1 Review of probability theory

1.1 Why probability theory?

- Probability theory provides mathematical framework for reasoning about prediction problems
- (Some alternatives: approximation theory, game theory, ...)
- Basic idea: regard quantities you are uncertain about (e.g., quantities you want to predict) as random variables defined on a probability space
- Starting from basic idea, can use probability theory to derive properties of optimal predictions, characterize uncertainty of error rate estimates, design and analyze learning algorithms, etc.

1.2 Probability spaces

- Goal: mathematical model for experiment with random outcomes

  (E.g., coin tosses, dice rolls, roulette wheel spins, ...)
- A (discrete) *probability space* (Ω, m) is comprised of a sample space Ω and a probability (mass) function m
  - *Sample space* Ω is the (finite or countable) set of possible outcomes
  - An *event* is a subset of Ω
  - Example: toss a coin
    - Possible outcomes: $\Omega = \{\mathsf{H}, \mathsf{T}\}$
    - "heads" $= \{\mathsf{H}\}$
    - "tails" $= \{\mathsf{T}\}$
    - ...
  - Example: toss a coin twice
    - Possible outcomes: $\Omega = \{\mathsf{TT}, \mathsf{TH}, \mathsf{HT}, \mathsf{HH}\}$
    - "both tails" $= \{\mathsf{TT}\}$
    - "at least one heads" $= \{\mathsf{TH}, \mathsf{HT}, \mathsf{HH}\}$
    - ...

- Example: roll a 6-sided die
  - Possible outcomes: $\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$
  - "odd" $= \{$⚀, ⚂, ⚄$\}$
  - "even" $= \{$⚁, ⚃, ⚅$\}$
  - "at most 3" $= \{$⚀, ⚁, ⚂$\}$
  - ...
- Example: repeatedly roll a 6-sided die and stop after seeing a "6"
  - Possible outcomes: $\Omega = \{$⚅, ⚀⚅, ⚁⚅, ⚂⚅, ⚃⚅, ⚄⚅, ⚀⚀⚅, ⚀⚁⚅, $\dots\}$
  - "one roll" $= \{$⚅$\}$
  - "two rolls" $= \{$⚀⚅, ⚁⚅, ⚂⚅, ⚃⚅, ⚄⚅$\}$
  - ...
- <u>*Probability (mass) function*</u> m is a function that assigns a real number to each outcome in Ω in a way that satisfies
  - $m(\omega) \geq 0$ for all $\omega \in \Omega$ (non-negativity), and
  - $\sum_{\omega \in \Omega} m(\omega) = 1$ (normalization)
- <u>*Probability of an event*</u> $E \subseteq \Omega$ in probability space (Ω, m) is

  $$\Pr(E) = \sum_{\omega \in E} m(\omega)$$

  (Notation unfortunately does not explicitly show (Ω, m))
- Sometimes we "abuse notation" by writing $m(E)$ to mean $\Pr(E)$
- Example: toss a fair coin twice
  - $m(\omega) = 1/4$ for every possible outcome ω

  $$\Pr(\text{tosses come up on same side}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

- Some events can be described in terms of other events using set theory
  - Union ("or")

    $$A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$$

  - Intersection ("and")

    $$A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$$

– Complement ("not")

$$A^{\mathsf{c}} = \{\omega \in \Omega : \omega \notin A\}$$

– Difference ("and not")

$$A - B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$$

(sometimes also written "$A \setminus B$"; same as $A \cap B^{\mathsf{c}}$)

– Example: roll a fair 6-sided die twice

 * $A =$ "first roll is even"
 * $B =$ "second roll is at most 3"
 * $A^{\mathsf{c}} =$ "first roll is odd"
 * $B^{\mathsf{c}} =$ "second roll is at least 4"
 * "first roll is even and second roll is at most 3"

$$A \cap B = \{\text{⚁⚀}, \text{⚁⚁}, \text{⚁⚂}, \text{⚃⚀}, \text{⚃⚁}, \text{⚃⚂}, \text{⚅⚀}, \text{⚅⚁}, \text{⚅⚂}\}$$

so

$$\Pr(A \cap B) = \frac{9}{36} = \frac{1}{4}$$

 * "first toss is even or second toss is at most 3"

$$\begin{aligned} A \cup B &= (A^{\mathsf{c}} \cap B^{\mathsf{c}})^{\mathsf{c}} \\ &= \Omega - (A^{\mathsf{c}} \cap B^{\mathsf{c}}) \\ &= \Omega - \{\text{⚀⚃}, \text{⚀⚄}, \text{⚀⚅}, \text{⚂⚃}, \text{⚂⚄}, \text{⚂⚅}, \text{⚄⚃}, \text{⚄⚄}, \text{⚄⚅}\} \end{aligned}$$

so

$$\Pr(A \cup B) = \frac{36 - 9}{36} = \frac{3}{4}$$

Q. Suppose a 6-sided die is weighted so that, for each $k \in \{1, 2, 3, 4, 5, 6\}$, the side showing k pips is k times as likely to show up as the side showing 1 pip. What is the probability that a roll of this die shows an even number of pips?

Q. Suppose A and B are events from a probability space such that $\Pr(A \cap B) = 1/4$, $\Pr(A^{\mathsf{c}}) = 1/3$, and $\Pr(B) = 1/2$. What is $\Pr(A \cup B)$?

1.3 Conditional probability

- Suppose, in an experiment described by probability space (Ω, m), you learn that an event E has occurred, but nothing else
  - Exact outcome ω may not yet be known to you
  - What probability space now models the experiment in light of the new information?
  - *Conditioning on E*: incorporating information that E occurred
- New probability space (Ω, m_E) with probability function defined in terms of original:

$$m_E(\omega) = \begin{cases} \dfrac{m(\omega)}{m(E)} & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E \end{cases}$$

  (We require $m(E) > 0$ in order to define m_E)
- Can check that m_E is a valid probability function on Ω

  (Normalization is ensured by the division by $\Pr(E)$)
- Notation: write $\Pr(F \mid E)$ for probability of event F in probability space (Ω, m_E), a.k.a. *probability of F conditioned on E*, a.k.a. *(conditional) probability of F given E*
- Example: roll a fair 6-sided die
  - $E = \{$⚁, ⚃, ⚅$\}$ = "even", $F = \{$⚁, ⚂, ⚄$\}$ = "prime"
  - Suppose you learn E occurred
    - Given this information, what is probability of F?

$$\Pr(F \mid E) = \sum_{\omega \in F} m_E(\omega) = \sum_{\omega \in F \cap E} \frac{m(\omega)}{\Pr(E)} = \frac{1/6}{1/2} = \frac{1}{3}$$

- Useful formula for conditional probability:

$$\Pr(F \mid E)\Pr(E) = \Pr(F \cap E)$$

- *Bayes' rule*: relates probabilities of event F before and after conditioning on information that event E occurs

$$\Pr(F \mid E) = \Pr(F) \times \frac{\Pr(E \mid F)}{\Pr(E)}$$

  - $\Pr(F \mid E)$ is probability of F after conditioning on information that E occurred
  - $\Pr(F)$ is probability of F in original probability space (before observing that E occurred)
  - Ratio $\Pr(E \mid F)/\Pr(E)$ is what relates these probabilities
    * Always non-negative, but can be zero (even if $\Pr(E) > 0$)
    * Whether it is more or less than 1 determines whether probability of F increases or decreases after incorporating information that E occurred

- Example: A casino has 100 identically-looking slot machines; but unbeknownst to you, the first 75 are "fair", and rest are "rigged". If you play on a "fair" machine, you are equally likely to win or lose. If you play on a "rigged" machine, you always lose.

  Suppose you enter the casino, pick a slot machine uniformly at random, play it once, and lose. Given this information, what is the probability that you played on a "rigged" machine?

  - Sample space: $\Omega = \{1, 2, \ldots, 100\} \times \{\mathsf{win}, \mathsf{lose}\}$
    (Other choices could also work)
  - Events of interest:
    * $R = \{(a, b) \in \Omega : 76 \leq a \leq 100\}$
    * $L = \{(a, b) \in \Omega : b = \mathsf{lose}\}$
  - Probabilities of interest:
    * $\Pr(R) = 25/100$, $\Pr(R^{\mathsf{c}}) = 75/100$
    * $\Pr(L \mid R) = 1$, $\Pr(L \mid R^{\mathsf{c}}) = 1/2$

* We also need $\Pr(L)$:

$$\begin{aligned}\Pr(L) &= \Pr(L \cap R) + \Pr(L \cap R^{\mathsf{c}}) \\ &= \Pr(L \mid R) \times \Pr(R) + \Pr(L \mid R^{\mathsf{c}}) \times \Pr(R^{\mathsf{c}}) \\ &= 1 \times \frac{25}{100} + \frac{1}{2} \times \frac{75}{100} \\ &= \frac{1}{4} + \frac{3}{8} = \frac{5}{8}\end{aligned}$$

– Using Bayes' rule:

$$\begin{aligned}\Pr(R \mid L) &= \Pr(R) \times \frac{\Pr(L \mid R)}{\Pr(L)} \\ &= \frac{1}{4} \times \frac{1}{5/8} \\ &= \frac{2}{5} = 40\%\end{aligned}$$

– Before playing the machine, 25% probability that picked machine is rigged; after playing machine and observing that you lost, the probability has increased to 40%

Q. In the casino example, suppose you play a randomly picked machine two times, and lose both times. What is probability that you picked a rigged machine, given this information?

Q. You repeatedly roll a fair 6-sided die and stop after seeing 6 pips face up. Suppose only even numbers of pips show up in all rolls. What is the probability that the number of rolls is 1, given this information?

1.4 Random variables

- *Random variable* X (on (Ω, m)) is a function that assigns a real number to each outcome in Ω

  – Facilitates quantitative analysis of experiments modeled by probability spaces

  – X defines probability space $(\mathbb{R}, p_X)$ with p_X defined by

$$p_X(x) = \Pr(X = x) = \Pr(\{\omega \in \Omega : X(\omega) = x\})$$

    * p_X is the probability (mass) function for X
    * Say p_X specifies the *probability distribution* of X
    * Sample space is set of real numbers $\mathbb{R}$, but probability function p_X takes values 0 outside of *range* of X

$$\operatorname{range}(X) = \{X(\omega) : \omega \in \Omega\}$$

    * So can also regard sample space as $\operatorname{range}(X)$
    * Shorthand: "$X = x$" means $\{\omega \in \Omega : X(\omega) = x\}$

  - Example: toss a fair coin three times
    * X = number of heads

$$\begin{aligned} X(\mathsf{TTT}) &= 0 & X(\mathsf{TTH}) &= 1 \\ X(\mathsf{THT}) &= 1 & X(\mathsf{THH}) &= 2 \\ X(\mathsf{HTT}) &= 1 & X(\mathsf{HTH}) &= 2 \\ X(\mathsf{HHT}) &= 2 & X(\mathsf{HHH}) &= 3 \end{aligned}$$

x	0	1	2	3
$p_X(x)$	1/8	3/8	3/8	1/8

    * Event "at least one heads" is also written as "$X \geq 1$"

$$\Pr(X \geq 1) = \frac{7}{8}$$

  - Example: roll a fair 6-sided die twice
    * X = number of pips from the first roll
    * Y = number of pips from the second roll
    * $Z = X + Y$

$$\begin{aligned} Z(\text{⚀⚀}) &= X(\text{⚀⚀}) + Y(\text{⚀⚀}) = 1 + 1 = 2 \\ Z(\text{⚀⚁}) &= X(\text{⚀⚁}) + Y(\text{⚀⚁}) = 1 + 2 = 3 \\ Z(\text{⚀⚂}) &= X(\text{⚀⚂}) + Y(\text{⚀⚂}) = 1 + 3 = 4 \\ &\text{etc.} \end{aligned}$$

- *Expectation* (a.k.a. *expected value*, *mean*, *average*) of random variable X in probability space (Ω, m)

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) m(\omega)$$

- Often more convenient to use equivalent formula

$$\mathbb{E}(X) = \sum_x x \Pr(X = x) = \sum_x x \, p_X(x)$$

(Summation is taken over $x \in \text{range}(X)$)

- Example: roll a fair 6-sided die
  - X = number of pips

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=1}^{6} x \, p_X(x) \\ &= 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} \\ &= \frac{21}{6} = 3.5\end{aligned}$$

- Example: toss a fair coin three times
  - X = number of heads

$$\begin{aligned}\mathbb{E}(X) &= \sum_{x=0}^{3} x \, p_X(x) \\ &= 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} \\ &= \frac{12}{8} = 1.5\end{aligned}$$

- If X is a random variable, and $Y = aX + b$ for some real numbers a and b, then

$$\mathbb{E}(Y) = \mathbb{E}(aX + b) = a \, \mathbb{E}(X) + b$$

- Caution: not all random variables have an expectation
  - Example: $p_X(x) = 1/x - 1/(x+1)$ for all positive integers x

• Beyond the expected value

- Random variables with same expected value can be very different
- Example:

* Toss fair coin 5 times; X = number of heads, $\mathbb{E}(X) = 2.5$
  - range$(X) = \{0, 1, 2, 3, 4, 5\}$
  - $|\{\omega \in \Omega : X(\omega) \in \{2, 3\}\}| = 20$
  - $|\{\omega \in \Omega : X(\omega) \in \{0, 1, 4, 5\}\}| = 12$
  - So values "close" to the expectation are more likely than those "far" from the expectation
* Roll a fair 6-sided die; Y = number of pips $- 1$, $\mathbb{E}(Y) = 2.5$
  - All possible values $\{0, 1, 2, 3, 4, 5\}$ of Y are equally likely, regardless of distance to the expectation
* X is less "spread out" than Y

– *Variance*: convenient measure of a random variable's "spread"

$$\mathrm{var}(X) = \mathbb{E}((X - \mu)^2)$$

where $\mu = \mathbb{E}(X)$

* The square-root of var(X)—called *standard deviation*—is roughly how much X deviates from μ on average
  - stddev$(X) = \sqrt{\mathrm{var}(X)}$
  - Caveat: $\sqrt{\mathbb{E}(X^2)}$ is not necessarily the same as $\mathbb{E}(\sqrt{X^2})$
* $\mathbb{E}(|X - \mu|)$ is exactly how much X deviates from μ on average, but less convenient to work with mathematically

– If X is a random variable, and $Y = aX + b$ for some real numbers a and b, then

$$\mathrm{var}(Y) = \mathrm{var}(aX + b) = \mathrm{var}(aX) = a^2 \,\mathrm{var}(X)$$

– There are many other "summary statistics" for random variables

Q. You repeatedly roll a fair 6-sided die and stop after seeing 6 pips face up. What is the expected number of rolls?

Q. If X is the number of heads in 5 tosses of a fair coin, and Y is number of pips shown in the roll of a fair 6-sided die, what are the variances of X and Y?

1.5 Multiple random variables

- If each of X and Y is a random variable (on (Ω, m)), then (2-dimensional) *random vector* $Z = (X, Y)$ is a $\mathbb{R}^2$-valued function on Ω given by

$$Z(\omega) = (X(\omega), Y(\omega))$$

  - Z defines probability space $(\mathbb{R}^2, p_Z)$ by

$$p_Z(x, y) = \Pr(X = x \wedge Y = y)$$

    p_Z is *joint probability function* for (X, Y)
  - Example: roll a fair 6-sided die 10 times
    * X = number of rolls with 6 pips
    * Y = number of rolls with 5 pips
  - Can generalize to n-tuples of random variables to get n-dimensional random vectors

- Useful fact: If X and Y are random variables (on (Ω, m)), then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

  i.e., expectation is additive

  - Example: roll a fair 6-sided die 10 times
    * X = number of rolls with 6 pips
    * Y = number of rolls with 5 pips
    * $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 5/3 + 5/3 = 10/3$
  - Generalizes to sums of n random variables $X_1, \ldots, X_n$

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$$

    and also linear combinations

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1 \mathbb{E}(X_1) + \cdots + a_n \mathbb{E}(X_n)$$

    i.e., expectation is linear

- Random variables X and Y are <u>*independent*</u> if, for all pairs of real numbers (x, y),

$$\Pr(X = x \wedge Y = y) = \Pr(X = x) \times \Pr(Y = y)$$

i.e.,

$$p_{(X,Y)}(x, y) = p_X(x) p_Y(y) \qquad \text{for all } (x, y)$$

  - Example: roll a fair 6-sided die
    * $X = \begin{cases} 1 & \text{if number of pips is at most 4} \\ 0 & \text{otherwise} \end{cases}$

$$p_X(0) = \frac{1}{3}, \quad p_X(1) = \frac{2}{3}$$

    * $Y = \begin{cases} 1 & \text{if number of pips is even} \\ 0 & \text{otherwise} \end{cases}$

$$p_Y(0) = p_Y(1) = \frac{1}{2}$$

    * Joint probability function

(x,y)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$p_{(X,Y)}(x,y)$	$1/6$	$1/6$	$1/3$	$1/3$

    * Check that this satisfies $p_{(X,Y)}(x, y) = p_X(x) p_Y(y)$ for all (x, y)
    * So X and Y are independent
    * Note: Here, X and Y are special kinds of random variables called <u>*indicator random variables*</u>—each one indicates whether or not a particular event occurs
    * Notation:

$$X = \mathbb{1}\{\text{number of pips is at most 4}\}$$
$$Y = \mathbb{1}\{\text{number of pips is even}\}$$

    * Distribution of an indicator random variable X is <u>*Bernoulli*</u>, written as $X \sim \text{Bernoulli}(\theta)$, where $\theta = \Pr(X = 1)$
  - A non-example: roll a fair 6-sided die 10 times

    * X = number of rolls with 6 pips
    * Y = number of rolls with 5 pips
    * $\Pr(X = 10 \wedge Y = 10) = 0$, yet

$$\Pr(X = 10) = \Pr(Y = 10) > 0$$

    * So X and Y are not independent

  - Generalizes to n random variables: $X_1, \ldots, X_n$ are *independent* if, for all n-tuples of real numbers $(x_1, \ldots, x_n)$,

$$\Pr(X_1 = x_1 \wedge \cdots \wedge X_n = x_n) = \Pr(X_1 = x_1) \times \cdots \times \Pr(X_n = x_n)$$

Q. Roll a fair 6-sided die; let X indicate if number of pips is at most 4, and let Y indicate if number of pips is even. Are X and Y independent?

Q. Toss a fair coin 10 times, and let X be the number times HTH appears as a substring of the outcome. What is the expected value of X? (Hint: Write X as a sum of 8 indicator random variables, and use the linearity of expectation.)

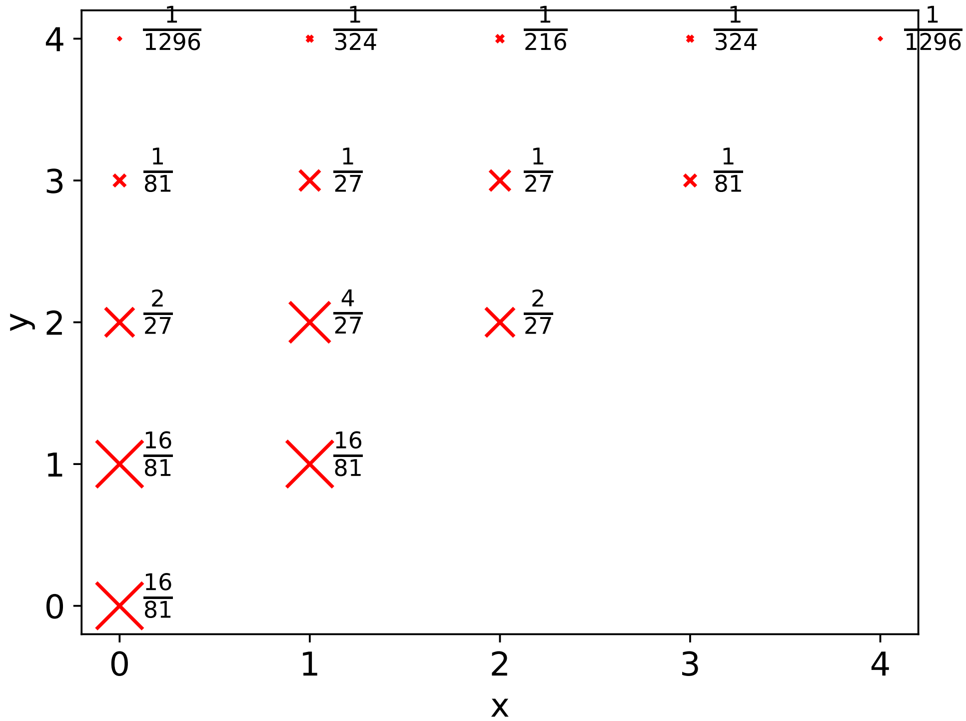
1.6 Dependence

- Random variables that are not independent are said to be *dependent*

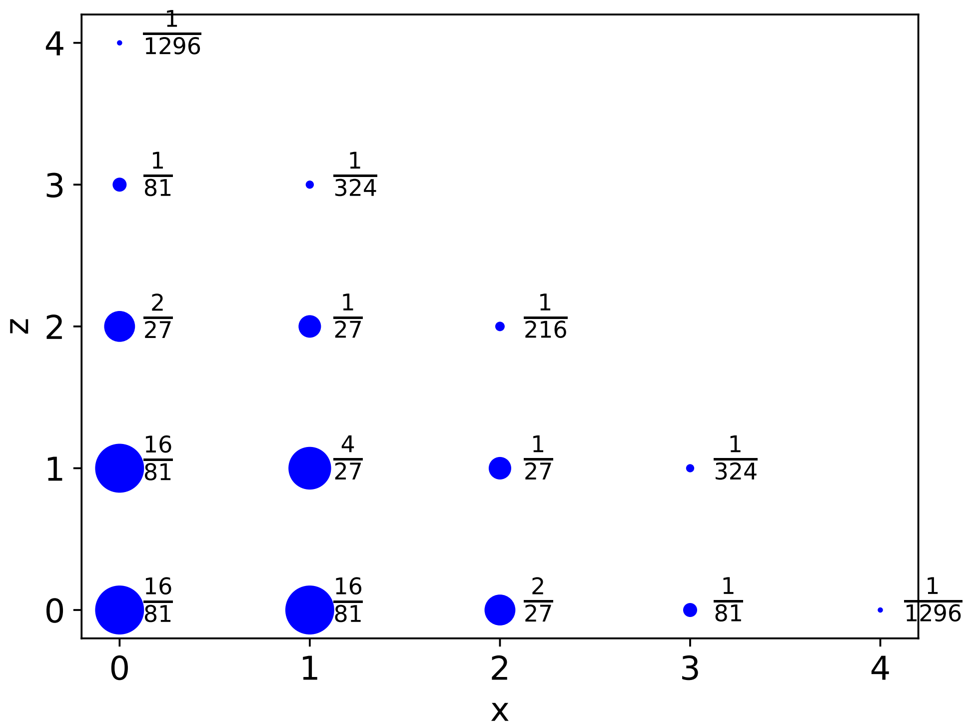
- Many different "types" of dependence

  - Example: Roll a fair 6-sided die n times; let X be the number of times a ⚃ comes up; let Y be the number of times a ⚃ or ⚄ comes up; let Z be the number of times a ⚅ comes up

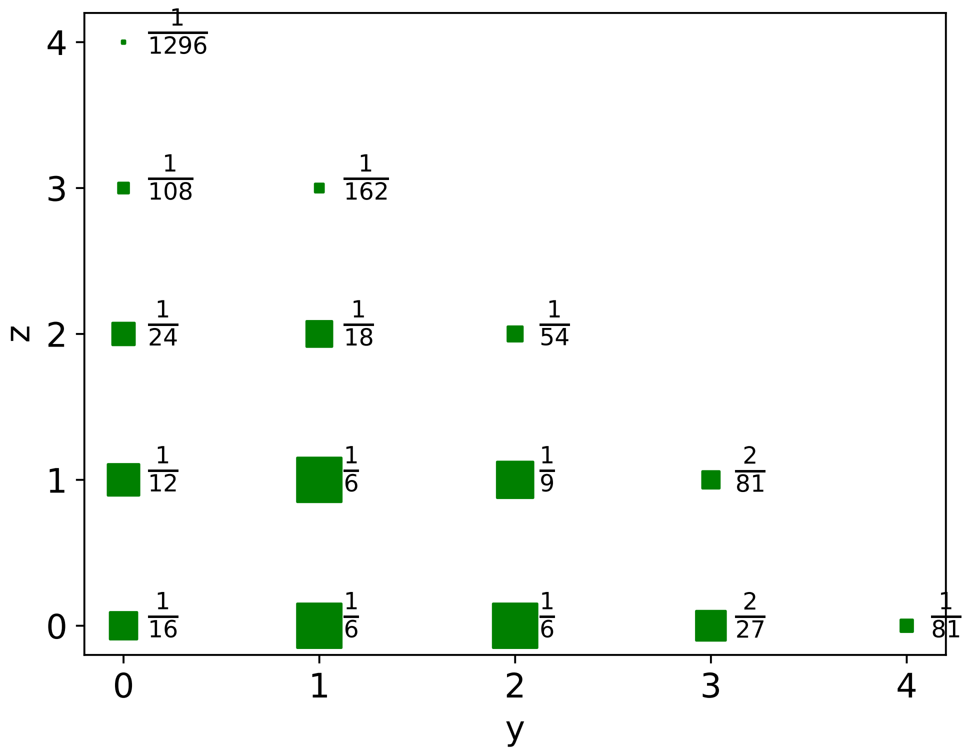
    * The larger X is, the larger Y must be
    * But the larger X or Y is, the smaller Z must be



y
x
0
1
2
3
4
1/1296
1/324
1/216
1/81
1/27
2/27
4/27
16/81



z
x
0
1
2
3
4
1/1296
1/81
1/324
2/27
1/27
1/216
16/81
4/27



- Say X and Y are *positively correlated* if $\mathbb{E}(XY) > \mathbb{E}(X)\,\mathbb{E}(Y)$
- In die rolling example with $n = 4$:

$$\begin{aligned} \mathbb{E}(XY) &= 4/3 \\ \mathbb{E}(X) &= 2/3 \\ \mathbb{E}(Y) &= 4/3 \\ \mathbb{E}(X)\,\mathbb{E}(Y) &= 8/9 \end{aligned}$$

  So X and Y are positively correlated
- Say X and Y are *negatively correlated* if $\mathbb{E}(XY) < \mathbb{E}(X)\,\mathbb{E}(Y)$
- Say X and Y are *uncorrelated* if $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$
- If X and Y are independent, then they are uncorrelated
  But converse is not necessarily true
- Example: toss a fair coin two times
  * X = number of heads
  * $Y = \begin{cases} 1 & \text{if first toss is heads and second toss is tails} \\ 0 & \text{if both tosses are the same} \\ -1 & \text{if first toss is tails and second toss is heads} \end{cases}$

  * $\mathbb{E}(X) = 1$, $\mathbb{E}(Y) = 0$, $\mathbb{E}(XY) = 0$
  * So X and Y are uncorrelated, but

$$\frac{1}{4} = \Pr(X = 0, Y = 0) \neq \Pr(X = 0) \times \Pr(Y = 0) = \frac{1}{4} \times \frac{1}{2} = \frac{1}{8}$$

- Also many different ways to "measure" dependence
  - *Covariance* between X and Y is

$$\begin{aligned} \operatorname{cov}(X, Y) &= \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\ &= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) \end{aligned}$$

  - For any constants a, b, c, d,

$$\operatorname{cov}(aX + b, cY + d) = ac \operatorname{cov}(X, Y)$$

  - *(Pearson's) correlation* between X and Y is

$$\operatorname{cor}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\operatorname{stddev}(X)\operatorname{stddev}(Y)}$$

  - In die rolling example with $n = 4$:

$$\begin{aligned} \operatorname{cov}(X, Z) &= -1/9, \quad \operatorname{var}(X) = \operatorname{var}(Z) = 5/9 \\ \operatorname{cor}(X, Z) &= -1/5 \end{aligned}$$

Q. If $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$, then what can you say about $\operatorname{cov}(X, Y)$?

Q. If $X = Y$, then what is the value of $\operatorname{cor}(X, Y)$?

Q. Is it possible to have $\operatorname{cor}(X, Y) > 1$? What about $\operatorname{cor}(X, Y) < -1$?

1.7 Marginal and conditional distributions

- Consider random variables X and Y (on (Ω, m))
- *Marginal distribution* of Y is the probability distribution given by

$$p_Y(y) = \Pr(Y = y) = \Pr(\{\omega \in \Omega : Y(\omega) = y\})$$

– Law of total probability:

$$p_Y(y) = \Pr(Y = y) = \sum_x \Pr(X = x \wedge Y = y) = \sum_x p_{(X,Y)}(x, y)$$

This process of summing $p_{(X,Y)}(x, y)$ over all possible values of X is called _marginalization_

- _Conditional distribution_ of Y given $X = x$ is probability distribution $p_{Y|X=x}$ given by

$$\begin{aligned} p_{Y|X=x}(y) &= \Pr(Y = y \mid X = x) \\ &= \frac{\Pr(Y = y \wedge X = x)}{\Pr(X = x)} \end{aligned}$$

- _Conditional expectation_ of Y given $X = x$ is

$$\mathbb{E}(Y \mid X = x) = \sum_y y \, p_{Y|X=x}(y)$$

– Example: roll a fair 6-sided die
  * $X = \mathbb{1}\{\text{number of pips is more than } 4\}$
  * $Y =$ number of pips
  * $\mathbb{E}(Y \mid X = 0) = 2.5$
  * $\mathbb{E}(Y \mid X = 1) = 5.5$
  * $Y' = \mathbb{1}\{\text{number of pips is even}\}$
  * $\mathbb{E}(Y' \mid X = 0) = 1/2$
  * $\mathbb{E}(Y' \mid X = 1) = 1/2$
  * $X' = \mathbb{1}\{\text{number of pips is more than } 3\}$
  * $\mathbb{E}(Y' \mid X' = 0) = 1/3$
  * $\mathbb{E}(Y' \mid X' = 1) = 2/3$

- Regard $Z = \mathbb{E}(Y \mid X)$ as a random variable in probability space $(\mathbb{R}, p_X)$

– $Z(x) = \mathbb{E}(Y \mid X = x)$

– Expected value of $\mathbb{E}(Y \mid X)$ is

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y \mid X)) &= \sum_x \mathbb{E}(Y \mid X = x)\, p_X(x) \\
&= \sum_x \sum_y y\, p_{Y \mid X=x}\, p_X(x) \\
&= \sum_x \sum_y y\, p_{(X,Y)}(x, y) \\
&= \sum_y y\, p_Y(y) \\
&= \mathbb{E}(Y)
\end{aligned}$$

This fact is called the *tower property* of conditional expectation

Q. Toss a fair coin two times; let

– $X =$ number of heads

– $Y = \begin{cases} 1 & \text{if first toss is heads and second toss is tails} \\ 0 & \text{if both tosses are the same} \\ -1 & \text{if first toss is tails and second toss is heads} \end{cases}$

For each $x \in \text{range}(X)$, what is the expected value of Y given $X = x$?

1.8 Continuous random variables

- So far, we have only considered *discrete random variables* (which have finite or countable ranges)

  – Probability distribution of random variable X can be specified either by its probability mass function p_X or by its *(cumulative) distribution function (cdf)* cdf_X

$$\text{cdf}_X(x) = \Pr(X \le x)$$

- A random variable is *continuous* if its distribution function is a continuous function

  – In some cases, these arise by starting with discrete distributions and taking an appropriate limit

  - In this class, we'll only discuss continuous random variables X whose distribution functions can be written as

    $$\mathrm{cdf}_X(x) = \int_{-\infty}^{x} p_X(u)\,\mathrm{d}u$$

    for a function p_X called the *(probability) density function (pdf)*

- Important example: uniform (on unit interval) random variable

  $$p_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

  - Notation: $X \sim \mathrm{Unif}([0,1])$
  - For any subinterval $I \subseteq [0,1]$, $\Pr(X \in I)$ is the length of the interval

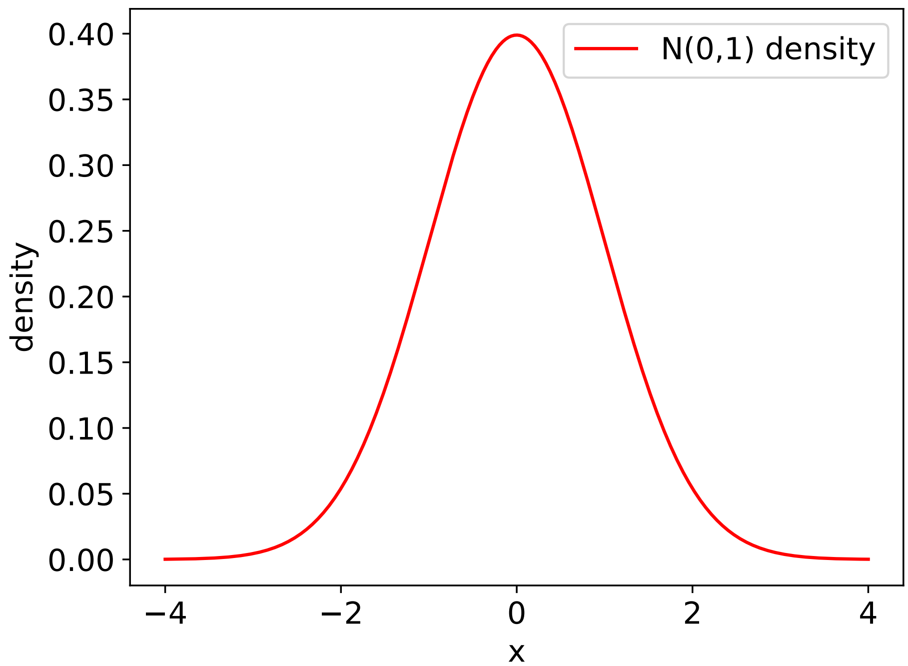
  Uniform (on unit square) random vector:

  $$p_{(X,Y)}(x,y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

  - Notation: $(X,Y) \sim \mathrm{Unif}([0,1]^2)$
  - Can verify that X and Y are independent, and each of X and Y has marginal distribution $\mathrm{Unif}([0,1])$

- Another important example: a *standard normal random variable* has density function

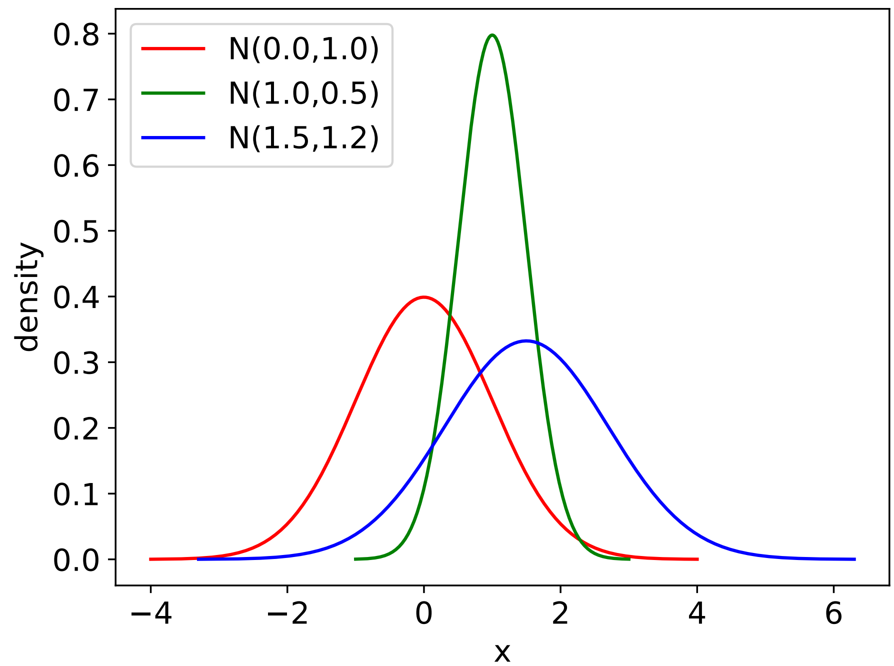
  $$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$



- More generally: a *normal random variable with mean μ and variance σ^2* has density function

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

  - Notation: "$X \sim \mathrm{N}(\mu, \sigma^2)$" means "$X$ is a random variable with density function ϕ_{μ,σ^2}"
  - Fact: If $X \sim \mathrm{N}(0, 1)$ and $Y = \mu + \sigma X$, then $Y \sim \mathrm{N}(\mu, \sigma^2)$ (Verify this using change-of-variable)



Q. What is the distribution function for $X \sim \mathrm{Unif}([0,1])$?

1.9 Two important theorems

- *Law of Large Numbers (LLN)*: If $X_1, X_2, \ldots$ is an infinite sequence of independent and identically distributed (i.i.d.) random variables with expectation μ, then

  $$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu$$

  as $n \to \infty$

  (We don't dwell upon the notions of convergence in this class)

- *Central Limit Theorem (CLT)*: If $X_1, X_2, \ldots$ is an infinite sequence of independent and identically distributed (i.i.d.) random variables with expectation μ and variance σ^2, then

  $$\frac{\sum_{i=1}^{n} X_i - \mu}{\sigma\sqrt{n}} \longrightarrow \mathrm{N}(0,1)$$

  as $n \to \infty$

2 Review of linear algebra

2.1 Why linear algebra?

- Many machine learning methods represent data as vectors of numbers
- Many methods for statistical analysis is based on linear algebraic ideas (e.g., linearity)
- Descriptions and analyses of many machine learning methods use linear algebraic notations and concepts

2.2 Euclidean spaces

- *Euclidean d-space*, denoted $\mathbb{R}^d$, is the d-dimensional generalization of three-dimensional physical space
- A *d-vector* $v \in \mathbb{R}^d$ is a d-tuple of real numbers

$$v = (v_1, \ldots, v_d)$$

  (We omit "d-" from "d-vector" when clear from context)

  - The i-th *component* (a.k.a. *entry*) of v is v_i

- Basic operations on d-vectors that produce d-vectors:
  - *Addition*: for $u, v \in \mathbb{R}^d$,

$$u + v = (u_1 + v_1, \ldots, u_d + v_d) \in \mathbb{R}^d$$

  - *Scalar multiplication*: for $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$,

$$cv = (cv_1, \ldots, cv_d) \in \mathbb{R}^d$$

- There is a special vector called the *zero vector* $0 = (0, \ldots, 0)$
  - Adding the zero vector to another vector v results in v
  - Scaling the zero vector by a real number c results in the zero vector

- The *norm* (a.k.a. *length*) of a vector $v \in \mathbb{R}^d$, denoted by $\|v\|$, is

$$\|v\| = \sqrt{v_1^2 + \cdots + v_d^2}$$

  - A *unit vector* is a vector with norm 1

- The *inner product* (a.k.a. *dot product*) between vectors $u, v \in \mathbb{R}^d$, denoted by $u^\mathsf{T} v$ (or $\langle u, v \rangle$), is

$$u^\mathsf{T} v = u_1 v_1 + \cdots + u_d v_d$$

  - Interpretation: $u^\mathsf{T} v = \|u\| \|v\| \cos(\theta)$ where θ is the "angle" between u and v
  - Note: $\|v\| = \sqrt{v^\mathsf{T} v}$

- *Cauchy-Schwarz inequality*: For any vectors $u, v \in \mathbb{R}^d$,

$$u^\mathsf{T} v \leq \|u\| \|v\|,$$

  with equality if and only if there is a real number $c \in \mathbb{R}$ such that $u = cv$

- Vectors $u, v \in \mathbb{R}^d$ are *orthogonal* if $u^\mathsf{T} v = 0$ (shorthand: "$u \perp v$")

  - A collection of vectors $v^{(1)}, \ldots, v^{(n)} \in \mathbb{R}^d$ is *orthogonal* if, for every $i \neq j$, $v^{(i)}$ and $v^{(j)}$ are orthogonal
  - A collection of vectors is *orthonormal* if it is orthogonal and every vector in the collection is a unit vector

- *Pythagorean theorem*: If $v^{(1)}, \ldots, v^{(n)}$ is an orthogonal collection of vectors, then

$$\|v^{(1)} + \cdots + v^{(n)}\|^2 = \|v^{(1)}\|^2 + \cdots + \|v^{(n)}\|^2$$

Q. Show that the only vector with length zero is the zero vector.

Q. Show that the *triangle inequality* holds: for any $u, v \in \mathbb{R}^d$,

$$\|u + v\| \leq \|u\| + \|v\|.$$

2.3 Linear dependence

- A *linear combination* of a finite collection of vectors $v^{(1)}, \dots, v^{(n)} \in \mathbb{R}^d$ is an expression that multiples each $v^{(i)}$ by a real number $c_i \in \mathbb{R}$, and then adds up the results:

$$c_1 v^{(1)} + \cdots + c_n v^{(n)}$$

  - A *non-trivial linear combination* of a finite collection of vectors $v^{(1)}, \dots, v^{(n)} \in \mathbb{R}^d$ is a linear combination $c_1 v^{(1)} + \cdots + c_n v^{(n)}$ where at least one of the $c^{(i)}$ is non-zero

- A collection of vectors is *linearly dependent* if there is a non-trivial linear combination of vectors from this collection that results in the zero vector

  - A collection of vectors that is not linearly dependent is said to be *linearly independent*

Q. Suppose unit vectors $v^{(1)}, \dots, v^{(n)}$ satisfy $|\langle v^{(i)}, v^{(j)} \rangle| \leq 1/n$ for all $i \neq j$. Show that these vectors must be linearly independent.

2.4 Subspaces, dimension, and bases

- The *span* of a collection of vectors is the set of all linear combinations of any subset of vectors from this collection

- A *subspace* $\mathcal{W}$ of $\mathbb{R}^d$ is a collection of vectors from $\mathbb{R}^d$ that is closed under addition and scalar multiplication and also contains the zero vector

  - $\mathbb{R}^d$ itself is a subspace of $\mathbb{R}^d$

- The *dimension* of a subspace $\mathcal{W}$, written $\dim(\mathcal{W})$, is the largest number k such that $\mathcal{W}$ contains a linearly independent set of k vectors

  - $\dim(\mathbb{R}^d) = d$

- A set of vector B from a subspace $\mathcal{W}$ is a *basis* for $\mathcal{W}$ if B is linearly independent and the span of B is $\mathcal{W}$

- Every basis for a subspace $\mathcal{W}$ has the same number of vectors, and that number is the dimension of the subspace
- It is often useful to order the vectors in a basis $B = (b^{(1)}, \dots, b^{(k)})$, and such an ordered set of vectors is called an *ordered basis*
- The *standard (coordinate) basis* for $\mathbb{R}^d$ is the ordered basis $(e^{(1)}, \dots, e^{(d)})$, where $e^{(i)}$ is the d-vector whose components are all zeros except for the i-th component, which has value one

2.5 Linear transformations and matrices

- A *linear transformation* $T \colon \mathbb{R}^d \to \mathbb{R}^k$ between the Euclidean spaces $\mathbb{R}^d$ and $\mathbb{R}^k$ is a function that satisfies the following two properties:
  - *Additivity*: $T(u + v) = T(u) + T(v)$ for any $u, v \in \mathbb{R}^d$
  - *Homogeneity*: $T(cv) = cT(v)$ for any $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$

  (Additivity & homogeneity = *linearity*)
- A $k \times d$ *matrix* A is a tableaux of kd real numbers arranged in k rows and d columns

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,d} \\ \vdots & \ddots & \vdots \\ A_{k,1} & \cdots & A_{k,d} \end{bmatrix}$$

  - The (i, j)-th *component* (a.k.a. *entry*) of A is $A_{i,j}$
  - We may regard A has an ordered collection of k-vectors $a^{(1)}, \dots, a^{(d)}$, one per column of A:

$$A = \begin{bmatrix} \uparrow & & \uparrow \\ a^{(1)} & \cdots & a^{(d)} \\ \downarrow & & \downarrow \end{bmatrix}$$

- The *(matrix-vector) product* of $k \times d$ matrix A and d-vector $x = (x_1, \dots, x_d)$, written Ax, is the linear combination of columns $a^{(1)}, \dots, a^{(d)} \in \mathbb{R}^k$ of A given by

$$x_1 a^{(1)} + \cdots + x_d a^{(d)}$$

  - Caution: a matrix-vector product Ax only makes sense if the number of columns of A equals the number of components of x

- The *(matrix-matrix) product* (a.k.a. *matrix multiplication*) of $k \times d$ matrix A and $d \times p$ matrix B, written AB, is the $k \times p$ matrix whose i-th column is the matrix-vector product of A and the i-th column of B

$$AB = \begin{bmatrix} Ab^{(1)} & \cdots & Ab^{(p)} \end{bmatrix}$$

  - Caution: a matrix-matrix product AB only makes sense if the number of columns of A equals the number of rows of B
  - Matrix-vector product can be viewed as a special case of matrix multiplication by pretending a d-vector is a $d \times 1$ matrix

- Matrix multiplication is associative (i.e., $A(BC) = (AB)C$) and distributive (i.e., $A(B + C) = AB + AC$), but not commutative (i.e., it is possible that $AB \neq BA$)

- The *transpose* of $k \times d$ matrix A, written $A^\intercal$ is the $d \times k$ matrix whose (i, j)-th component is $A_{j,i}$

  - What is the "meaning" of $A^\intercal$?
  - For every $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^k$, we have $\langle Ax, y\rangle = \langle x, A^\intercal y\rangle$
  - Relates certain angles in $\mathbb{R}^d$ to certain angles in $\mathbb{R}^k$

- Special matrix-matrix product: *outer product* of k-vector $u \in \mathbb{R}^k$ and d-vector $v \in \mathbb{R}^d$, written $uv^\intercal$ (where v is treated as $d \times 1$ matrix, so $v^\intercal$ is $1 \times d$ matrix, a.k.a. *row vector*)

$$uv^\intercal = \begin{bmatrix} u_1 \\ \vdots \\ u_k \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_d \end{bmatrix} = \begin{bmatrix} u_1 v_1 & \cdots & u_1 v_d \\ \vdots & \ddots & \vdots \\ u_k v_1 & \cdots & u_k v_d \end{bmatrix}$$

  (Result is a $k \times d$ matrix)

Q. Show that, for any $k \times d$ matrix M, the transformation $T \colon \mathbb{R}^d \to \mathbb{R}^k$ given by $T(v) = Mv$ is a linear transformation.

Q. Show that, for any linear transformation $T \colon \mathbb{R}^d \to \mathbb{R}^k$, there is a $k \times d$ matrix M such that $T(v) = Mv$ for all $v \in \mathbb{R}^d$.

2.6 Orthogonal complements and projectors

- For a k-dimensional subspace $\mathcal{W}$ of $\mathbb{R}^d$, the *orthogonal complement* of $\mathcal{W}$, written $\mathcal{W}^\perp$, is the set of all vectors v that are orthogonal to every vector in $\mathcal{W}$

$$\mathcal{W}^\perp = \{v \in \mathbb{R}^d : v \perp w \text{ for all } w \in \mathcal{W}\}$$

  - Sometimes write "$v \perp \mathcal{W}$" to mean "$v \perp w$ for all $w \in \mathcal{W}$"
  - $\mathcal{W}^\perp$ is also a subspace of $\mathbb{R}^d$

- For every $k \times d$ matrix A:

  - *Column space* of A, denoted $\mathsf{CS}(A)$, is the span of columns of A
  - *Nullspace* of A, denoted $\mathsf{NS}(A)$, is all $x \in \mathbb{R}^d$ such that $Ax = 0$
  - *Row space* of A is $\mathsf{CS}(A^\intercal)$; *left nullspace* of A is $\mathsf{NS}(A^\intercal)$
  - $\mathsf{CS}(A^\intercal)$ and $\mathsf{NS}(A)$ are subspaces of $\mathbb{R}^d$
  - $\mathsf{CS}(A)$ and $\mathsf{NS}(A^\intercal)$ are subspaces of $\mathbb{R}^k$
  - *Rank* of A is $\dim(\mathsf{CS}(A))$ and is also equal to $\dim(\mathsf{CS}(A^\intercal))$
  - $\operatorname{rank}(A) + \dim(\mathsf{NS}(A)) = d$, and $\operatorname{rank}(A) + \dim(\mathsf{NS}(A^\intercal)) = k$
  - $\mathsf{CS}(A^\intercal)$ and $\mathsf{NS}(A)$ are orthogonal complements of each other
  - $\mathsf{CS}(A)$ and $\mathsf{NS}(A^\intercal)$ are orthogonal complements of each other

- A *projection operator* (a.k.a. *projector*) $P \colon \mathbb{R}^d \to \mathbb{R}^d$ is a linear transformation satisfying *idempotency*, i.e., $P(v) = P(P(v))$ for all $v \in \mathbb{R}^d$

- For any subspace $\mathcal{W}$ of $\mathbb{R}^d$, there is a projector $P \colon \mathbb{R}^d \to \mathbb{R}^d$, called the *orthogonal projector* (a.k.a. *orthoprojector*) to $\mathcal{W}$, such that

$$P(v) \in \mathcal{W} \quad \text{and} \quad v - P(v) \in \mathcal{W}^\perp$$

  - If $\mathcal{W} = \mathsf{CS}(A)$ for a $k \times d$ matrix A, then $P(v) = AA^\dagger v$ for all $v \in \mathbb{R}^d$, where $A^\dagger$ is Moore-Penrose pseudoinverse of A
    - $*$ If $\operatorname{rank}(A) = d$, then $P(v) = A(A^\intercal A)^{-1}A^\intercal$