Math 4460 Course Notes

August 24, 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

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

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

Traditionally, $\pi_{\mathcal{X}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$, defined by mapping $(x, y) \mapsto x$ (read: $\pi_{\mathcal{X}}(x, y) := x$), is taken to be "easy, efficient, or cheap" to evaluate or sample while $\pi_{\mathcal{Y}} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Y}$, defined by mapping $(x, y) \mapsto y$, is computationally expensive, expensive otherwise, difficult for other reasons, or altogether infeasible. The original space $\mathcal{X} \times \mathcal{Y}$ is itself inaccessible, except for some (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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$. 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 $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(y = f(x)|x) \neq 1$.

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 $(x_i, y_i) \sim_{iid} \mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$ 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 $\prod_{j=1}^{m} \mathbb{P}_{\mathcal{X} \times \mathcal{Y}}$. More on this later.

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

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

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

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

1.3 Zeroth Assignments

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

The purpose of the assignment is threefold:

1. Gain initial exposure to the *structure* of machine learning code, including object oriented programming and the typical methods included.

2. Shake off any residual rust using Python.

3. To gain deeper appreciation for the *aim* of building model $\tilde{y} : \mathcal{X} \to \mathcal{Y}$ as in diagram (6), and metrics that illustrate success.

1.4 Probability

1.4.1 Context

Recall the Standard Diagram

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

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

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

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

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 $\mathbb{P}_{\mathcal{X}} = \int d\mathbb{P}_{\mathcal{X}}(x)$, and define expectation $\mathbb{E}(f)$ of a random variable $f : \mathcal{X} \to \mathbb{R}$ as a *Lebesgue* integral $\mathbb{E}(f) := \int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$. We kick this lecture off by emphasizing that probability has nothing to do with randomness...yet. When we return to admitting randomness into our lexicon, it will be as a *result*. For the moment, we forget any association between probability and chance, stochasticity, randomness, or any other (for now) anathema word affiliated with the notion of uncertainty.

1.4.2 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

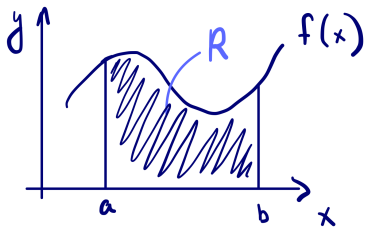

Figure 1: Area

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

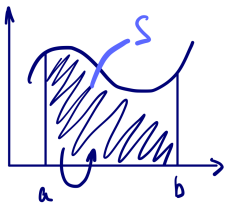

Figure 2: Volume

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)[2] the power set of $\mathcal{X}$ to the closed interval $[0,1]$ satisfying the following two properties:

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

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

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

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

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

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

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

You should check that countable additivity implies *finite* additivity $\mathbb{P}_{\mathcal{X}}\left(\sum_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} \mathbb{P}_{\mathcal{X}}(A_j)$. You will need the fact that $\mathbb{P}_{\mathcal{X}}(\varnothing) = 0$, which itself is implied by conditions 1. and 2. Indeed, $\mathcal{X} = \mathcal{X} \sqcup \bigsqcup_{j=2}^{\infty} \varnothing$. 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 architecture constructing measure depends on the fact that some structure must be placed on the set of subsets of $\mathcal{X}$ which are suitable for measurement; in particular, that such a set comprises a so-called σ-algebra, is not necessarily (and in many cases in fact is not) the entire power set $2^{\mathcal{X}} = \{A \subset \mathcal{X}\}$, and so on. We pay lip-service to this nuance, but fret little over the possibility that we will accidentally run across both a measure $\mathbb{P}_{\mathcal{X}}$ and subset $A \subset \mathcal{X}$ for which $\mathbb{P}_{\mathcal{X}}(A)$ does not (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 $\int_{[a,b]} d\mathbb{P}_{\mathcal{X}}(x)$, 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

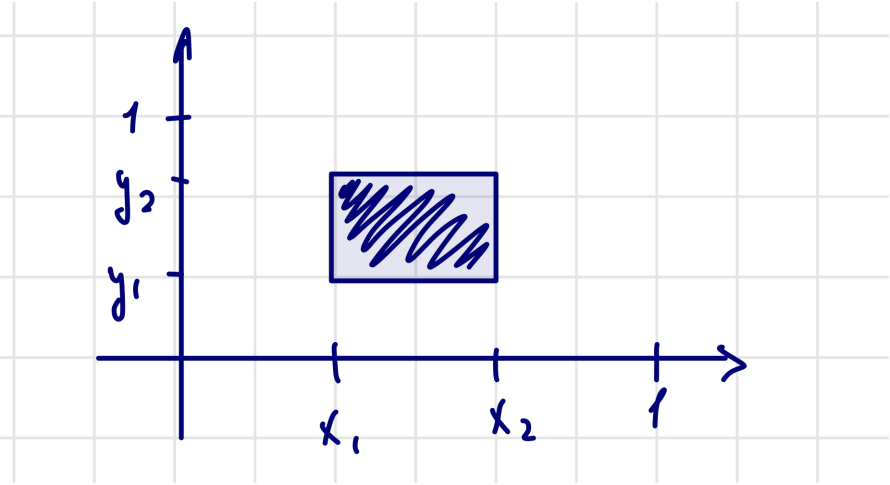

Figure 3: The area measure

stated earlier since the events or subsets of the space X, $\{0\}$ and $\{1\}$ are disjoint and the sum of their probabilities is 1:

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

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 $\mathbb{R}$.

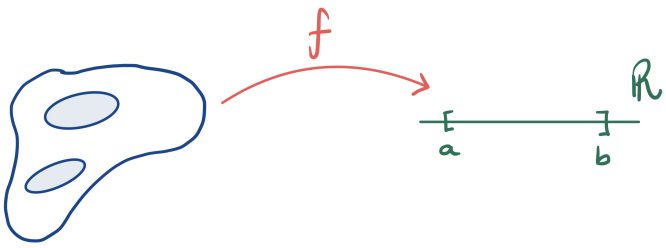

Figure 4: Random variable f

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

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 $a_1, \ldots, a_k$. Then we define the *Lebesgue* integral of f, denoted $\int_{\mathcal{X}} f(x) d\mathbb{P}_{\mathcal{X}}(x)$, to be

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

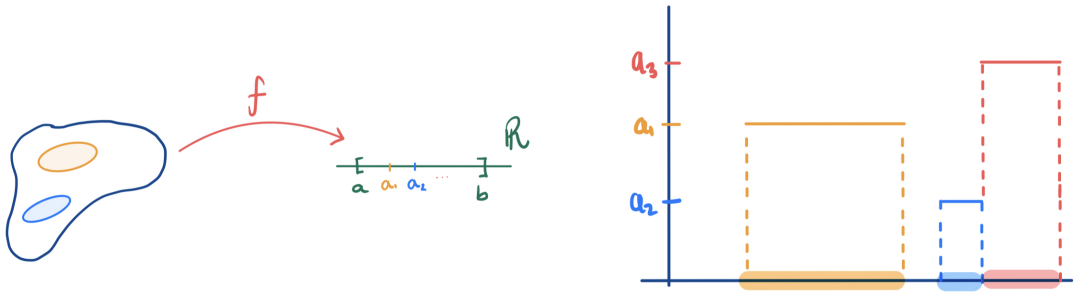

Figure 5: Simple random variable f

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.

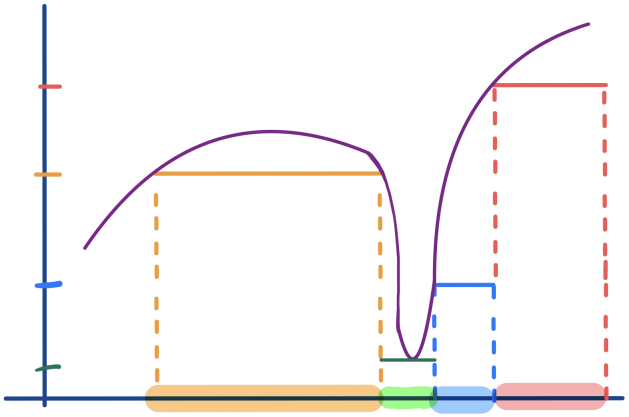

Figure 6: Approximation via simple functions

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

1. To articulate the fact that the extension from discrete to continuous random variables may *seem* confusing, and that is due in part to the nauseating head-spinning move from Lebesgue to who knows (but usually Riemann) integration without even lip-service paid to the fact that expectation of discrete r.v.s itself is a non-trivial extension of our conceptual apparatus.
2. Ease of notation: $d\mathbb{P}_{\mathcal{X}}$ always makes sense and makes immediately obvious what our measure is. In other words, I don't always want to say: suppose that the density of a probability space exists, and anyway it might not and that doesn't matter!, for $\int_A d\mathbb{P}_{\mathcal{X}}(x)$ makes sense regardless of whether we can integrate (Riemann-wise) as $\int \varphi(x)dx$ (for density $\varphi(x)$). Related: the

notation $d\mathbb{P}_{\mathcal{X}}$ collapses the distinction between continuous and discrete random variables. The distinction, in my view, is convoluted and confusing, especially when we have mixed discrete-continuous nonsense going on (e.g. in the case of binary classification $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$). Of course, successfully working with the collapse requires an comfort with the abstraction, and that may by itself be initially confusing as well.

Remark 2.4. Observe also that 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

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

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

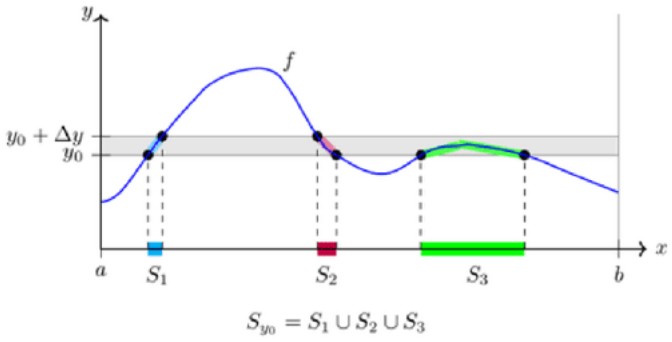

Figure 7: Visualizing Lebesgue Integration

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

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

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

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

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

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

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

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

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

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

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

Definition 2.6. Let $(\mathcal{X}, \mathbb{P}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$ be probability spaces. A *map* $f : (\mathcal{X}, \mathbb{P}_{\mathcal{X}}) \to (\mathcal{Y}, \mathbb{P}_{\mathcal{Y}})$ *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

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

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

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

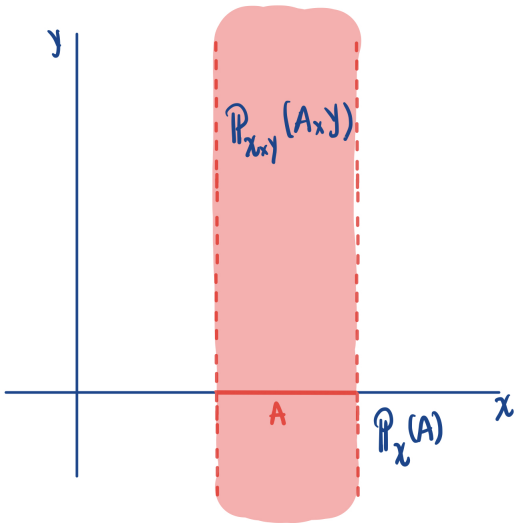

Figure 8: Geometry of Marginalization

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.

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. 8).

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 on 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:

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

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

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

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* 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 2.6. 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).

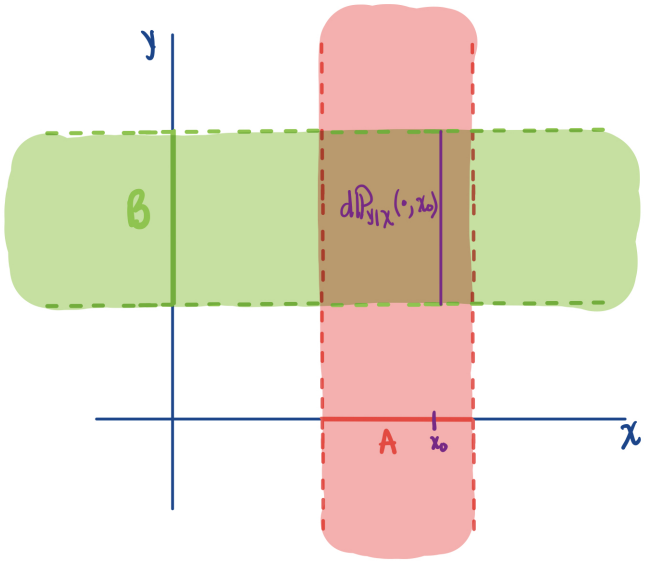

Figure 9

In math, hand-waiving can sometimes be dangerous. On the other hand, it can sometimes allow us to think reasonably about, and operationalize our intuition of, notions whose formalism is "beyond the scope of this course," all with the aim of performing computations and symbolic manipulations. If you are interested in more rigorously making sense of $d\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$, I recommend diving into the measure theory.[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 value.

Now for the geometric intuition: in multivariable calculus, to integrate a function $f(x, y)$, $\int_R f$, 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 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

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

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 $\left(\mathcal{X}^m, \mathbb{P}_{\mathcal{X}^m}\right)$ is *independent* if

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

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

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

Remark 2.7. Be careful with this definition: from spaces $(\mathcal{X}_j, \mathbb{P}_{\mathcal{X}_j})$, the independent measure $¶_{\prod\mathcal{X}_j}$ in (15) is not the only possible measure on joint space $\prod_j\mathcal{X}_j$!