Informal Notes for Mathematics for Machine Learning
Imperial College London

Lecture 1: Formalizing ML Problem Settings

Lecturer: Matthew Wicker

1 Course Introduction

This course, Mathematics for Machine Learning 70015, sets out with the goal of helping students develop a mathematical fluency that allows them to transform the information received in previous introductory ML courses into usable knowledge base that can be used and adapted to the current cutting-edge machine learning problems. In other words, where other courses have taught you *what* modelling choices you can make in ML, this course aims to help you understand the *why* one makes a given decision. This course has been restructured from previous years. I have also included the addition of these more-or-less comprehensive lecture notes, BUT **these lecture notes are a first pass and are just a rough-draft. I urge you to check for type-os and to reference the listed reference materials for a more complete picture.** The course is s how this structured as follows:

1. In the first week, we will discuss different machine learning settings and briefly go over brush up on skills you hopefully acquired in previous courses in the context of formulating linear models. Additionally, in the first week, I will prompt students to give me their desires and goals for the course which will inform some of the optional material we cover.

2. In the second week, we will touch on a few critical concepts in optimization such as automatic differentiation, convergence, and convexity.

3. In the third and fourth weeks we will dive into the world of probability to take a probabilistic perspective on the machine learning models we develop.

4. In the fifth week we will turn the probability study up to 11 and focus on Bayes theorem. We will not only cover the rich mathematical framework that Bayes affords us, but also we will do practical experiments on the strengths and limitations of predictive uncertainty.

5. In week six and seven we will combine everything we have learned thus far. First in a study of principle component analysis which lies neatly at the intersection of all of the mathematical skills we will have developed. Secondly, we will spend one week covering two advanced topics. In particular covering more advanced topics in optimization that should help you understand modern NN training and adversarial examples and we will also cover a student-selected topic.

2 Learning Objectives for this Lecture

By the end of our first lecture you should be familiar with the standard formulation of a key machine learning problem settings. In addition, we review some of the basic notation that will be crucial to navigating this course. At the bottom of these lecture notes are a collection of problems that test prerequisite knowledge that will be crucial to the rest of the course. If you struggle with these, please seek out a TA in the first instance and then myself so that we can ensure you do not fall behind.

3 Machine Learning Problem Settings

We will start here with a focus on supervised learning settings. Below we discuss the two primary supervised learning settings (classification and regression). Both of these settings will be our primary focus throughout the course; however, we will also provide some details on the unsupervised learning setting which can be a useful thing to keep in mind during the course and is something we will touch on in our final weeks.

3.1 Supervised learning

Supervised learning settings assume that we have access input-output pairs, (x, y), where $x \in \mathbb{R}^n$ is our input or feature space and $y \in \mathbb{R}^m$ is our output space. It is typically assumed that we have many such pairs comprising a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{d}$. When d (the number of data-points) is very small, we call the task *few-shot* learning. However, for this course we will typically assume we have an abundance of data. How we learn to reproduce y given x depends heavily on the semantics of the task which in turn greatly effect the modelling decisions we can/should make. Two such semantically distinct settings are classification and regression.

3.1.1 Classification

Classification settings generally aim to map inputs $x \in \mathbb{R}^n$ to a discrete output y, where $y \in 1, ..., C$, with C being the number of classes. Another way of writing the co-domain (meaning the output space of our model) is as $\mathbb{Z}^C$, the group with C elements. It is generally assumed that the classes in this group are distinct and mutually exclusive i.e., that an image contains either a dog or a cat, but not both. The setting in which one is predicting multiple classes simultaneously is called multi-label classification. Throughout the course, we will focus on the multi-class (not multi-label) setting.

3.1.2 Regression

A regression task is typically defined to be when the co-domain of our model is continuous or ordered. Various extensions of this basic problem can arise, such as having high-dimensional inputs,

outliers, non-smooth responses, etc. As we go through the course, we will see many such examples of important complications in regression setting. For now, let us simply highlight that we take any machine learning problem with a continuous co-domain will be referred to as regression. A similar, but distinct, setting involves having an ordered output (this is known as *ordinal regression*) for example, we may have a model that predicts grades i.e., $\{A, B, C, D, F\}$. which have distinct elements but a clear semantic structure that is more aligned with regression than classification.

3.2 Unsupervised Learning

In unsupervised learning the critical difference is the lack of targets y. In this setting the objective is typically to discover some structure underlying the data. This can take the form of clustering or density estimation. We will touch on this in the final weeks of the course when we study principle component analysis and some more current, open research questions.

4 Problem Settings: Underlying Mathematics

Now that we have covered a few different key learning settings, we should take a step back and ask ourselves critical questions about these formulations. For example, we have supposed the existence of a dataset $\{(x^{(i)}, y^{(i)})\}_{i=1}^{d}$ (assuming supervised learning). But what exactly is this mathematical object? How can we analyze and think about this information in a way that informs our downstream model choices? In the case of the dataset, it is often useful to think of the data as a sample from an unknown joint distribution $P(x, y)$. Similarly, consider the output of a neural network trained in medical image recognition classifies a patient scan as containing a malignant tumor, should we just assume this is the case? What if we wanted our model to tell us how certain it was that this was the case? Again, we would model the output using probability theory. The key tools needed to reason about a sample from an unknown distribution or about uncertainty come from probability and statistics. These areas will continue to be critical throughout the course and so it is important that we start with a solid understanding of how to mathematically manipulate and reason about probability distributions.

4.1 Random Variables

Formally speaking, a random variable is a measurable function from a sample space onto a measurable space. Until we get to lectures 6 and 7, we will simply consider a random variable X (r.v.'s will be denoted using uppercase or bold text and will be specifically identified as random variables) to be a function from a defined sample-space onto the unit interval $[0, 1]$. We denote the probability that a random variable X is set to a particular value from the sample space as $p(X = x)$ as $p(x)$. When the sample space is a continuous space, we call the function p a *probability mass function* or p.m.f. Where the sample space is discrete p is called a *probability density function* or p.d.f. Students who may be rusty on probability theory may consult Section 2.2 of Kevin Murphy's wonderful book on probabilistic machine learning. Below, we will cover some of the key distributions we will employ throughout the course.

4.1.1 Discrete Random Variables

Bernoulli and Multinoulli Perhaps the most well-known discrete random variable is a Bernoulli random variable whose outcome can take one of two values. That is, we have a binary sample space (e.g., a coin can land on *heads* or *tails*). Such a distribution has a single parameter θ and we say that event 0 hppens with probability θ and event 1 happens with probability $1 - \theta$, by the law of total probability.

Of course, there are many cases in which we do not have a two outcome sample space, for example the rolling of a dice. In this case, we need a multinoulli distribution. Traditionally, the parameter of a multinoulli $\theta \in \mathbb{R}^s$ is a vector where $s > 2$ is a natural number representing the number of outcomes in our sample space. of course $\sum_{i=0}^{s-1} \theta_i = 1$ is necessary to ensure that the multinoulli represents a proper probability mass function. We will see such multinoulli distributions crop up as the output of classification models.

The binomial and multinomial [1]

Suppose we toss a coin n times. Let $X \in 0, ..., n$ be the number of heads. If the probability of heads is θ, then we say X has a binomial distribution, written as $X \sim Bin(n, \theta)$. The p.m.f. is given by

$$\text{Bin}(k|n, \theta) := \binom{n}{k} \theta^k (1 - \theta)^{nk}$$

where,

$$\binom{n}{k} := \frac{n!}{(n-k)!k!}$$

is the number of ways to choose k items from n (this is known as the binomial coefficient, and is pronounced n choose k). This distribution has the following mean of θ and variance of $n\theta(1\theta)$.

On the other hand, we have the multinomial distribution which models the distribution of outcomes when sampling from a multinoulli distribution. To model this we could ask about the probability of a certain set of outcomes. For example, if I were to roll a four sided die $n = 20$ times, what is the probability that I would have one 1, ten twos, 8 threes, and 1 four? We would encode these outcomes into the vector $x = \langle 1, 10, 8, 1 \rangle$ and then we could compute the probability as follows:

$$\text{Mu}(x|n, \theta) := \binom{n}{x_1, \ldots, x_K} \prod_{j=0}^{K} \theta_j^{x_j}$$

where K is the number of outcomes in the sample space (in our case 4) and $\binom{n}{x_1, \cdot, x_K}$ is the multinomial coefficient. We will not often use these equations explicitly, but they ought to be relatively familiar to students in order to properly navigate some exercises.

Test your understanding: Assume, $n = 1$ in both of the above distributions. How does the expression change? What familiar distribution does this give us?

[1] We have taken our exposition of these distributions directly from Kevin Murphy's textbook on probabilistic machine learning.

The empirical distribution Given a set of data, $D = \{x^{(0)}, \ldots, x^{(N-1)}\}$, we define the empirical distribution, also called the empirical measure, as follows:

$$p_{\text{emp}}(x) := \frac{1}{N} \sum_{i=1}^{N-1} \delta_{x_i}(A)$$

where $\delta_x(A)$ is the Dirac measure, defined by:

$$\delta_x(A) = \begin{cases} 0, & \text{if } x \notin A \\ 1, & \text{if } x \in A \end{cases} \tag{1}$$

In general, one can also associate weights with each element of an empirical distribution, i.e., $p_{\text{emp}}(x) := \frac{1}{N} \sum_{i=1}^{N-1} w_i \delta_{x_i}(A)$ as long as each $w_i < 1$ and $\sum_{i=1}^{N-1} w_i = 1$.

4.1.2 Continuous Random Variables

The most widely used distribution in statistics and machine learning is the Gaussian or normal distribution. Its probability density function (pdf) is given by:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Here $\mu = \mathbb{E}[X]$ is the mean (and mode), and $\sigma^2 = \mathbb{V}[X]$ is the variance. $\sqrt{2\pi\sigma^2}$ is the normalization constant needed to ensure the density integrates to 1. We write $X \sim N(\mu, \sigma^2)$ to denote that $p(X = x) = N(x|\mu, \sigma^2)$. If $X \sim N(0, 1)$, we say X follows a standard normal distribution. We will often talk about the precision of a Gaussian, by which we mean the inverse variance: $\lambda = \frac{1}{\sigma^2}$. A high precision means a narrow distribution (low variance) centered on μ. Note that, since this is a pdf, we can have $p(x) > 1$. To see this, consider evaluating the density at its center, $x = \mu$. We have $N(\mu|\mu, \sigma^2) = (\sigma\sqrt{2\pi})^{-1} \exp(0)$, so if $\sigma < \frac{1}{\sqrt{2\pi}}$, we have $p(x) > 1$.

The cumulative distribution function (cdf) of the Gaussian is defined as:

$$\Phi(x; \mu, \sigma^2) = \int_{-\infty}^{x} N(z|\mu, \sigma^2)\, dz$$

This integral has no closed-form expression, but is built into most software packages. In particular, we can compute it in terms of the error function (erf):

$$\Phi(x; \mu, \sigma^2) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

where $z = \frac{x-\mu}{\sigma}$, and

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-t}^{t} \exp(-t^2)\, dt$$

The Gaussian distribution is the most widely used distribution in statistics. And we will observe these reasons a bit further in the course when we cover central limit theorem and hypothesis testing in the context of model testing and validation.

4.2 Independently and Identically Distributed

4.2.1 Independence

In the context of statistical analysis, independence refers to the idea that the values or outcomes of one observation in a dataset do not depend on or influence the values of any other observation. Mathematically, let $X_1, X_2, \ldots, X_n$ be a set of random variables representing a sample of n observations. These random variables are said to be independent if, for any subset of observations $X_{i_1}, X_{i_2}, \ldots, X_{i_k}$ (where $1 \leq i_1 < i_2 < \ldots < i_k \leq n$), the joint probability distribution factorizes:

$$P(X_{i_1} = x_{i_1}, X_{i_2} = x_{i_2}, \ldots, X_{i_k} = x_{i_k}) = P(X_{i_1} = x_{i_1}) \cdot P(X_{i_2} = x_{i_2}) \cdot \ldots \cdot P(X_{i_k} = x_{i_k})$$

This property is crucial when dealing with real-world data, as it allows us to model each observation as an independent realization of some underlying process. For example, when rolling a fair six-sided die multiple times, the outcomes of each roll are independent events.

4.2.2 Identically Distributed

The "identically distributed" part of i.i.d. refers to the idea that all random variables in the sample follow the same probability distribution. That is, they share the same probability density function (pdf) or probability mass function (pmf), depending on whether the data is continuous or discrete.

Mathematically, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables, it means that they all have the same pdf or pmf, denoted as $f_X(x)$, and the cumulative distribution function (cdf) $F_X(x)$.

$$f_{X_1}(x) = f_{X_2}(x) = \ldots = f_{X_n}(x) = f_X(x)$$

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x)$$

This assumption simplifies statistical analysis since we can treat each observation in the sample in the same way, knowing they all come from the same distribution.

4.2.3 Significance

The concept of i.i.d. samples is foundational in statistics, as it underpins many statistical techniques, including hypothesis testing, confidence intervals, and regression analysis. When we assume that our data is i.i.d., we can make stronger statistical inferences and draw more reliable conclusions about the underlying population.

In future, we will delve deeper into the mathematics behind i.i.d. samples, explore the central limit theorem, discuss its implications in practical data analysis, and demonstrate its application in various statistical methods.

5 Mathematics of Linear Models

Linear models are the work horse of statistics and (supervised) machine learning. When augmented with kernels or other forms of basis function expansion, it can model also non- linear relationships. And when the Gaussian output is replaced with a Bernoulli or multinoulli distribution, it can be used for classification. The first few weeks of this course we will focus almost exclusively on these models, and they will certainly appear on your exam.

5.1 Linear Regression Model

Linear regression is exactly what the name suggests: performing regression with a linear model. We assume a supervised setting where we have access to a dataset $\mathcal{D}$, as detailed above. More than just fitting the labels y, we assume some Gaussian noise. Assuming domain of $\mathbb{R}^n$ and a one dimensional co-domain we can write our model as $f(x) = x^\top \theta$ and we will model our noise as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Thus our full model can be written as:

$$\hat{y}^{(i)} = x^{(i)\top}\theta + \epsilon$$

The key idea behind learning is to find θ such that $\hat{y}^{(i)} \approx y^{(i)}$. This will be the primary topic of the next lecture.

5.2 Logistic Regression Model

In logistic regression, we have a labelled dataset where the codomain is $\mathbb{Z}^2$, we will denote the negative class with 0 and 1 for positive class. The logistic regression model aims to model the probability of the positive class, denoted as $P(Y = 1)$, where Y is the binary outcome variable. This probability is expressed as follows:

$$P(Y = 1|x^{(i)}) = \frac{1}{1 + e^{-(\theta_0 + \sum_{j=1}^{M} \theta_j x_j^{(i)})}}$$

Here, $\langle \theta_0, \theta_1, \ldots, \theta_M \rangle$ are the logistic regression coefficients that we aim to estimate. The logistic function, denoted by $e^{-(\beta_0 + \sum_{j=1}^{M} \beta_j x_j^{(i)})}$, transforms the linear combination of features into a normalized value between 0 and 1 which is inturn interpreted as a probability.

6 Reference Material

- Chapter 1 of Kevin Murphy, Machine Learning: A probabilistic perspective. We didn't cover 1.4 which may be of interest.

A Warm-up Exercises

To start, here are some exercises which test knowledge which is assumed in the course.

A.1 Probability Theory

We assume that you are familiar with probability theory up to the Computing 2nd year `50008` *Probability & Statistics* course. Here are some questions to serve as a refresher. Students who are not familiar with this background should refer to the notes of `50008` *Probability & Statistics* or relevant chapters of [**mml**]. **We recommend you look at these questions when/before the course starts**. If you need a refresher, or if you do not know the notation, refer to the `50008` *Probability & Statistics* notes, or discuss with a TA.

1. $P(A) \geq 0$
2. $P(\Omega) = 1$
3. $P(A \cup B) = P(A) + P(B)$ if $P(A \cap B) = 0$

Question 1 (Set Theory and Probability)**.** Using the three axioms of probability show that

a. Write down the sample space of a dice. In your notation, use the set A to denote the event of a 3 or 4 occurring. What is the complement of A, denoted $\neg A$? $\{1,2,3,4,5,6\}$ · $A = \{3,4\}$, $\bar{A} = \{1,2,5,6\}$

b. For a problem about lengths, we have a sample space $\Omega = [0, 1]$. For $A = (0.3, 0.4]$, what is $\neg A$? $[0, 0.3]$, $(0.4, 1]$

c. $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$

d. $\mathbb{P}(\varnothing) = 0$, where $\varnothing$ is the empty set

e. $0 \leq \mathbb{P}(A) \leq 1$

f. $A \subseteq B \implies \mathbb{P}(A) \leq \mathbb{P}(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

g. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

h. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_i \subseteq A_{i+1} \forall i$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} \mathbb{P}(A_i)$$

Hint: Use axiom 3. *: The emphasis of this course isn't on these kinds of details, even though this should be doable with 1st-year calculus.

i. For two mutually exclusive events A, B, what is $\mathbb{P}(A \cup B)$?

Question 2 (Independent events)**. Independent events don't come up as much as independent random variables, so it's ok to just follow this answer, rather than spending lots of time on it.** When tossing two coins (where we care about the order), we have a sample space $\Omega = \{HH, HT, TH, TT\}$.

a. What outcomes are contained in the event that corresponds to the the first coin being heads? We denote the event E_{1H}, and others similarly.

b. If you assume that all outcomes have equal probability, show that E_{1H} and E_{2T} are independent.

c. If you assume that E_{1H} and E_{2H} are independent and 0.5 each, show that all outomes must have equal probability.

Question 3 (Random Variables)**.** Consider throwing two fair dice.

a. What is the sample space for all outcomes that you can get from throwing two dice? We specify the probability of each outcome to be the same.

b. Define two random variables A, B which map the outcome to the face value on each die respectively. Find the probability mass function for A from the probability on outcomes. The answer will work from the definition of a random variable, but you will probably intuitively get the right answer as well.

c. Show that A and B are independent.

d. Define the random variable $C = A + B$. Derive the probability mass function of C.

Question 4 (Continuous Random Variables)**.** Consider the random variable X with a probability density $p(x) = C \cdot x$ when $x \in [0, 1]$ and 0 elsewhere.

a. Calculate C.

b. Calculate $\mathbb{P}(0.3 \leq X \leq 0.75)$.

c. Calculate $\mathbb{P}(X \in [0.3, 0.75] \cup [0.8, 0.9])$.

d. Calculate $\mathbb{E}_X[X]$, $\mathbb{E}_X[X^2]$, $\mathbb{V}_X[X]$.

Check your answers by performing numerical integration, e.g. in Python.

Question 5 (Joint Discrete Random Variables)**.** Consider two random variables A, C, where A is the outcome of one die, and C gives the sum of A and the sum of another die B.

a. From intuition, write a table of $\mathbb{P}(C = c|A = a)$, which we use to denote the probability of C taking the value c, if we know that A has taken the value a.

b. Write a table of $\mathbb{P}(C = c, A = a)$. To help you think it through, consider a tree of outcomes that can occur. This helps illustrate independence between outcomes, which helps you figure out when you can multiply probabilities.

c. From the values in the table $\mathbb{P}(C = c, A = a)$ find $\mathbb{P}(2 \leq C \leq 4)$ and $\mathbb{P}(2 \leq C \leq 4, 2 \leq A \leq 4)$.

We will cover conditional probability more later, but for now just think it through.

Question 6 (Multivariate Integration)**.** Consider two continuous random variables X, Y with joint density $p(x, y) = C \cdot (x^2 + xy)$ when $x \in [0, 1]$ and $y \in [0, 1]$, and 0 elsewhere.

a. Find C.

b. Find $\mathbb{P}(0.3 \leq X \leq 0.5)$.

c. Find $\mathbb{P}(X < Y)$. Perform the integration twice in both orders, once integrating over x first, once by integrating over y first.

d. **Bonus:** Convince yourself that you know how to do this for $p(x, y, z) = C \cdot (x^2 + xyz)$ as well.

Check your answers by performing numerical integration, e.g. in Python.

Question 7 (Statistics Terminology)**.** Recall the following statistical terminology.

a. What is a statistic? f^n computed from data

b. What is an estimator? f^n that estimates a parameter

c. What is a consistent estimator? correct when $n \to \infty$

d. What is a sample? outcome for a RV.

Question 8 (Vector notation)**.** We define the probability density on the vector $\mathbf{x} \in \mathbb{R}^3$ with all elements $0 \leq x_k \leq 1$ as

$$p(\mathbf{x}) = \frac{1}{C}(x_1^2 + x_1 x_2 + x_2^2 + 2x_2 x_3)\,. \qquad (2)$$

$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$

Put this into notation that only uses $\mathbf{x}$ as a single whole vector.

Question 9 (Noise conditional independence)**.** Consider the probability of the data in linear regression, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs $\mathrm{X} \in \mathbb{R}^{D \times N}$ where $\mathrm{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$:

$$p(\mathbf{y}|\boldsymbol{\theta}, \mathrm{X}) = \mathcal{N}\big(\mathbf{y}; \boldsymbol{\theta}^\intercal X, \sigma^2 \mathrm{I}\big) \qquad (3)$$

Show that all y_ns are independent, for a fixed setting of the parameters $\boldsymbol{\theta}$ and given inputs X.

Question 10 (Maximum likelihood revision)**.** For a Gaussian distribution with mean μ and variance σ^2.

a. Derive the probability distribution for N iid draws.

b. Derive the maximum likelihood estimator for the mean μ and variance σ^2.

Question 11 (Maximum likelihood and minimum loss)**.** Show that the solution to the Maximum Likelihood estimator for linear regression is the same as the minimum squared loss estimator.

Question 12 (MML 5.1-5.3)**.** This is revision. Compute the derivatives for w.r.t. x for

a. $f(x) = \log(x^4)\sin(x^3)$

b. $f(x) = (1 + \exp(-x))^{-1}$

c. $f(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

A.2 Linear Algebra

Question 13 (Dot product)**.** Compute $\boldsymbol{x}^\top \boldsymbol{y}$ where $\boldsymbol{x} = (1, -2, 5, -1)^\top$ and $\boldsymbol{y} = (0, 4, -3, 7)^\top$.

Question 14 (Matrix product)**.** Compute $\boldsymbol{y} = A\boldsymbol{x}$ as well as the ℓ_2 norm of $\boldsymbol{x}$ and $\boldsymbol{y}$, where

$$A = \begin{pmatrix} -1 & 4 & 7 & 2 \\ 3 & -2 & -1 & 0 \\ 5 & 3 & 0 & -1 \end{pmatrix}, \quad \boldsymbol{x} = (-3, 2, 1, 3)^\top.$$

Question 15 (Basis)**.** Which of the following set of vectors are basis for $\mathbb{R}^2$?

1. $\{(1, 1), (1, 0)\}$
2. $\{(2, 4), (3, -1)\}$
3. $\{(1, -1), (0, 2), (2, 1)\}$
4. $\{(2, -1), (-2, 1)\}$
5. $\{(0, 3)\}$

Question 16 (Span of vectors)**.** Which of the following points are within the span of $\{(-1, 0, 2), (3, 1, 0)\}$?

1. $(0, 1, 1)$
2. $(1, 1, 4)$
3. $(2, 1, 1)$
4. $(-3, 4, 2)$
5. $(0, 0, 0)$

Question 17 (Rotation matrix in $\mathbb{R}^2$)**.** What is the 2×2 matrix that rotates all the non-zero vectors in $\mathbb{R}^2$ by $45°$ counter-clockwise?

Question 18 (Linear equations)**.** Given the following system of linear equations:

$$\begin{aligned} x + 2y &= 2 \\ 3x + 2y + 4z &= 5 \\ -2x + y - 2z &= -1 \end{aligned}$$

Answer the following questions:

a Writing this system in a matrix form $A\boldsymbol{x} = \mathbf{b}$ with $\boldsymbol{x} = (x, y, z)^\top$. What are A and $\mathbf{b}$?

b Solve this system, or show that the solution does not exist.

c What is the rank of A?

Question 19 (Eigen decomposition)**.** Consider a matrix $A \in \mathbb{R}^{d\times d}$ and assume it has an eigen decomposition of $A = Q\Lambda Q^{-1}$ where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$. When A is symmetric we also have $Q^{-1} = Q^\top$. Answer the following questions:

a. If A is symmetric, show that $\boldsymbol{x}^\top A\boldsymbol{x} \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b. Show that $Tr(A) = \sum_{i=1}^d \lambda_i$ where $Tr(A)$ is the trace of A.

c. Show that $det(A) = \prod_{i=1}^d \lambda_i$ where $det(A)$ is the determinant of A.

d. Why an entry λ_i in the diagonal matrix Λ is one of the solutions for the equation $A\boldsymbol{q} = \lambda\boldsymbol{q}$, $\boldsymbol{q} \neq \mathbf{0}$?

Question 20. This week we discussed and defined several machine learning problem settings. In our lecture we talked about self-driving cars. We discussed object recognition as classification and steering prediction as regression. As an open research question I would like graduate students to (first without researching) think of how to formulate the route planning problem setting. Fully flesh the aspect of a model we discussed: input space, output space, loss function, and objective. Then do some research and check how close you got to what is currently state-of-the-art.

Question 3 (Random Variables)**.** Consider throwing two fair dice.

a. What is the sample space for all outcomes that you can get from throwing two dice? We specify the probability of each outcome to be the same.

b. Define two random variables A, B which map the outcome to the face value on each die respectively. Find the probability mass function for A from the probability on outcomes. The answer will work from the definition of a random variable, but you will probably intuitively get the right answer as well.

c. Show that A and B are independent.

d. Define the random variable $C = A + B$. Derive the probability mass function of C.

a. (1,1) (1,2) (1,3) (1,4) (1,5) (1,6)
(2,1)
⋮
(6,1) - - - (6,6)

b. $P(A=1) = \frac{1}{6} = P(A=2) = P(A=3) = \cdots$ same for B

$P_A(x) = \frac{1}{6}$, $x \in \{1,2,3,4,5,6\}$; 0 else

$P_B(x) = \frac{1}{6}$, $x \in \{1,2,3,4,5,6\}$; 0 else

c. $P(A \cap B) = P(A)P(B)$ if indep

$P(A)P(B) = \frac{1}{36}$, $x \in \{1\text{-}6\}$, $P(A \cap B) = \frac{1}{36}$ ∴ indep

d. $C = A + B$

no. of ways	2	3	4	5	6	7	8	9	10	11	12
	1	2	3	4	5	6	5	4	3	2	1

$P_C(x) = \frac{1}{36}$, $x = \{2,12\}$
$\frac{1}{18}$, $x = \{3,11\}$
$\frac{1}{12}$, $x = \{4,10\}$
$\frac{1}{9}$, $x = \{5,9\}$
$\frac{5}{136}$, $x = \{6,8\}$
$\frac{1}{6}$, $x = \{7\}$

Question 4 (Continuous Random Variables)**.** Consider the random variable X with a probability density $p(x) = C \cdot x$ when $x \in [0, 1]$ and 0 elsewhere.

a. Calculate C.

b. Calculate $\mathbb{P}(0.3 \leq X \leq 0.75)$.

c. Calculate $\mathbb{P}(X \in [0.3, 0.75] \cup [0.8, 0.9])$.

d. Calculate $\mathbb{E}_X[X], \mathbb{E}_X[X^2], \mathbb{V}_X[X]$.

a. $p_X(x) = \begin{cases} Cx & , x \in [0,1] \\ 0 & , \text{else} \end{cases}$

$\int_0^1 Cx\,dx = 1 \quad : \quad \left[\frac{C}{2}x^2\right]_0^1 = \frac{C}{2} = 1 \quad \rightarrow \quad C = 2$

b $\int_{0.3}^{0.75} 2x\,dx = \quad \left[x^2\right]_{0.3}^{0.75} = 0.4725$

c. $\int_{0.7}^{0.9} 2x = \quad \left[x^2\right]_{0.7}^{0.9} = 0.16$

$+ 0.4725 = 0.6325$

d. $E_X[X] = \int x f_X(x)\,dx \rightarrow \int_0^1 2x^2\,dx = \left[\frac{2}{3}x^3\right]_0^1 = \frac{2}{3}$

$E_X[X^2] = \int_0^1 2x^3 = \left[\frac{1}{2}x^4\right]_0^1 = \frac{1}{2}$

$\mathrm{Var}[X] = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$

Question 6 (Multivariate Integration)**.** Consider two continuous random variables X, Y with joint density $p(x, y) = C \cdot (x^2 + xy)$ when $x \in [0, 1]$ and $y \in [0, 1]$, and 0 elsewhere.

a. Find C.

b. Find $\mathbb{P}(0.3 \leq X \leq 0.5)$.

c. Find $\mathbb{P}(X < Y)$. Perform the integration twice in both orders, once integrating over x first, once by integrating over y first.

d. **Bonus:** Convince yourself that you know how to do this for $p(x, y, z) = C \cdot (x^2 + xyz)$ as well.

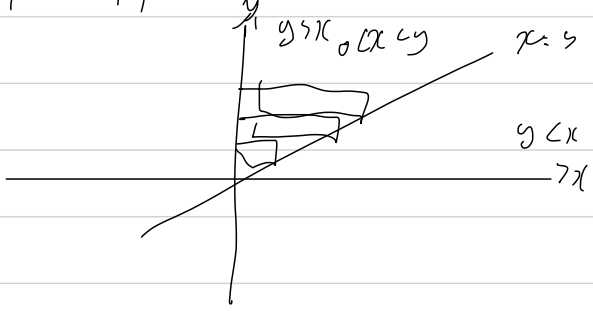
a. $C\int_0^1\int_0^1 x^2 + xy\,dx\,dy = C\int_0^1 \left[\frac{x^3}{3} + \frac{1}{2}x^2 y\right]_0^1 dy = C\int_0^1 \frac{1}{2}y + \frac{1}{3}\,dy = C\left[\frac{1}{4}y^2 + \frac{1}{3}y\right]_0^1$

$= \frac{1}{4} + \frac{1}{3} = \frac{7}{12}C = 1 \quad C = \frac{12}{7}$ //

b. Marginal: $f_X(x) = \frac{12}{7}\int_0^1 x^2 + xy\,dy = \frac{12}{7}\left[x^2y + \frac{1}{2}xy^2\right]_0^1 = \frac{12}{7}\left(x^2 + \frac{1}{2}x\right)$

$\frac{12}{7}\int_{0.3}^{0.5} x^2 + \frac{1}{2}x\,dx$. $\left[\frac{x^3}{3} + \frac{x^2}{4}\right]_{0.3}^{0.5} = \frac{109}{875}$

c. $P(X<Y)$



$\frac{12}{7}\int_0^1 \int_0^y x^2 + xy\,dx\,dy$

$= \left[\frac{x^3}{3} + \frac{1}{2}yx^2\right]_0^y = \frac{1}{3}y^3 + \frac{1}{2}y^3 = \frac{5}{6}y^3$

$\frac{12}{7}\int_0^1 \frac{5}{6}y^3\,dy = \frac{12}{7}\left[\frac{5}{24}y^4\right]_0^1 = \frac{5}{24} \times \frac{12}{7} = \frac{5}{14}$

Question 18 (Linear equations)**.** Given the following system of linear equations:

$$\begin{aligned} x + 2y &= 2 \\ 3x + 2y + 4z &= 5 \\ -2x + y - 2z &= -1 \end{aligned}$$

Answer the following questions:

a Writing this system in a matrix form $A\boldsymbol{x} = \mathbf{b}$ with $\boldsymbol{x} = (x, y, z)^\top$. What are A and $\mathbf{b}$?

b Solve this system, or show that the solution does not exist.

c What is the rank of A?

a $A \begin{bmatrix} x \\ y \\ z \end{bmatrix} = b$

$$\begin{bmatrix} 1 & 2 & 0 \\ 3 & 2 & 4 \\ -2 & 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix}$$

b $\det(A) = \begin{vmatrix} 2 & 4 \\ 1 & -2 \end{vmatrix} - \begin{vmatrix} 3 & 4 \\ -2 & -2 \end{vmatrix} = -8 - 2 = -10$

mat. of minors $\begin{bmatrix} -8 & 2 & 7 \\ -4 & -2 & 5 \\ 8 & 4 & -14 \end{bmatrix}$ → cofactor $\begin{bmatrix} -8 & -2 & 7 \\ 4 & -2 & -5 \\ 8 & -4 & -4 \end{bmatrix}$ → transpose $\begin{bmatrix} -8 & 4 & 8 \\ -2 & -2 & -4 \\ 7 & -5 & -4 \end{bmatrix}$

$A^{-1} = \frac{1}{12} \begin{bmatrix} -8 & 4 & 8 \\ \cdots & \cdots & \cdots \end{bmatrix}$

full rank as det ≠ 0 ; idp cols.

Question 19 (Eigen decomposition)**.** Consider a matrix $A \in \mathbb{R}^{d \times d}$ and assume it has an eigen decomposition of $A = Q\Lambda Q^{-1}$ where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$. When A is symmetric we also have $Q^{-1} = Q^\top$. Answer the following questions:

a. If A is symmetric, show that $\boldsymbol{x}^\top A\boldsymbol{x} \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^{d \times 1}$ if and only if $\lambda_i \geq 0$ for all $i = 1, ..., d$.

b. Show that $Tr(A) = \sum_{i=1}^{d} \lambda_i$ where $Tr(A)$ is the trace of A.

c. Show that $det(A) = \prod_{i=1}^{d} \lambda_i$ where $det(A)$ is the determinant of A.

d. Why an entry λ_i in the diagonal matrix Λ is one of the solutions for the equation $A\boldsymbol{q} = \lambda\boldsymbol{q}$, $\boldsymbol{q} \neq \mathbf{0}$?

a. A symmetric

$$x^\top A x = x^\top Q \Lambda Q^{-1} x = x^\top Q \Lambda Q^\top x = (Q^\top x)^\top \Lambda (Q^\top x)$$