Math 4460 Course Notes

October 14, 2024

Contents

1 Lecture 1

1.1 Administrativa

1.1.1 Canvas

Assignments, starter code, and data will be housed in canvas. Please check weekly for newly posted assignments.

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

$$\begin{tikzcd}
\mathcal{X} \times \mathcal{Y} \arrow[r, "\pi_{\mathcal{Y}}"] \arrow[d, "\pi_{\mathcal{X}}"] & \mathcal{Y} \\
\mathcal{X} \arrow[ur, dashed, "\tilde{y}"'] &
\end{tikzcd} \tag{1}$$

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 (finite) *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 (arbitrary) $(x, y) \in \mathcal{X} \times \mathcal{Y}$. 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 often that $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = f(x)|x) \neq 1$ for any function $f : \mathcal{X} \to \mathcal{Y}$.

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 high 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 $(x_i, y_i) \sim_{\text{iid}} \mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$ which *means* that the data "set" is a *point* in the (higher dimensional) 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 $\displaystyle\prod_{j=1}^{m} \mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$. 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 $\tilde{y}(x) = \displaystyle\sum_{j=0}^{n} a_j x^j$ may have arbitrarily large degree n. Or a fully connected neural network with many nodes and many layers. And so on. More generally, the (full) 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 Assignment

Programming Assignment You may find the first programming assignment under pa0 on Canvas,[1] along with starter code and data. I suggest you follow the tutorial at Real Python real python[2] 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 a model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ as in diagram (1), and metrics that illustrate success.

1.4 Probability

1.4.1 Context

Recall the Standard Diagram

$$\begin{tikzcd}
\mathcal{X} \times \mathcal{Y} \arrow[r, "\pi_{\mathcal{Y}}"] \arrow[d, "\pi_{\mathcal{X}}"] & \mathcal{Y} \\
\mathcal{X} \arrow[ur, dashed, "\tilde{y}"'] &
\end{tikzcd} \tag{2}$$

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.

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 $\mathbb{P}_{\mathcal{X}} = \displaystyle\int d\mathbb{P}_{\mathcal{X}}(x)$, and define expectation $\mathbb{E}(f)$ of a random variable $f : \mathcal{X} \to \mathbb{R}$ as a *Lebesgue* integral

[1] I index pythonically.

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

$\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$. I hesitate in bringing in the Lebesgue integral, if it weren't for that expectation of discrete random variables just *is* defined as a Lebesgue integral. We shall bite the bullet and lean on the concept in discussing arbitrary random variables as well.

We kick this lecture off by asserting that probability has nothing to do with randomness. . . at least not yet. When we return to admitting randomness into our lexicon, it will be as a *result*. For now, we forget any association between probability and chance, stochasticity, randomness, or any other (for now) anathema word affiliated with the notion of uncertainty. If you find yourself tempted to rely on the association, make an honest attempt to circle back and fight the temptation.

1.4.2 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 return to this momentarily.

Example: Area

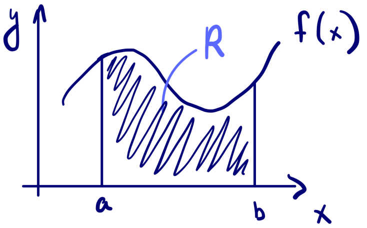


Figure 1: Area

In our notation area A of region R Is $A(R) = \int_R dA$. If you want to import calculus intuition, when we interpret dA as a differential area element, we may express it elsewise as the *product* $dA = f(x)dx$, 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 and we recover the standard form $A(R) = \int_a^b f(x)dx$.

Example: Volume

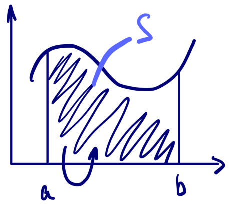


Figure 2: Volume

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 Formalism: Axiomatizing Measure

Now we formalize what we mean by 'probability measure.' The first thing to say 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)[3] 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 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.

[3] 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.

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, $\mathcal{X} = \mathcal{X} \sqcup \bigsqcup_{j=2}^{\infty} \varnothing$. Also check a very useful second fact, the Union Bound: $\mathbb{P}(\cup_{j\in\mathbb{N}} A_j) \leq \sum_{j\in\mathbb{N}} \mathbb{P}(A_j)$. You may find it worthwhile to verify that $A \subset B$ implies that $\mathbb{P}_{\mathcal{X}}(A) \leq \mathbb{P}_{\mathcal{X}}(B)$.

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 scaffolding around 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 (and fundamentally cannot) 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 such 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 also 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 $\displaystyle\int_{[a,b]} d\mathbb{P}_{\mathcal{X}}(x)$, 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 $\mathcal{X} = \mathcal{Y} = [0,1]$, and define $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}([x_1,x_2]) \times [y_1,y_2]) := (x_2 - x_1) \cdot (y_2 - y_1)$. This example is a preliminary look into independence as $\mathbb{P}_{\mathcal{X}}([x_1,x_2]) = x_2 - x_1$, $\mathbb{P}_{\mathcal{Y}}([y_1,y_2]) = y_2 - y_1$ so the measure on product space as defined is also $\mathbb{P}_{\mathcal{X}}([x_1,x_2])\mathbb{P}_{\mathcal{Y}}([y_1,y_2])$ We will return to this example when we discuss independence, which we'll will want to situate as a notion which properly at home in high dimension: independence really is a high dimensional phenomenon and 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:

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

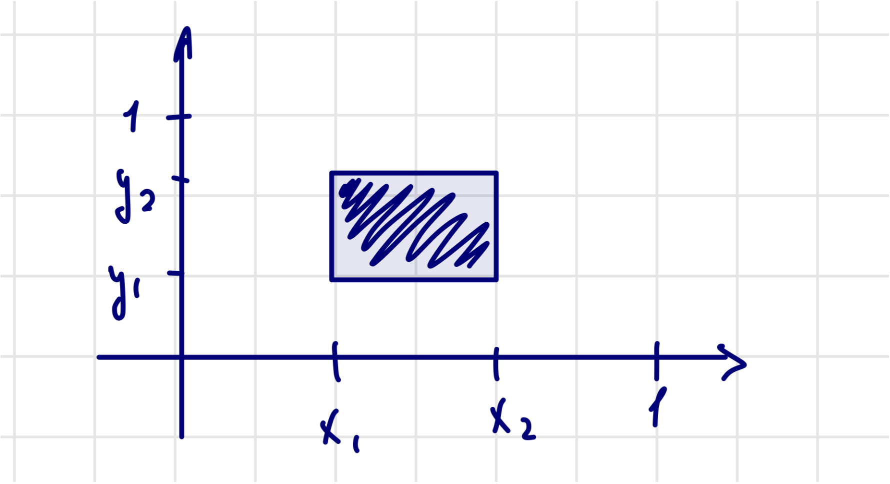


Figure 3: The area measure

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 said that probability is a theory of measure, which really means that it's a theory of integration, and integration deals with numbers. So far, we've said nothing about the nature of the set $\mathcal{X}$, and we don't need to. What we do need is a way to associate numerical 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 $\mathbb{R}$.

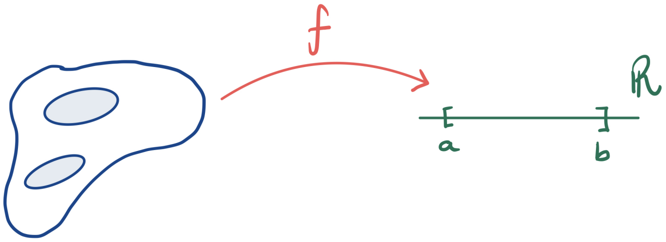


Figure 4: Random variable f

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. that it is a true (probability) measure. In essence, we require a condition known as 'measurability,' 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 and after pausing to make the quick remark will proceed to not appreciating it: the supposition that "everything is measurable" (now including functions) 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 $\mathcal{X}$, so really everything comes down to the notion of measurability of events in the first place!)

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

Remark 2.3. While we have in definition 2.1 defined a random variable to be a *function*, we will often (also) call the codomain $(\mathbb{R}, \mathbb{P}_{\mathbb{R}})$ with its induced measure a 'random variable,' without concern for the domain which produced it.

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*. We will do this in steps, 1. first for discrete random variable (taking finitely many values) and 2. by extension to arbitrary random variables.

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

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

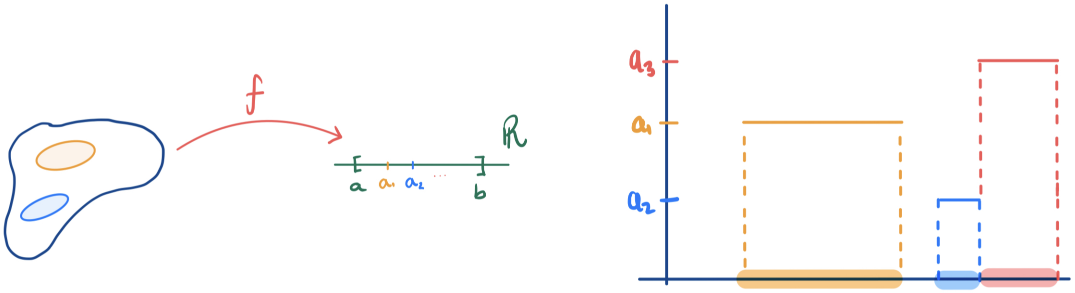


Figure 5: Simple random variable f

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. Before doing so, observe 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 2.4. 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:

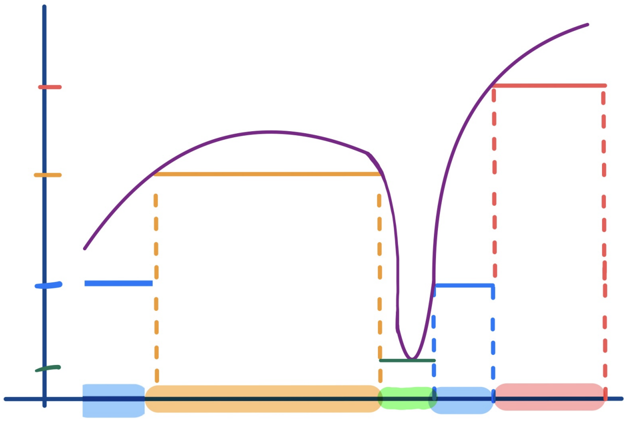


Figure 6: Approximation via simple functions

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.

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 2.5. Observe also that the notation $\mathbb{E}(f)$ 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

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

For more on the Lebesgue integral, consider notes by Ikhlas Adi, 2017.

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 $\int_{-\infty}^{\infty} f(x)\rho(x)dx$.

You can even treat the combo $\rho(x)dx$ *as* $d\mathbb{P}_{\mathcal{X}}(x)$.

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 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 $f : \mathcal{X} \to \mathbb{R}$ be a random variable 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 2.6. 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* $f : (\mathcal{X}, \mathbb{P}_{\mathcal{X}}) \to (\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$ *of probability spaces* 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 $B \subseteq \mathcal{X}$.

With expectation (integration) in hand, we now consider various kinds of decompositions of measure.

2.3 Joint Measures

A joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X} \times \mathcal{Y}})$ is a probability space whose base set is a product of (other) sets $\mathcal{X}, \mathcal{Y}$. In principle, you could imagine that each individually has its own measure $\mathbb{P}_{\mathcal{X}}$, $\mathbb{P}_{\mathcal{Y}}$, but do not need to suppose so apriori. Still: we obtain separate measure by marginalization (section 2.4); indeed, there are some relations among the joint measure $\mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$ and $\mathbb{P}_{\mathcal{X}}$, $\mathbb{P}_{\mathcal{Y}}$. Before diving into what they are, keep in mind that a joint measure $\mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$ is *not*, in general, constructed from measures $\mathbb{P}_{\mathcal{X}}$, $\mathbb{P}_{\mathcal{Y}}$. Therefore, you should think of the joint measure (simply and primarily) as a measure on the joint space; any connection to the marginal spaces is secondary.

2.4 Marginalization

We now turn to induced probability measures on the standard diagram

$$\begin{tikzcd}
\mathcal{X} \times \mathcal{Y} \arrow[r, "\pi_{\mathcal{Y}}"] \arrow[d, "\pi_{\mathcal{X}}"] & \mathcal{Y} \\
\mathcal{X} \arrow[ur, dashed, "\tilde{y}"'] &
\end{tikzcd} \tag{6}$$

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.

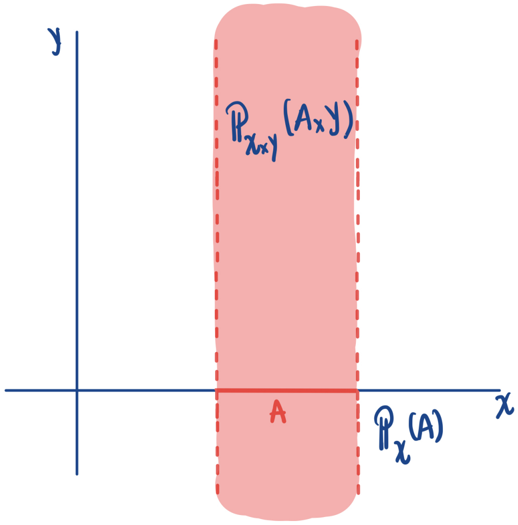

y
$\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(A\times\mathcal{Y})$
A
$\mathbb{P}_{\mathcal{X}}(A)$
x

Figure 7: Geometry of Marginalization

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. To measure event $A \subset \mathcal{X}$ downstairs, you look at the tube $A \times \mathcal{Y}$ upstairs and measure *it* (fig. 7).

2.5 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 $\mathbb{P}_{\mathcal{X}}(A) \neq 0$.

The proviso in this definition raises a question of what conditional probability means when $\mathbb{P}_{\mathcal{X}}(A) = 0$. We define it implicitly, with notation $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ that you will eventually become comfortable manipulating mechanically:

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

For random variable $f : \mathcal{X} \times \mathcal{Y}$ we then have

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

Starting with (9), you should read this as: the inner integral $\int_{B|A} d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ defines a random variable *on* $\mathcal{X}$, of which we may evaluate expectation with respect to measure $\mathbb{P}_{\mathcal{X}}$. What is the random variable? Whatever it is, it makes the equality (9)! At the level of formalism, this implicit definition may not seem very illuminating, but momentarily we will interpret it geometrically in a manner reminiscent of concepts from multivariable calculus which you *are* already familiar with.

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 2.7. There were many questions on interpreting the critter $A|B$. In pictures I draw the rectangle $A \times B$. It is fine to think of $A|B$ pictorially as $A \times B$, but you must 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).

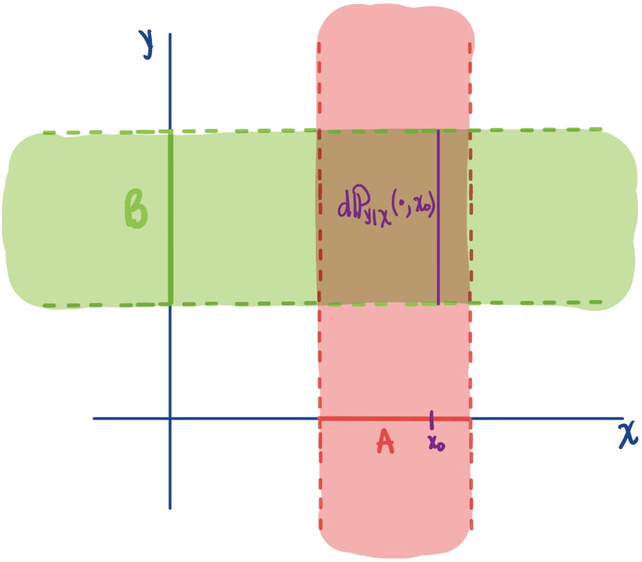


Figure 8

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.[4]

The level of rigor we need is: we may fluidly manipulate integral expressions involving joint probability, relying on our multivariable intuition from calculus that decomposes high dimensional integration as sequence of one-dimensional integrals. What changes with probability, is that the (measure used for the) inner one-dimensional integral may depend on the outer variable.

Now for the geometric intuition: in multivariable calculus, to integrate $\int_R f$ a function $f(x,y)$ over some region $R \subset \mathcal{X} \times \mathcal{Y}$ (imagine e.g. that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, so this function $f : \mathbb{R}^2 \to \mathbb{R}$), we decompose

[4] See e.g. the Radon-Nikodym Theorem.

the high-dimensional integral into two *one*-dimensional integrals $\int_{\mathcal{X}}\int_{\mathcal{Y}} f(x,y)dy\,dx$, i.e. we integrate the function $f(x,y)$ with respect to y, holding x fixed, then integrate the what is left with respect to x. That means we slice $\mathbb{R}^2$ at $\{x\} \times \mathbb{R}$, integrate in $\mathbb{R}$ (y's copy of $\mathbb{R}$), which returns a *value*, then integrate those values again with respect to $\mathbb{R}$ (x's copy of $\mathbb{R}$).

2.6 Independence

Forget, for 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 one's first typical encounter with 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, etc. Use your geometric intuition liberally!

Definition 2.8. In general, we say that $(\mathcal{X}^m, \mathbb{P}_{\mathcal{X}^m})$ 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

$$\mathbb{P}_{\prod \mathcal{X}_j} = \prod_{j=1}^m \mathbb{P}_{\mathcal{X}_j}. \tag{15}$$

Precisely speaking the measure on the left measures events in $\mathcal{X} := \prod \mathcal{X}_j$ while each component measure on the right only measures events in $\mathcal{X}_j$. The solution is: A rectangle $A = \prod A_j$ in $\mathcal{X}$ is a product of individual events $A_j \subset \mathcal{X}_j$, which each measure $\mathbb{P}_{\mathcal{X}_j}$ can measure.

Remark 2.8. Be careful with this definition: from spaces $(\mathcal{X}_j, \mathbb{P}_{\mathcal{X}_j})$, the independent measure $\mathbb{P}_{\prod \mathcal{X}_j}$ in (15) is not the only possible measure on joint space $\prod_j \mathcal{X}_j$!

2.7 Data

We will discuss the connection between data and probability more later, but for now assert an easy way to interpret marginalization and conditioning in terms of data. Imagine a spreadsheet with n columns. Data in the first column has three unique values, 'a,' 'b,' and 'c.' Marginalization, in terms of data, simply corresponds to deleting columns, e.g. to obtain the marginal for column 1, we delete (or ignore) columns 2-n. Conditioning, by contrast, corresponds to deleting rows. Conditioned on column 1 data equaling (say) 'a' amounts to looking only at those rows (all columns 2-n included) at which the column 1 entry is 'a.'

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, continuing geometric interpretation for $A \times B$, $A|B$ and $B|A$ as "sets" in $\mathcal{X} \times \mathcal{Y}$ (remark 2.7).

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

Ultimately, our purpose in going through this rigmarole is for computation. I would like you to feel natural 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 already comfortable 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. This example illuminates why we were destined to "fail" on the 0th programming assignment. We start with $(\mathcal{X} \times \mathcal{Y} = \mathbb{R}^k \times \{0,1\}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}} = \mathbb{P}_{\mathcal{X}}\mathbb{P}_{\mathcal{Y}})$ independent. Suppose that we have "optimal" model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$. We revisit through a computation what optimality means. Computing accuracy, which is the measure of event $\{\tilde{y} = y\} = \{(x,y) \in \mathcal{X} \times \mathcal{Y} : \tilde{y}(x) = y\}$, we have:

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\tilde{y} = y) &= \mathbb{E}(\mathbb{1}_{\tilde{y}(x)=y}) \\
&:= \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=y} \, d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x,y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}|\mathcal{X}} \mathbb{1}_{\tilde{y}(x)=y} \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=y} \, d\mathbb{P}_{\mathcal{Y}}(y) d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{1}_{\tilde{y}(x)=0}(1-p) + \mathbb{1}_{\tilde{y}(x)=1} p \, d\mathbb{P}_{\mathcal{X}}(x) \\
&= \int_{\mathcal{X}} \max\{p, 1-p\} d\mathbb{P}_{\mathcal{X}}(x) \\
&= \max\{p, 1-p\} \int_{\mathcal{X}} d\mathbb{P}_{\mathcal{X}}(x) \\
&= \max\{p, 1-p\}.
\end{aligned} \tag{16}$$

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, reminder, 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 we've written it such that we may 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 in pa0, you should have had something pretty close to p for almost all x. The reason is that the data was generated from an independent joint measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}} = \mathbb{P}_{\mathcal{X}}\mathbb{P}_{\mathcal{Y}}$.[5] It turns out that your model was pretty much optimal!

[5] Note that independence here is different than when we say we *sampled* data e.g. $x_1, \ldots, x_m \sim_{\text{iid}} \mathbb{P}$ *independently* (and identically distributed) from some distribution, see section 3.2.

Remark 3.1. In a machine learning context, we assume or at least hope that our joint measure on input-output is *not* independent. If it is, that renders the task of building a model *from* input to output effectively pointless: just look at the labeled data and choose uniformly what's best.

Next we interpret independence for data.

3.2 Data

Referring to the standard diagram (1), 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 $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big) \sim_{\text{iid}} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ is (sampled) *independent and identically distributed* if $S \in (\mathcal{X}\times\mathcal{Y})^m$ 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.

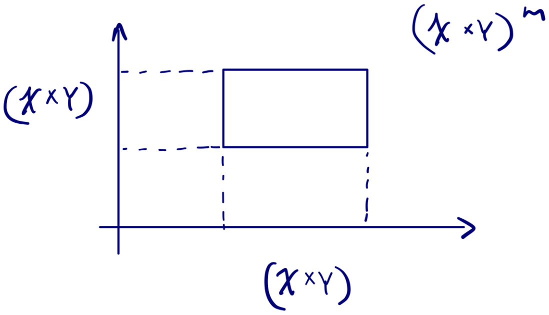


Figure 9

Remark 3.2. 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 related to your measure $\mathbb{P}_{\mathcal{X}}$. One way of handwaiving this relation 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 $\mathbb{P}_{\mathcal{X}}(A)$. In other words, we are saying in English that $\mathbb{E}(\mathbb{1}_{x\in A}) = \mathbb{P}_{\mathcal{X}}(A)$. Since we generally do not evaluate $\mathbb{P}_{\mathcal{X}}$ directly, and $x \in A$ or $x \notin A$ simply, we need a more data-centric way to talk about probability, one which ties with our understanding of "randomness," which up to this point we've barely mentioned! The connection arises (primarily) from the Law of Large Numbers. First a quick review of concentration bounds.

3.3 Concentration Inequalities

Concentration inequalities are sort of the heartbeat (for guarantees) in ML. They are omnipresent and make many important results work. We enumerate a few here and will return to them again later in the semester when we discuss *probably approximately correct* (PAC) learnability.

Proposition 3.1 (Markov)**.** Suppose that $(\mathcal{X} \subset \mathbb{R}^{\geq 0}, \mathbb{P}_{\mathcal{X}})$ 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 equality by linearity of integration. Since $x \geq 0$, the second term bounds

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

which follows from the fact that $x \geq t$ on $x \in [t, \infty)$. The first integral $\int_{[0,t)} x d\mathbb{P}_{\mathcal{X}}(x) \geq 0$, so we have $\mathbb{E}(x) \geq \int_{[t,\infty)} t d\mathbb{P}_{\mathcal{X}}(x) = t \cdot \int_{[t,\infty)} d\mathbb{P}_{\mathcal{X}}(x)$. Interpreting the integral as probability and solving proves the result. □

Proposition 3.2 (Chebyshev)**.** Let $(\mathcal{X} \subset \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ be a probability space with mean $\mu_{\mathcal{X}}$ and 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. Apply proposition 3.1 to $\{|x - \mu_{\mathcal{X}}| > t\} = \{(x - \mu_{\mathcal{X}})^2 > t^2\}$ □

The next two are more advanced; we introduce them here so that when we see them later in the semester it will be for a second time.

Proposition 3.3 (Hoeffding)**.** Let $(\mathcal{X} \subset [0, 1], \mathbb{P}_{\mathcal{X}})$ be a probability space with mean $\mathbb{E}(x) = \mu_{\mathcal{X}}$. Then

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|\frac{1}{m}\sum_{j=1}^{m} x_j - \mu_{\mathcal{X}}\right| > \epsilon\right) \leq 2e^{-2m\epsilon^2}. \tag{17}$$

The Law of Large Numbers (LLN), which we recall in section 3.3.1, provides a guarantee that sample mean is close to expectation, in the limit. There are various ways of cashing this out (Strong Law, Weak Law, etc.), but Hoeffding is not an asymptotic result: eq. (17) holds *for all* $m \in \mathbb{N}$. Of course, for cases where $\mathcal{X} \subset [0, 1]$ (or more generally, is bounded) Hoeffding readily proves LLN by taking the limit $\mu \to \infty$.

Finally, we present Glivenko-Cantelli, a result which guarantees that an empirical cumulative distribution function (ECDF) is close to the CDF.

Theorem 3.1 (Glivenko-Cantelli)**.** Let $(\mathcal{X} = \mathbb{R}, \mathbb{P}_{\mathcal{X}})$ bea 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

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

The statement of this theorem is phenomenal: no matter what $\mathbb{P}_{\mathcal{X}}$ is, you can articulate precise conditions for satisfaction of a precision specification (ε) with arbitrarily high confidence $(1 - \delta)$ provided m is "sufficiently large." And again... entirely independent of $\mathbb{P}_{\mathcal{X}}$. Sometimes, one must pause to marvel at the beauty of mathematics.

There is a stronger form of this result which replaces the right hand side of (18) with the right hand side of (17).[6]

[6] See the DKW(M) inequality https://en.wikipedia.org/wiki/Dvoretzky–Kiefer–Wolfowitz_inequality. Massart (the 'M') was responsible for the tight coefficient of 2.

3.3.1 Law of Large Numbers (LLN)

The version of LLN we provide is the *Weak Law*, which is sufficient for our purposes. The Strong Law denotes a different mode of convergence ('almost surely' as opposed to the Weak Law's 'in probability').

Theorem 3.2. 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 mean $\mu_{\mathcal{X}} := \mathbb{E}(x) = \int_{\mathcal{X}} x d\mathbb{P}_{\mathcal{X}}(x)$ and variance $\sigma_{\mathcal{X}}^2 := \mathbb{E}((x - \mu_{\mathcal{X}})^2)$ are finite. Define empirical mean $s_m : \mathcal{X}^m \to \mathbb{R}$ as a random variable by $(x_1, \ldots, x_m) \mapsto \frac{1}{m}(x_1 + \ldots + x_m)$. Then for any $\varepsilon > 0$,

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

Proof. Without loss of generality suppose that $\mu_{\mathcal{X}} = 0$.[7] By Chebyshev's inequality (proposition 3.2),

$$\begin{aligned}
\mathbb{P}_{\mathcal{X}^m}(|s_m - \mu_{\mathcal{X}}| > \varepsilon) &= \mathbb{P}_{\mathcal{X}^m}\left((x_1 + \ldots + x_m)^2 > m^2\varepsilon^2\right) \\
&\le \frac{\mathbb{E}\left[\left(\sum_{i=1}^m x_i\right)\left(\sum_{j=1}^m x_j\right)\right]}{m^2\varepsilon^2} \\
&= \frac{\sum_{i=j=1}^m \mathbb{E}(x_i^2) + \sum_{i \neq j}^m \mathbb{E}(x_i)\mathbb{E}(x_j)}{m^2\varepsilon^2} \\
&= \frac{m\sigma_{\mathcal{X}}^2}{m^2\varepsilon^2} \xrightarrow{m \to \infty} 0,
\end{aligned}$$

where we use independence and mean-zero to eliminate $\mathbb{E}(x_i x_j) = 0$ for $i \neq j$. □

Remark 3.3. The statement of LLN measures the set of point $(x_1, \ldots, x_m) \in \mathcal{X}^m$ that are at least ε-far from sample mean $s_m^{-1}(\mu_{\mathcal{X}})$.

With LLN under our belt, we may finally interpret randomness into probability: to say that $(x_1, \ldots, x_n) \sim_{iid} \mathbb{P}_{\mathcal{X}}$ (at 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. That is, for any event $A \subset \mathcal{X}$, consider random variable $\mathbb{1}_{x \in A} := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else.} \end{cases}$ Then

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

As the convergence in this expression is "in probability," we first fix $\varepsilon > 0$; then for any $\delta > 0$, there is $M_\delta > 0$ so that

$$\mathbb{P}_{\mathcal{X}^m}\left(\left|\frac{1}{m}\sum_{j=1}^m \mathbb{1}_{x_j \in A} - \mathbb{P}_{\mathcal{X}}(A)\right| > \varepsilon\right) < \delta \text{ whenever } m > M_\delta. \tag{19}$$

There is a procedure to mechanize this result. Suppose you've set $\varepsilon, \delta > 0$ and (happen to) know M_δ satisfying (19). Sample $x_1, \ldots, x_m \sim_{iid} \mathbb{P}_{\mathcal{X}}$ with $m > M_\delta$ and evaluate the random variable $\mathbb{1}_{|s_m(x_1,\ldots,x_m) - \mathbb{P}_{\mathcal{X}}(A)| > \varepsilon}$. The answer will be zero or one. Most of the time it should be zero. If you do this many times,[8] and take the empirical average of *these* results, the answer should be less than δ.

[7] Think about why applying the transformation $x \mapsto x - \mu_{\mathcal{X}}$, inducing mean zero random variable, does not affect the LLN result.

[8] As $\mathbb{1}_{|s_m(x_1,\ldots,x_m) - \mathbb{P}_{\mathcal{X}}(A)| > \varepsilon}$ is a $[0, 1]$ bounded random variable, proposition 3.3 applies and you can in fact identify what 'many' needs to be to satisfy the subsequent 'should be.'

3.4 Monkeys on a Keyboard: 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 a random variable 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 (despite density being greatest inside)!

We run a computation to more concretely illustrate the point.

Example 3.2. Consider random variable $(\mathcal{X} = [0, 1], \mathbb{P}_{\mathcal{X}})$ 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 $1 > \mathbb{P}_{\mathcal{X}}(A) \ge 1 - \varepsilon$ has high probability. The "hypercube" $A^m \subset \mathcal{X}^m$ in high dimension has measure

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

where the first equality follows by independence. As long as $1 > \varepsilon > 0$, $\lim\limits_{m\to\infty}(1 - \varepsilon)^m = 0$.

As a practical example, imagine tasking some monkey to button mash on a keyboard in perpetuity. Assuming independence and identical distribution of key strokes, the probability of hitting any character (or punctuation) is roughly $1/30$, and therefore the probability of matching a length ℓ sequence of characters is $1/30^\ell$ (independence!). The complete works of Shakespeare have, let's estimate, roughly $4,000,000$ characters. Thus, the probability that a monkey fails to type out the complete works of Shakespeare, in one iteration of $4,000,000$ key strokes, is about $1 - (1/30)^{4000000}$ (fairly close to 1). In one go, it will likely not type out the complete works. But if it keeps trying, eventually it will.

4 Lecture 4

4.1 Setting: Detection / Binary Classification

We are ready to discuss a first instance of machine learning problem: binary classification, where $\mathcal{Y} = \{0, 1\}$. Recalling (1), the task is to construct *model* $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ so that $\tilde{y}(x) \approx y$ for $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$-most $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Alternatively, we describe this goal in functional form as: $\tilde{y} \circ \pi_{\mathcal{X}} \approx \pi_{\mathcal{Y}}$ in expectation. Pretty much all supervised learning takes this form, and all that we've done at this point is impose a constraint on codomain $\mathcal{Y}$, namely that it is a discrete set consisting of two elements.

The supervised learning problem for binary classification is typically solved in two steps:

1. construct *score* function (which here we'll denote by) $\tilde{y} : \mathcal{X} \to \mathbb{R}$ (or to closed interval $[0, 1]$), and
2. *threshold* score to obtain discrete output $\hat{y}_t : \mathcal{X} \to \mathcal{Y}$ defined by $\hat{y}_t(x) := \mathbb{1}_{\tilde{y}(x)\geq t}$.

We'll return more extensively to step 1. soon enough. For now, suppose that we have a score function $\tilde{y}$ and correspondingly, on labeled data set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, a set of scores

$$\{(\tilde{y}(x_1), y_1), \ldots, (\tilde{y}(x_m), y_m)\}.$$

4.2 Interpreting the Score

The first point to make is that we would like the scores to separate classes. Namely, if we group scores by class $S_j := \{\tilde{y}(x_k) : y_k = j\}$ for $j = 0, 1$, you should expect a histogram of each to correspondingly separate:

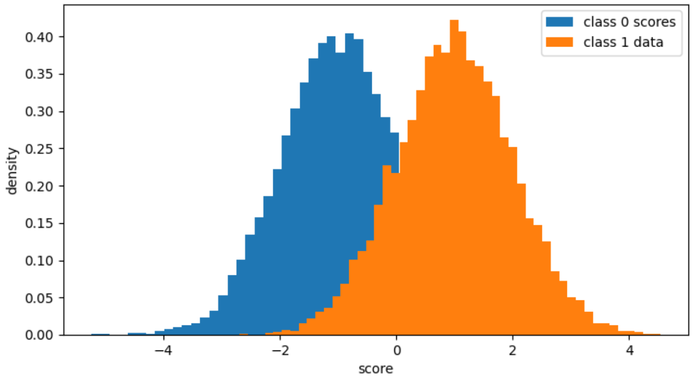


Figure 10: Conditional Histograms

Indeed, the score should serve as some sort of indicator of class! (Here we recall the asymmetry between sampling from "input" space $\mathcal{X}$ and from $\mathcal{Y}$: the score is evaluable as a function only of input.) The more these two histograms separate, in principle, the better the score. Of course, the histogram represents data we already have, and what you'd like to know is that a histogram of scores on new data looks similar. In probability speak, the normalized histograms approximate *conditional densities* f_j, for $j = 0, 1$, according to which

$$\int_{[a,b]} f_j(s)ds = \mathbb{P}_{\mathcal{X}|\mathcal{Y}}([a, b]|y = j).$$

In a perfect world, both 1. the densities would well-approximate the histograms and 2. the densities would be strongly separated. A measure of separation is given e.g. by the *Kolmogorov-Smirnov* (KS) score, namely:

$$ks := \sup_{t\in\mathbb{R}} |\mathbb{P}_{\mathcal{X}|\mathcal{Y}}(s \leq t|y = 0) - \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(s \leq t|y = 1)|. \tag{20}$$

Understand geometrically why $ks \approx 1$ corresponds to (near) total separation while $ks \approx 0$ corresponds to (near) total overlap.

Example Consider a typical use-case where a bank must decide whether to extend a loan to an applicant based on information they provide at the time of application. With $\mathcal{Y} = \{0, 1\}$, let $y = 0$ correspond to 'application denied' and $y = 1$ to 'loan approved.' Let $\mathcal{X} = \mathbb{R}^k$ and suppose that each component denotes some measurable attribute, e.g. x^0 debt to income ratio, x^1 income, x^2 age, x^3 years of employment and so on.[9][10] You can imagine that the projection $\pi_0 : \mathcal{X} \to \mathbb{R}$ returning debt-to-income (dti) would provide a fairly decent predictor of whether a loan will (should) be extended. In this case, technically our label denotes something more like an answer to the question 'will they repay their loan?'

Consider a significantly harder example, where input data in $\mathcal{X} = \mathbb{R}^{3\times k\times k}$ are images and $y = 1$ corresponds to 'Waldo is in the image.' In *this* case, π_0 returning the red value of the first pixel will unlikely be a particularly good score for distinguishing labels.

4.2.1 Optimal Threshold

Now, the score function $\tilde{y} : \mathcal{X} \to \mathbb{R}$ induces a measure $\mathbb{P}_{\mathbb{R}\times\mathcal{Y}}$ on $\mathbb{R} \times \{0, 1\}$ by $\mathbb{P}_{\mathbb{R}\times\mathcal{Y}}([a, b] \times \{j\}) := \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}\big(\tilde{y}^{-1}([a, b]) \times \{j\}\big)$. We've seen induces measures, though it's worth recalling them in this context, which is slightly more motivated to a problem at hand. While the measure in simpler space $\mathbb{R} \times \mathcal{Y}$ is *defined* in terms of the one in more complex space $\mathcal{X} \times \mathcal{Y}$, we will compute in the simpler space, in the hopes that class separation is easier to see there. If this sounds counterintuitive keep in mind that this is half the whole point of using random variables (as maps to $\mathbb{R}$).

Given such a distribution of scores, there remains the question how to predict?, i.e. how to associate output label y to given input point x with score $\tilde{y}(x)$. We feel that scores on "one side" should belong to one label, and indeed, thresholding the score provides half an answer: let label predictor $\hat{y}_t := \mathbb{1}_{\tilde{y}(x)\geq t}$ segment the space: to the left of t belongs one label, and to the right the other. Still, what should t be? To answer this question we must first identify what our objective is. One reasonable objective is to maximize accuracy (or minimize error) $\hat{y}_t(x) = y$ over data $(x, y) \in \mathcal{X} \times \mathcal{Y}$. It isn't quite right to say this should hold *for all* (x, y)—in particular, because there may be pairs (x, y), (x, y') with $y \neq y'$—on which account we merely ask for equality to hold for *most* (x, y).[11] Measuring 'most' requires a measure ($\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$), and so we ask for approximate equality for 'most in probability' or what amounts to the same, in expectation for suitably defined random variable ($\mathbb{E}(\mathbb{1}_{\hat{y}_t(x)=y})$). Maximizing expected accuracy is equivalent to minimizing expected error, and for the sake of uniformity with future convention, we'll stick to minimizing error.

We compute: for fixed $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}(\mathbb{1}_{\hat{y}_t(x)\neq y}) &= \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\hat{y}_t(x)\neq y}\, d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x, y) \\
&= \int_{\mathcal{Y}} \int_{\mathcal{X}|\mathcal{Y}} \mathbb{1}_{\hat{y}_t(x)\neq y}\, d\mathbb{P}_{\mathcal{X}|\mathcal{Y}}(x|y)\, d\mathbb{P}_{\mathcal{Y}}(y) \\
&= \sum_{y\in\mathcal{Y}} \int_{\mathcal{X}|\mathcal{Y}} \mathbb{1}_{\hat{y}_t(x)\neq y}\, d\mathbb{P}_{\mathcal{X}|\mathcal{Y}}(x|y)\mathbb{P}_{\mathcal{Y}}(y) \\
&= \mathbb{P}_{\mathcal{Y}}(y = 0) \int_t^{\infty} f_0(s)ds + \mathbb{P}_{\mathcal{Y}}(y = 1) \int_{-\infty}^{t} f_1(s)ds.
\end{aligned} \tag{21}$$

In the last line, we recall that $\mathbb{1}_{\hat{y}_t(x)\neq y} = 1$ means that $\hat{y}_t(x) \neq y$, which happens when either $\tilde{y}(x) \geq t$ and $y = 0$ *or* $\tilde{y}(x) \leq t$ and $y = 1$. It's at this step we sub in the measure $\mathbb{P}_{\mathbb{R}\times\mathcal{Y}}$, whose conditional w.r.t. y is given by integrating the corresponding density f_j; again this is one reason for working with scores: you can integrate in $\mathbb{R}$ perhaps a lot easier than you can in some wonky space $\mathcal{X}$.

Let's interpret limiting cases. When $t \to -\infty$, $\int_{\mathbb{R}} f_0 = 1$ and the error becomes $\mathbb{P}_{\mathcal{Y}}(y = 0)$. Similarly, when $t \to \infty$, we have $\int_{\mathbb{R}} f_1 = 1$ and get $\mathbb{P}_{\mathcal{Y}}(y = 1)$. (What does this mean in terms of what

[9] Superscript denotes component, not an exponent, which we use instead of subscript because that typically indexes the data sample.

[10] Note that while available data *may* in principle be usable for decisioning, there may be countervailing considerations prohibiting its use, regulation around consumer finance providing an excellent case in point.

[11] Think of concrete examples. Two applicants for a loan may have the same debt-to-income ratio and yet different outcomes of repayment. In other words, the input-output relation between input data (dti) and output label (repayment) is not functional. That might be a problem if we didn't have probability!

our model predicts?)

To find optimal t^* minimizing (21), take the derivative (w.r.t. t) and set to zero. By the Fundamental Theorem of Calculus, we obtain

$$t^* = \arg_{t\in\mathbb{R}} -\mathbb{P}_{\mathcal{Y}}(y=0)f_0(t) + \mathbb{P}_{\mathcal{Y}}(y=1)f_1(t) = 0.$$

When $\mathbb{P}(y=0) = \mathbb{P}(y=1)$ (both labels are equally likely), a candidate for the optimal point occurs when both conditional densities are equal. Considering fig. 10, does this make sense?

4.3 Constructing the Score

The procedure we outlined for finding t^* to threshold a score (and obtain hard predictions) is a microcosm of the same procedure we employ for *constructing* score $\tilde{y} : \mathcal{X} \to \mathbb{R}$. In that case, we looked for the minimum $\min_{t\in\mathbb{R}} \mathbb{E}(\mathbb{1}_{\tilde{y}_t \neq y})$ of an expectation for a (one-parameter) parameterized random variable. In this case, we similarly execute optimization $\min_{\tilde{y}:\mathcal{X}\to\mathbb{R}} \mathbb{E}(\ell_{\tilde{y}})$ for a random variable $\ell_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, called a *loss* function, parameterized (this time) by model $\tilde{y}$. Two things to note: first, we have yet to specify what the loss function is (there will be many options). Secondly, our search is no longer over $\mathbb{R}$ but instead a function space, namely the collection of maps $\mathcal{X} \to \mathbb{R}$. This space is much larger, and therefore the problem signif harder!

Still, let's take a stab at solving it. Suppose that you want score to reflect (something like) likelihood, so that higher scores (near 1) correspond to $y = 1$ and lower scores (near 0) correspond to $y = 0$. Then a perfectly fine loss function may be $\ell_{\tilde{y}}(x,y) := (\tilde{y}(x) - y)^2$.[12] To find optimal (call it) $y^* : \mathcal{X} \to \mathbb{R}$, we compute

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

and this time condition the other way to get

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

Minimizing the inner integral $\int_{\mathcal{Y}|\mathcal{X}} (\tilde{y}(x) - y)^2 d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)$ *pointwise in* x will minimize the entire expectation. Therefore, we fix x and consider the optimization problem:

$$\min_{\tilde{y}(x)\in\mathbb{R}} \int_{\mathcal{Y}|\mathcal{X}} (\tilde{y}(x) - y)^2 \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x).$$

The advantage of *this* expression is that we've once again reduced the problem to an optimization in $\mathbb{R}$ (and swept under the rug whether a map $\mathcal{X} \to \mathbb{R}$ defined by $y^*(x) = \arg\min_{v\in\mathbb{R}} \int_{\mathcal{Y}|\mathcal{X}} (v - y)^2 d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)$ is legit, i.e. measurable). Computing the conditional expectation by expanding the integrand, we obtain

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

Recall that the conditional integral integrates w.r.t. the "free variable" y, not conditioned one x; that allows e.g. pulling $\tilde{y}(x)$ outside the integral. We rinse and repeat steps from thresholding, take the derivative of the right hand side with respect to $\tilde{y}(x)$[13] and set to zero to obtain

$$2\tilde{y}(x) \underbrace{\int_{\mathcal{Y}|\mathcal{X}} d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)}_{1} - 2 \underbrace{\int_{\mathcal{Y}|\mathcal{X}} y \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)}_{\mathbb{E}(y|x)} = 0,$$

which implies that $y^*(x) = \mathbb{E}(y|x)$. For this chosen loss function, the task is totally solved, and if such were the case we'd basically be done with ML, we could go home and move on to bigger and harder problems. There's one catch: we typically do not know the measure $\mathbb{P}$ (either $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ or $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$)!

[12] Notice when $\tilde{y}(x) \in \{0,1\}$ this loss function is identical to the one we used for thresholding.

[13] With x fixed, we're looking for the value $v \in \mathbb{R}$ to which we'll assign $\tilde{y}(x)$. So if you prefer, expand $\int_{\mathcal{Y}|\mathcal{X}} (v - y)^2 d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)$, take derivative w.r.t. v, and set to zero.

4.4 Metrics

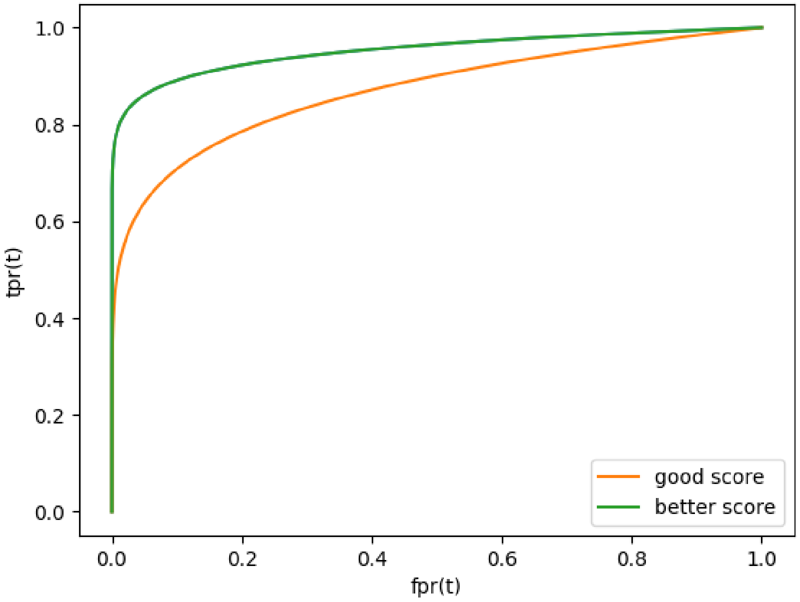
The performance of binary classifier $\hat{y}_t : \mathcal{X} \rightarrow \mathcal{Y}$ may be evaluated w.r.t. the classifier itself ($\hat{y}_t$) or w.r.t. the score $\tilde{y} : \mathcal{X} \rightarrow [0,1]$ which produces it. The ks score in (20), e.g., is an example of the latter. Something like accuracy $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(\hat{y}_t = y)$ is an example of the former. Both kinds of metrics are important to understand. We enumerate a few more here.

When we are looking at score-induced measure $\mathbb{P}_{\tilde{\mathcal{Y}}\times\mathcal{Y}}([a,b] \times \{j\}) = p_j \int_a^b f_j(t)dt$ we interpret concretely how to compute the corresponding metric.

True Positive Rate True positive rate (tpr) and false positive rate (fpr) are conditional probabilities, conditioned on truth $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$: $\mathrm{tpr}(t) := \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\hat{y}_t = 1|y = 1)$ and $\mathrm{fpr}(t) := \mathbb{P}_{\mathcal{X}|\mathcal{Y}}(\hat{y}_t = 1|y = 0)$. Note that the event being measured is $\left\{(x,y) \in \mathcal{X} \times \mathcal{Y} : \hat{y}_t(x) = 1\, y = 1\backslash 0\right\}$. This is why we are typically fast and loose with induced measures: we could have written e.g. $\mathbb{P}_{\tilde{\mathcal{Y}}|\mathcal{Y}}$ instead, keeping in mind how induced measures are defined. For score-induced measure tpr and fpr are $\int_t^\infty f_j(t)dt$, for $j = 1, j = 0$, respectively. Notice that class (im)balance p_j does not appear in tpr/fpr.[14]

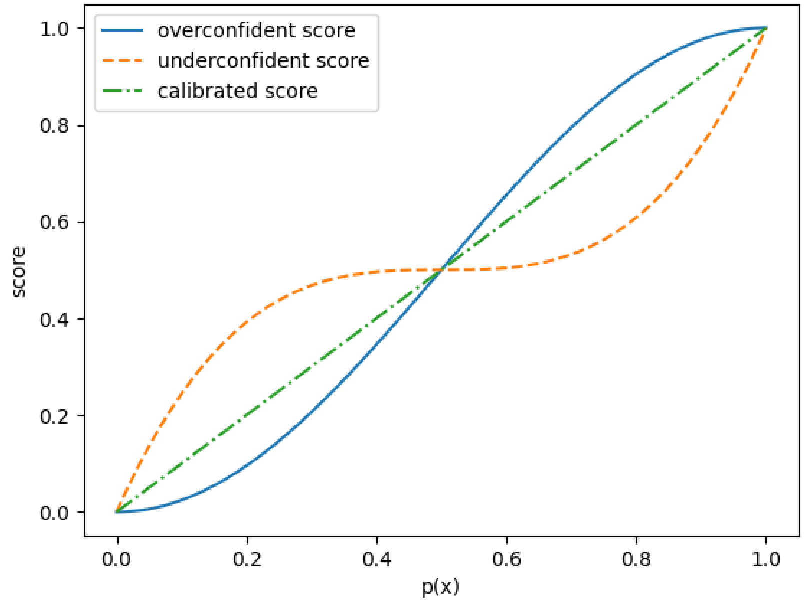
Precision Precision measures efficiency of a positive predictions: $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|\hat{y}_t(x) = 1)$. For score-induced measure, $\mathrm{prec}(t) = \dfrac{p_1 \int_t^\infty f_1(t)dt}{\int_t^\infty p_0 f_0(t) + p_1 f_1(t)dt}$. Notice that the denominator is the marginal $\mathbb{P}_{\mathcal{X}}(\hat{y}_t = 1)$.

ROC-AUC Parametrizing threshold $t \in [0,1]$ defines a curve $t \mapsto (\mathrm{fpr}(t), \mathrm{tpr}(t))$ as indicated in figure below While the point $(0,1)$ (meaning fpr $= 0$, tpr $= 1$) is the ideal, it will typically be difficult for a model to realize this. When one does, however, that automatically sets the optimal threshold: for less fortunate models (such as the ones indicated in the plot), judgment or other considerations must adjudicate whether a smaller threshold is better than a larger one; the latter will have lower false positive rate (good), but also lower true positive rate (bad).

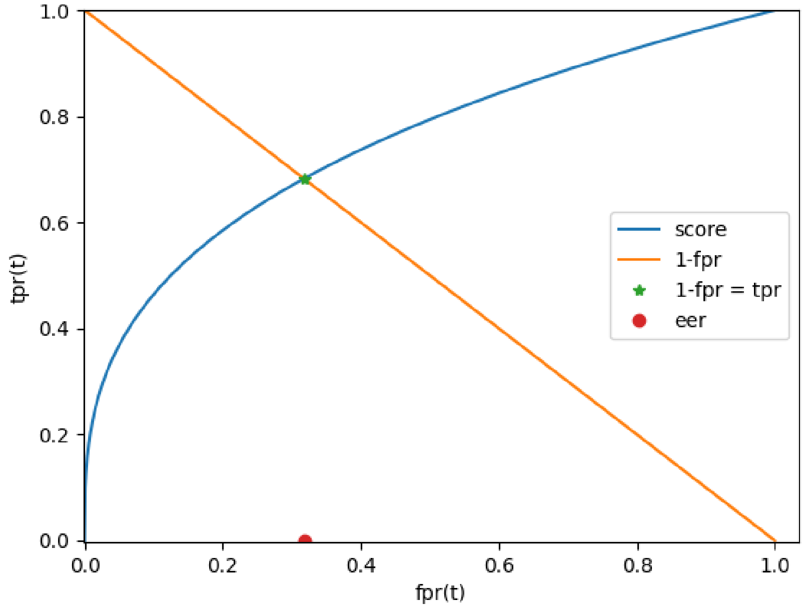


[14] The integral to ∞ is to account for scores which live in $\mathbb{R}$: when $\tilde{y}(X) \subset [0,1]$, $f_j(t) = 0$ for $t > 1$¿1.

Calibration Insofar as we expect score $\tilde{y}$ to separate classes, we can also desire that it reflect a conditional probability $p(x) := \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = 1|x)$. To the extent that $\tilde{y} \approx p(x)$, we say the model is *calibrated*.



Equal Error Rate Finally, we define equal error rate as $\mathrm{eer} := \arg_{\mathrm{fpr}}\{1 - \mathrm{fpr} - \mathrm{tpr} = 0\}$; see the figure below.



5 Lecture 5

5.1 Introduction

We recast the supervised learning problem in the context of the standard diagram (1) 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 especially 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* leads to an *algorithm* for finding the best approximating function $y^* : \mathcal{X} \to \mathcal{Y}$.

The ingredients for making sense of optimal model 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 for geometric concepts.

We start by introducing the notion of inner product, which you may think of heuristically as a means for algebraicizing 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. Let's go.

5.2 Inner Product Spaces

Definition 5.1. 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 $(\mathcal{V}, \mathbb{R})$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ is said to be an *inner product space*, denoted $(\mathcal{V}, \mathbb{R}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$.

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 if you need to). 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 5.1. 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," 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 5.1 and example 5.1 part (3), then we respond by saying: don't fret, the apparent conflict has easily remediable (if annoying) patchwork.

Example 5.1. 1. Let $\mathcal{V} = \mathbb{R}^n$ be n-dimensional Euclidean space 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. Let $(\mathcal{V} = \mathcal{C}([a,b],\mathbb{R}),\mathbb{R})$ be collection of continuous functions $\mathcal{C}([a,b],\mathbb{R}) := \{f : [a,b] \to \mathbb{R} : f \text{ is continuous}\}$ on a compact domain. Then $\langle\cdot,\cdot\rangle : \mathcal{V}^2 \to \mathbb{R}$ 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 consider probability space $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ and let $\mathcal{V} = \{f : \mathcal{X} \to \mathbb{R} : \mathbb{E}(f^2) < \infty\}$ be the 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 5.1 are all related. Indeed they are, and one should work out for themselves a space $\mathcal{X}$ and measure $\mathbb{P}_{\mathcal{X}}$ 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 elements 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 $||v + w|| \leq ||v|| + ||w||$; see, e.g., a short computation which cites Cauchy-Schwarz.

Norm induces a metric. A norm $|| \cdot || : \mathcal{V} \to \mathbb{R}$ on vector space $\mathcal{V}$ induces a metric $d : \mathcal{V}^2 \to \mathbb{R}$ by $(v, w) \mapsto ||v - w||$. Verification is straightforward; you rely on that addition (subtraction) makes sense in $\mathcal{V}$.

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

Definition 5.2. 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 5.3. 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 5.2. 1. Related to example 5.1, 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. You do not need to know this.

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.

5.3 Orthogonality

Definition 5.4. 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

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

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

Theorem 5.1. (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 $v = \mathrm{h} + \mathrm{h}^\perp$ where $\mathrm{h} \in \mathcal{H}$ and $\mathrm{h}^\perp \in \mathcal{H}^\perp$. (Uniqueness means for another representation $v = \mathrm{k} + \mathrm{k}^\perp$ where $\mathrm{k} \in \mathcal{H}$ and $\mathrm{k}^\perp \in \mathcal{H}^\perp$ that $\mathrm{k} = \mathrm{h}$ and $\mathrm{k}^\perp = \mathrm{h}^\perp$.) Moreover, $\mathrm{h} = \arg\min_{\mathrm{h}' \in \mathcal{H}} \|\mathrm{h}' - v\|$.

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

Proposition 5.1. (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 5.2. (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

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

□

5.4 Orthogonal Projection on Hilbert Subspaces

Let's proceed with the two following useful result:

Proposition 5.3. 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 $\mathrm{h}^* \in \mathcal{H}$ such that

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

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

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

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

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

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[15] 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 $v - h^* \perp h^* - h$. 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 $\min_{h \in \mathcal{H}} \|v - h\|$. □

[15] Note that this is not the same 'closed' that appears in theorem 6.1!

6 Lecture 6

We prove Hilbert Projection in section 6.1 and move on to uses of the orthogonality condition to solve our optimization problem in section 6.2. The proof is given for completeness and may safely be skipped.

6.1 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 $\mathcal{H} = \big\{\mathcal{X} \times \mathcal{Y} \to \mathcal{Y} \in \mathcal{V} : f = \tilde{y} \circ \pi_{\mathcal{X}}\big\}$
Now, let's prove Hilbert projection theorem. Let us recall the statement of the theorem.

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 $v = h' + \bar{h}$ 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 5.2 (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 5.3 implies that $h^* = \arg\min\limits_{h \in \mathcal{H}} \|v - h\|^2$. 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, $h^* = h', \bar{h} = h^\perp$. □

6.2 Optimal Predictor, Solved

With theory under our belt, let's return to the optimization problem. We reiterate that in order to use the Orthogonality Principle, we must articulate what our spaces are ($\mathcal{H} \subset \mathcal{V}$). We stated at the outset (beginning of section 5) that $\mathcal{Y} = \mathbb{R}$, and that the space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$ is a joint probability space. While we were very specific to fix $\mathcal{Y}$, we will be more permissive with $\mathcal{X}$: there is little need to specify this space apriori, but you may think $\mathbb{R}^k$, some categorical set $\{1, \ldots, \ell\}$, or some combination of both $\mathbb{R}^k \times \prod_j^n \{1, \ldots, \ell_j\}$. The ambient vector space $\mathcal{V}$ will be the collection of random variables on $\mathcal{X} \times \mathcal{Y}$ (i.e. maps $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$) with finite second moment ($\mathbb{E}(f^2)$), and the inner product will be as defined in example 3 in section 5.2. All that is left to specify is the subspace $\mathcal{H} \subset \mathcal{V}$.

In each of the following examples, we will consider different subspaces. When we first introduced the standard diagram (2) in section 1, we stated that we wanted a *function* $y^* : \mathcal{X} \to \mathcal{Y}$ best approximating y. Of course, stated as *this*, it doesn't quite make sense: y^* is a function (of $\mathcal{X}$), while $y \in \mathcal{Y}$ is a point. Probability helps us clarify: we want a function $y^* : \mathcal{X} \to \mathcal{Y}$ which evaluates as $y^*(x)$ on $\mathcal{X}$ to approximate the point $y \in \mathcal{Y}$, and this for $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$-"most" $(x, y) \in \mathcal{X} \times \mathcal{Y}$. And the diagram helps us rigorously line things up: the function y^* composes with function $\pi_{\mathcal{X}}$ (the latter of which on $\mathcal{X} \times \mathcal{Y}$) to give us a function $y^* \circ \pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$; the projection onto the second factor $\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$ also defines such a map and hence both $y^* \circ \pi_{\mathcal{X}}$ live in the same space $\mathcal{V}$ and therefore can be compared (distance measured) by use of the inner product on $\mathcal{V}$. More importantly, the space of functions f on $\mathcal{X} \times \mathcal{Y}$ satisfying our spec that $f = \tilde{y} \circ \pi_{\mathcal{X}}$ for some function $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ actually defines a subspace $\mathcal{H} \subset \mathcal{V}$, $\mathcal{H} := \{f \in \mathcal{V} : f = \tilde{y} \circ \pi_{\mathcal{X}} \text{ for some } \tilde{y} : \mathcal{X} \to \mathcal{Y}\}$. You can check that assertion on your own.

Last note before diving into the examples, one of the reasons for considering different spaces touches upon computational limitations: in each of the following, we will solve *mathematically* the optimization problem. Which is one step. But in practice you never even get the mathematical solution; rather, you approximate *it* using data or whatever. And the richer (larger) your space $\mathcal{H}$ is may require more data or more computation or something of that ilk to get. Therefore, you collect a few working examples in your toolkit, and upon presentation with a problem figure out what level of complexity in your subspace is needed to fit the bill.

6.2.1 Optimal Function $\mathcal{X} \to \mathcal{Y}$

With $\mathcal{H} = \{f : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} : f = \tilde{y} \circ \pi_{\mathcal{X}}\}$, we seek $\tilde{y} \circ \pi_{\mathcal{X}} \in \mathcal{H}$ best approximating $\pi_{\mathcal{Y}}$. We already computed in section 4.3 that $y^* \in \arg\min_{\tilde{y}} \mathbb{E}\left(\tilde{y}(x) - y\right)^2$, is the conditional expectation $y^*(x) = \mathbb{E}(\mathcal{Y}|\mathcal{X})$. That computation was done in the context of scoring for binary classification, but actually works in general for regression. We can also solve with the Orthogonality Principle. It states that

$$\mathbb{E}\left((y^* \circ \pi_{\mathcal{X}} - \pi_{\mathcal{Y}}) \cdot \tilde{y} \circ \pi_{\mathcal{X}}\right) = 0, \text{ for all } \tilde{y} \circ \pi_{\mathcal{X}} \in \mathcal{H} \tag{22}$$

Let's expand the equation above using conditional expectation:

$$\begin{aligned}
\mathbb{E}\left((y^* - y)\tilde{y}\right) &= \int_{\mathcal{X}\times\mathcal{Y}} (y^* - y(x)) \cdot \tilde{y}(x) d\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}(x, y) &(23)\\
&= \int_{\mathcal{X}} \int_{\mathcal{Y}|\mathcal{X}} (y^*(x) - y) \cdot \tilde{y}(x) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) d\mathbb{P}_{\mathcal{X}}(x) &(24)\\
&= \int_{\mathcal{X}} \tilde{y}(x) \left(\int_{\mathcal{Y}|\mathcal{X}} (y^*(x) - y) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)\right) d\mathbb{P}_{\mathcal{X}}(x) &(25)
\end{aligned}$$

In the last line, we pull $\tilde{y}(x)$ of the the inner expectation because that integral is with respect to y.

Now, we make a conjecture: *if* we could find $y^* \in \mathcal{H}$ that makes the inner integral equal to zero for all x, i.e.

$$\int_{\mathcal{Y}|\mathcal{X}} (y^*(x) - y) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = 0, \tag{26}$$

then the whole integral is zero and *therefore* we know this y^* satisfies (22).

Let's just suppose that such y^* exists, and see what happens. We have

$$
\begin{aligned}
& \int_{\mathcal{Y}|\mathcal{X}} (y^*(x) - y) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = 0 && (27)\\
\Rightarrow\quad & \int_{\mathcal{Y}|\mathcal{X}} y^*(x) d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) = \int_{\mathcal{Y}|\mathcal{X}} y \, d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x) && (28)\\
\Rightarrow\quad & y^*(x) \underbrace{\int_{\mathcal{Y}|\mathcal{X}} d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y|x)}_{=1} = \mathbb{E}(y|x) && (29)\\
\Rightarrow\quad & y^*(x) = \mathbb{E}(y|x) && (30)
\end{aligned}
$$

The form y^* takes supposing it exists demonstrates, under reasonable measurability assumptions (which we've just stipulated to hold), *that* it does.

6.2.2 Optimal Constant

In section 6.2.1, we considered the extremely large space of *all* functions from $\mathcal{X} \to \mathcal{Y}$. Now we consider the other extreme, "functions" which are constant: $\mathcal{H}_0 := \{\underline{c} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y} : \underline{c}(x, y) \equiv c \text{ with } c \in \mathbb{R}\}$.

To find optimal $\underline{c}^*$ we apply Orthogonality Principle (OP):

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

At this point, "forget the math" and try checking your intuition. This result translates as follows: if we say that we want to use *no* information from $\mathcal{X}$ whatsoever and have a single number the "best" captures label y, with such crude control, the result tells us to take the (marginal) mean. If you didn't know the Orthogonality Principle, this is likely what you would have guessed anyway.

6.2.3 Optimal Line: Linear Regression

Now consider the class of linear functions $\mathcal{H}_1 = \{a_0 + a_1(\cdot) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} | \, a_0, a_1 \in \mathbb{R}\}$. The be clear, an element $a_0 + a_1(\cdot) \in \mathcal{H}_1$ maps $(x, y) \mapsto a_0 + a_1 x$. To find $a_0^\star + a_1^\star(\cdot) \in \mathcal{H}_1$ optimally approximating $\pi_y : \mathcal{X} \times \mathcal{Y} \xrightarrow{\mathcal{Y}}$ we (like a machine) use the orthogonality principle which tells us that $(a_0^\star + a_1^\star(\cdot)) \circ \pi_\mathcal{X} - \pi_y \perp (b_0 + b_1(\cdot)) \circ \pi_\mathcal{X}$ for all $b_0 + b_1(\cdot) \in \mathcal{H}_1$, that is:

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

In particular, 'for all' implies 'for *two* specific cases':

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

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

These two cases gives the following system of equations:

$$\begin{cases} a_0^* + a_1^* \mathbb{E}(x) = \mathbb{E}(y) \\ a_0^* \mathbb{E}(x) + a_1^* \mathbb{E}(x^2) = \mathbb{E}(xy) \end{cases} \qquad (31)$$

which can be written in matrix form

$$\begin{pmatrix} 1 & \mathbb{E}(x) \\ \mathbb{E}(x) & \mathbb{E}(x^2) \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} \mathbb{E}(y) \\ \mathbb{E}(xy). \end{pmatrix}$$

$$\begin{pmatrix} a_0^* \\ a_1^* \end{pmatrix} = \begin{pmatrix} 1 & \mathbb{E}(x) \\ \mathbb{E}(x) & \mathbb{E}(x^2) \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{E}(y) \\ \mathbb{E}(xy). \end{pmatrix}$$

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

$$a_0^* + ... + a_n^* x^n = \pi_{H_n}(y) \tag{32}$$

and you will solve the general problem in a worksheet.

Now, writing eq. (31) in matrix form suggests, as is true, that a_j^* are the unknowns, but also, as is not [true], that $\mathbb{E}(x)$, $\mathbb{E}(xy)$, etc. are known. Properly speaking we need to know measure $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ to compute expectations (though for the marginals you can get away with less, e.g. $\mathbb{P}_{\mathcal{X}}$ for $\mathbb{E}(x)$), and typically we simply do not. But there's a nice thing about means: they are a single value "representation" or "approximation" or "summary" or whatever else you want to call it of the otherwise very sophisticated thing we call 'the measure' $\mathbb{P}$. And we have things at our disposal for approximating *it*: a mean and *empirical* mean—i.e. the mean $\frac{1}{m}\sum x_j$ obtained from *data*—are close, by the Law of Large Numbers. Of course, 'close' is qualified (and quantified) by things like concentration bounds, but the general point is more data gives better estimates, and you can use data to pretend like you actually do know the "knowns" in eq. (31).

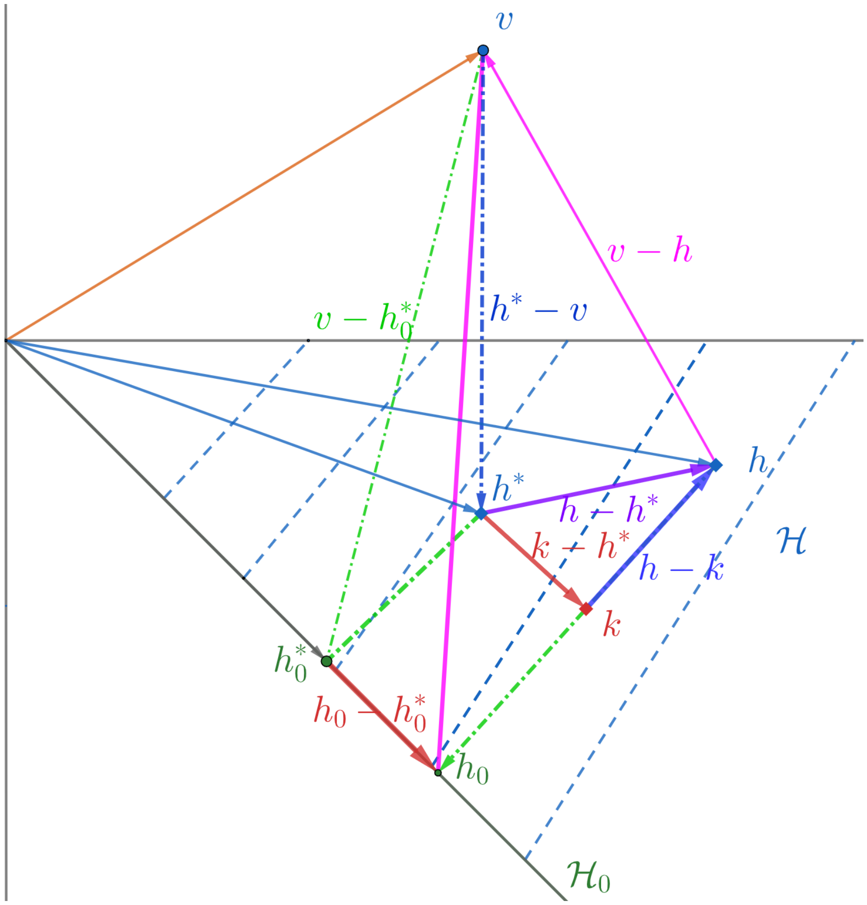


Figure 11: Bias-Variance in Hilbert Space

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

6.3 Bias-Variance: First Glance

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

$$h^* = \arg\min_{h\in\mathcal{H}} \|\nu - h\|^2 \tag{33}$$

$$h_0^* = \arg\min_{h\in\mathcal{H}_0} \|\nu - h_0\|^2 \tag{34}$$

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

$$h_0 := \pi_{\mathcal{H}_0}(h) \tag{35}$$

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

$$\|\nu - h^*\| \leq \|\nu - h_0^*\|. \tag{36}$$

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

$$\|\nu - h_0^*\|^2 = \|\nu - h^*\|^2 + \|h^* - h_0^*\|^2 \tag{37}$$

because $h_0 \in \mathcal{H}$ (implying that $\nu - h^* \perp h^* - h_0^*$). In practice, however we rarely obtain h^* or h_0^* when searching for optimal approximation of ν in $\mathcal{H}$ or $\mathcal{H}_0$; instead we get approximations $\tilde{h}$ or $\tilde{h}_0$, respectively, of *them*. And because $\mathcal{H}$ is a larger space than $\mathcal{H}_0$, you may imagine that the residual in approximations $\|h^* - \tilde{h}\|$ is larger (more wiggle room for error) than the error of approximation $\|h_0^* - \tilde{h}_0\|$ in $\mathcal{H}_0$. That imagination is all very heuristic, and relies on the geometry illustrated in fig. 11, but at the very least provides a first pass expression of the idea that larger complexity hypothesis class is good in on respect ($\|\nu - h^*\|$ becomes smaller), and potentially bad in another ($\|h^* - \tilde{h}\|$ may be larger).

7 Lecture 7

7.1 Introduction

We take a step back to collect what we have and muse on what's missing. We learned the projection theorem, which provides a concrete (computable) way to construct a model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ for regression $\mathcal{Y} = \mathbb{R}$ and when our measure of performance is given by $\ell_{\tilde{y}}(x, y) = (\tilde{y}(x) - y)^2$. The scenario for which the Orthogonality Principle "works" is highly specific, namely, we must be in a regression setting and the measure of performance for model correctness must be squared error, which abstractly summaries our requirement for a linear (vector) space[16] with an *inner product*. In general, our working notion of model performance—what we will call *loss*—need not come from or otherwise be in any way related to an inner product. While we used the inner product to algebraicize geometric notions, you should turn this around for concrete grounding 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.) In other words, Orthogonality gave us a connection between geometry and algebra, which is helpful for intuition, and now we will be using the algebra exclusively; but keep the geometry in your head for help.

We take a moment to review notation: $(x, y) \in \mathcal{X} \times \mathcal{Y}$ generally denotes a point in the joint space. But we will see the letter '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. Finally, 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.

7.2 Machine Learning: Beyond Hilbert Projection

Much of the setting has been set: we have joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X} \times \mathcal{Y}})$, and recall the standard diagram (1). Our goal is to construct *model* $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ for which $\tilde{y}(x) \approx y$ for $\mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$-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 $\|\cdot\|$ was a metric induced by norm induced by inner product $\langle \cdot, \cdot \rangle := \mathbb{E}(fg)$. 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 an 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 predictive functions $\mathcal{H} := \{\mathcal{X} \to \mathcal{Y}\}$ (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 (what we call) the *hypothesis class* $\mathcal{H}$ from the get-go.

Once we have our hands on *a* model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$, we'll need some way to evaluate how well $\tilde{y}$ "does" w.r.t. arbitrary labeled data point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, which we express by saying there is some random variable $\ell_{\tilde{y}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which we call the *loss function*, mapping $(x, y) \mapsto \ell_{\tilde{y}}(x, y) \in \mathbb{R}$. 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 more precisely, we want $\mathbb{E}(\ell_{\tilde{y}})$ to be small. The norm induced from inner product in Hilbert projection suggests a loss function: $\ell_{\tilde{y}}(x, y) := (\tilde{y}(x) - y)^2$, but by no means is the only one possible. We'll see others.

Now, in this case our task is to solve optimization problem

$$\min_{\tilde{y} \in \mathcal{H}} \mathbb{E}(\ell_{\tilde{y}}). \tag{38}$$

What's the "variable"?: The model! So you may think of the loss ℓ as inducing a *map*

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

[16] Which in this case recall is $\mathcal{V} : \{\mathcal{X} \times \mathcal{Y} \to \mathcal{Y} = \mathbb{R} :$ w finite second moment$\}$ the set of random variables on $\mathcal{X} \times \mathcal{Y}$.

from models to the set of random variables,[17] returning a random variable $\ell_{\tilde{y}}$ 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

$$y^* \in \arg\min_{\tilde{y}\in\mathcal{H}} \mathbb{E}(l_{\tilde{y}}),$$

or in other words finding model y^* which realizes equality in (38): $\mathbb{E}(\ell_{y^*}) = \min_{\tilde{y}\in\mathcal{H}} \mathbb{E}(\ell_{\tilde{y}})$.

You should think of supervised machine learning as glorified 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 *only* with fitting data as such but fitting the source, whence data comes. It amounts to fitting the measure (distribution)! Hence how ml provides a concrete setting for understanding probability: you really cannot talk about machine learning without it.

Collecting ingredients, we have:

1. Joint probability space $(\mathcal{X} \times \mathcal{Y}, \mathbb{P}_{\mathcal{X}\times\mathcal{Y}})$.

2. Data $S = \big((x_1, y_1), \dots, (x_m, y_m)\big) \in (\mathcal{X} \times \mathcal{Y})^m$, 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: $(x_1, y_1), \dots, (x_m, y_m) \sim_{\text{iid}} \mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$ (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 is incorrect. It's both 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 $\mathcal{H} \subsetneq \{\mathcal{X} \times \mathcal{Y}\}$. With the Orthogonality Principle, we allowed all models. Moving forward we won't. It's simply too hard to do anything if we consider the set of all possible predictors.

4. Loss function $\ell_{(\cdot)} : \mathcal{H} \to \{\mathcal{X} \times \mathcal{Y} \to \mathbb{R}\}$, and finally

5. Objective: to minimize $\mathbb{E}(\ell_{\tilde{y}})$ ranging over all models $\tilde{y} \in \mathcal{H}$.

Now we can get to work. We state the question thus: is there some *algorithm*

$$\hat{y}_{(\cdot)} : \bigcup_{m\in\mathbb{N}} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H} \tag{39}$$

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. And, if you already knew what the optimal model y^* is, there'd be no need to use *data* to construct one. 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 not, don't worry. We'll return to 'guarantee good model in probability' later in the semester.

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

$$\mathbb{E}(\ell_{\hat{y}_S}) - \mathbb{E}(\ell_{y^*}), \tag{40}$$

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

$$\underbrace{\mathbb{E}(\ell_{\hat{y}_S}) - \mathbb{E}(\ell_{y^*_{\mathcal{H}}})}_{\text{estimation error}} + \underbrace{\mathbb{E}(\ell_{y^*_{\mathcal{H}}}) - \mathbb{E}(\ell_{y^*})}_{\text{approximation error}} \tag{41}$$

where $y^*_{\mathcal{H}} \in \arg\min_{\tilde{y}\in\mathcal{H}} \mathbb{E}(\ell_{\tilde{y}})$. 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

[17] Be careful: when going over Hilbert Projection, the codomain here was $\mathcal{V}$ of which $\mathcal{H}$ was a subspace. That suggests we're looking for a map $\mathcal{H} \to \mathcal{V}$. Not quite: the random variable $\ell_{\tilde{y}}$ is not related to the map $\pi_{\mathcal{Y}}$, really, in any way.

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.

Before analyzing performance, let's get straight on how we use data to concoct a model in the first place! With the Orthogonality Principle, we saw one way: estimate various means. We will take a somewhat different approach now.

7.3 Introduction to Empirical Risk Minimization

The (pretty much) *one algorithm* for solving the ML problem is Empirical Risk Minimization. Recall that our definition (39) for $\hat{y}_{(\cdot)}$ takes in an arbitrary dataset S and outputs model $\hat{y}_S \in \mathcal{H}$.

Definition 7.1. Let $S \in \left(\mathcal{X} \times \mathcal{Y}\right)^m$ be a data sequence. We define *empirical risk minimization* as

$$\hat{y}_S := \arg\min_{\tilde{y} \in \mathcal{H}} e_S(\ell_{\tilde{y}}). \tag{42}$$

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

$$e_S(\ell_{\tilde{y}}) := \frac{1}{|S|} \sum_{(x,y) \in S} \ell_{\tilde{y}}(x, y) \tag{43}$$

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

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

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 central questions.

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

Much of the remainder of this semester will be focused on addressing the first two questions. We may come back to the third at the end of the semester when we discuss PAC-learnability.

8 Lecture 8

We ended last class by introducing *empirical risk minimization* (erm), the generic "algorithm" wherein we construct model $y^*_{\mathcal{H}}$ by taking the function in $\mathcal{H}$ which minimizes—not our actual objective $\mathbb{E}(\ell_{(\cdot)})$, but its proxy—empirical risk $e_S(\ell_{(\cdot)}) := \frac{1}{m}\sum_{j=1}^{m} \ell_{(\cdot)}(x_j, y_j)$ for data $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. It is a weird thing to call empirical risk minimization an "algorithm," for as defined it's simply a function, from data space $(\mathcal{X} \times \mathcal{Y})^{\omega}$ to hypothesis class space $\mathcal{H}$, taking data S and spitting out model $\hat{y}_S \in \mathcal{H}$. In fact, we stated that we will need a method for actually obtaining $\hat{y}_S$, and this properly is "the" algorithm. We discuss the algorithm now.

8.1 Hypothesis Class

Before getting into the algorithm details proper, however, we pay lip-service to the second task enumerated when introducing erm, namely the need to select a hypothesis class $\mathcal{H} \subset \{\mathcal{X} \to \mathcal{Y}\}$ of functions. We will discuss this task in more depth later in the semester, but for now we shall suppose that $\mathcal{H}$ is always a collection of functions parameterized by a finite set of values. For example, a function in the set of polynomials $\mathbb{R}[x] = \bigcup_{n \in \mathbb{N}} \left\{ \sum_{j=0}^{n} a_j x^j :, \, a_j \in \mathbb{R} \right\}$ is parameterized by coefficients $(a_0, \ldots, a_n)$. However, polynomials—without qualification—are not finitely parameterizable. A *given* polynomial $p(\cdot) = \sum_j^n a_j (\cdot)^j$ is, but the class itself is not.

Example 8.1. Consider $\mathcal{H}_n = \left\{ \sum_{j=0}^{n} a_j x^j : a_j \epsilon \mathbb{R} \right\}$ polynomials of degree $\leq n$.

In this case, we can say that

$$\mathcal{H} \cong \mathbb{R}^{k+1} \tag{44}$$

with mapping

$$p := \sum_{j=0}^{k} a_j x^j \mapsto \left(a_0, a_1, ..., a_k\right) \in \mathbb{R}^{k+1}$$

Thus, empirical risk, which defines a function

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

$$p \to e_S(p)$$

may simply be interpreted as a multivariable function.

Note that we are ambiguous on the nature of "isomorphism." We are assuming apriori little algebraic structure in $\mathcal{H}$ (we've dropped the Hilbert space requirement, since $\mathcal{Y}$ need not be $\mathbb{R}$, and $\ell_{(\cdot)}$ need not come from inner product).

We pause to interpret the foregoing discussion pictorially, see fig. 12. In this case we display at the top the map $e_S : \mathcal{H} \to \mathbb{R}$. In this figure $\mathcal{H} \cong \mathbb{R}^2 = span(a_0, a_1)$, but feel free to imagine generalizations: this is only a *picture*. We've presented the level sets in colors, namely in the domain, points on a curve of the same color map to the same value in $\mathbb{R}$ under e_S. The axes a_0 and a_1 parameterizing $\mathcal{H}$ correspond to *coefficients* of polynomial $a_0 + a_1 x$, which defines function $a_0 + a_1(\cdot) : \mathbb{R} \to \mathbb{R}$ (apparently $\mathcal{X} = \mathcal{Y} = \mathbb{R}$). We have also plotted *data* and ostensibly the line of best fit, as well as others corresponding to specially demarcated points in $\mathcal{H}$, all with the same intercept a_0, and different slopes.

With graphical interpretation by way of introduction, we now introduce the algorithm, gradient descent, we shall use to find our point of interest (in the picture the dark green point in $\mathcal{H}$ or equivalently dark green line in $\mathbb{R}^2$).

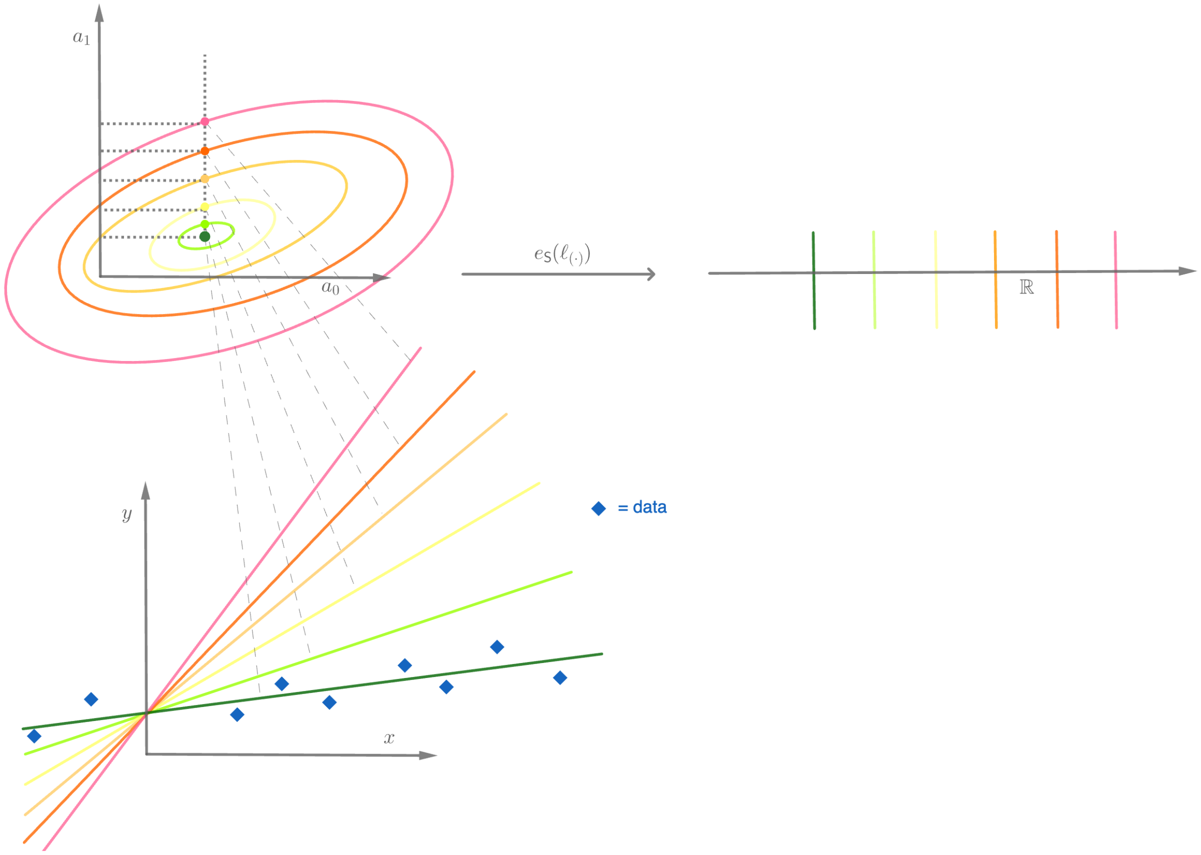


Figure 12: $\mathcal{H} \cong \mathbb{R}^k$

8.2 One-Dimensionalizing the Optimization Problem

We introduce an iterative method for finding approximation of $\min_{h\in\mathcal{H}} e_S(\ell_{\tilde{y}})$. The first step is to convert optimization in $\mathcal{H} \cong \mathbb{R}^K$, already made tractable by virtue that $\mathcal{H}$ is finite dimensional, into one of ordinary one-dimensional calculus. Specifically, we will iteratively specify (sufficiently) smooth curve $\gamma : \mathbb{R} \to \mathcal{H} \cong \mathbb{R}^k$ by defining point-evaluations $\gamma(t_1), \gamma(t_2), \dots, \gamma(t_k), \dots$ that hopefully in the limit $\lim\limits_{k\to\infty} \gamma(t_k) = \arg\min\limits_{h\in\mathcal{H}} e_S(\ell_h)$. We aren't going to define γ for *all* times $t \in \mathbb{R}$ but rather use the idea of discretely "meandering in space" ($\mathcal{H}$) and sampling values ($e_S(\ell_{\gamma(t_j)}) \in \mathbb{R}$) to cook up the next point t_{j+1} to sample. This all sounds somewhat convoluted, but we'll see that it resolves the optimization task into a procedure which is rather simple to implement.

Let's characterize what we'd like this curve γ to satisfy, supposing we knew it already. If we compose empirical risk e_S with the curve γ, we obtain real-valued function $\eta := e_S(\ell_{(\cdot)}) \circ \gamma : \mathbb{R} \to \mathbb{R}$. We wish that at some point $t^* \in \mathbb{R}$, $\eta(t^*) = \min_{h\in\mathcal{H}} e_S(\ell_h)$. Careful on what this means, we are not saying *only* that single-variable real-valued function η achieves *its* minimum but also that its minimum coincides with the minimum of $e_S(\ell_{(\cdot)})$. Apriori, there is no reason to suppose it should since the curve $\gamma(\ R) \subsetneq \mathcal{H}$, unless we somehow stipulate γ so that it does include the global minimum in $\mathcal{H}$.

Supposing that $\eta(t^*) = e_S(\ell_{h^*}) = \min_{h\in\mathcal{H}} e_S(\ell_h)$, we know from calculus that $\frac{d}{dt}\Big|_{t=t^*} \eta(t) = 0$, which by the chain rule means that

$$\frac{d}{dt}\Big|_{t=t^*} e_S \circ \gamma(t) = \nabla e_S(\gamma(t^*)) \cdot \dot{\gamma}(t^*) = 0, \tag{45}$$

where $\dot{\gamma}$ is the velocity tangent vector, and ∇e_S is the gradient of e_S which we know is orthogonal to level sets. We illustrate this scenario in fig. 13, with circular level sets and grey curve. The grey arrow is its velocity $\dot{\gamma}(t_0)$ at some time t_0 and the dark red arrow is the (*negative of* the) gradient $\nabla e_S(\gamma(t_0))$.[18] The purple curve is another curve.

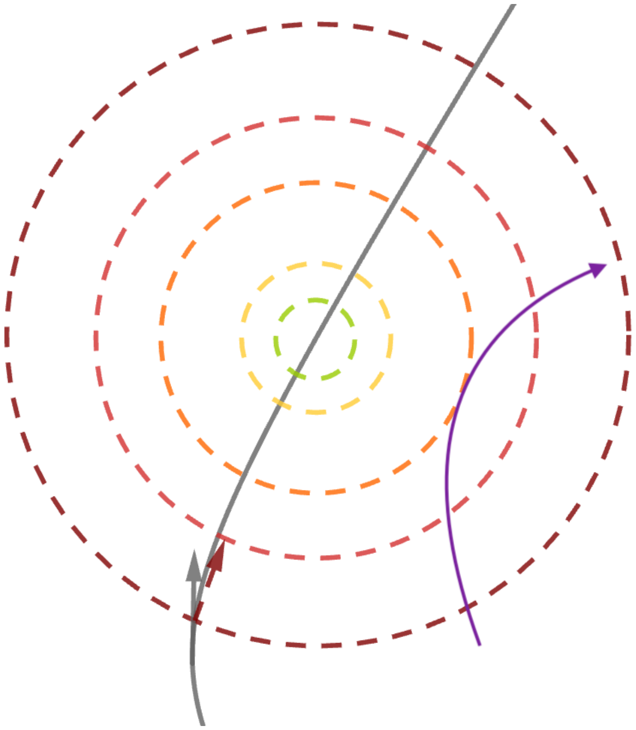


Figure 13: Components of $\frac{d}{dt}\eta$

The product $e_S(\gamma(t^*)) \cdot \dot{\gamma}(t^*)$ in eq. (45) is a dot product, on account of which there are three ways it could be zero:

1. $\nabla e_S(\gamma(t^*)) = 0$
2. $\dot{\gamma}(t^*) = 0$
3. $\dot{\gamma}(t^*) \perp \nabla e_S(\gamma(t^*))$

Keeping in mind that our task is to design γ, we comment on these options. For the first, the gradient "is what it is," we have no control over it. Of course, wherever $\nabla e_S = 0$ is a candidate minimum. We would like to find it. Option 2 is to be avoided, if possible. As $\dot{\gamma}$ is velocity of the curve, this condition states that the curve has stopped moving. Since we are interested in probing the space $\mathcal{H}$ by means of γ, this condition is supremely unhelpful, and easily avoided as long as we demand e.g. that γ have constant speed $||\dot{\gamma}|| \equiv 1$. Condition 3. is tricky and also to be avoided: the velocity of curve is orthogonal to gradient means that the velocity is tangent to the level set. In other words, while this point *may* be a local minimum for η (see the purple curve in fig. 13), it is not a min of $e_S(\cdot)$: the fault is on the choice of curve.

Thus, we want γ to more closely resemble the grey curve than purple one, and have specified a few conditions to approximately ensure that it does: the curve should have non-zero speed and its velocity vector should not be tangent to level sets (orthogonal to the gradient). In fact, we can say more: because $\dot{\gamma} \cdot \nabla e_S = \|\dot{\gamma}\|\|\nabla e_S\| \cos\theta$, where θ denotes the angle between $\dot{\gamma}$ and ∇e_S. If we mandate constant speed $||\dot{\gamma}|| \equiv 1$, then this expression is simply $\|\nabla e_S\| \cos\theta$.

This discussion lays the ground work for the iterative algorithm, to which we now turn.

[18] The gradient orthogonal to level set points in the direction of steepest *increase*, and therefore should be pointed in the opposite direction.

8.3 Gradient Descent

Consider first-order Taylor approximation of

$$\eta(t+dt) = \eta(t) + \eta'(t)dt + o(dt) \tag{46}$$

Supposing $dt \approx 0$ is small, you can "ignore it" (little-oh let's us ignore it in a mathematically appropriate way).[19] Our task is to define points $\gamma(t_0), \gamma(t_1), \dots, \gamma(t_k), \dots$ so that $\lim_{k\to\infty} \gamma(t_k) = \eta^* := \arg\min_{h\in\mathcal{H}} e_S(\ell_h)$. The times t_k will simply be $t_0 + k \cdot dt$, so really the task is to define what point (in $\mathcal{H}$) γ shall be at these points.

Considering eq. (46), a good candidate would be to make $\eta(t+dt) < \eta(t)$: the next point should be smaller than the last and by as much as possible. Fixing dt as some constant, the only control we have is over $\eta'(t)$, which we saw from the last section is just $\nabla e_S(\gamma(t))\cos(\theta)$, where recall θ is the angle between gradient ∇e_S and $\dot{\gamma}$. This expression isolates the *only* piece of control we have, namely the angle θ. If we wish to make $\eta(t) - \eta(t+dt)$ as large as possible, therefore, then our best bet is to make $\cos(\theta)$ as *small* (or largely negative) as possible: $\cos(\theta*) = -1$, achieved for $\theta^* = \pi$. In other words, the direction of optimal (fastest) decrease is for $\dot{\gamma}$ to be aligned with gradient, but pointed in the opposite direction.

With this, we define gradient update $h_{k+1} := h_k - \delta\nabla e_S(\ell_{h_k})$ for some small $\delta > 0$, and define the *gradient descent* algorithm as 'repeat this gradient update' a bunch of times:

Algorithm 1 Gradient Descent

```
Input: Initial guess $\alpha_0 \in \mathbb{R}^k \cong \mathcal{H}$
Output: Final $\alpha^*$
  Choose fixed $\epsilon > 0$ (some tolerance)
  Pick some $\delta > 0$ (the "learning rate")
  Set $k = 0$
  while $\|\nabla_\alpha e_S(\ell_{\alpha_k})\| \geq \epsilon$ do
      Set $\alpha_{k+1} \leftarrow \alpha_k - \eta\nabla_\alpha e_S(\ell_\alpha)$
      Set $k \leftarrow k+1$
      Compute $\nabla_\alpha e_S(\ell_{\alpha_k})$
  end while
  return $\alpha^* \leftarrow \alpha_k$
```

Apriori there's no reason to assume that the algorithm halt. In the absence of guarantees, we may opt for an alternative loop in which we define a specific number of steps to update the gradient. There are numerous variants and more sophisticated versions of this algorithm; we start with the simple one to start. Guarantees *do* exist for convergence, performance, etc., but they rely on assumptions (e.g. convexity) which often do not exist. The more sophisticated methods exist precisely to mitigate issues. We may come across some later in the semester.

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)
```

[19] https://en.wikipedia.org/wiki/Big_O_notation

```
    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
```

8.4 Examples

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

Example 8.2 (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 8.3 ((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:

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