition in x_js defining event in $\mathcal{X}^{2m}$ is identical—bracketing factors—to condition in x'_js defining the analogous event, and therefore their measures are the same. The final line follows by marginalizing out the residual (extra) factors of $\mathcal{X}^m$.

Next, we decompose the last probability on right hand side of eq. (25) using law of total probability ($\mathbb{P}_{\mathcal{X}\times\mathcal{S}} = \mathbb{P}_{\mathcal{X}}\mathbb{P}_{\mathcal{S}|\mathcal{X}}$)/expectation $\left(\int_{\mathcal{X}\times\mathcal{S}} = \int_{\mathcal{X}}\int_{\mathcal{S}|\mathcal{X}}\right)$:

$$\mathbb{P}_{\mathcal{X}^m\times\mathcal{S}^m}\left(\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \frac{\varepsilon}{4}\right) = \int_{\mathcal{X}^m} \mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m}\left(\left|\frac{1}{m}\sum_{j=1}^{m} s_j\left(\mathbb{1}_{x_j \leq t}\right)\right| > \frac{\varepsilon}{4}|\bar{x}\right) d\mathbb{P}_{\mathcal{X}^m}(\bar{x}). \tag{26}$$

Up to this point $t \in \mathbb{R}$ has been arbitrary, fixed. But the statement of Glivenko-Cantelli considers supremum over $t \in \mathbb{R}$. The bound, therefore, over original event $\sup_{t\in\mathbb{R}} |F(t) - F_m(t)| > \varepsilon$ is in terms of

$$\int_{\mathcal{X}^m} \mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m} \left(\sup_{t\in\mathbb{R}} \left| \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) \right| > \frac{\varepsilon}{4} | \overline{x} \right) d\mathbb{P}_{\mathcal{X}^m}(\overline{x}).$$

Isolating the inner probability

$$\mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m} \left(\sup_{t\in\mathbb{R}} \left| \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) \right| > \frac{\varepsilon}{4} | \overline{x} \right),$$

we note that $\overline{x}$ is fixed in the conditional, so the sequence $(\mathbb{1}_{x_1 \le t}, \ldots, \mathbb{1}_{x_m \le t}) \in \{0,1\}^m$ takes at most $m+1$ values as t ranges over $\mathbb{R}$. Therefore the event (in $\mathcal{S}^m|\mathcal{X}^m$)

$$\left\{ \left| \sup_{t\in\mathbb{R}} \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) \right| > \frac{\varepsilon}{4} | \overline{x} \right\} \subseteq \bigcup_{t\in\{t_0,\ldots,t_m\}} \left\{ \left| \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) | \overline{x} \right| > \frac{\varepsilon}{4} \right\},$$

to which we may apply the union bound in probability:[4]

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m} \left(\left| \sup_{t\in\mathbb{R}} \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t\}} \right) \right| > \tfrac{\varepsilon}{4} | \overline{x} \right) &\le \sum_{t\in\{t_0,\ldots,t_m} \mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m} \left(\left| \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) \right| > \frac{\varepsilon}{4} | \overline{x} \right) \\
&= (m+1) \mathbb{P}_{\mathcal{S}^m|\mathcal{X}^m} \left(\left| \frac{1}{m} \sum_{j=1}^{m} s_j \left(\mathbb{1}_{x_j \le t} \right) \right| > \tfrac{\varepsilon}{4} | \overline{x} \right) \\
&\le 2(m+1) e^{-\frac{m(\varepsilon/4)^2}{2}} \\
&= 2(m+1) e^{\frac{-m\varepsilon^2}{32}}
\end{aligned}$$

In the third line, we cite Hoeffding's bound. When we substitute this back and calculate the outer integral in equation 26, and conclude, piecing everything together, that

$$\mathbb{P}_{\mathcal{X}^m} \left(|F(t) - F_m(t)| > \varepsilon \right) \le 8(m+1) e^{\frac{-m\varepsilon^2}{32}} \int_{\mathcal{X}^m} d\mathbb{P}_{\mathcal{X}^m}(x) = 8(m+1) e^{\frac{-m\varepsilon^2}{32}}$$

as desired. □

[4] We reiterate that s_j are the only variables in this expression; we are conditioning on $\mathcal{X}^m$ and therefore holding $x \in \mathcal{X}^m$ fixed; as an event we may write e.g. the left hand side explicitly as $\left\{ \overline{s} \in \mathcal{S}^m : \left| \sup_{t\in\mathbb{R}} \frac{1}{m} \sum_{j=1}^{m} \sigma_j \left(\mathbb{1}_{x_j \le t} \right) \right| \right\}$

5 Lecture 5

Our setting: we have joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ and recall the standard diagram eq. (6). Our goal is to construct *model* $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ for which $\tilde{y}(x) \approx y$ for "most" pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Probability makes sense (/ precisifies) 'most.'

Once we have our hands on *a* model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$, we would like some way to *measure* (not in the measure-theoretic sense!) how well $\tilde{y}$ "does" on arbitrary labeled data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which we capture by saying there is some random variable $l_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we call the *cost function*, mapping $(x, y) \mapsto l_{\tilde{y}}(x, y) \in \mathbb{R}$. Often or usually this random variable will be nonnegative and we'd like it to be as small as possible on as many points as possible. Said differently, we want $\mathbb{E}(l_{\tilde{y}})$ to be small.

What is the "variable?" The model! So you can think of cost l as inducing a *map*

$$l_{(\cdot)} : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}\times\mathcal{Y}},$$

$\tilde{y} \to l_{\tilde{y}}$, where $\{\tilde{y} : \mathcal{X} \times \mathcal{Y}\}$ and $\{l : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$ ie the cost maps from the set of models to the set of random variables, returning a random variable $l_{\tilde{y}}$ for each specified model $\tilde{y} \in \mathcal{X} \to \mathcal{Y}$. And thus stated we cast *the* problem of supervised machine learning (sml) as finding an optimal model $y^* \in \arg\min_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \mathbb{E}(l_{\tilde{y}})$. Note that our objective is stated "in expectation." Crudely you can think of sml as curve fitting, and I have no problem with this plebian perspective as long as you distinguish fitting the data in your hands from the data "not at your immediate disposal." Supervised machine learning deals not with fitting data as such but fitting the source, whence data comes. It amounts to fitting the measure! Hence how ml provides a concrete setting for understanding probability: you really cannot talk about ml without it.

Now, let's consider a binary classification: $\mathcal{Y} = \{0, 1\}$.
Example of the cost function:

1. $l_{\tilde{y}}(x, y) = \mathbb{1}_{\tilde{y}(x)\neq y}$
2. $l_{\tilde{y}}(x, y) = (\tilde{y}(x) - y)^2$

To solve binary classification problem, where $\mathcal{Y} = \mathbb{R}$ and $\mathbb{P}_{\mathcal{Y}}(0, 1) = 0$ ie $\mathbb{P}_{\mathcal{Y}}(\{0, 1\}) = 1$:

1. Construct score function $\tilde{y} : \mathcal{X} \to [0, 1]$
2. Obtain classification threshold ie set $t \in \mathbb{R} : \bar{y} : \mathcal{X} \to \{0, 1\} : x \to \bar{y}(x) = \mathbb{1}_{\tilde{y}(x)\geq t}$

The objective of our optimization problem is the following:

$$\min_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \mathbb{E}(l_{\tilde{y}}) = \int_{\mathcal{X}\times\mathcal{Y}} l_{\tilde{y}}(x, y) d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x, y)$$

We are going to consider cost function $l_{\tilde{y}}(x, y) = (\tilde{y}(x) - y)^2$ and find $y^* = \arg\min_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \mathbb{E}(l_{\tilde{y}})$

$$\mathbb{E}(l_{\tilde{y}}) = \int_{\mathcal{X}\times\mathcal{Y}} (\tilde{y}(x) - y)^2 \, d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}|\mathcal{X}} (\tilde{y}(x) - y)^2 \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) d\mathbb{P}_{\mathcal{X}}(x)$$

We need to minimize globally by minimizing pointwise in $\mathcal{X}$ ie we need to define $\tilde{y}(x) \in \mathcal{X}$ $\forall x \in \mathcal{X}$.

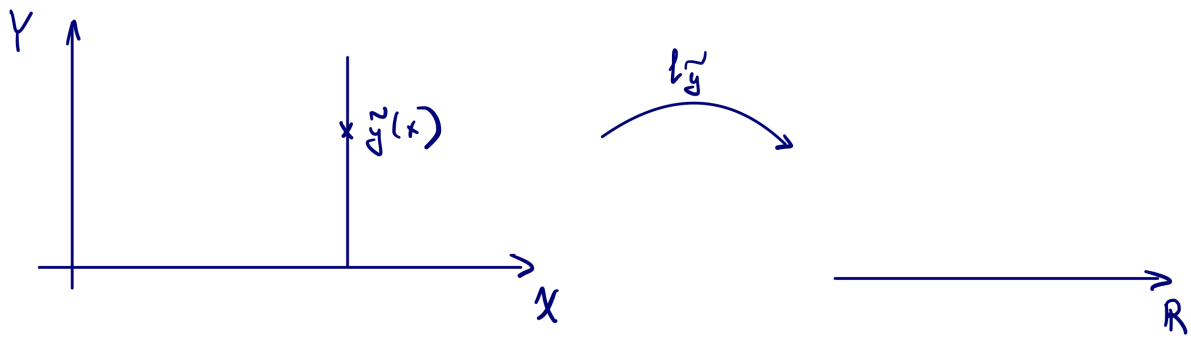

Figure 13

$$\begin{aligned}\min_{\tilde{y}(x)\in\mathcal{Y}} \int_{\mathcal{Y}|\mathcal{X}} (\tilde{y}(x)-y)^2 \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) &= \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y=0|x)(\tilde{y}(x))^2 + \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y=1|x)(\tilde{y}(x)-1)^2 \\ &= (1-p(x))\tilde{y}(x)^2 + p(x)(\tilde{y}(x)-1)^2\end{aligned}$$

To find an optimal $y^*(x) \in \mathcal{Y}$ we need to differentiate w.r.t. $\tilde{y}(x)$ and set that derivative to 0.

$$\begin{aligned}&\frac{d((1-p(x))\tilde{y}(x)^2+p(x)(\tilde{y}(x)-1)^2)}{d(\tilde{y}(x))} = (1-p(x))2\tilde{y}(x) + p(x)2(\tilde{y}(x)-1) = 0 \\ &2\tilde{y}(x) - 2p(x)\tilde{y}(x) + 2p(x)\tilde{y}(x) - 2p(x) = 0 \\ &\tilde{y}(x) - p(x) = 0 \\ &\tilde{y}(x) = p(x)\end{aligned}$$

Therefore,

$$y^* : \mathcal{X} \to \mathcal{Y} : y^*(x) = \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y=1|x) = \mathbb{E}(\mathcal{Y}|\mathcal{X}) \text{ is optimal.}$$

In general,

$$y^*(x) = \int_{\mathcal{Y}|\mathcal{X}} y \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)$$

But we have a problem: we typically don't know true measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ (or $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$).

Now, consider $\mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\tilde{y}(x)|y)$, where $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ induces a measure $\mathbb{P}_{\mathcal{Y}^2}$ s.t. $\mathbb{P}_{\mathcal{Y}^2}(\tilde{y} \in [a,b], y) = \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\tilde{y}(^{-1}([a,b])), y)$. Denote $F_0(t) = \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\tilde{y}(x) \le t|y=0)$ and $F_1(t) = \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\tilde{y}(x) \le t|y=1)$.

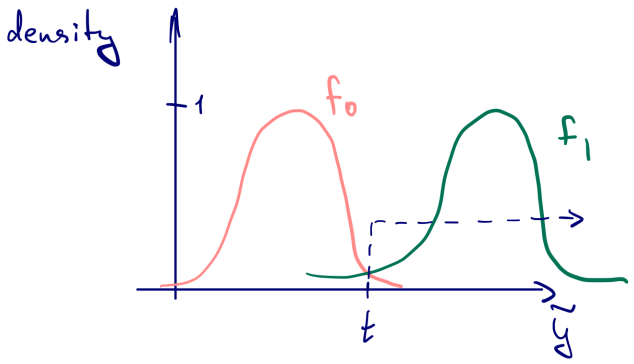

Figure 14

We consider various metrics for evaluating score $\tilde{y} \in [0,1]$:

1. ks score $F_y(t)\mathbb{P}_{\tilde{y}|y}(\tilde{y} \leq t|y)$ and set ks $:= \sup_{t\in\mathbb{R}} |F_0(t) - F_1(t)|$
   Ideally, classifier $\tilde{y}$ separates class data perfectly.

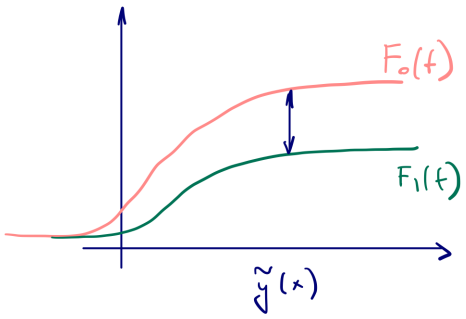

Figure 15

2. Consider the score itself. We obtain a classifier by tresholding ie $\tilde{y}_t(x) := \mathbb{1}_{\tilde{(y)}(x)\geq t}$. Then,

$$\int_t^\infty f_y(s)ds = \begin{cases} \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\tilde{y}_t = 1|y = 1) = tpr(t), \\ \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\tilde{y}_t = 1|y = 0) = fpr(t) \end{cases} \tag{27}$$

   We want tpr to be as big as possible and fpr to be as small as possible.
   Key point: true/false positive rate parametrized by the score range ie tpr/fpr: $\tilde{y} \to [0,1]$.
   We also have a notation for precision: $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|\tilde{y}_t = 1)$.

3. Parametrization on (27) induces ROC/AUC curves.

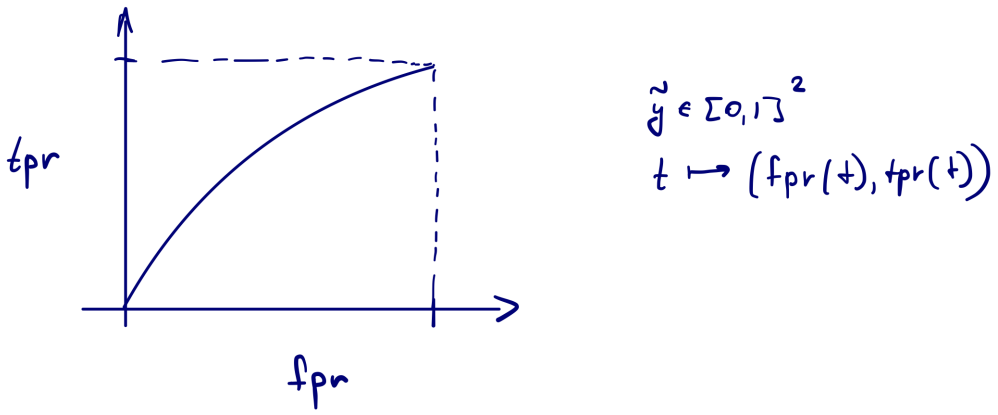

Figure 16

Gold standard: $\exists t^* : (\mathrm{fpr}(t^*), \mathrm{tpr}(t^*)) = (0, 1)$.

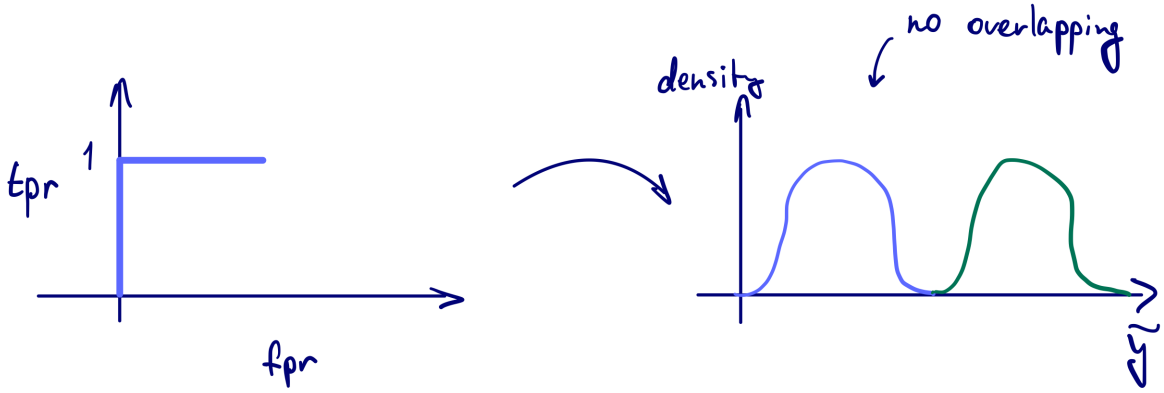

Figure 17

4. Typically think of $\tilde{y}$ as a likelihood score ie likelihood of belonging to a certain class. To make it rigorous:
$$y^*(x) = \mathbb{P}_{y|\mathcal{X}}(y = 1|x) =: p(x) \tag{28}$$
But it's not necessarily true (was true in pa0):

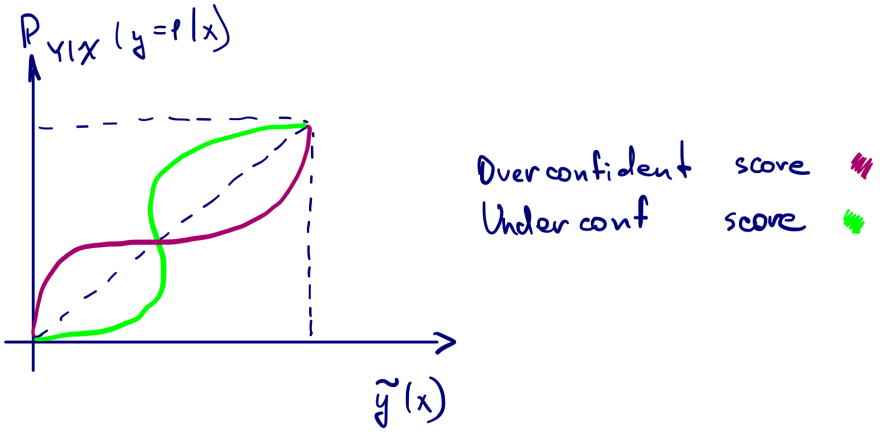

Figure 18

If (28) is satisfied we say that $\tilde{y}$ is calibrated.

5. Equal error rate is defined in ws3. Finally, following from that, we have

$$\mathbb{P}_{y|x}(y=1|x) = \frac{\mathbb{P}_y(y=1)f_1(x)}{\mathbb{P}_y(y=0)f_0(x)+\mathbb{P}_y(y=1)f_1(x)}$$

6 Lecture 6

6.1 Introduction

We recast the supervised learning problem in the context of the standard diagram (6) as a problem of approximating $\pi_{\mathcal{Y}}$ using a function

$$\tilde{y} \circ \pi_{\mathcal{X}} \in \mathcal{H} := \left\{ h : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} : \; h = \tilde{y} \circ \pi_{\mathcal{X}} \right\}.$$

It turns out that concepts from linear algebra and functional analysis are particularly suitable for finding optimally approximating function $y^* \circ \pi_{\mathcal{X}} \in \mathcal{H}$ when $\mathcal{Y} = \mathbb{R}$ and the notion of approximation is defined by "square distance." In particular, the *Hilbert Projection Theorem* provides an *algorithm* for approximating the best approximating function $y^* : \mathcal{X} \to \mathcal{Y}$.

The ingredients for making sense of optimal y^* require completing the following steps:

1. identifying appropriate vector space $\mathcal{V} \subset \{f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}\}$ and subspace $\mathcal{H} \subset \mathcal{V}$,
2. identifying an appropriate notion of distance $d(f, g)$ between two elements $f, g \in \mathcal{V}$
3. identifying an appropriate notion of projection and orthogonality, the latter of which extensionally defines (at the very least identifies with) the former. More to the point: orthogonality is a computably verifiable property which may then be used for recognizing (and finding!) projection. More on this later.
4. Items 2. and 3. hint at the need for an *inner product*, which will provide the algebraic backdrop of geometric concepts.

We start by introducing the notion of inner product, which you may think of heuristically as a mechanism through which to algebraicize many geometric notions. The inner product will be used to define a norm which is a notion of "length" (a measure!), and a norm will be used to define "distance" (another "measure!"), and finally—and importantly—inner products will provide a notion of orthogonality.

Hilbert projection relies on having a Hilbert space, which is a vector space with an inner product that defines a distance or metric. A Hilbert space is a vector space with inner product "complete" w.r.t. in the metric induced by the norm induced by the inner product. More on this later.

6.2 Inner Product Spaces

Definition 6.1. Let $(\mathcal{V}, \mathbb{R})$ be a real vector space. A map $\langle \cdot, \cdot \rangle : \mathcal{V}^2 \to$ sending $(v, w) \mapsto \langle v, w \rangle$ is said to be an *inner product* if this map satisfies the following three properties:

1. (linearity) $\langle cv + v', w \rangle = c\langle v, w \rangle + \langle v', w \rangle$ for $v, v', w \in \mathcal{V}$ and $c \in \mathbb{R}$,
2. (symmetry) $\langle v, w \rangle = \langle w, v \rangle$ for $v, w \in \mathcal{V}$
3. (positivity) $\langle v, v \rangle \geq 0$ with equality iff $v = 0$.

A real-vector space equipped with an inner product is said to be an *inner product space*.

A traditional definition of inner product on a real vector space may pay lip-service to *bilinearity*, or linearity in *both* factors. Indeed, symmetry together with linearity in the first factor combine to imply linearity in the second factor (check!). The definition we've given is not the most general one; of course you can have vector spaces over other fields, and that will slightly modify the defining properties. We will not need this added generality.

Remark 6.1. There is a subtlety with the third axiom, as we will see in the third example following. It is possible for $f \neq g$ to satisfy $\langle f, g \rangle = 0$, but this is a measure theoretic peculiarity which says: disagreement may occur between f and g, but such is "almost unobservable" and so for all intents and purposes we may say that f"="g. This is a nuance which you may ignore if you wish, but if you find yourself scratching your head on some edge case incongruity between the third property in definition 6.1 and example 6.1 part (3), then we respond by saying: don't fret, the apparent conflict has easily remediable (if annoying) patchwork .

Example 6.1. 1. Let $\mathcal{V} = \mathbb{R}^n$ and define inner product $\langle v, w\rangle := \sum_{j=1}^{n} v_j w_j$ as the dot product, for $v, w \in \mathcal{V}$. One may readily check that the dot product satisfies the properties of inner product.

2. $(\mathcal{V} = \mathcal{C}([a,b],\mathbb{R}),\mathbb{R})$ where $\mathcal{C}([a,b],\mathbb{R}) := \{f : [a,b] \to \mathbb{R} : f \text{ is continuous}\}$ and $\langle\cdot,\cdot\rangle : \mathcal{V}^2 \to \mathbb{R}$ defined by $\langle f, g\rangle \int_a^b f(x)g(x)dx$ is an inner product. Verification basically comes down to linearity of integration.

3. Now $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ and $\mathcal{V} = \{f : \mathcal{X} \to \mathbb{R} : \mathbb{E}(f^2) < \infty\}$ space of random variables with finite second moment. Then one may readily check that $\langle\cdot,\cdot\rangle : \mathcal{V}^2 \to \mathbb{R}$ defined by $\langle f, g\rangle := \mathbb{E}(fg) = \int_{\mathcal{X}} f(x) \cdot g(x) d\mathbb{P}_{\mathcal{X}}(x)$ defines an inner product.

Recalling from calculus that Σ behaves much the same as $\int$, we may intuit that examples 6.1 are all related. Indeed they are, and one should work out for themselves a measure $\mathbb{P}_{\mathcal{X}}$ on $\mathbb{R}$ to make them match (+ some normalization). (You may think of $\mathbb{R}^n$ as "sets of functions from the n-element set $\{1,\dots,n\}$ to $\mathbb{R}$" or element in this set simply as point values at $1,\dots,n$.)

Inner product defines a norm. We set notation and define $\|v\| := \sqrt{\langle v, v\rangle}$. This notation is highly suggestive of a norm, and the axioms defining inner product almost confirm that it is. However, one must still check that the triangle inequality holds, namely that $\|v + w\| \leq \|v\| + \|w\|$, a short computation which cites Cauchy-Schwarz.

Norm defines a metric. A norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ defines a metric $d : \mathcal{V}^2 \to \mathbb{R}$ by $(v, w) \mapsto \|v - w\|$. Verification is straightforward.

We are on the verge of presenting the main theorem, but there's one more technical concept we need to introduce: completeness. Before we delve into that, let's establish the primary mathematical structure that holds our focus in this chapter.

Definition 6.2. Let $(\mathcal{V}, \mathbb{R}, \langle,\rangle)$ be inner product space. We say that $\mathcal{V}$ is a *Hilbert space* if $\mathcal{V}$ is complete with respect to the norm-induced metric $d : \mathcal{V}^2 \to \mathbb{R}$.

Now, a couple of definitions:

Definition 6.3. 1. A sequence $\{v_n\}$ in the metric space $(\mathcal{V}, d)$ is called a *Cauchy sequence* if for every $\epsilon > 0$ we can find an index $n_0 \in \mathbb{N}$, such that for any pair of indices $m, n > n_0$, $d(x_m, x_n) < \epsilon$.

2. A metric space $(\mathcal{V}, d)$ is considered *complete* if every Cauchy sequence within it converges to a point within $\mathcal{V}$. In such a case, it is also said that the metric d is complete in V.

Example 6.2. 1. Related to example 6.1, the metric d induced by the dot product is the standard Euclidean metric, for which the space $\mathcal{V} = \mathbb{R}^n$ is complete.

2. $(\mathcal{V} = \mathcal{C}([a,b],\mathbb{R}),\mathbb{R})$ with the induced metric is not complete. In fact, the sequence of functions

$$f_n(x) := \begin{cases} 0 & \text{if } -1 \leq x \leq 0 \\ nx & \text{if } 0 < x \leq 1/n \\ 1 & \text{if } 1/n < x \leq 1 \end{cases}$$

is a Cauchy sequence that converges to a step function, which is not in $\mathcal{V}$. However, the supremum metric (in this case, the maximum by Weierstrass's theorem) equip $\mathcal{V}$ as a complete metric space. Note that this metric is not generated by an inner product, which means that $\mathcal{V}$ does not constitute a Hilbert space but, rather, a Banach space.

3. The space $\mathcal{V} = \{f : \mathcal{X} \to \mathbb{R} : \mathbb{E}(f^2) < \infty\}$ is a complete metric space with the induced metric. You may see this case as the "completion" of the example above.[5]

[5]Being a bit technical, the space of infinitely differentiable compactly supported functions $\mathcal{C}_0^\infty(X)$ is a subset of $\mathcal{C}(X)$ which in turn, is a subset of $L^p(X)$. The latter, is complete with the L^p-norm ($p = 2$, our case of interests), and $\mathcal{C}_0^\infty(X)$ is dense on this space.

6.3 Orthogonality

Definition 6.4. Let $(\mathcal{V}, \mathbb{R}, \langle , \rangle)$ be inner product space. We say that $v, w \in \mathcal{V}$ are *orthogonal*—denoted $v \perp w$—if $\langle v, w \rangle = 0$. For subspace $\mathcal{H} \subset \mathcal{V}$, we define the orthogonal complement

$$\mathcal{H}^{\perp} := \{h' \in \mathcal{V} : h' \perp \mathcal{H} \text{ i.e. } h' \perp h \, \forall \, h \in \mathcal{H}\}$$

The following theorem decomposes a Hilbert space into a closed subspace and its orthogonal complement. The result is closely related to projection.

Theorem 6.1. (Hilbert Projection) Let $(\mathcal{V}, \mathbb{R}, \langle , \rangle)$ be Hilbert space, and $\mathcal{H} \subset \mathcal{V}$ a closed subspace. Then $\mathcal{V} = \mathcal{H} \oplus \mathcal{H}^{\perp}$. This means for any $v \in \mathcal{V}$, we may uniquely write $v = h + h^{\perp}$ where $h \in \mathcal{H}$ and $h^{\perp} \in \mathcal{H}^{\perp}$. (Uniqueness means for another representation $v = k + k^{\perp}$ where $k \in \mathcal{H}$ and $k^{\perp} \in \mathcal{H}^{\perp}$ that $k = h$ and $k^{\perp} = h^{\perp}$.) Moreover, $h = \arg\min_{h' \in \mathcal{H}} \|h' - v\|$.

Before that, let's start with some background we will use through next section.

Proposition 6.1. (Pythagorean Theorem) Let $(\mathcal{V}, \mathbb{R}, \langle \cdot, \cdot \rangle)$ be an inner product space and suppose that $v \perp w$ (recall that this means $\langle v, w \rangle = 0$). Then $\|v - w\|^2 = \|v\|^2 + \|w\|^2$

Proof. Expanding $\|v - w\|^2 := \langle v - w, v - w \rangle$ and using the orthogonality condition we have:

$$\langle v - w, v - w \rangle = \langle v, v \rangle - 2\langle v, w \rangle + \langle w, w \rangle = \langle v, v \rangle + \langle w, w \rangle = \|v\|^2 + \|w\|^2$$

□

Proposition 6.2. (Parallelogram Law) Let $(\mathcal{V}, \mathbb{R}, \langle , \rangle)$ be an inner product space. If $w, v \in \mathcal{V}$ then $2(\|v\|^2 + \|w\|^2) = \|v + w\|^2 + \|v - w\|^2$.

Proof. From

$$\|v + w\|^2 = \langle v + w, v + w \rangle = \langle v, v \rangle + 2\langle v, w \rangle + \langle w, w \rangle = \|v\|^2 + 2\langle v, w \rangle + \|w\|^2$$

and

$$\|v - w\|^2 = \langle v - w, v - w \rangle = \langle v, v \rangle - 2\langle v, w \rangle + \langle w, w \rangle = \|v\|^2 - 2\langle v, w \rangle + \|w\|^2$$

adding up together we have

$$\|v + w\|^2 + \|v - w\|^2 = 2(\|v\|^2 + \|w\|^2).$$

□

6.4 Orthogonal Projection on Hilbert Subspaces

Let's proceed with the two following useful result:

Proposition 6.3. Let $(\mathcal{V}, \mathbb{R}, \langle , \rangle)$ be an inner product space $\mathcal{H} \subseteq \mathcal{V}$ subspace and $v \in \mathcal{V}$. Suppose there is an $h^* \in \mathcal{H}$ such that

$$h^* \in \arg\min_{h \in \mathcal{H}} \|v - h\|$$

Then $v - h^* \in \mathcal{H}^{\perp}$ where $\mathcal{H}^{\perp} = \{h' \in \mathcal{V} : h' \perp h \text{ for all } h \in \mathcal{H}\}$. Conversely, suppose there is $h^* \in \mathcal{H}$ such that $v - h^* \in \mathcal{H}^{\perp}$. Then,

$$h^* = \arg\min_{h \in \mathcal{H}} \|v - h\|$$

Proof. Let $h^* \in \arg\min_{h \in \mathcal{H}} \|v - h\| \neq \emptyset$ (nonempty by assumption). We must show that $\langle v - h^*, h \rangle = 0$ for arbitrary $h \in \mathcal{H}$, i.e. that $v - h^* \in \mathcal{H}^{\perp}$. Given $h \in \mathcal{H}$, we define smooth map real-variable function

$$\begin{aligned} f : \mathbb{R} &\longrightarrow \mathbb{R} \\ t &\longmapsto f(t) := \|v - (h^* - th)\|^2. \end{aligned}$$

Since $h^* \in \mathcal{H}$, $h^* - th \in \mathcal{H}$ for all $t \in \mathbb{R}$ ($\mathcal{H}$ is a subspace implies $\mathcal{H}$ is closed[6] under addition and scalar multiplication). Thus, by definition of h^*, $\|v - h^*\| \leq \|v - \tilde{h}\|$ for all $\tilde{h} \in \mathcal{H}$ so that f obtains a minimum at $t = 0$. On the other hand,

$$f(t) = \|v - h^* + th\|^2 = \langle v - h^* + th, v - h^* + th\rangle = \langle v - h^*, v - h^*\rangle + 2t\langle v - h^*, h\rangle + t^2\langle h, h\rangle$$

by bilinearity and symmetry of the inner product. As a function of t, f has a critical point in $t = 0$ (because is a minimum), so

$$0 = f'(0) = 2\langle v - h^*, h\rangle + 2t\langle h, h\rangle\big|_{t=0} = 2\langle v - h^*, h\rangle$$

which implies $\langle v - h^*, h\rangle = 0$.

In the other direction, suppose that there is $h^* \in \mathcal{H}$ such that $\langle v - h^*, h\rangle = 0$ for all $h \in \mathcal{H}$. Taking $h \in \mathcal{H}$, $h^* - h \in \mathcal{H}$ so that $\langle v - h^*, h^* - h\rangle = 0$. Then, by the Pythagorean theorem

$$\|v - h\|^2 = \|v - h^* + h^* - h\|^2 = \|v - h^*\|^2 + \|h^* - h\|^2 \geq \|v - h^*\|^2$$

equality iff $h = h^*$. Then h^* is the unique minimizer in $\arg\min_{h \in \mathcal{H}} \|v - h\|$. □

[6] Note that this is not the same 'closed' that appears in theorem 7.1!

7 Lecture 7

7.1 Hilbert Projections Theorem

Recall our setting: $(\mathcal{V}, \mathbb{R}, \langle , \rangle : \mathcal{V}^2 \to \mathbb{R})$, where $\mathcal{V} = \{f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} : \mathbb{E}(f^2) < \infty\}$ and $\mathcal{H} = \{\mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \in \mathcal{V} : f = \tilde{y} \circ \pi_{\mathcal{X}}\}$
Now, let's prove Hilbert projection theorem. Let us recall the statement of the theorem.

Theorem 7.1. (Hilbert Projection) Let $(\mathcal{V}, \mathbb{R}, \langle , \rangle)$ be Hilbert space, and $\mathcal{H} \subset \mathcal{V}$ a closed subspace. Then $\mathcal{V} = \mathcal{H} \oplus \mathcal{H}^{\perp}$. This means for any $v \in \mathcal{V}$, we may uniquely write $v = h^* + h^{\perp}$ where $h^* \in \mathcal{H}$ and $h^{\perp} \in \mathcal{H}^{\perp}$. (Uniqueness means for another representation $v = h' + \bar{h}$ where $h' \in \mathcal{H}$ and $\bar{h} \in \mathcal{H}^{\perp}$ that $h' = h^*$ and $\bar{h} = h^{\perp}$.) Moreover, $h = \arg\min_{h' \in \mathcal{H}} \|h' - v\|$.

Proof. Let $v \in \mathcal{V}$ and set $\delta := \inf_{h \in \mathcal{H}} \|v - h\|^2$, $h \in \mathcal{H}$. Then $\exists$ a sequence $\{h_j\}_{j=1}^{\infty} \in \mathcal{H}$ such that $\lim_{j \to \infty} \|v - h_j\| \geq \delta$. Note, $\mathcal{H}$ is closed in $\mathcal{V}$ and $\mathcal{V}$ is complete. Therefore, $\mathcal{H}$ is complete. Now, we want to show that $\{h_j\}_{j=1}^{\infty}$ is Cauchy. Since

$$v - h_j - (v - h_i) = h_i - h_j$$

and

$$v - h_j + (v - h_i) = 2v - (h_i + h_j) = 2\left(v - \frac{h_i + h_j}{2}\right)$$

we may write, using proposition 6.2 (Parallelogram Law),

$$2(\|v - h_j\|^2 + \|v - h_i\|^2) = \|h_i - h_j\|^2 + \left\|2\left(v - \frac{h_i + h_j}{2}\right)\right\|^2$$

which is equivalent to

$$\|h_i - h_j\|^2 = 2\|v - h_j\|^2 + 2\|v - h_i\|^2 - 4\left\|\left(v - \frac{h_i + h_j}{2}\right)\right\|^2.$$

Because $\frac{h_i + h_j}{2} \in \mathcal{H}$, and therefore $\|v - \frac{h_i + h_j}{2}\| \geq \delta$ we have that

$$0 \leq \|h_i - h_j\|^2 \leq 2\|v - h_j\|^2 + 2\|v - h_i\|^2 - 4\delta^2 \xrightarrow{i,j \to 0} 0.$$

Hence, $\{h_j\}_{j=1}^{\infty}$ is a Cauchy sequence. Completeness of $\mathcal{V}$ implies the existence of $h^* \in \mathcal{V}$ such that $h_j \to h^*$ as $j \to \infty$ and closedness of $\mathcal{H}$ implies $h^* \in \mathcal{H}$. Then, set $h^{\perp} = v - h^*$. By the way we defined $\{h_j\}$, the previous sentence means $h^* \in \arg\min_{h \in \mathcal{H}} \|v - h\|^2$, and proposition 6.3 implies that $h^* = \arg\min_{h \in \mathcal{H}} \|v - h\|^2$. Therefore, $v - h^* \in \mathcal{H}^{\perp}$ by orthogonality principle. Now, we are gonna prove uniqueness.
Suppose $h', \bar{h}$ as above. Then

$$v = h^* + h^{\perp} = h' + \bar{h},$$

then

$$h^* + h^{\perp} - h' - \bar{h} = 0.$$

In particular, $h^* - h' = \bar{h} - h^{\perp}$, where both $h^* - h', \bar{h} - h^{\perp} \in \mathcal{H}$. Note, $\mathcal{H}^{\perp} \cap \mathcal{H} = \emptyset$. Thus, $h^* = h', \bar{h} = h^{\perp}$. □

Moving forward, let's take a look at some examples.

7.2 Optimal Predictor, Revisited

Consider $\mathcal{H} = \{f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} : f = \tilde{y} \circ \pi_{\mathcal{X}}\}$. Now we want to use function $y \in \mathcal{H}$ approximate

$$\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \tag{29}$$

$$(x, y) \to y \tag{30}$$

We saw $y^* \in \arg\min_{\tilde{y}} \mathbb{E}\,(\tilde{y}(x)-y)^2$, where $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ ie $y^*(x) = \mathbb{E}(\mathcal{Y}|\mathcal{X})$. With HP, let $y^* \circ \pi_{\mathcal{X}}$ realize $\arg\min_{\tilde{y}} \mathbb{E}\,(\tilde{y}(x)-y)^2$. Then, by OP (orthogonality principle), we need only to find $y^* \in \mathcal{H}$ such that

$$\mathbb{E}\,((y^* \circ \pi_{\mathcal{X}} - \pi_{\mathcal{Y}})\tilde{y} \circ \pi_{\mathcal{X}}) = 0, \text{ for all } \tilde{y} \circ \pi_{\mathcal{X}} \in \mathcal{H} \tag{31}$$

In fact, we could expand the equation above and convert it to conditional probability form.

$$\begin{aligned}
\mathbb{E}\,((y^*-y)\tilde{y}) &= \int_{\mathcal{X}\times\mathcal{Y}} (y^*-y(x))\tilde{y}(x)\,d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) && (32)\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}|\mathcal{X}} (y^*(x)-y)\tilde{y}(x)\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)\,d\mathbb{P}_{\mathcal{X}}(x) && (33)\\
&= \int_{\mathcal{X}} \tilde{y}(x)\left(\int_{\mathcal{Y}|\mathcal{X}} (y^*(x)-y)\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)\right) d\mathbb{P}_{\mathcal{X}}(x) && (34)
\end{aligned}$$

If we could find $y^* \in \mathcal{H}$ that makes the inner integral always equal to zero, which is

$$\int_{\mathcal{Y}|\mathcal{X}} (y^*(x)-y)\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = 0, \text{ for all } y^* \in \mathcal{H}, \tag{35}$$

then we know this y^* satisfies (31). Assume y^* exists. We have

$$\begin{aligned}
& \int_{\mathcal{Y}|\mathcal{X}} (y^*(x)-y)\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = 0 && (36)\\
\Rightarrow\quad & \int_{\mathcal{Y}|\mathcal{X}} y^*(x)\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = \int_{\mathcal{Y}|\mathcal{X}} y\,d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) && (37)\\
\Rightarrow\quad & y^*(x)\int_{\mathcal{Y}|\mathcal{X}} d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = \mathbb{E}(y|x) && (38)\\
\Rightarrow\quad & y^*(x) = \mathbb{E}(y|x) && (39)
\end{aligned}$$

Therefore, when we consider all possible approximation in $\mathcal{H}$, $\mathbb{E}(y|x)$ is actually optimal.Therefore, $y^* \circ \pi_{\mathcal{X}} - \pi_{\mathcal{Y}} \in \mathcal{H}^{\perp}$ ie $y^* \circ \pi_{\mathcal{X}} \in \arg\min_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \|\tilde{y} \circ \pi_{\mathcal{X}} - \pi_{\mathcal{Y}}\|^2$.

7.2.1 Optimal among constants

Consider $\mathcal{H}_0 = \{\underline{c} : \mathcal{X}\times\mathcal{Y} \to \mathcal{Y} : \underline{c}(x,y) \equiv c \text{ for some } c \in \mathbb{R}\}$.

Does the estimate $\underline{c}^*$ belongs to $\mathcal{H}$ of $\pi_{\mathcal{Y}}$. To find it we are going to apply Orthogonality Principle (OP):

$$\begin{aligned}
&\mathbb{E}((c^*-\pi_{\mathcal{Y}})c) = 0\ \forall c \in \mathbb{R}\\
&c\int_{\mathcal{X}\times\mathcal{Y}} (c^*-y)\,d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) = 0\\
&c^*\int_{\mathcal{X}\times\mathcal{Y}} d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} y\,d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y)\\
&c^*\int_{\mathcal{X}\times\mathcal{Y}} d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) = \int_{\mathcal{Y}} y\,d\mathbb{P}_{\mathcal{Y}}(y)\\
&c^* = \mathbb{E}(\mathcal{Y})
\end{aligned}$$

7.2.2 Linear Regression

Consider $\mathcal{H}_1 = \{a_0 + a_1 x : \mathcal{X}\times\mathcal{Y} \to \mathbb{R} | a_0, a_1 \in \mathbb{R}\}$. To find $a_0^\star + a_1^\star \in \mathcal{H}_1$ optimally approximating $\pi_y : \mathcal{X}\times\mathcal{Y} \xrightarrow{\mathcal{Y}}$ we will use the orthogonality principle $\langle a_0^\star + a_1^\star - \pi_y, b_0 + b_1 x\rangle = 0$ for all $b_0 + b_1 x \in \mathcal{H}_1$, that is to say

$$\mathbb{E}(a_0^* + a_1^* x - y)(b_0 + b_1 x) = 0\ \forall b_0, b_1 \in \mathbb{R}$$

In particular,

$$\mathbb{E}(a_0^* + a_1^* x - y) = 0 \text{ for } b_0 = 1, b_1 = 0$$
$$\mathbb{E}((a_0^* + a_1^* x - y)x) = 0 \text{ for } b_0 = 0, b_1 = 1$$

Thus, we get the following system of equations.

$$\begin{cases} a_0^* + a_1^* \mathbb{E}(\mathcal{X}) = \mathbb{E}(\mathcal{Y}) \\ a_0^* \mathbb{E}(\mathcal{X}) + a_1^* \mathbb{E}(\mathcal{X}^2) = \mathbb{E}(\mathcal{X}\mathcal{Y}) \end{cases} \tag{40}$$

The system can be rewritten in matrix form

$$\begin{pmatrix} 1 & \mathbb{E}(\mathcal{X}) \\ \mathbb{E}(\mathcal{X}) & \mathbb{E}(\mathcal{X}^2) \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\mathcal{Y}) \\ \mathbb{E}(\mathcal{X}\mathcal{Y}). \end{pmatrix}$$
$$\begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} 1 & \mathbb{E}(\mathcal{X}) \\ \mathbb{E}(\mathcal{X}) & \mathbb{E}(\mathcal{X}^2) \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}(\mathcal{Y}) \\ \mathbb{E}(\mathcal{X}\mathcal{Y}). \end{pmatrix}$$

In general, for $H_n = \left\{ \sum_{j=0}^{n} a_j x^j : a_j \epsilon \right\}$, the Orthogonality principle provides a procedure for finding $a_0^*, a_1^*, ..., a_n^*$, such that:

$$a_0^* + ... + a_n^* x^n = \pi_{H_n}(y) \tag{41}$$

So, we have found a way to deal with the computation of $\mathcal{P}_{\mathcal{X}\times\mathcal{Y}}$ without knowing a measure. Now, the question is the following: which class of functions is better? A richer hypothesis class $\mathcal{H}_1 \supset \mathcal{H}_0$ gives a better approximator. Why wouldn't we use the most complex $\mathcal{H}$ possible?

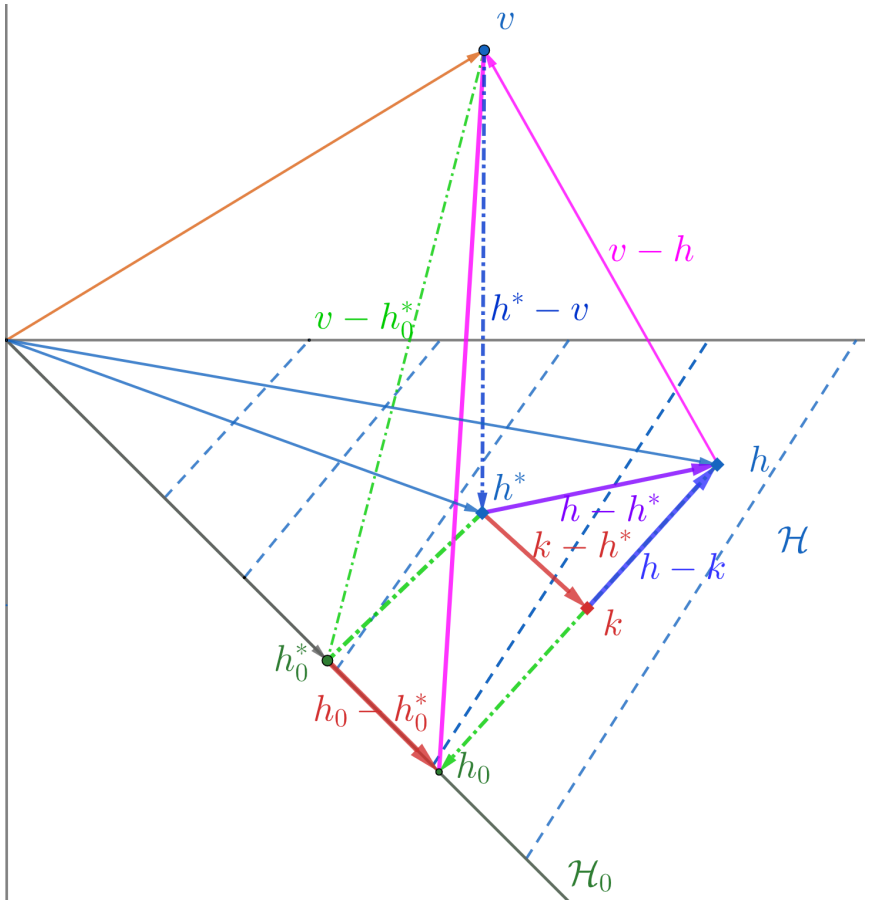

Figure 19: Bias-Variance in Hilbert Space

7.3 Bias-Variance: First Glance

Let $v \in \mathcal{V}$ and consider projections

$$h^* = \arg\min_{h \in \mathcal{H}} \|v - h\|^2 \tag{42}$$

$$h_0^* = \arg\min_{h \in \mathcal{H}_0} \|v - h_0\|^2 \tag{43}$$

onto subspaces $\mathcal{H} \supset \mathcal{H}_0$, respectively. For arbitrary $h \in \mathcal{H}$, set

$$h_0 := \pi_{\mathcal{H}_0}(h) \tag{44}$$

We are interested in the relationship between $\|v - h\|$ and $\|v - h_0\|$; which vector better approximates v? Of course, we know that h^* is a better approximation of v than h_0^*, since $\mathcal{H}_0 \subset \mathcal{H}$. Mathematically, this means that

$$\|v - h^*\| \leq \|v - h_0^*\|. \tag{45}$$

As shown in the Figure 19, by Pythagorean Theorem, we have

$$\|v - h\|^2 = \|v - h^*\|^2 + \|h^* - h\|^2 \tag{46}$$

$$\|v - h_0\|^2 = \|v - h_0^*\|^2 + \|h_0^* - h_0\|^2 \tag{47}$$

However, since $h_0^* - h_0 = \pi_{\mathcal{H}_0}(h^* - h)$ we have

$$\|h^* - h\|^2 \geq \|h_0^* - h_0\|^2. \tag{48}$$

Thus, while $\|v - h^*\|^2 \leq \|v - h_0^*\|^2$, $\|h^* - h\|^2 \geq \|h_0^* - h_0\|^2$ and therefore $\|v - h\|$ and $\|v - h_0\|$ are not comparable. This observation describes the famed bias–variance tradeoff, which summarizes as: a richer hypothesis class $\mathcal{H} \subset \mathcal{V}$ is better in one respect but not necessarily in another.

8 Lecture 8

8.1 Introduction

The purpose of this lecture is to collect and orient where we are. We give our first steps towards PAC-learnability passing through empirical risk minimization. We learned the projection theorem, which provides a concrete way to construct a model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ when $\mathcal{Y} = \mathbb{R}$. In this case, we call the supervised learning problem *regression*. The scenario for which projection "works" is highly specific, namely, we must be in a regression setting and the measure of performance for model "correctness" must be the mean squared error, which alternatively cashes out as: we must have a linear (vector) space[7] with an *inner product*. In general, our working notion of model performance—what we will define as *loss*—need not come from or be in any way related to an inner product. We interpreted an inner product as a mechanism for algebraicizing geometric notions, and you should turn this around when making sense of new concepts: supposing our loss *were* to come from an inner product, what is the corresponding geometry which encodes the particular phenomenon we're considering (e.g. bias-variance, optimality, etc.)

Note on notation: $(x, y) \in \mathcal{X} \times \mathcal{Y}$ will generally denote a point in the joint space. But we will see y floating around with many decorations, and each one means a different thing. Undecorated y is a *point* or element of $\mathcal{Y}$; by contrast, $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ will usually denote a *map*, which takes as input *points* $x \in \mathcal{X}$ and outputs *points* $\tilde{y}(x) \in \mathcal{Y}$. We will also use $\hat{y}_{(\cdot)} : \bigcup_{m \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ to denote an *algorithm* which takes as input a data *sequence* $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$ and outputs a *model* $\hat{y}_S \in \mathcal{H}$ (the latter of which takes as input points in $\mathcal{X}$ and returns points in $\mathcal{Y}$). And so on. The reason for being uber persnickety about what maps what to what is because y, $\tilde{y}$, and $\hat{y}_{(\cdot)}$ are all different kinds of animals; y^* will usually denote a map $\mathcal{X} \to \mathcal{Y}$, like $\tilde{y}$, except be special in that it's optimal in some respect. All of this is convention, and you should be careful to read the context in which these particular instances are being used to ensure they match in those scenarios the identity we're claiming they do here.

8.2 Machine Learning: Beyond Hilbert Projection

Our setting is the same: we have joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ and recall the standard diagram eq. (6). Our goal is to find *model* $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ for which $\tilde{y}(x) \approx y$ for "most" pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Probability makes sense (/ precisifies) 'most.' Before, we expressed our problem as finding model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ minimizing the expression $\|\pi_{\mathcal{Y}} - \tilde{y} \circ \pi_{\mathcal{X}}\|$ where $\|\cdot\|$ defines metric induced by norm induced by inner product $\langle f, g\rangle := \mathbb{E}(fg)$. In general, $\mathcal{Y}$ need not be $\mathbb{R}$ (in which case all the vector space stuff may disappear) and even if it is, our measure of model performance need not be one induced by inner product. And if there's no inner product, there's no obvious geometry.

Furthermore, with Hilbert Projection, we worked inside the "full" space of maps $\mathcal{H} := \{\mathcal{X} \to \mathcal{Y}\}$ (and as an afterthought, considered implications on restricted classes $\mathcal{H}_0 \subsetneq \mathcal{H}$) but going forward, we will need to explicitly set out our *hypothesis class* from the get-go.

Once we have our hands on *a* model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$, we would like some way to *measure* (not in the measure-theoretic sense!) how well $\tilde{y}$ "does" on arbitrary labeled data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which we capture by saying there is some random variable $l_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we call the *loss function*, mapping $(x, y) \mapsto l_{\tilde{y}}(x, y) \in \mathbb{R}$. Often or usually this random variable will be nonnegative and we'd like it to be as small as possible on as many points as possible. Said differently, we want $\mathbb{E}(l_{\tilde{y}})$ to be small. The norm induced from inner product in Hilbert projection suggests a loss function: $l_{\tilde{y}}(x, y) := (\tilde{y}(x) - y)^2$. We'll see others.

What is the "variable"? The model! So you can think of loss l as inducing a *map*

$$l_{(\cdot)} : \mathcal{H} \to \{\mathcal{X} \times \mathcal{Y} \xrightarrow{f} \mathbb{R}\}$$

from models to the set of random variables, returning a random variable $l_{\tilde{y}}$ for each specified model $\tilde{y} \in \mathcal{H}$. And thus stated we cast *the* problem of supervised machine learning (sml) as finding an optimal model $y^* \in \arg\min_{\tilde{y} \in \mathcal{H}} \mathbb{E}(l_{\tilde{y}})$. Note that our objective is stated "in expectation." Crudely you

[7] Which recall is $\mathcal{V} : \{\mathcal{X} \times \mathcal{Y} \to \mathcal{Y} = \mathbb{R}\}$ the set of random variables on $\mathcal{X} \times \mathcal{Y}$.

can think of sml as curve fitting, and there is no problem with this plebian perspective as long as you distinguish fitting the data in your hands from the data "not at your immediate disposal." Supervised machine learning deals not with fitting data as such but fitting the source, whence data comes. It amounts to fitting the measure! Hence how ml provides a concrete setting for understanding probability: you really cannot talk about ml without it.

Collecting ingredients, we have:

1. Joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$.
2. Data $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$, where probability space $(\mathcal{X} \times \mathcal{Y})^m$ has independent measure $\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m} = \big(\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}\big)^m$. In the literature, you'll often see it written as: $(x_1, y_1), \ldots, (x_m, y_m) \sim_{\mathtt{iid}} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ (or even just: iid from a distro, which amounts to the same). I call the data set a data *sequence* because it's a *point* living in some product space $(\mathcal{X} \times \mathcal{Y})^m$. People call it a set because you can usually get away without the rigor. 'Usually'. But it's wrong. It's wrong technically, and it's misleading. If there's any other fun descriptor to communicate the egregiousness of calling it a set, feel free to insert here.
3. Hypothesis class $\mathcal{H} \subsetneq \{\mathcal{X} \times \mathcal{Y}\}$. In Hilbert Projection, we allowed all models. Now we won't. It's simply too hard to do anything if we consider the set of all possible maps.
4. Loss function $l_{(\cdot)} : \mathcal{H} \to \{\mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$, and finally
5. Objective, minimize $\mathbb{E}(l_{\tilde{y}})$ ranging over all models $\tilde{y} \in \mathcal{H}$.

Now we can get to work.

We state our problem thus: is there some *algorithm*

$$\hat{y}_{(\cdot)} : \bigcup_{m\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$

which takes an arbitrary data set S and outputs model $\hat{y}_S \in \mathcal{H}$? Here's an algorithm: pick your favorite model $\tilde{y}_0 \in \mathcal{H}$ and no matter what the data is, set $\hat{y} \equiv \tilde{y}_0$ (this means: $\hat{y}_S = \tilde{y}_0$ for every data S). Well that's silly, there's no connection here to how good $\tilde{y}_0$ is. So we want some conditions that say something to the effect of: as the length of the data sequence grows, the probability that the resulting model is good also grows. Something like that. If something smells like Glivenko-Cantelli, good, and if it doesn't, please review Glivenko-Cantelli.

In abstraction, we can usually figure out what an optimal $y^* \in \arg\min_{\tilde{y}\in\mathcal{H}} \mathbb{E}(l_{\tilde{y}})$ is. The name of the game is to decompose the expectation on the right using LTE (law of total probability, which is a fancy way of saying: remember what conditional probability is definition 2.7) and examining the inner integral, as a function of x.

Once we have a model $\hat{y}_S$, we'll want to consider deviation from optimality:

$$\mathbb{E}(l_{\hat{y}_{\tilde{y}}}) - \mathbb{E}(l_{y^*}), \tag{49}$$

where here optimality is global optimality $y^* \in \arg\min_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \mathbb{E}(l_{\tilde{y}})$, and we decompose this deviation using our celebrated 'add zero' trick:

$$\underbrace{\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*_{\mathcal{H}}})}_{\text{estimation error}} + \underbrace{\mathbb{E}(l_{y^*_{\mathcal{H}}}) - \mathbb{E}(l_{y^*})}_{\text{approximation error}}, \tag{50}$$

where $y^*_{\mathcal{H}} \in \arg\min_{\tilde{y}\in\mathcal{H}} \mathbb{E}(l_{\tilde{y}})$. Which doesn't really do much except help us conceptualize what might be going on, moving forward. Approximation is an intrinsic property of the hypothesis space and tells us how good is even possible. Our algorithm likely won't return optimal even in $\mathcal{H}$, so estimation error tells us how far we are from optimality-in-$\mathcal{H}$. This error is data/algo-dependent.

8.3 Towards PAC-learnability

We have a hypothesis class, a subset of maps so that $\mathcal{H} \to \{f : \mathcal{X} \to \mathcal{Y}\}$. We want an algorithm:

$$\hat{y}_{(\cdot)} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$$

which maps $(S \to \hat{y}_S)$ This algorithm takes a data sequence and returns a model. In addition, we also want: where "$\approx$" means $\mathbb{E}(l_{\hat{y}_S}) \approx \mathbb{E}(l_{y^*})$ or that the expected loss of the model is approximately equal to the expected loss of the optimal. Notice that the domain of the algorithm is any finite data sequence; the input is data, and the subsequent output is a model where different data will net you a different model.

Can we come up with any guarantees? Given $\epsilon > 0$ and $\delta \in (0, 1)$, can we ensure that

$$\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m}\left(\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*}) > \epsilon\right) < \delta?$$

where $\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*})$ indicates that the difference in expected values refers to the set in S (data) that fulfills this condition of the difference being $> \epsilon$. This is the deviation from the optimal loss.

Our setting: a joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$, hypothesis class of maps $\mathcal{H} \subsetneq \{f : \mathcal{X} \to \mathcal{Y}\}$, and loss function, a random variable, $l_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for each model $\tilde{y} \in \mathcal{H}$. We would like to design an algorithm

$$\hat{y}_{(\cdot)} : \bigcup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$$

so that deviation of model performance from optimality

$$\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*})$$

is "small" "in probability." We'll usually denote y^* implicitly by the following $\mathbb{E}(l_{y^*}) := \inf_{\tilde{y}:\mathcal{X}\to\mathcal{Y}} \mathbb{E}(l_{\tilde{y}})$ even if such a map $y^* : \mathcal{X} \to \mathcal{Y}$ doesn't exist. We have two scarequotes: the first, "small," deals with the difference between $\mathbb{E}(l_{\hat{y}_S})$ and $\mathbb{E}(l_{y^*})$ while the second, "in probability," deals with measurement of $S \subset (\mathcal{X} \times \mathcal{Y})^m$, where $m = |S|$. Since we didn't explicitly state what the measure on $(\mathcal{X} \times \mathcal{Y})^m$ is, it must mean we take the only measure we have, $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ and measure the product space with this measure, using independence $\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m} = \left(\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}\right)^m$. Of course, we also want to connect an intuition—which has undergirded much of what we've seen with concentration inequalities and asymptotic statements (like lln)—that more data is better. In other words, we would expect something like:

$$\lim_{|S|\to\infty} \mathbb{E}(l_{\hat{y}_S}) = \mathbb{E}(l_{y^*}).$$

We will precisify this desire by characterizing the hypothesis class.

Definition 8.1. We say that a hypothesis class $\mathcal{H} \subsetneq \{X \to Y\}$ is PAC learnable (Probably Approximately Correct) if there is an algorithm $\hat{y}_{(\cdot)} : \bigcup_{m \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ and a map $\mu : (0, 1) \times (0, 1) \to \mathbb{N}$ such that $m = |S| > \mu(\epsilon, \delta)$ implies

$$\mathbb{P}_{(\mathcal{X}\times\mathcal{Y})^m}\left(\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*}) > \epsilon\right) < \delta \tag{51}$$

In words, suppose that $\hat{y}_{(\cdot)}$ is the algorithm; PAC learnability encodes a guarantee that your data will likely produce a well-performing (good) model.[8]

8.4 Introduction to Empirical Risk Minimization

The *one algorithm* for solving the ML problem is Empirical Risk Minimization (almost always). Recall that our definition for $\hat{y}_{(\cdot)}$ takes in an arbitrary dataset S and outputs model $\hat{y}_S \in \mathcal{H}$.

[8] Choose the richest hypothesis class such that estimation may be controllably guaranteed.

Definition 8.2. Let $S \in (\mathcal{X} \times \mathcal{Y})^m$ be a data sequence. We define *empirical risk minimization* as

$$\hat{y}_S := \arg\min_{\tilde{y} \in \mathcal{H}} e_S(\tilde{y}). \tag{52}$$

where $e_S(\cdot) : \mathcal{H} \to \mathbb{R}$ is defined by as:

$$e_S(\tilde{y}) := \frac{1}{|S|} \sum_{(\mathcal{X} \times \mathcal{Y}) \in S} l_{\tilde{y}}(x, y) \tag{53}$$

Or if we write data explicitly as $S = \big((x_1, y_1), ..., (x_m, y_m)\big)$, then

$$e_S(\tilde{y}) = \frac{1}{m} \sum_{j=1}^{m} l_{\tilde{y}}(x_j, y_j)$$

We hope two things: (1) Hope that empirical estimate is a *good one* with respect to empirical performance, and (2) hope it generalizes (beyond data). Having defined "the" algorithm which will serve as the golden standard for what we do moving forward, we are left with three thematic questions.

1. How do we actually "do" ERM, i.e. operationally construct or find $\arg\min_{\tilde{y} \in \mathcal{H}} e_s(l_{\tilde{y}})$?
2. How do we select hypothesis class $\mathcal{H}$?
3. Can we put guarantees or rigorous bounds on deviation of model performance from optimality $\mathbb{E}(l_{\hat{y}_S}) - \mathbb{E}(l_{y^*})$?

9 Lecture 9

We present and prove the celebrated Universal Approximation (UA) Theorem, which states that one-layer neural networks express continuous function on a compact domain. The most well-known proof was given in [1], and numerous others have since followed, in particular [2] which mathematically is similar to the one given here, and [3], having similar visuals. While the original proof in [1] is elegant and crisp, it relies on fairly sophisticated tools from analysis (Riesz Representation and Hahn-Banach, e.g.). As such, appreciation for its succinctness may initially evade nevertheless eager and enthusiastic students with less matured mathematical backgrounds.

UA can be shown by elementary techniques that make apparent the expressivity of neural networks. We warm up in section 9.1 with a proof that piecewise constant functions express continuous ones, and claim as preview that the argument is conceptually similar to that of UA. We detour through partitions of unity in section 10, using UA as excuse for introducing an otherwise powerful and versatile tool from geometry that makes possible global constructions from a collection of local data. Then in section 11, we prove the Universal Approximation Theorem in steps. We first show that linear spaces of partitions of unity express continuous functions, and then that a class of functions that expresses partitions of unity thereby expresses continuous functions. We finally show that one layer neural networks express a class of partitions of unity. We conclude with an algorithm for constructing such a network and connect this algorithm with some plots interpreting a (properly parametered) pair of nodes as approximate of bump function in partition of unity.

9.1 A Preliminary Result

We review a fact from calculus that piecewise constant functions are expressive of continuous ones. First a definition:

Definition 9.1. We say a class $\mathcal{H}$ of functions expresses (another) class of functions $\mathcal{G}$ if for each $g \in \mathcal{G}$ and $\varepsilon > 0$, there is $h_{g,\varepsilon} \in \mathcal{H}$ for which $\|g - h_{g,\varepsilon}\|_\infty < \varepsilon$.

This definition slightly generalizes dense subsets: a class $\mathcal{H}$ may express $\mathcal{G}$ without satisfying any containment relations.

Proposition 9.1. The class

$$\mathcal{H} = \bigcup_{m\in\mathbb{N}} \left\{ \tilde{f}(x) = \sum_{j=1}^{m} \alpha_j \mathbb{1}_{x\in I_j} \,:\, \alpha_j \in \mathbb{R}, I_j \text{ an open, closed, or half open interval in } [0,1] \right\}$$

of piecewise constant functions expresses $\mathcal{C}([0,1])$.

Proof. Let $f : [0,1] \to \mathbb{R}$ be continuous. For fixed $\varepsilon > 0$ and $x \in [0,1]$ there is $\delta_x > 0$ for which

$$|f(x) - f(x')| < \varepsilon \text{ whenever } |x - x'| \le \delta.$$

Thus, there is open cover $\{B_{\delta_x}(x)\}_{x\in[0,1]}$ of $[0,1]$ and therefore finite subcover $\{B_{\delta_{x_j}}(x_j)\}_{j=1}^m$. Select $x'_1 < \ldots < x'_{m-1}$ so that $x'_j \in B_{\delta_{x_j}}(x_j) \cap B_{\delta_{x_{j+1}}}(x_{j+1})$. For completeness, set $x'_0 = 0 \in B_{\delta_{x_1}}(x_1)$ and $x'_{m+1} = 1 \in B_{\delta_{x_m}}(x_m)$.

We define

$$\tilde{f}(x) := \sum_{j=1}^{m} f(x_j)\mathbb{1}_{x\in I_j} \tag{54}$$

where

$$I_j := \begin{cases} [x'_{j-1}, x'_j) & \text{if } j < m \\ [x'_{m-1}, 1] & \text{else.} \end{cases}$$

It is easy to see that $\|f - \tilde{f}\|_\infty < \varepsilon$: for $x \in I_j \subset [0,1]$,

$$|f(x) - \tilde{f}(x)| = |f(x) - f(x_j)| < \varepsilon,$$

because $x \in I_j$ implies that $|x - x_j| < \delta_{x_j}$. □

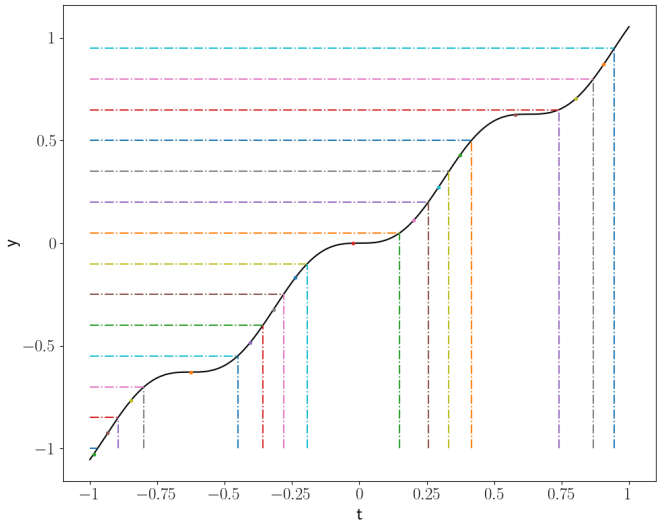

Figure 20: Constructing piecewise constant approximator

The plot in fig. 20 crudely reflects the argument in proposition 9.1: slicing the codomain in ε-bins, we obtain a partition of domain into bins each which may serve as the I_js. On I_j, the value may be any sampled point $f(x)$ for $x \in I_j$. We denote evaluation at the midpoint with a small dot. An argument for universal approximation of one-layer neural networks is conceptually very similar: on (sufficiently) small intervals, we approximate a function value using a portion of the network, which portion itself approximates the indicator $\mathbb{1}_{x \in I_j}$ as a bump function, with almost zero tails outside the interval.

10 Lecture 10

[TO DO]

10.1 Partitions of Unity

Partitions of unity allow construction of global objects using locally defined data. Philosophically, they define state-dependent convex combinations, and are typically used to specify which local data contributes where in a global construction. The fundamental component for partitions of unity in $\mathbb{R}$ is a smooth step function $\rho : \mathbb{R} \to [0,1]$ which surjectively maps onto the codomain.

Definition 10.1. We define a *smooth step* function $\rho : \mathbb{R} \twoheadrightarrow [0,1]$ to be a monotonic, surjective, smooth map.

A smooth step is much like a sigmoidal, except that we do not require of the latter that it obtain its limit. A pair of smooth steps ρ_0, ρ_1 induce a smooth bump function β having compact support as follows: supposing that $\rho_0 \geq \rho_1$, then $\beta := \rho_0 - \rho_1$ has the property that $\beta = 0$ when $\rho_0 = 0$ or $\rho_1 = 1$. If $\rho_0^{-1}(1) \cap \rho_1^{-1}(0) \neq \varnothing$, then β also realizes 1 at some value $x_1 \in \rho_0^{-1}(1) \cap \rho_1^{-1}(0)$, and the set of functions $\left\{\beta_\ell := (1-\beta)\mathbb{1}_{x \leq x_1}, \beta, \beta_r := (1-\beta)\mathbb{1}_{x \geq x_1}\right\}$ then defines a partition of unity.

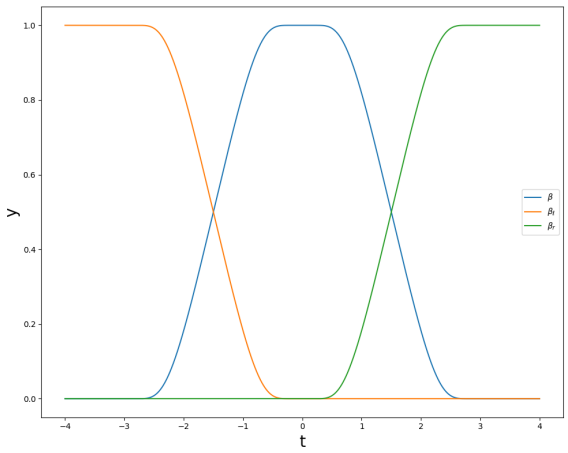

Figure 21: Partition of unity

The formal definition is as follows ([4, §13]):

Definition 10.2. Let $\mathcal{O} = \{U_\alpha\}_{\alpha \in \mathcal{A}}$ be an open cover of $\mathbb{R}$, so that $\mathbb{R} = \bigcup_{\alpha \in \mathcal{A}} U_\alpha$. A smooth *partition of unity* $\{\rho_\alpha : \mathbb{R} \to [0,1]\}_{\alpha \in \mathcal{A}}$ is a collection of smooth functions satisfying

1. $\{\operatorname{supp}(\rho_\alpha)\}_{\alpha \in \mathcal{A}}$ is locally finite, and
2. $\sum_{\alpha \in \mathcal{A}} \rho_\alpha \equiv 1$.

When $\operatorname{supp}(\rho_\alpha) \subset U_\alpha$ for each α, we say that the partition of unity is *subordinate to* $\mathcal{O}$.

In the example above, the partition of unity is subordinate to $\mathcal{O} = \{(-\infty, x_1), U_1, (x_1, \infty)\}$ with $U_1 \supset \beta^{-1}((0,1])$. The local finiteness condition of supports implies that $\rho_\alpha(x) \neq 0$ for only finitely many α at each point $x \in \mathbb{R}$. The second condition is the state-dependent convex combination: for each $x \in \mathbb{R}$, $\sum_{\alpha \in \mathcal{A}} \rho_\alpha(x) = 1$. In general, partitions of unity exist ([4], [5]). In other words, given any open cover $\mathcal{O}$ of $\mathbb{R}$, there is partition of unity subordinate to $\mathcal{O}$.

We will see (lemma 11.3) that we can construct smooth step functions to approximate sigmoidals, and reverse the process to construct an approximator from neural networks for a suitable step, and therefore of partition of unity. (See [5] or [4] for more details on the construction.)

Here is a use case: we want to construct a continuous or smooth approximating function with finitely many function evaluations. To this end, we put a bump function about each evaluated point x_j and multiply by the function value $f(x_j)$. We isolate a notion:

Definition 10.3. For fixed partition of unity $\{\rho_\alpha : \mathbb{R} \to [0,1]\}_{\alpha \in A}$, we call the vector space

$$\left\{ \sum_{\alpha \in A} c_\alpha \rho_\alpha : c_\alpha \in \mathbb{R} \right\}_{\alpha \in A}$$

generated by partition a *linear space of (the) partition of unity*.

By local finiteness, this collection is indeed a linear space. Our strategy will be: identify the relevant open cover, cite existence of partition of unity, and reason about a suitable element in the corresponding linear space of this partition. With this preamble, we may now move on to the Universal Approximation Theorem.

11 Lecture 11

11.1 Universal Approximation

Theorem 11.1. Let $\mathcal{H}_n = \left\{ \sum_{j=1}^{n} c_j \sigma(w_j(\cdot) + b_j) + d : \mathbb{R} \to \mathbb{R} : c_j, w_j, b_j, d \in \mathbb{R} \right\}$ and $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$ with $\sigma : \mathbb{R} \to \mathbb{R}$ sigmoidal. Then $\mathcal{H}$ expresses $\mathcal{C}([0,1])$.

We prove in three lemmas.

11.2 Lemma 1

Lemma 11.1. Linear spaces of partitions of unity (definition 10.3) express $\mathcal{C}([0,1])$.

Proof. Suppose without loss of generality that $f([0,1]) \subseteq [0,1]$. Because f is continuous, for each $x \in [0,1]$, there is $\delta_x > 0$ for which

$$|f(x) - f(x')| < \varepsilon/2 \text{ whenever } |x - x'| < \delta_x. \tag{55}$$

The collection of δ_x induces open cover $\mathcal{O} := \left\{ (x - \delta_x, x + \delta_x) \right\}_{x \in [0,1]}$ of $[0,1]$ and therefore also a partition of unity

$$\left\{ \rho_{x'} : [0,1] \to [0,1] \right\}_{x' \in A} \tag{56}$$

subordinate to $\mathcal{O}$ for finite $A \subset [0,1]$.[9]

We claim that

$$\tilde{f}_{pre}(\cdot) := \sum_{x' \in A} \rho_{x'}(\cdot) f(x') \tag{57}$$

$\varepsilon/2$-approximates f, i.e. that $\|f - \tilde{f}_{pre}\|_\infty < \varepsilon/2$. Indeed, for $t \in [0,1]$,

$$\begin{aligned}
|f(t) - \tilde{f}_{pre}(t)| &= \left| \sum_{x' \in A} f(x') \rho_{x'}(t) - f(t) \right| \\
&= \left| \sum_{x' \in A} (f(x') - f(t)) \rho_{x'}(t) \right| \\
&\leq \sum_{x' \in A} \rho_{x'}(t) |f(x') - f(t)| \\
&= \sum_{|t - x'| < \delta_{x'}} \rho_{x'}(t) |f(x') - f(t)| \\
&\leq \sum_{|t - x'| < \delta_{x'}} \rho_{x'}(t) \cdot \varepsilon/2 \\
&= \tfrac{\varepsilon}{2} \sum_{x' \in A} \rho_{x'}(t) \\
&= \tfrac{\varepsilon}{2}.
\end{aligned} \tag{58}$$

The equality in the second line follows because $\sum_{x'} \rho_{x'}(t) = 1$, and the equality in the fourth line because $\text{supp}(\rho_{x'}) \subset (x' - \delta_{x'}, x' + \delta_{x'})$. □

11.3 Lemma 2

Having shown that arbitrary functions may be arbitrarily-well approximated by linear combinations of (elements of a) partitions of unity, we now show that approximations of such well-approximate any function.

[9] We do note require A to be finite for this argument, but will for the next (lemma 11.2). Since $[0,1]$ is compact, there is no harm supposing apriori that the cover is finite. We also do not need $f([0,1]) \subset [0,1]$, but as f is continuous the range is bounded. Containment in $[0,1]$ is also convenient for lemma 11.2.

Lemma 11.2. Suppose that function space $\mathcal{H}$ expresses (a linear space of) partitions of unity. Then $\mathcal{H}$ expresses $\mathcal{C}([0,1])$.

We state and show this explicitly as the argument will be useful for construction.

Proof. We start with expression in eq. (57) and suppose that $\{\rho_{x'} : [0,1] \to [0,1]\}_{x' \in A}$ is partition of unity subordinate to $\mathcal{O}$ (eq. (56)). Because $[0,1]$ is compact, we suppose without loss of generality that A is finite, say $A = \{x'_1, \ldots, x'_m\}$. For each $x' \in A$ there is by assumption $\nu_{x'} \in \mathcal{H}$ for which $\|\rho_{x'} - \nu_{x'}\|_\infty < \frac{\varepsilon}{2m}$. We then define

$$\tilde{f}(\cdot) := \sum_{x' \in A} f(x')\nu_{x'}(\cdot), \tag{59}$$

and show that $\tilde{f}$ ε-approximates f.

Indeed,

$$|f(t) - \tilde{f}(t)| \leq |f(t) - \tilde{f}_{pre}(t)| + |\tilde{f}_{pre}(t) - \tilde{f}(t)| \leq \frac{\varepsilon}{2} + |\tilde{f}_{pre}(t) - \tilde{f}(t)|,$$

by lemma 11.1, and we are left to showing that $|\tilde{f}_{pre}(t) - \tilde{f}(t)| < \varepsilon/2$.

Computing,

$$\begin{aligned} |\tilde{f}(t) - \tilde{f}_{pre}(t)| &= \left| \sum_{x' \in A} f(x')\rho_{x'}(t) - \sum_{x' \in A} f(x')\nu_{x'}(t) \right| \\ &= \sum_{x' \in A} f(x') \left| \rho_{x'}(t) - \nu_{x'}(t) \right| \\ &\leq \sum_{x' \in A} |\rho_{x'}(t) - \nu_{x'}(t)| \\ &\leq \sum_{x' \in A} \frac{\varepsilon}{2m} \\ &= \frac{\varepsilon}{2}. \end{aligned} \tag{60}$$

The first inequality follows by assumption that $f([0,1]) \subseteq [0,1]$. □

The final ingredient needed is to realize lemma 11.2 with one-layer neural networks, namely *that* neural networks express partitions of unity. Because smooth steps (definition 10.1) form a skeleton of partitions of unity, it suffices to show that one layer neural networks express smooth steps.

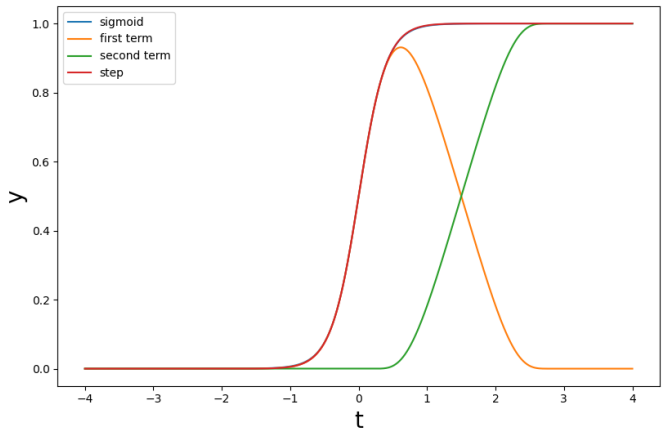

Figure 22: Example construction of lemma 11.3

11.4 Lemma 3

Lemma 11.3 (From σ to ρ)**.** Let $\sigma : \mathbb{R} \to [0,1]$ be a sigmoidal and $\varepsilon > 0$. There is smooth step $\rho : \mathbb{R} \twoheadrightarrow [0,1]$ (definition 10.1) for which

$$\|\rho - \sigma\|_\infty < \varepsilon.$$

This lemma guarantees that sigmoidals of the form $\sigma(w(\cdot)+b)$ express smooth steps, and moreover that the linear space of such sigmoidals express smooth bumps, and therefore partitions of unity. Conversely, there is a class of smooth steps that expresses said sigmoidals.

Proof. Define open cover

$$\mathcal{O} = \left\{ I_0 := \sigma^{-1}((0,\varepsilon)), I_\sigma := \sigma^{-1}((\varepsilon/2, 1-\varepsilon/2)), I_1 := \sigma^{-1}((1-\varepsilon,1)) \right\},$$

and let $\{\beta_0, \beta_\sigma, \beta_1\}$ be a smooth partition of unity subordinate to $\mathcal{O}$. We then define

$$\rho := \sigma \cdot \beta_\sigma + \beta_1 = \sum_{k \in \{0,\sigma,1\}} k\beta_k,$$

and conclude that $|\rho(x) - \sigma(x)| < \varepsilon$. □

Using partition from fig. 21, we illustrate this construction with a plot: the red graph is orange times blue + green in fig. 22.

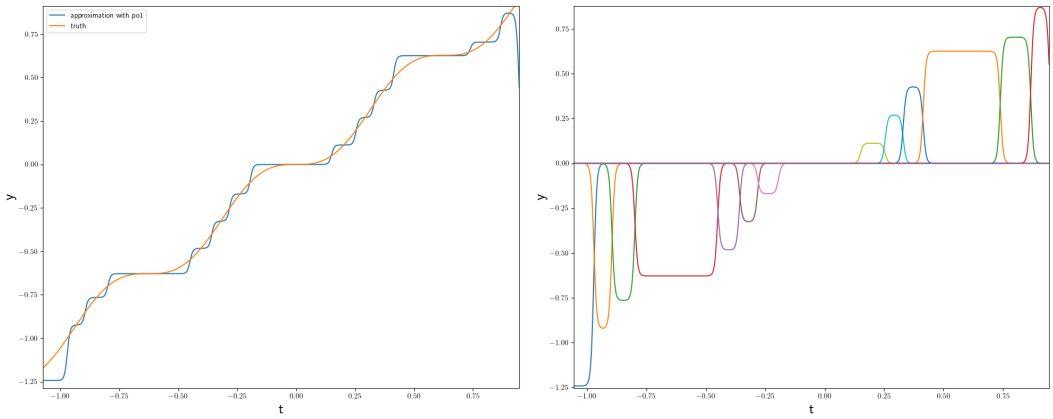

Figure 23: $\varepsilon = 0.15$

We collect ingredients for proof of the main theorem.

Proof of theorem 11.1. By lemma 11.3, neural networks express partitions of unity, and therefore by lemma 11.2 express $\mathcal{C}([0,1])$. □

The arguments readily translate into a recipe for *constructing* an approximating neural network. Starting with piecewise constant approximation in eq. (54), we successively define smooth bumps. The network will have $2m$ nodes, defined as

$$\nu_{2m} := \sum_{j=1}^{m} f(x_j) \underbrace{\left(\sigma(w_j(\cdot)+b_j) - \sigma(w_{j+1}(\cdot)+b_{j+1})\right)}_{\text{bump approximation}}.$$

For $j = 1$, set $b_1 := \sigma^{-1}(1-\varepsilon)$. Because $x_j' \in B_{\delta_{x_j}}(x_j) \cap B_{\delta_{x_{j+1}}}(x_{j+1})$, we choose w_{j+1} and b_{j+1} satisfying the following: $\sigma\big(w_{j+1}(x_{j+1} - \delta_{x_{j+1}}) + b_{j+1}\big) < \varepsilon$ and $\sigma\big(w_{j+1}(x_j + \delta_{x_j}) + b_{j+1}\big) > 1-\varepsilon$. If σ is strictly increasing then there is a linear system of equations

$$\begin{pmatrix} x_{j+1} - \delta_{x_{j+1}} & 1 \\ x_j + \delta_{x_j} & 1 \end{pmatrix} \begin{pmatrix} w_{j+1} \\ b_{j+1} \end{pmatrix} = \begin{pmatrix} \sigma^{-1}(\varepsilon) \\ \sigma^{-1}(1-\varepsilon) \end{pmatrix}$$

which may be solved to find these parameters.

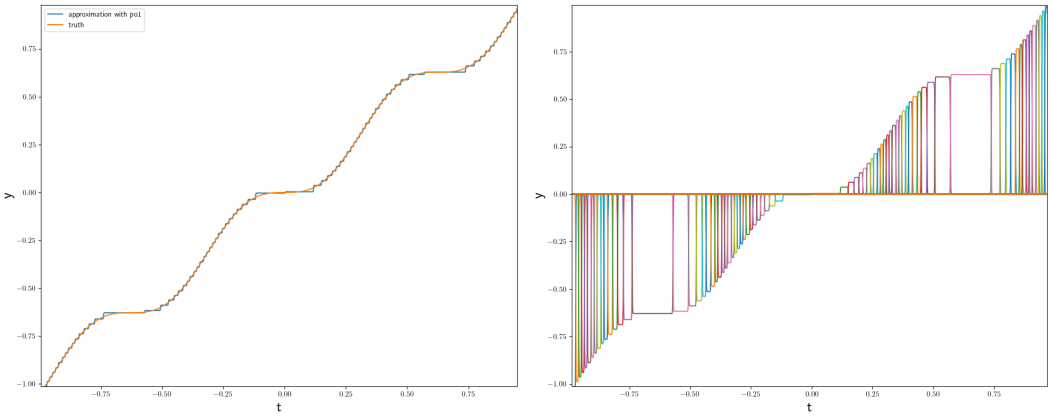

Figure 24: $\varepsilon = 0.025$

12 Lecture 12

12.1 Construction of a piecewise function

Example 12.1. Let $\mathcal{C}_{fin} := \mathcal{C}(\mathbb{R}) \cap L_1$ is "learnable" in $\mathcal{H} = \{\text{piecewise constants}\}$, where $\mathcal{C}(\mathbb{R}) = \{\text{continuous functions on } \mathbb{R}\}$, $L_1 = \{f : \mathcal{X} \to \mathbb{R} : \int_{\mathcal{X}} |f| d\mathbb{P}_{\mathcal{X}}(x) < \infty\}$

WLOG: suppose that $f : \mathcal{X} \to \mathbb{R}^{\geq 0}$ and fix $\epsilon > 0$. Since $f \in L_1$ there is some $T > 0$ for which

$$\int_{|t|>T} f(x) d\mathbb{P}_{\mathcal{X}}(x) < \frac{\epsilon}{2} \tag{61}$$

Note, (61) defines part of $\tilde{f}_S : \tilde{f}_S(t) = 0$ for $|t| > T$ and

$$\int_{\mathbb{R}} f d\mathbb{P} = \int_{[-T,T]} f d\mathbb{P} + \int_{\mathbb{R}\setminus[-T,T]} f d\mathbb{P}.$$

We want to show that we may guarantee a model $\tilde{f}_S \in \mathcal{H}$ with $\mathbb{E}(|f - \tilde{f}_S|) < \epsilon$ as long as $|S|$ is sufficiently large.

Now, since f is continuous on a compact interval $[-T, T]$, it's uniformly continuous on it ie

$$\exists \delta > 0 \text{ such that } |x - x'| < \delta \implies |f(x) - f(x')| < \frac{\epsilon}{2} \text{ on } [-T, T]$$

Let $\tilde{f}(x) = \sum_{j=1}^{k} f(x_j) \mathbb{1}_{x \in I_j}$, where $x_j \in I_j$. Also, note length$(I_j) < \delta$. Then,

$$\begin{aligned}
\mathbb{E}(|f - \tilde{f}|) &= \int_{\mathbb{R}} |f - \tilde{f}| = \int_{\mathbb{R}\setminus[-T,T]} |f - \tilde{f}| + \int_{[-T,T]} |f - \tilde{f}| \\
&= \int_{|t|>T} |f(x)| d\mathbb{P}_{\mathcal{X}}(x) + \int_{[-T,T]} |f(x) - \tilde{f}(x)| d\mathbb{P}_{\mathcal{X}}(x) \\
&< \frac{\epsilon}{2} + \int_{\bigsqcup_{j=1}^{k} I_j} |f(x) - \tilde{f}(x)| d\mathbb{P}_{\mathcal{X}}(x) < \frac{\epsilon}{2} + \sum_{j=1}^{k} \int_{I_j} |f(x) - \tilde{f}(x)| d\mathbb{P}_{\mathcal{X}}(x) \\
&= \frac{\epsilon}{2} + \sum_{j=1}^{k} \int_{I_j} |f(x) - \tilde{f}(x)| d\mathbb{P}_{\mathcal{X}}(x) \leq \frac{\epsilon}{2} + \sum_{j=1}^{k} \int_{I_j} \frac{\epsilon}{2} d\mathbb{P}_{\mathcal{X}}(x) = \frac{\epsilon}{2} + \frac{\epsilon}{2} \sum_{j=1}^{k} \int_{I_j} d\mathbb{P}_{\mathcal{X}}(x) = \frac{\epsilon}{2} + \frac{\epsilon}{2} \mathbb{P}\big([-T, T]\big) < \epsilon.
\end{aligned}$$

Now, to guarantee construction of $\tilde{f}$ let $S = (x_1, ..., x_l) \backsim \text{iid } \mathbb{P}_{\mathcal{X}^l}$ to show $\mathcal{P}\big(\text{ empty bin } I_j\big) \xrightarrow{l \to \infty} 0$.

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}^l}\big(\cup_{j=1}^{k} \{I_j \cap S\} = \emptyset\big) &\leq \sum_{j=1}^{k} \mathbb{P}_{\mathcal{X}^l}\big(S \cap I_j = \emptyset\big) \\
&= \sum_{j=1}^{k} \mathbb{P}_{\mathcal{X}^l}\big(x_1 \notin I_j, x_2 \notin I_j, , ..., x_l \notin I_j\big) = \sum_{j=1}^{k} \prod_{i=1}^{l} \big(1 - \mathbb{P}_{\mathcal{X}^i}(I_j)\big) \\
&\leq k \max_{i=1...k} \big(1 - \mathcal{P}_{\mathcal{X}}(I_i)\big)^l \xrightarrow{l \to \infty} 0.
\end{aligned}$$

12.2 Introduction to Optimization

Our goal is to find

$$y^*_{\mathcal{H}} \in \arg\min_{\tilde{y} \in \mathcal{H}} \mathbb{E}(l_{\tilde{y}}), \text{ where} \tag{62}$$

$$l_{(\cdot)} : \mathcal{H} \to \mathbb{R}^{\mathcal{X} \times \mathcal{Y}}$$

$$\tilde{y} \to l_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$

We want $\mathbb{E}(l_{\hat{y}_S})$ from (62) to be small. To computationally solve this problem we need a way to search in $\mathcal{H}$. Thus, we need to make/choose space $\mathcal{H}$ in which we may do calculus.

12.3 Empirical Risk Minimization

$$\hat{y}_{(\cdot)} : \bigcup_{m \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$
$$S \to \hat{y}_S$$

Let us define $\hat{y}_S \in \arg\min_{\tilde{y} \in \mathcal{H}} e_S(\tilde{y})$, where $e_s(\cdot) : \mathcal{H} \to \mathbb{R}$ is defined by as:

$$e_s(\tilde{y}) := \frac{1}{|S|} \sum_{(x,y) \in S} l_{\tilde{y}}(x, y),$$
$$S = \big((x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\big)$$

The question can be phrased as how to solve $\arg\min_{\tilde{y} \in \mathcal{H}} e_S(\tilde{y})$. For us to really understand this question, we need to convert the problem space $\mathcal{H}$ into a finite-dimensional space. Then, the problem becomes a finite-dimensional optimization problem.

Example 12.2. Consider $\mathcal{H} = \left\{\sum_{j=0}^{n} a_j x^j : a_j \epsilon \mathbb{R}\right\}$ ie polynomials of degree $\leq n$.

Then,

$$\mathcal{H} \cong \mathbb{R}^{k+1}$$
$$p := \sum_{j=0}^{k} a_j x^j \to (a_0, a_1, ..., a_k) \in \mathbb{R}^{k+1}$$

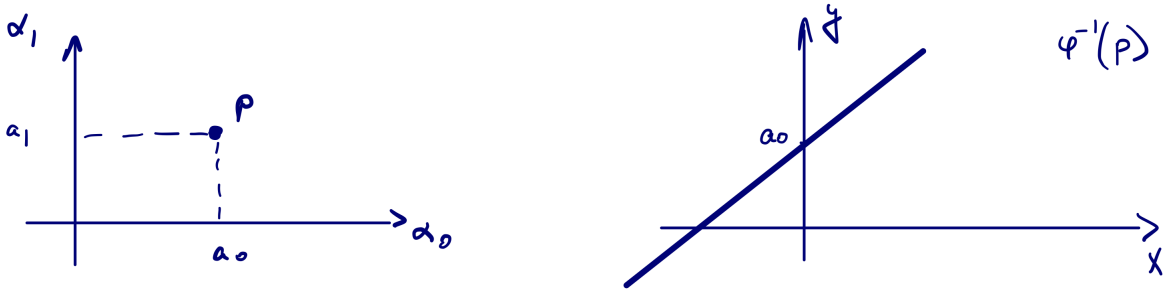

Figure 25: Point p defining polynomial $p \in \mathcal{H}$

And thus,

$$e_S : \mathcal{H} \cong \mathbb{R}^{k+1} \to \mathbb{R}$$
$$p \to e_S(p)$$

is simply a multivariable function.

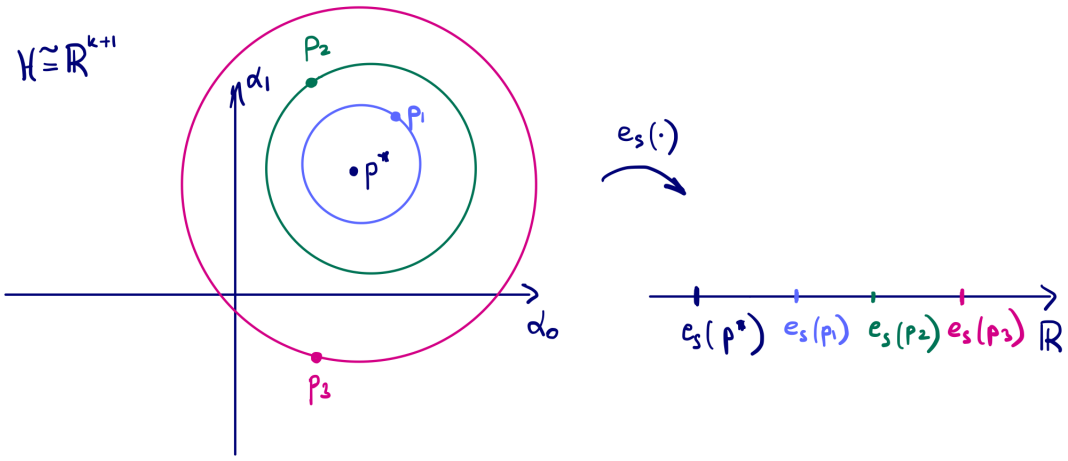

Figure 26: Level sets and mapping to empirical error

corresponds directly to residuals in $\mathcal{X} \times \mathcal{Y}$ space

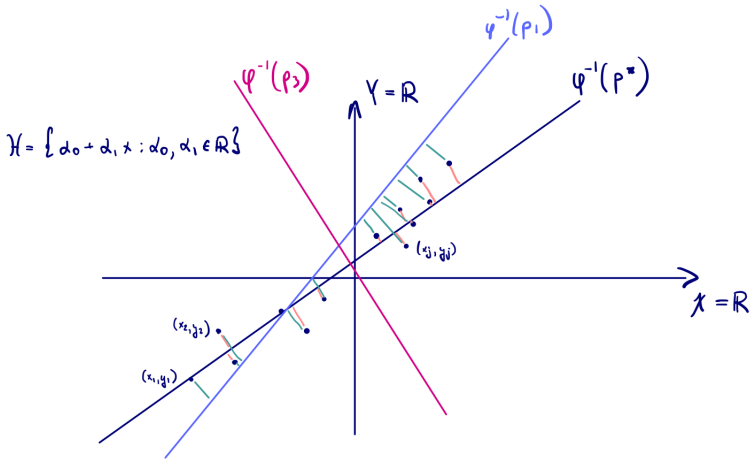

Figure 27: Function space

$$\mathbb{R} \xrightarrow{\gamma} \mathcal{H} \xrightarrow{e_S} \mathbb{R}$$
$$e_S \cdot \gamma : \mathbb{R} \to \mathbb{R}$$

So, composition is a single variable function. Thus, our goal is to construct a curve $\gamma : \mathbb{R} \to \mathcal{H}$:

$$\lim_{t\to\infty} \gamma(t) = p^*, \text{ where}$$

$$p^* \in \arg\min_{p \in \mathbb{R}^k} e_S(p)$$

Now, suppose $\gamma(t^*) = p^*$ for some $t^* < \infty$.

Recall: $f : \mathbb{R} \to \mathbb{R}$ differential and x^* is a local min / max $\Rightarrow f'(x^*) = 0$.

Then,

$$\frac{d}{dt}|_{t=t^*} = e_S \circ \gamma(t) = 0$$

and

$$\frac{d}{dt} e_S \circ \gamma = \nabla e_S \cdot \dot{\gamma}, \text{ where}$$

$\dot{\gamma}$ is the velocity tangent vector, and ∇e_S is the gradient orthogonal to level sets.

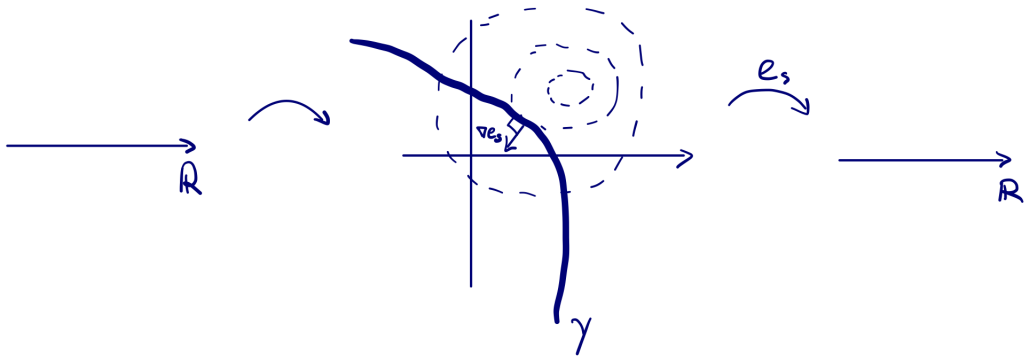

Figure 28: Map γ and the empirical error $e_s(\cdot)$

Now, we want to identify conditions for which the derivative is equal to 0. There are 3 possible conditions:

1. $\nabla e_s = 0$
2. $\dot{\gamma} = 0$
3. $\dot{\gamma} \perp \nabla e_s$

Condition 1. is good since we would like to find p where the empirical error of p is minimized.

Condition 2. is bad, because $\dot{\gamma} = 0$ essentially means we stop moving or searching for the optimal choice. Therefore, we are going to ignore this condition

Condition 3. is tricky. Recall our goal is to arrive at the optimal 'location' where the derivative of $\nabla e_S \cdot \dot{\gamma} = 0$. However, when $\dot{\gamma} \perp \nabla e_s$ we are not necessarily at the optimal p even if the derivative is equal to 0 (because $\nabla e_s \neq 0$). The information that Condition 3. provides is the direction of the 'right' path namely the direction of ∇e_s when $\dot{\gamma} \perp \nabla e_s$ (28).

13 Lecture 13

In this session, we delve into convexity providing contextual information alongside illustrative results.

13.1 Convexity: three scenarios

Convexity can be found in three different contexts: sets, functions, and optimization problems. To begin our exploration, let's focus on the first of them: sets.

Definition 13.1. A set $S \subseteq \mathbb{R}^d$ is convex if for any two points $x, y \in \mathbb{R}^d$, $\alpha x + (1-\alpha)y \in S$ for $\alpha \in (0,1)$.

Basically, we're saying that for any two points inside the set, the segment between the two points fall entirely inside the set. Examples of convex sets includes: (1) planes, (2) polyhedros (aka intersection of half-spaces), (3) spheres, etc.

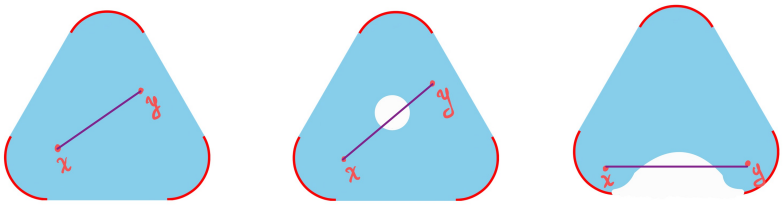

Figure 29: Left: Illustration of a convex set. Right: Segments for which the definition of convexity is not met.

Definition 13.2. For a function $f : \mathbb{R}^d \to \mathbb{R}$ we define:

$$\text{graph}(f) := \{(x, r) : f(x) = r\}$$

$$\text{epigraph}(f) := \{(x, r) : f(x) \leq r\}$$

We say the function f is convex if its epigraph (i.e. the region above the graph of the function) is a convex set.

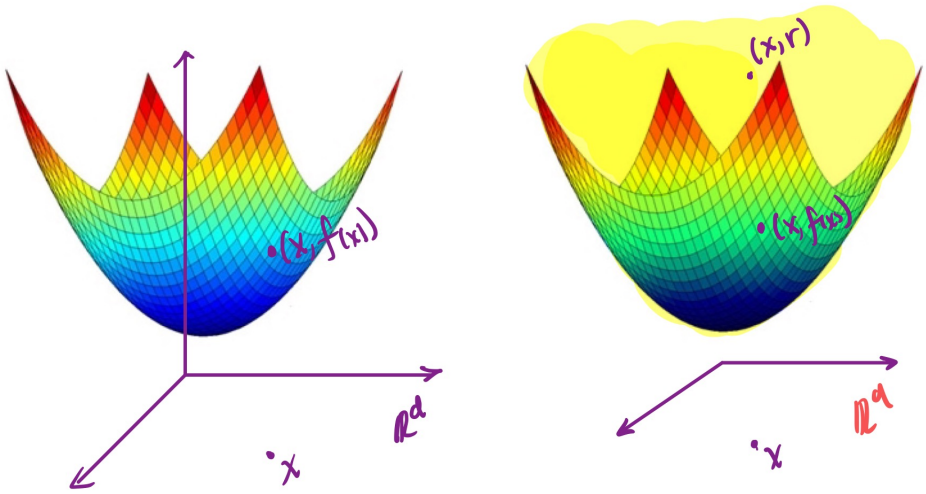

Figure 30: Graph (left) and epigraph (right) of a convex function.

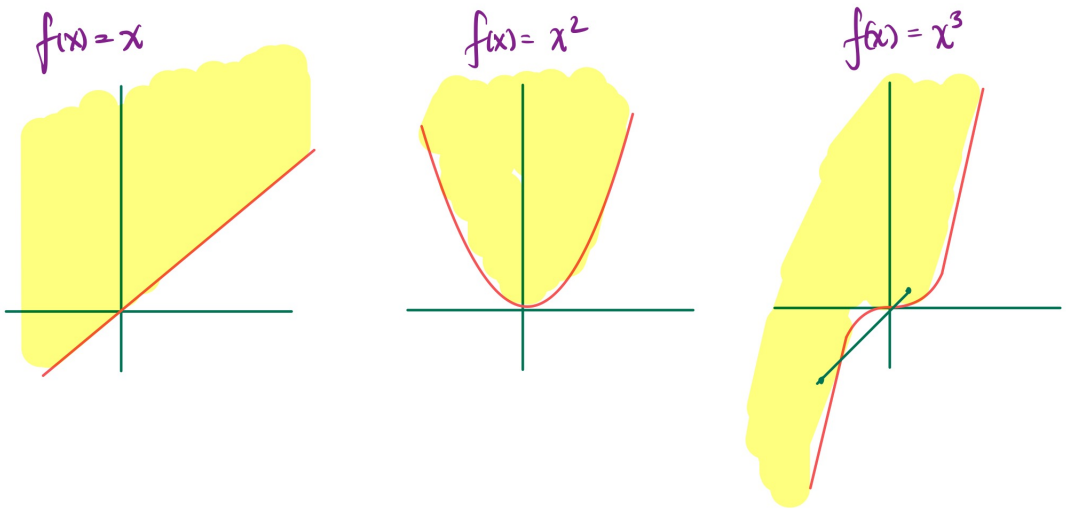

Figure 31: "Left: Convex epigraphs depicted in two figures. Right: Segment for which the definition of convexity is not satisfied.

Definition 13.3. An optimization problem

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

with constraints $h_i(x) = 0$, $i = 1, \ldots$, $g_j(x) \leq 0$, $j = 1, \ldots$, is said to be convex if the objective function $f(x)$ is convex, the functions g_j are convex and h_i are linear.

Although we mention here the definition of a convex optimization problem, it won't have later use through this lecture.

Principle of Duality Duality is the ability to view a mathematical concept from two different perspectives: a primal one, and a dual one.

Let's see what duality means in the first two context where convexity appears, sets and functions.

Dual definition of sets The definition of convexity for sets, that we have is internal. We are given two points inside the set and the requirement of the segment between the points is inside the set too. The dual (external) definition, is that instead of taking a point inside the set, we take an hyperplane that supports the set, that means that is tangential in such a way that the set falls entirely in the positive half space defined by the hyperplane.

Definition 13.4. A set is said to be convex if, when you consider the intersection of positive half spaces determined by all the supportive hyperplanes, you recover the set itself.

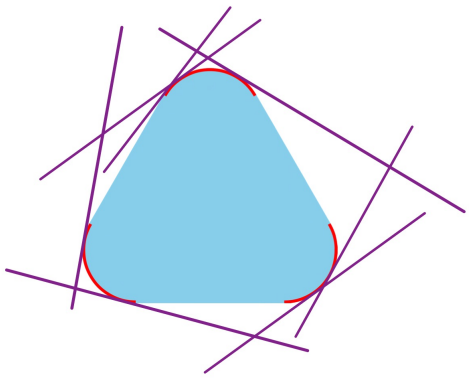

Figure 32: Convex set depicted as the intersection of potentially infinite half-spaces (inequalities).

According with this dual definition, convex sets are defined possibly by infinitely many linear inequalities (a polyhedro with possibly "infinitely" many faces).

Dual definition of functions Convexity allows to extrapolate local behaviour of a function (null gradient -local- equivalent to global minima). Local behaviour of a function is described via Taylor approximation for which particularly in the linear case, tells you that for every point in the graph of the function (assuming differentiability) its neighbors have a good approximation with respect to the tangent space associated to the tangent point (tangent hyperplane). This can be described by the formula

$$f(y) \approx f(x) + \nabla f(x)(y - x), \quad \forall x, y \in \mathrm{dom}(f) \subseteq \mathbb{R}^d.$$

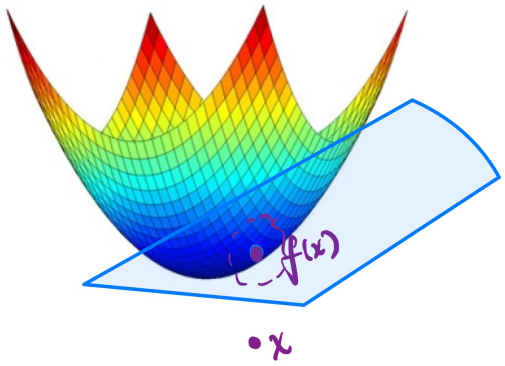

Figure 33: Tangent hyperplane at the point $(x, f(x))$.

Definition 13.5. A (differentiable) function is convex, if its graph is always above its associated tangent hyperplane, i.e.

$$f(y) \leq f(x) + \nabla f(x)(y - x), \quad \forall x, y \in \mathrm{dom}(f) \subseteq \mathbb{R}^d. \tag{63}$$

Consequences of this definition are huge. Let's assume that you find the point $x^* \in \mathrm{dom}(f)$ for which $\nabla f(x^*) = 0$. The hyperplane associated to that point is horizontal, and because convex functions will always be above tangent hyperplanes, this means that the point x is actually a global minimizer. We see that by replacing $\nabla f(x^*) = 0$ on equation (63) to get $f(y) \leq f(x^*) \forall y \in \mathrm{dom}(f) \subseteq \mathbb{R}^d$.

13.2 Gradient: geometric interpretation

In this section, we are going to review about what the gradient is. We start with a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, for which we all know that partial derivatives with respect to a point $x \in \mathbb{R}^d$ can be computed, so after putting them together in a vector, we call to that the gradient of f in x. This is

$$\nabla f(x) = (\partial_{x_1} f(x), \ldots, \partial_{x_d} f(x)) \in \mathbb{R}^d, \; x \in \mathbb{R}^d.$$

Let's work out what's the geometric content of this vector. Let's start with a curve $c : \mathbb{R} \to \mathbb{R}^d$, $c(t) := (c_1(t), \ldots, c_d(t))$ continuously differentiable. By the chain rule,

$$\frac{d}{dt} f(c(t)) = \partial_{x_1} f(c(t)) c_1'(t) + \cdots + \partial_{x_d} f(c(t)) c_d'(t) = \langle \nabla f(c(t)), \dot{c}(t) \rangle$$

where $\dot{c}(t)$ means the velocity vector of the particle at a time t.

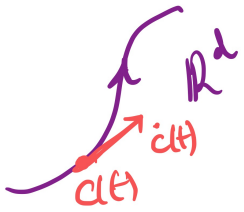

Figure 34: Curve in configuration space $\mathbb{R}^d$ along with its velocity vector..

This perspective of the gradient allows us for a familiar interpretation of the gradient, in the sense that as a vector, points to the direction of the steepest ascend of the function. To see this, let's consider level sets $\{f = a\}$ of f. Given $x \in \mathbb{R}^d$ and v in the $d-1$-dimensional tangent space of $\{f = a\}$ at x. Let's define the curve $c : \mathbb{R}^d \to \{f = a\}$ such that

$$c(0) = x, \quad \dot{c}(0) = v$$

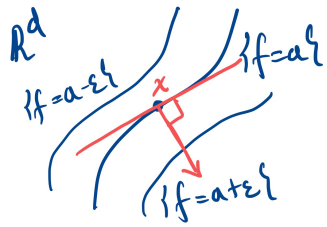

Figure 35: The gradient visualized as an orthogonal vector to the level sets, indicating the ascending direction.

Then,

$$0 = \frac{d}{dt}a = \frac{d}{dt}f(c(0)) = \langle \nabla f(c(0)), \dot{c}(0)\rangle = \langle \nabla f(x), v\rangle$$

i.e.

$$\nabla f(x) \perp v$$

so the gradient is always orthogonal to the level sets.

13.3 Some equivalences

Definition 13.6. A function $F : \mathbb{R}^d \to \mathbb{R}^d$ is monotone if $\langle F(x) - F(y), x - y\rangle \geq 0$.

Proposition 13.1. Let $f : \mathbb{R}^d \to R$ differentiable. Then, f is convex if and only if $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is monotone.

Proof. Assume f is convex. Then, by the dual definition of convex functions, for every $x, y \in \mathbb{R}^d$

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

but also

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

Adding both inequalities and using bilinearity of the dot product we get

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

which is equivalent to

$$0 \leq \langle \nabla f(x) - \nabla f(y), x - y\rangle.$$

Conversely, let's define $g(t) := f(x + t(y - x)) = f(ty + (1-t)x)$ for $t \in [0, 1]$. Then

$$g'(t) = \lim_{h \to 0} \frac{f(ty + (1-t)x + h(y-x)) - f(ty + (1-t)x)}{h} = \langle \nabla f(ty + (1-t)x), y - x \rangle$$

interpreting the limit as a directional derivative of f in the direction of the vector $y - x$. Note that

$$g(0) = f(x), \;\; g(1) = f(y), \;\; g'(0) = \langle \nabla f(x), y - x \rangle.$$

By the Fundamental Theorem of Calculus

$$f(y) = g(1) = g(0) + \int_0^1 g'(t)dt = g(0) + g'(0) + \int_0^1 \left(g'(t) - g'(0) \right) dt$$

where the integrand satisfies

$$g'(t) - g'(0) = \langle \nabla f(ty + (1-t)x) - \nabla f(x), y - x \rangle = \frac{1}{t} \langle \nabla f(x + t(y-x)) - \nabla f(x), x + t(y-x) - x \rangle \geq 0$$

by monotonicity of the gradient function. Then

$$f(y) \geq g(0) + g'(x) = f(x) + \langle \nabla f(x), y - x \rangle.$$

Therefore, f is convex. □

Now, let's make some remarks on equivalences for convex functions. When f is not differentiable, we have the following equivalence:

Proposition 13.2. The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, if and only if for every $x, y \in \mathbb{R}^d$ the inequality

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

is satisfied.

Proof. Exercise. Immediate via the definition of epigraph. □

When the function has more regularity than just differentiability, there is an additional equivalence:

Proposition 13.3. A twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex, if and only if the Hessian $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^d$.

Proof. Exercise. □

14 Lecture 14

This chapter delves into Gradient Descent algorithm and discusses the general structure of the ML pipeline, provide examples and some numerical considerations.

14.1 Gradient Descent

Let's recall from Figure 27 and Figure 28 that our goal is to construct $\gamma : \mathbb{R} \to \mathcal{H}$ such that σ "finds" p^*, where

$$\mathbb{R} \xrightarrow{\gamma} \mathcal{H} \xrightarrow{e_S} \mathbb{R}, \qquad e_S \cdot \gamma : \mathbb{R} \to \mathbb{R}$$

Taking *time*-derivative and equating to zero, we get to the relation

$$\dot{\gamma} \cdot \nabla e_S = 0$$

Let's set $\|\dot{\gamma}\| = 1$. Then

$$0 = \dot{\gamma} \cdot \nabla e_S = \|\dot{\gamma}\| \|\nabla e_S\| \cos\theta = \|\nabla e_S\| \cos\theta$$

Thus, $\theta = \pi$ induces the direction of steepest descent. This suggests the algorithm:

Algorithm 1 Gradient Descent

Input: Initial guess $\alpha_0 \in \mathbb{R}^k$
Output: final α
  Choose fixed $\epsilon > 0$ (the desired maximum error)
  Pick some $\eta > 0$ (the learning rate) which is ~ 0, and also can be dynamic
  set $k = 0$
  Compute $\nabla_\alpha e_s(\tilde{y}_\alpha)$
  while $\|\nabla_\alpha e_s(\tilde{y}_\alpha)\| \geq \epsilon$ **do**
    Set $\alpha_{k+1} \leftarrow \alpha_k - \eta \nabla_\alpha e_s(\tilde{y}_\alpha)$
    Set $k \leftarrow k + 1$
    Compute $\nabla_\alpha e_s(\tilde{y}_\alpha)$
  end while
  return $\alpha \leftarrow \alpha_k$

Note that, a priori there's no reason to assume that the algorithm halt. In such a case, we can have an alternative loop, for which we define a specific number of steps to update the gradient. Note also that at halting, ∇e_S might be big.

However, without resorting to a convexity argument, there's a possibility that the algorithm won't converge to a minimum. To address this issue, we can consider employing adaptive learning rates, prioritize specific metrics for optimization evaluation, pick coherent (contextual) values for initialization, etc. Remember that optimization is not easy in practice.

In terms of code, the general ML pipeline can be summarized as follows:

```
class Model:
    def __init__(hyperparameters):
        set as attributes these hyperparameters
        set initial parameters
        self.p = random list of coefficients

    def train(input,labels):
        for j in range(self.num_steps):
            self.train_step(x_data,y_data)

    def train_step(x_data,y_data):
        y_pred = self.forward(x_data)
        grad = self.backward(y_data,y_pred)
        self.update_grad(grad)
```

```
def forward(x_data):
    y_pred = model(x_data)

def backward(y_data,y_pred):
    loss = compute_loss(y_data,y_pred)
    grad = compute_grad(y_pred,loss)

def upgrade_grad(grad)
    self.p = self.p - self.lr*grad
```

Note that p here, corresponds to Example 12.2.

14.2 Examples

Let's go now to some explicit calculations for gradients.

Example 14.1 (Linear Regression). Let $\mathcal{H} = \{\sum_{j=0}^{k} \alpha_j x^j : \alpha_j \in \mathbb{R}\}$ and $l_{\tilde{y}(x,y)} = (\tilde{y}(x) - y)^2$. To get $\nabla_\alpha e_s(\tilde{y})$ we start by computing $\nabla_\alpha l_{\tilde{y}_\alpha}(x,y)$ and then average over data. This is

$$\nabla_\alpha l_{\tilde{y}_\alpha}(x,y) = 2(\tilde{y}(x) - y)\nabla_\alpha \tilde{y}_\alpha(x)$$

where the last term

$$\nabla_\alpha \tilde{y}_\alpha(x) = \nabla_\alpha \left(\sum_{y=1}^{k} \alpha_j x^j \right) = \sum_{y=0}^{k} \nabla_\alpha \alpha_j x^j = \left(\partial_{\alpha_0} \sum_{y=0}^{k} \alpha_j x^j, \ldots, \partial_{\alpha_k} \sum_{y=0}^{k} \alpha_j x^j \right) = (1, x, x^2, \ldots, x^k).$$

Now, given a dataset $\mathcal{S} = ((x_1, y_1)...(x_m, y_m))$, we have that

$$\nabla_\alpha e_{\tilde{y}_s}(x,y) = \frac{1}{m} \sum_{y=1}^{m} \nabla_\alpha l_{\tilde{y}_\alpha}(x,y) = \frac{2}{m} \sum_{j=1}^{m} (\tilde{y}(x_j) - y_j)(x_j^0, x_j^1, ...x_j^n)$$

Example 14.2 ((Classification)). Let's consider the probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ where $X = \mathbb{R}$, $Y = [0,1]$, and $\mathbb{P}_Y(y = (0,1)) = 0$. We are curious if for the hypothesis class and the loss function

$$\mathcal{H}_n = \{\frac{1}{1 + e^{-wx+b}} : w, b \in \mathbb{R}\}, \quad l'_{\tilde{y}}(x,y) = -\log(\tilde{y}(x))^y (1 - \tilde{y})^(1 - y)$$

whether one would favor l or l'. Is it l' convex in the parameters? Recall $y^*(x) = \mathbb{P}_{y|x}(y = 1|x)$ is the optimal classifier. What are the gradients previously computed but for l'?

Scaling. Consider the scenario where our objective is to explore models approximating e^x in the following manner:

$$\sum_{j=0}^{k} \frac{1}{j!} x^j$$

These models, backed by theoretical guarantees (convergent power series), are reliable approximations. The hypothesis class is such that $\mathcal{H} \cong \mathbb{R}^{k+1}$ but the region in which the parameters live has the following representation:

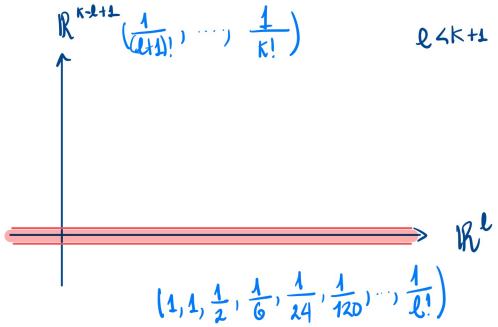

Figure 36

What might be a potential problem when we perform gradient descent over this set of parameters? How do we expect the gradient to be like? Think about sensitivity.

References

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," Mathematics of control, signals and systems, vol. 2, no. 4, pp. 303–314, 1989. 42

[2] T. Chen, H. Chen, and R.-w. Liu, "A constructive proof and an extension of cybenko's approximation theorem," in Computing Science and Statistics, pp. 163–168, Springer New York, 1992. 42

[3] M. A. Nielsen, Neural Networks and Deep Learning. 2018. 42

[4] L. Tu, An Introduction to Manifolds. Springer, 2011. 44

[5] J. Lee, Introduction to Smooth Manifolds. Springer, 2003. 44