EngSci 760: Decision Making and Modelling Under Uncertainty

Oliver Maclaren (oliver.maclaren@auckland.ac.nz)

Overview

Part I: **Decision** topics

- Basic concepts
- Risk, probability, utility
- Finite sample decisions
  - Empirical risk
  - Statistical decision theory

Part II: **Modelling** topics

- Probability, graphical models, independence
- Causal interpretations of graphical models
- Stochastic process models (esp. Markov)
- Simulation and estimation

Part I

Decision-making under uncertainty

Basic concepts

Overview

Decision theory is a multidisciplinary field, broadly concerned with formalising and analysing ways of **making the 'best' or 'optimal' decision given available information and uncertainties**.

Decision theory informs and is informed by

- Philosophy
- Economics
- Statistics
- Psychology
- Operations research
- Machine learning/artificial intelligence

etc!

Note: Some of you may have seen elements of decision theory in ENGSCI / STATS 255. This is *not assumed* here. We start from the beginning as well as consider different aspects.

Scope

We will look at both

(1) **'Simple' decision theory**

and

(2) **'Statistical' (or 'empirical') decision theory**

The latter extends and incorporates simple decision theory and adds in the availability of **sample data** to inform decisions. This is relevant for engineering, machine learning etc. Due to the lack of sample data, the former is sometimes called **no-data** decision making (though we do have some information!).

People

Some key contributors include

- The early work on **utility** in the 18th and 19th centuries by **Bernoulli** (addressing the *St. Petersburg paradox*), and **Bentham** and **Mill** (giving rise to *utilitarianism*);
- The work of **von Neumann and Morgenstern** (1947) on **expected utility** for *simple decision theory*;
- **Wald**'s (1950) work on **statistical decision functions** for *statistical decision theory*,

along with others such as **Vapnik**, **Fisher**, **Efron**, **Manski**, **Huber**, **Dawid**, **Shapiro** and more from a range of areas such as machine learning, statistics, econometrics, operations research etc, who developed related **empirical risk** (or *empirical decision theory*) ideas.

Individuals or groups?

The term 'decision theory' usually refers to **single-person** decision making rather than **group** decision making (e.g. voting etc.)

Intermediate between these is the study of **games** (game theory) where, e.g., **multiple individuals** independently (simultaneously or in-turn) make decisions to achieve typically **conflicting** goals while not knowing exactly what their 'opponent' will do.

We will mostly study **individual** decision making, though will also consider games in which an individual **plays against 'nature'**.

Normative or descriptive?

Approaches to decision theory can be classified as either

- **Normative**: about how people **should** decide, 'in principle', or
- **Descriptive**: about how people **do** decide, 'in practice',

though in reality it is typically a **mix**.

Essentially though, we study **normative**, i.e., **idealised**, decision making here, i.e. we study **algorithms/principles** for **'optimal' decision making** that we would ideally like to follow … even if we don't in practice!

Meaning of uncertainty

Within what we are calling **simple decision theory** (c.f. statistical or empirical decision theory given sample data), a key element is **uncertainty**. In this context the following types of uncertainty are usually distinguished:

- **Certainty**: decisions in a **deterministic nature** setting, i.e. optimisation with no probabilities or ignorance
  - E.g. linear programming or other function optimisation problems. (See previous part of course).
  - 'Pay-off' only depends on your decision
- **Risk**: decisions in the presence of **known probability distributions**, i.e. given a known probability model over possible outcomes a decision needs to be made despite the *risk* associated with the probability distribution.
  - E.g. flip a fair ($\mathbb{P}(H) = 0.5, \mathbb{P}(T) = 0.5$) coin. Guess correctly win \$1000, guess incorrectly lose \$500, refuse to play win/lose nothing. What to do?
  - 'Pay-off' for given decision depends on randomly drawn 'state of nature' (see later).
  - 'Known unknowns'.
- **Ignorance**: decisions in the presence of **unknown probabilities**, a set of probabilities, or no probabilities but unknown 'state of nature'.
  - E.g. same as coin example but $\mathbb{P}(H) = p$, where now p is not known.
  - 'Pay-off' for given decision depends on unknown 'state of nature'. Don't know what determines state of nature.
  - 'Unknown unknowns'

Aside: more on the meaning of uncertainty

Many people read 'uncertainty' as inherently meaning **involving probability**.

However, in (simple) decision theory there is a long history of taking 'uncertainty' to mean what we call **ignorance** (unknown probability distribution) rather than **risk** (known probability distribution).

Bayesians treat risk and ignorance the same – using probability – but others don't.

We will treat **uncertainty** as an **umbrella term** that may refer to **either risk or ignorance**.

Uncertainty and statistical/empirical decision theory

As mentioned, **statistical/empirical decision theory** adds to the simple decision theory setting by allowing for the availability of **sample/empirical data** before the decision is made.

In general, statistical decision theory involves *both* **risk** and **ignorance**. For example, we don't know the true θ, say θ_0, in the probability model $\mathbb{P}(X;\theta)$ (i.e. we don't know which distribution we are making decisions under) but we *do have* a **sample** realisations $X = x$ generated using the true θ_0.

That is, we have a realisation from $X \sim \mathbb{P}(X;\theta_0)$ but don't know θ.

Example decision: *decide* on an **estimate** $\hat{\theta}$ of θ_0 by using the sample data. In this context, the **decision rule** $\hat{\theta}(X)$, which says what to do for any realisation of the random data X, is called an **estimator**, and the result for a given realisation x, i.e. $\hat{\theta}(x)$, is called an **estimate**.

As hinted at, **Bayesians** introduce an additional element (a *prior*) to reduce statistical decision theory back to one of pure **risk**, while non-Bayesians (can) think of it as a **game** between you and 'nature' involving **both risk and ignorance**.

Simple decision theory: minimal formulation

The key elements of **simple decision theory** are:

- **States of nature**
  - Well-defined 'outcomes *of nature*': **Nature's choice**.
  - Here, have **risk** or **ignorance** concerning which will occur (we don't control).
  - Probabilites (under risk) may depend on acts (see later).
- **Acts/decisions**
  - **Your** choices/actions: you have control over
  - Should be definite enough to determine an overall outcome (see below) for any given state of nature
- **Outcomes** (for given decision and state of nature)
  - **How things turned out**, given your decision and the state of nature that resulted, in the ways you care about.
  - Outcome $=$ $f(\text{Your choice}, \text{Nature's choice})$, i.e. outcome $= f(\text{decision}, \text{state of nature})$.
  - Simple, general version is **just to take it as the pair (decision, state of nature)**...
  - **Not the value** of the outcome, just 'how things turned out'.

The last point above is clarified by...

Simple decision theory: utility

The final ingredient of simple decision theory is

- **Utility** (of an outcome),
  - A function utility $= u(\text{outcome})$.
  - The **value** *you* place on the outcome of your decision given the state of nature that resulted.
  - We often just write $u(\text{outcome}(\text{decision}, \text{state of nature}))$ as $u(\text{decision}, \text{state of nature})$ (see Parmigiani & Inoue, 2009 and Schervish et al., 1990 for discussion of this)
  - It is **not in general monetary value!**
    - How much is \$1 worth to a hungry person with no money? To a millionaire?
  - It is a **numerical representation** of **your** (the decision maker's) **relative preference for outcomes**.

It's important to emphasise that utility is really just a **convenient way to represent preferences for outcomes, and preferences are key**. In particular, we want the utility function u, defined here for 'outcomes', to satisfy:

$$a \text{ is } \textit{preferred} \text{ to } b, \text{written } a \succ b$$
$$\text{if and only if}$$
$$\text{The } \textit{utility} \text{ of } a \text{ is greater than the utility of } b, \text{written } u(a) > u(b)$$

Note the second inequality is a normal inequality for numbers as **utility is numerical**, while the first 'fancy inequality' is an abstract relation defined on outcomes. This is what gives utility...utility: numbers are easier to work with!

Preferences over outcomes and utility scales

The idea then is that minimal (but still controversial, e.g., how reasonable is the total ordering below in all cases?) assumptions on preferences allow various **representation theorems** that tell you how to represent your (pre-existing) preferences **numerically** with a utility function.

Example *minimal* assumptions on preferences for options/outcomes a, b, c might be:

$$\forall a, b: \ a \succ b \text{ or } a \prec b \text{ or } a \approx b \text{ [totally ordered]}$$
$$\forall a, b, c: \ a \succ b \text{ and } b \succ c \text{ implies } a \succ c \text{ [transitive]}$$

where $\succ$ represents the relation 'is preferred to' and $\approx$ represents that you are 'indifferent' between the options. Even these can be controversial!

There are a few other consistency conditions such as not being able to both prefer a to b and prefer b to a etc that we won't bother to cover (see Resnik).

In general, there are different **types of utility scale** depending on the assumptions on the **structure of your preferences**. For example...

Ordinal preferences

Having **ordinal preferences** means that **relative preferences** make sense but **relative differences** don't mean anything.

E.g.

"bad", "ok", "good"

can be represented numerically equally well by both

1, 2, 3

and

5, 20, 30.

There are some simple decision rules (e.g. maximin utility) that work for ordinal preferences, while others (e.g. minimax regret, expected utility) require more structure/stronger assumptions on preferences. This leads to...

Interval (cardinal) preferences

Having **interval preferences** (also called **cardinal preferences**) means that we care about **relative differences (i.e. intervals) in utility**. In particular, interval preferences require that, for all options x, y, z, w,

$$|u(x) - u(y)| > |u(z) - u(w)|$$

means that the preference for x over y is stronger than the preference for z over w.

We will consider what this means exactly in more detail in the tutorial but for now, note that the rankings from above (for ordinal utilities) are not equivalent from the point of view of interval/cardinal preferences.

Expected utility and von Neumann and Morgenstern

In their classic work in 1947, von Neumann and Morgenstern (VNM) devised assumptions on preferences specifically for decisions under **risk** (see later).

These correspond to a type of interval preference scale and lead to the further idea of using **expected utility** for decisions under **risk**.

We will cover this in more detail next lecture, but the basic idea is that we can value a decision under *risk* by the *expected* utility of outcomes, as calculated using a *given* probability distribution.

Decision tables under ignorance

Under **ignorance** (uncertainty but no known probabilities) we have the following three key tables (illustrated for four states of nature and four decisions for simplicity).

Outcome table

	state of nature 1	state of nature 2	state of nature 3	state of nature 4
decision 1	$o(d_1, s_1)$	$o(d_1, s_2)$	...	...
decision 2	$o(d_2, s_1)$	$\ddots$		
decision 3	$\vdots$			
decision 4	$\vdots$			

where '$o(d_i, s_j)$' stands for the *outcome* from decision i under state of nature j,

Utility function over outcomes (as table)

	outcome 1	outcome 2	outcome 3	outcome 4
utility	$u(o_1)$	$u(o_2)$	...	...

Utility table

	state of nature 1	state of nature 2	state of nature 3	state of nature 4
decision 1	$u(o(d_1, s_1))$	$u(o(d_1, s_2))$	...	...
decision 2	$u(o(d_2, s_1))$	$\ddots$		
decision 3	$\vdots$			
decision 4	$\vdots$			

Decision tables under ignorance: straight to utility table

As discussed, it is often easier to consider the 'outcome' just to be (d_i, s_j) and hence we just need the utility table in the form:

Utility table: (d, s)-form

	state of nature 1	state of nature 2	state of nature 3	state of nature 4
decision 1	$u(d_1, s_1)$	$u(d_1, s_2)$	...	...
decision 2	$u(d_2, s_1)$	$\ddots$		
decision 3	$\vdots$			
decision 4	$\vdots$			

From now we will generally just work with these (d, s)-form tables.

Decision tables with risk

When we are under conditions of **risk**, i.e. uncertain states occuring with **known probabilities**, we have the extra ingredient:

$$\mathbb{P}(s_j; d_i)$$

which stands for the probability of state j 'given' state i. We have scare quotes around 'given', and have used ';' rather than | (though won't always), as this may not be (but can be) a conditional probability in the formal sense (which, according to the standard Kolmogorov approach requires probabilities over the d_i), rather it is just the probability of state j when decision i is made. This leads to:

Probability-weighted utility table

	state of nature 1	state of nature 2	state of nature 3	state of nature 4
decision 1	$u(d_1, s_1)$ $\mathbb{P}(s_1; d_1)$	$u(d_1, s_2)$ $\mathbb{P}(s_2; d_1)$	... ...	... ...
decision 2	$u(d_2, s_1)$ $\mathbb{P}(s_1; d_2)$	$\ddots$ $\ddots$		
decision 3	$\vdots$ $\vdots$			
decision 4	$\vdots$ $\vdots$			

Note on decision-dependent probabilities

In Savage's (1954) contribution to statistical decision theory, 'Foundations of Statistics', he assumed probabilities of states are independent of acts (decisions), i.e. $\mathbb{P}(s; d) = \mathbb{P}(s)$ for all states and decisions.

For some problems this requires complex definitions of states. An alternatvie is to use 'conditional' or decision-dependent probabilities of states (as in previous table). This was the approach taken in Jeffrey's (1965) 'Evidential Decision Theory'.

> We take this 'conditional' approach as default, i.e. that probabilities of states can depend on decisions taken.

See discussion in Resnik (1987) and later in this course.

Utility and loss

Given a **utility function**, we can define a **loss function** as the negative of utility:

$$\text{loss}(\text{outcome}) = -\text{utility}(\text{outcome})$$

Or, using the (d, s) form of outcomes,

$$\text{loss}(d, s) = -\text{utility}(d, s)$$

for decision d, state of nature s.

We can then aim to **minimise loss instead of maximise utility** (same result). Loss functions rather than utility functions are common in e.g. statistics and machine learning.

We write the loss function as

$$\text{loss}(d, s) = l(d, s)$$

Regret

It may be the case that a decision problem always has a non-zero i.e. **unavoidable loss** (negative utility), **even if we knew the state of nature before making our decision** (i.e. 'given perfect information').

If the decision **must** be made, for example, some such as Savage, have argued that we should measure the **actual loss for a given decision and state of nature relative to the minimum achievable loss for that state of nature**, i.e. relative to the **best retrospective decision** after learning the state of nature (or best decision as given by an 'oracle').

This is called the **regret**, and is typically used under conditions of **ignorance** rather than risk:

$$\text{regret}(d, s) = \text{loss}(d, s) - \min_{d|s}(\text{loss}(d, s))$$

where $\min_{d|s}$ represents that we are minimising for a fixed s (state of nature) while varying d (decision), leading to the best case decision for that state.

Note that regret is in a sense a sort of 'meta' loss function requiring access to the whole 'elementary' loss table – in essence the 'outcome' considered by a decision maker includes both (d, s) and an **entire possible elementary utility table** to define the best possible decision for each state. Loss functions and regret functions **give the same decision when:**

$$\min_{d} \text{loss}(d, s) = 0, \text{ for each } s$$

i.e. **when the best possible decision for any given state (perfect information) has zero loss**. In this context we call a loss function a **regret loss function**.

Savage noted that Wald often assumed this implicitly, and Savage helped popularise the use of regret for non-Bayesian decision making (though he himself advocated for the Bayesian approach).

Conversion from absolute to relative loss (regret)

We can always convert a loss table to a regret table as illustrated on this simple example:

Loss table, loss(d, s), with best decision for each state indicated

	state of nature 1	state of nature 2	state of nature 3
decision 1	1	0	6
decision 2	3	4	5
$\min_{d\|s}$	1	0	5

Using:

$$\text{regret}(d, s) = \text{loss}(d, s) - \min_{d|s}(\text{loss}(d, s))$$

we have:

Regret table

	state of nature 1	state of nature 2	state of nature 3
decision 1	0	0	1
decision 2	2	4	0
$\min_{d\|s}$	0	0	0

Now each column has at least one zero.

Note on regret and risk

Regret is typically used under **ignorance** (unknown probabilities). Under **risk**, i.e. **known probabilities** over states of nature, we could also consider the regret table **weighted by probabilities**, just like with expected utility.

However, I have (personally) only seen this done under the special (Savage!) case that **state probabilities are independent of decisions**.

- In *this* case the **minimum expected regret decision is the same as the minimum expected loss decision** (different loss values, same decision).
- Hence, in this setting at least, you might say there is no real benefit to using regret under risk (c.f. ignorance) compared to using expected loss.
- However (and again in this setting), it can be useful as a measure of the **expected value of perfect information** (see e.g. Jeffrey, 1965, and below).

In terms of loss and regret, recall that the regret is defined as:

$$\text{regret}(d, s) = \text{loss}(d, s) - \min_{d|s}(\text{loss}(d, s))$$

Under risk (with constant state probabilities, say) we have the expected regret:

$$\text{ER}(d) = \mathbb{E}_{\mathbb{P}_s}\left[\text{regret}(d, s)\right] = \mathbb{E}_{\mathbb{P}_s}\left[\text{loss}(d, s)\right] - \mathbb{E}_{\mathbb{P}_s}\left[\min_{d|s}(\text{loss}(d, s))\right]$$

which can be considered the **expected opportunity loss**, and is just calculated by weighting the regret table by the relevant probabilities. The **minimum** of this gives the **expected value of perfect information** or how much value (reduction in loss) 'perfect' (oracle) information provides relative to your best decision under uncertainty.

Decision rules

So...given a decision table, **how do we make a decision?** We use a **decision rule** (but how to we *decide* what decision rule to use?! See Resnik on the issue of 'second order decisions').

For **ignorance** we have

- Either **maximin utility** or (equivalently) **minimax loss**
- Either **maximin relative utility** or (equivalently) **minimax regret** (relative loss)

These will give the same decisions if max utility = 0 (min loss = 0) but not in general (see Resnik).

For **risk** we have

- Either **max expected utility** or (equivalently) **min expected loss**
- Either **max expected relative utility** or (equivalently) **min expected regret** (relative loss).

These will give the same decisions if state probabilities are independent of decisions but not in general (see Jeffrey 1965 or Barnett 1999).

Example: loss (-utility) form

Lets consider an example. First consider ignorance (no known probabilities).

Loss table

	state of nature 1	state of nature 2	state of nature 3
decision 1	1	0	6
decision 2	3	4	5

Regret table

	state of nature 1	state of nature 2	state of nature 3
decision 1	0	0	1
decision 2	2	4	0

Example: Minimax absolute loss

Loss table

	s_1	s_2	s_3	max loss
d_1	1	0	6	6
d_2	3	4	5	5
minimax loss				5 (d_2)

Example: Minimax relative loss (regret)

Regret table

	s_1	s_2	s_3	max loss
d_1	0	0	1	1
d_2	2	4	0	4
minimax regret				1 (d_1)

Notes:

- Different decisions! Pros and cons to each.
- Minimax regret decisions are **usually less pessimistic** than minimax loss
- Minimax regret implies an interval loss scale while absolute loss only requires an ordinal scale (see Resnik/tutorial).

Example: Minimum expected loss/regret

Suppose we also have the following decision-dependent probability table, along with the loss and regret tables from before:

Probability table

	state of nature 1	state of nature 2	state of nature 3
decision 1	0.5	0.2	0.3
decision 2	0.1	0.0	0.9

Note that **each row** sums to one and that the expected loss or regret value of a given decision d is given by

$$\mathrm{EL/ER}(d) = \mathbb{E}_{\mathbb{P}(s;d)}\left[f(d,s)\right] = \sum_s f(d,s)\mathbb{P}(s;d)$$

for the table entries (either loss or regret) $f(d,s)$ and for discrete tables (can generalise to continuous functions).

For **each decision**, we can calculate the expected absolute loss and the expected relative loss (regret).

(Note again though that regret with decision-dependent probabilities is uncommon, possibly because e.g. it implies the oracle decision is evaluated under the probability distribution associated with the actual decision).

This gives (exercise!) a **minimum expected loss** of 2.3 (decision 1) and **minimum expected regret** of 0.2 (decision 2). These can be different decisions as the probabilities are decision-dependent.

Example: decision-dependent/independent probabilities

Redo the example with decision-independent state probabilities of

Probability table

	state of nature 1	state of nature 2	state of nature 3
decision 1	0.5	0.2	0.3
decision 2	0.5	0.2	0.3

You should get the **same decision, though different numerical values for loss and regret**. This relationship, between minimum expected loss and miniumum expected regret, holds in general for decision-independent state probabilities.

Question: what is the expected value of perfect information for this problem?

Summary of decision rules

The most important decision rules we've seen (in loss/regret form rather than utility form) are:

Minimax loss

$$\min_d \left[\max_{s|d} [l(d,s)] \right]$$

Minimax regret

$$\min_d \left[\max_{s|d} \left[l(d,s) - \min_{d|s} l(d,s) \right] \right]$$

Minimum expected loss

$$\min_d \left[\mathbb{E}_{\mathbb{P}_{s;d}} [l(d,s)] \right]$$

We will examine minimum expected loss more in the next lecture.

Decisions under risk

Overview

Now we consider decision-making under **risk** in more detail, where we take 'risk' to mean we have **known probabilities over states of nature**.

We first (briefly!) recap some basic ideas of **probability** and **expectation**, and then consider how we can go from **utility** to **expected utility**.

Note: we will return to probability modelling in more detail later!

Probability

What is probability? There are multiple 'concrete' **interpretations** of the **same formal mathematical theory** (e.g. Kolmogorov's axioms), e.g.

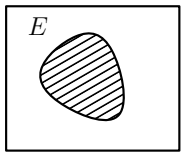
- Classical (ratios of cases)
- Frequency (limit of relative frequencies)
- Propensity ('tendency' to produce an observable frequency of outcomes)
- Subjective (degree of belief or betting behaviour).

See e.g. Resnik (1987), Gillies (2012), Hacking (2006) etc.

There are also **multiple** (equivalent) **formalisations** as well! E.g.

- Kolmogorov, in terms of **sets and subsets**
- Boole, Peirce, Bernstein, Keynes, Cox, in terms of **(logical) propositions**
- Ramsey, De Finetti, in terms of **betting behaviour**
- Huygens, in terms of **expectation** operations

Events vs proposition language:



Probability: standard formalisation (Kolmogorov)

In the standard formalisation of probability theory we have

1. A **sample space** S (or Ω) of all possible 'elementary outcomes' s (i.e. $s \in S$).
2. A collection of **events**, where each event E is a subset of the sample space, $E \subseteq S$.
3. A **probability function** $\mathbb{P}$ (or 'measure') which assigns a 'probability' $\mathbb{P}(E)$ to each **event** $E \subseteq S$,

where

$$0 \leq \mathbb{P}(E) \leq 1, \text{ for all } E \subseteq S$$
$$\mathbb{P}(S) = 1$$
$$\mathbb{P}(\emptyset) = 0, \text{ where } \emptyset \text{ is the empty set}$$
$$\text{If } A, B \subseteq S \text{ are mutually exclusive, i.e., } A \cap B = \emptyset, \text{ then}$$
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

(See other courses/books/later for more detail).

Note the set $\leftrightarrow$ proposition relationship:

$$\mathbb{P}(E) \leftrightarrow \mathbb{P}(s \in E) \leftrightarrow \mathbb{P}(\text{'event } E \text{ occurs'})$$

Example

Toss a coin twice. Then

$$S = \{HH, HT, TH, TT\}$$

$$E = \text{ “first coin is heads”} = \{HH, HT\} \subseteq S$$

Fair coin, independent tosses:

$$\mathbb{P}(\{s\}) = 0.25 \text{ for all } s \in S$$

Note that need to use subset $\{s\}$ so that we have an event. See ‘random variables’ for how to make this easier.

Conditional probability

For $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

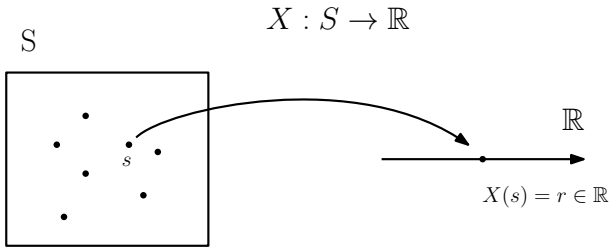
where first equality is the **definition** and second is **Bayes’ theorem**, which follows from the definition of conditional probability and $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$.

Note: using **Bayes’ theorem** is not the same as e.g. doing **Bayesian statistics** or **Bayesian decision theory**.

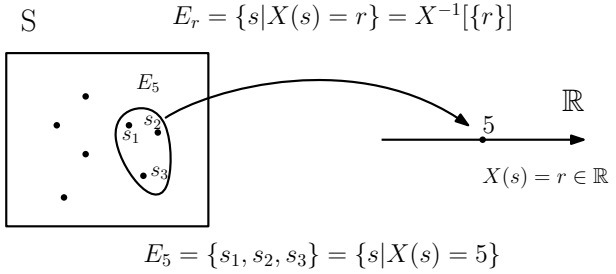
Random variables

Random variables **assign numbers to outcomes** and allow us to **define events** as sets of **outcomes with the same assigned number**.

A random variable is thus (formally) actually a **function from outcomes to numbers**:



Particular values of random variables define events:



We define

$$\mathbb{P}(X = x) = \mathbb{P}(E_x)$$

Expectation of random variables

For random variables X, X_1, X_2, expectation $\mathbb{E}$ satisfies

1. If $X > 0$, $\mathbb{E}[X] > 0$
2. If c is a constant, $\mathbb{E}[cX] = c\mathbb{E}[X]$
3. $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$
4. $\mathbb{E}[1] = 1$

where the '1' inside the last expectation is the constant random variable equal to the number 1 for all outcomes and the '1' on the right hand side is the scalar number 1. These axioms define $\mathbb{E}$ as a **linear operator on random variables**.

Can even **define probability starting from expectation!** For event A, define

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A]$$

where $\mathbb{1}_A$ is the **indicator random variable** for event A:

$$\mathbb{1}_A(s) = \begin{cases} 1 \text{ if } s \in A \\ 0 \text{ otherwise} \end{cases}$$

Example

As before, consider

$$S = \{HH, HT, TH, TT\}$$

$X(s)$ = "number of heads in $s \in S$".

So X values are $\{2, 1, 0\}$ and

$$\mathbb{P}(X = 2) = \mathbb{P}(\{HH\})$$
$$\mathbb{P}(X = 1) = \mathbb{P}(\{HT, TH\})$$
$$\mathbb{P}(X = 0) = \mathbb{P}(\{TT\})$$

where the left-hand sides are the random variable versions of the event-based right-hand sides.

Computing expectations for discrete random variables

For **discrete random variables** the expectation becomes a **weighted average**.

A discrete random variable is defined by its probability mass function

$$\mathbb{P}(X = x_i) = p_i,\ i = 1, ..., M$$

where $\sum_i p_i = 1$, and the expectation of the random variable is given by

$$\mathbb{E}[X] = \sum_{i=1}^{M} p_i x_i.$$

Sample expectations

Suppose we have N realisations (samples) of the **discrete** random variable X, generating a sequence of realised values

$$\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N$$

where each $\mathrm{x}_i \in \{x_1, x_2, ..., x_M\},\ i = 1, ..., N$ is a (possibly repeated) value from the original sample space of X. We can then use the **sample relative frequencies** to get a sample approximation to $\mathbb{E}$ and take

$$\begin{aligned} p_i &\approx f_i \\ &= \text{ observed relative frequency of occurrence of outcome } x_i \\ &= \frac{\#\text{ times outcome } x_i \text{ occurred}}{\#\text{ 'trials'}} \end{aligned}$$

Then

$$\mathbb{E}[X] \approx \sum_{i}^{M} f_i x_i$$

This gives the average across the *original sample space*, $\{x_1, x_2, ..., x_M\}$, based on the frequencies. We can also get the same result by weighting each (possibly repeated) *sample realisation* value x_i by $1/N$, hence averaging along the *sample sequence*:

$$\mathbb{E}[X] \approx \sum_{i}^{N} \frac{1}{N} \mathrm{x}_i = \frac{1}{N} \sum_{i}^{N} \mathrm{x}_i.$$

Here $\mathrm{x}_i \in \{x_1, x_2, ..., x_M\}$ is the value of the ith *sample realisation*, and there may be repeated *sample space* values, i.e. we may have $\mathrm{x}_i = \mathrm{x}_j$ for $i \neq j$.

The same ideas generalise to **continuous** random variables, we just have vanishing probability of getting repeated values.

Utility revisited: expected utility rule?

How to we justify/get to expected utility? Following von Neumann and Morgenstern, **the key idea** is to go from:

Preferences over **outcomes**

$\leftrightarrow$

Utility of **outcomes**

to:

Preferences over **'lotteries'** over outcomes

$\leftrightarrow$

Utility of **'lotteries'** = **Expected utility** of outcomes under lottery

Let's consider this process in more detail.

Lotteries

Lotteries (also called 'prospects') consist of

- A collection of **outcomes** (or 'prizes')
- A **probability** for each outcome

E.g.

Example lottery L:

	win	lose	draw
probability	0.6	0.3	0.1

We can start from *two* (and *one*) outcome lotteries and build up **compound** lotteries.

Define

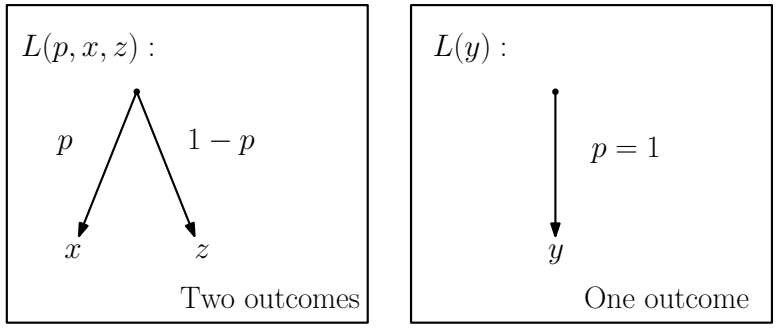
$$L(p, x, y)$$

to mean a **lottery** where you recieve outcome (prize) x with probability p and outcome y with probability $1 - p$.

A **single outcome** lottery, or a lottery with $p = 1$, just means you are guaranteed to receive the only possible outcome, say x. We can then write the lottery as $L(x)$.

Each outcome can hence be identified with a 'degenerate' lottery.

Lottery diagrams and utility



Question: how do we **value** (assign utilities to) **lotteries** (c.f. outcomes).

E.g. what would you do if offered a choice between the above two lotteries themselves.

Expected utility: sketch/motivation of VNM

To value (i.e. assign a utility value to) a lottery, we begin by assuming we have utilities over **outcomes**, say

$$u(x) > u(y) > u(z)$$

where these might represent e.g.

- x is trip to Wellington
- y is trip to Christchruch
- z is trip to Dunedin

VNM then started with the following steps:

1. Define utility of the **lottery** with **guaranteed outcome** y as the utility of the associated outcome, $u(L(y)) = u(y)$
2. Given y above, ask **"for what probability p would you have no preference between (be indifferent between) $L(p, x, z)$ and $L(y)$?"**

A **key assumption** of VNM is that for **any** x, y, z outcomes such that $u(x) > u(y) > u(z)$, we can **always** construct a lottery with some probability p such that $L(p, x, z)$ and $L(y)$ are **valued the same** (assigned the same utility).

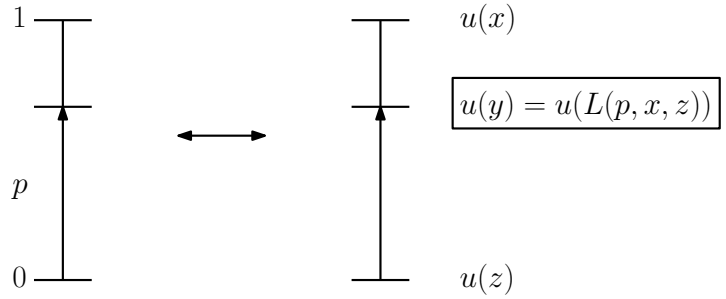
Along with other axioms, VNM **derived the expected utility rule** for assigning utility to **lotteries** (see e.g. Resnik for more details/axioms).

VNM axioms and interval scale

The VNM axioms (assumptions) lead to an **interval** utility scale, where **relative differences** in utilities matter, not just their order.

Such scales are unique up to a **positive linear transformation** of utility, i.e. $u' = au + b$ where u' is the new utility, u the old utility, and a and b are constants with $a > 0$.

We can visualise the resulting scale as



probability of lottery	⟷	relative degree of preference between lottery outcomes

showing that the **probability** you would accept for the lottery is tied to your relative degree of preference for outcomes.

'Derivation'

From the picture above we can argue (see e.g. Resnik for full details)

$$\begin{aligned} u(L(p,x,z)) &= u(y) \text{ (indifferent)} \\ &= u(z) + p[u(x) - u(z)] \text{ (location on interval)} \\ &= pu(x) + (1-p)u(z) \\ &= \mathbb{E}[u(\text{outcome})] \text{ (expected utility of outcomes under lottery)} \end{aligned}$$

i.e.

$$u(L(p,x,z)) = \mathbb{E}[u(\text{outcomes})] \text{ under lottery}$$

Utility of lottery = expected utility of outcomes under lottery!

VNM summary

von Neumann and Morgenstern proved that **if** people's **preferences** between **lotteries** satisfy a particular set of axioms, **then** the **utility** of a **lottery** is given by the **expected utility** of the **outcomes** under that lottery.

- This is called the 'expected utility theorem'
- It is a **representation** theorem, showing how to **represent** one thing (preference structure over lotteries) in terms of another (expected utility under lottery)

General lotteries

Same idea as simple lotteries.

Can define

$$L(\mathbb{P}, \mathcal{Z})$$

where $\mathbb{P}$ is a probability distribution defined over the outcome set $\mathcal{Z}$.

Then we have, under the VNM axioms,

$$u(L(\mathbb{P}, \mathcal{Z})) = \mathbb{E}_{\mathbb{P}}[u(z)]$$

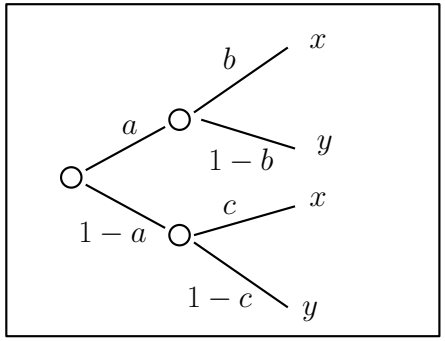
where $u(z)$ is the utility of the outcome z.

Compound lotteries

According to VNM assumptions we can evaluate **compound** lotteries, i.e. lotteries where the outcomes are lotteries etc, according to **probability trees**.

E.g.

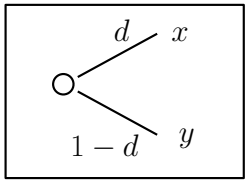
$$L_1 = L(a, L(b, x, y), L(c, x, y))$$



$$\equiv$$

$$L_2 = L(d, x, y)$$

where $d = ab + (1 - a)c$



i.e.

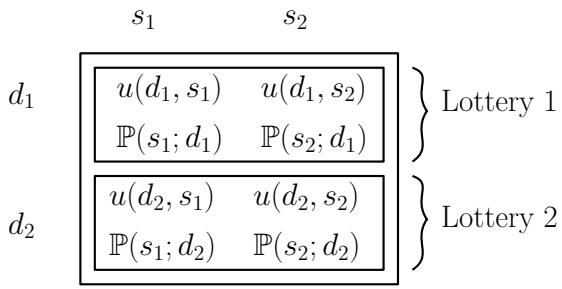
$$u(L_1) = u(L_2)$$

Application to decision theory??

After that detour, we see that VNM tell us how to **value lotteries** and hence enable us to **choose between them** according to the principle that **we should choose the option we value the most, i.e. assign the highest utility to**.

In the context of **decision theory**, **each possible decision defines a lottery** and we value decisions by **valuing the associated lottery**. We can then choose the decision to take by **choosing the possible decision with highest utility**.

E.g.



and

$$u(d_1) = u(\text{lottery 1}) = \mathbb{E}[u(d_1, s)]$$
$$u(d_2) = u(\text{lottery 2}) = \mathbb{E}[u(d_2, s)]$$

We then choose d_1 if $u(d_1) > u(d_2)$ etc.

By choosing the **highest expected utility decision** we are simply **choosing the lottery we prefer**! Again, this is a **representation theorem**, representing pre-existing preferences.

Utility of a decision problem

We have seen that the value of a decision under risk is given (under VNM conditions) by the expected utility of the associated lottery.

We can also define the value of a **set of possible decisions**, i.e. the **value of a decision problem**, as the maximum utility of available decisions

$$u(\{d_1, d_2, ...\}) = \max_{d_i \in \{d_1, d_2, ...\}} u(d_i)$$

This also helps when considering **multistage decisions under risk**, which comes up in **dynamic programming** (see next module):

- The process known as **'averaging out and folding back'** (Raiffa, 1968) in decision theory is equivalent to (Bellman's) **dynamic programming** in the presence of risk.
- These problems involve **decision trees** (see Resnik handout from Lecture 1) with both **decision nodes** and **chance (risk) nodes**.
- Briefly, **decision nodes** are valued according to the above (max utility over possible decisions), while **chance (risk) nodes** are evaluated according to expected utility.
- Solving 'backwards' from the final outcomes (which we can value immediately) allows us to value the multistage decision problem and determine the best **policy**, which is a **'vector-valued decision' of all the 'elementary' decisions** we make during the multistage process.

Again, you will see these sorts of ideas more in later modules!

Problems

A) Suppose we have three outcomes x, y, z with $u(x) = 1, u(y) = c, u(z) = 0$, for some c between 0 and 1.

   Suppose you are indifferent between a lottery $L(p, x, z)$ and y for a given p. Determine c, i.e. the utility of y, in terms of p.

B) Calculate the expected utility of the compound lottery:

$$L(0.6, L(0.5, 0.4), 1)$$

C) Solve problem 1 in Resnik 4-1 (attached) about the two horses Ace and Jack.

D) Complete any unfinished examples from Lecture 1!

E) Solve the tutorial problems associated with the first two lectures (see Canvas).

Statistical decision theory: basic concepts

Overview

So far we have considered what we've called **simple decision theory** where we are strictly considering either **risk** – here meaning we are given known probability distributions over outcomes of nature (states of nature) – **or** a scenario of **ignorance** – here meaning there is no information about the probabilities of the states of nature.

Statistical decision theory considers an intermediate case where there is an element of **both risk and ignorance**.

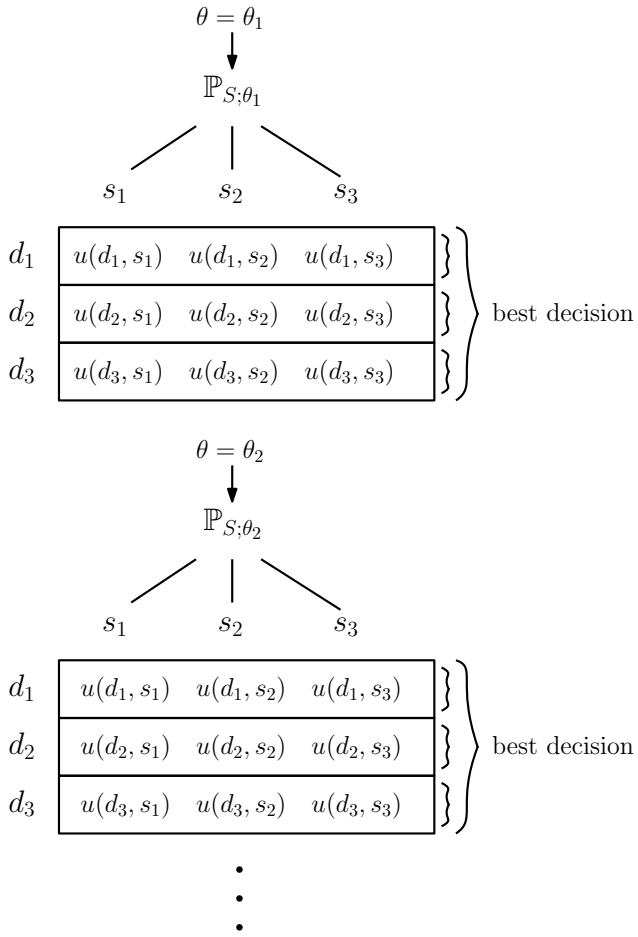
For example, we might know that the outcomes of nature follow **one of a given family of distributions** but we **don't know which one**. We have **ignorance about the risk**!

We will see how much further we can get in just this setting (not very!) before considering the **other additional component** of statistical decision theory: **empirical data**. For example we have **samples** (realisations) from an **unknown distribution** for X depending on parameter θ:

$$X \sim \mathbb{P}_{X;\Theta}, \text{ where } \theta \text{ is unknown}$$
$$\mathrm{x}_0 \text{ is the realised value of } X.$$

First: No data statistical decision theory

In the 'no data' setting of statistical decision theory we can imagaine that we now have **multiple possible decision tables**:

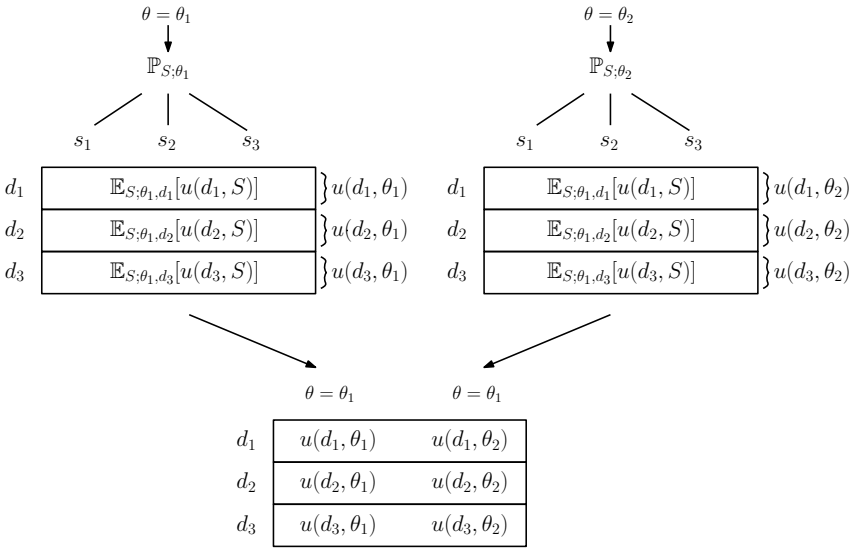


But which table??

We **don't know** θ, which we might call **'nature's parameter'**.

Now, **in statistical decision theory θ itself is often called the 'state of nature'** from the beginning, though in terms of 'simple decision theory' this is more like a distribution or lottery over states (outcomes) of nature.

From this point of view, a better term might be **'Nature's distribution'** or **'Nature's lottery'**. However, we can also see why the term state of nature can be applied by recalling that we can value – **assign utilities too** – lotteries using their **expected utility**. This effectively **reduces an unknown distribution to an unknown state of nature** in a **new decision table**:



Wald decision table

In **Wald-style** statistical decision theory we usually start from ('reduced') tables like we just obtained, i.e.

	$\theta = \theta_1$	$\theta = \theta_2$
d_1	$u(d_1, \theta_1)$	$u(d_1, \theta_2)$
d_2	$u(d_2, \theta_1)$	$u(d_2, \theta_2)$
d_3	$u(d_3, \theta_1)$	$u(d_3, \theta_2)$

Now, unless we have a **distribution over distributions**, i.e. over the θ values, we can't apply expected utility again! This leads to some options:

- Minimax
- Change the loss or the decisions available
- Assume/come up with a distribution over distributions (a *prior* in the Bayesian approach)
- Use some data if we have it!

Before considering more sophisticated approaches to the above Wald-style tables, let's go back to the 'pre-reduction' form and consider the **simplest** way we might use **empirical data**, leading to…

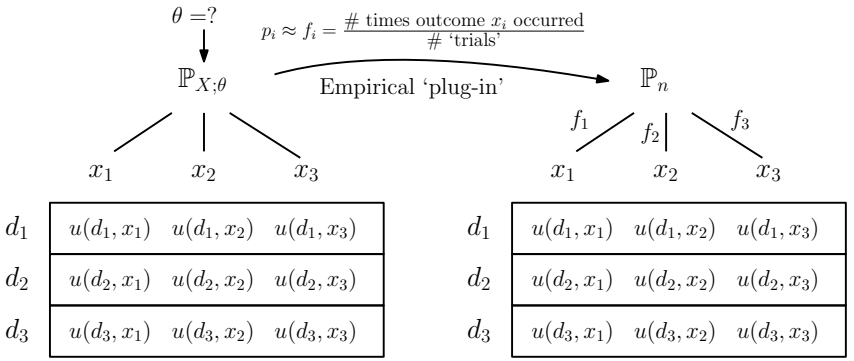
'Naive' empirical approach

Suppose we have a sample of IID (independent and identically-distributed) data representing measurements of our 'outcomes of nature', here denoted X, sampled from **Nature's actual distribution**, i.e.

$$X \sim \mathbb{P}_{X;\theta_{\text{true}}},$$
$$\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N \text{ is a sequence of realised values of } X$$

where each x_i takes a value in the original sample space, e.g. $\{x_1, ..., x_M\}$.

We can then consider the **empirical distribution** over **states (outcomes) of nature** X giving a problem of '**decision-making under empirical risk**".



True decision table, under 'risk'...
if θ was known (which it isn't!)

Decision table under empirical risk

Notes:
- This clearly requires either a repeatable experiment or a repeatable simulation model etc.
- The continuous version will just have frequencies $1/n$ on the sample data points and zero elsewhere, assuming each continuous value only occurs once.

Decision-making under empirical risk

This leads to the idea of **solving the decision problem under empirical risk** (known probabilities = empirical relative frequencies) and **hoping the result is similar to solving the true decision problem**. The essential idea of this approach has been invented/proposed/discovered in many fields under different names:

- **Analog/as-if/plug-in** estimation and decision making in statistics and econometrics, along with related ideas such as **maximum likelihood**, **bootstrap**, **M-estimation** and **minimum score** estimation (Fisher, Efron, Manski, Huber, Dawid).
- **Empirical risk minimisation** in machine learning (Vapnik etc)
- **Sample average approximation** (SAA) in operations research (Shapiro etc)

Along with many more!

A simple setting where this type of decision problem is particularly relevant is **statistical estimation** problems where the goal is to 'decide' on a **best summary of a probability distribution**.

Importantly, even though we will ultimately 'plug-in' our empirical approximation, giving a 'decision under empirical risk' problem, it is **important conceptually to first consider the 'ideal' problem that we would solve if the risk (true distribution) was known, as this is the problem that we wish to approximate**.

We can call this our **target decision problem** and the solution the **target decision**, borrowing from the statistical concept of a **target parameter**.

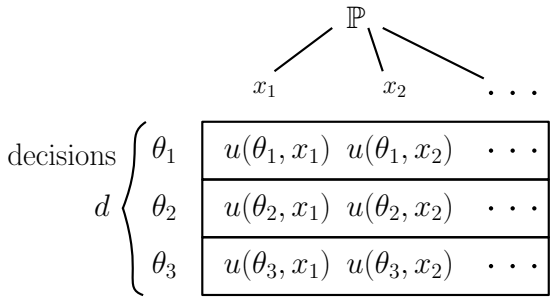
Defining target decisions in statistics: Population parameters

Imagine we had 'infinite samples', i.e. knew the distribution over state outcomes exactly – a case of **known risk**.

We haven't 'parameterised' this, it is just **some given distribution** $\mathbb{P}$ over the state outcomes, here denoted by x for the outcomes and X for the associated random variable.

We want to 'decide' on a **single number summary** of this distribution/the associated random variable. This clearly **'loses information'** about the variability of the distribution/random variable – there is a 'loss' from just giving a single number.

These single number summaries of distributions are often called **population/summary/descriptive parameters**. Leads to the (typically *continuous* equivalent of the) below 'table':



A common example loss for given outcome x is the quadratic loss:

$$l(\theta, x) = \frac{1}{2}(\theta - x)^2$$

Best 'decision' according to minimum expected loss:

$$\min_\theta \mathbb{E}_X[l(\theta, X)] = \min_\theta \mathbb{E}_X[\frac{1}{2}(\theta - X)^2]$$

Solution to target problem

Consider

$$\min_\theta \mathbb{E}_X[\frac{1}{2}(\theta - X)^2]$$

Note that $\mathbb{E}_X[\frac{1}{2}(\theta - X)^2]$ is a deterministic function of θ as the randomness has been 'averaged out'. So this has the form

$$\min_\theta f(\theta)$$

where $f(\theta) = \mathbb{E}_X[\frac{1}{2}(\theta - X)^2]$. Assuming the minimum is given by the usual 'first-order' condition, we can hence solve

$$\frac{d}{d\theta} f(\theta) = \frac{d}{d\theta}\mathbb{E}_X[\frac{1}{2}(\theta - X)^2] = 0.$$

Now, we would have to think carefully about what's going on if this involved differentiating with respect to the random variable (see e.g. stochastic calculus!) but here we are differentiating with respect to a non-random quantity and it holds (under 'standard conditions') that

$$\frac{d}{d\theta}\mathbb{E}_X[l(\theta, X)] = \mathbb{E}_X[\frac{\partial}{\partial\theta} l(\theta, X)]$$

i.e. **we can 'move the derivative inside the expectation'** (another view: write the expectation as an integral and use the Leibniz integral rule).

This means we have

$$\begin{aligned}\frac{d}{d\theta}\mathbb{E}_X[\frac{1}{2}(\theta - X)^2] &= \mathbb{E}_X[\frac{\partial}{\partial\theta}\left(\frac{1}{2}(\theta - X)^2\right)] \\ &= \mathbb{E}_X[\theta - X] \\ &= 0,\end{aligned}$$

which implies, as expectation is **linear** and θ is a constant

$$\theta = \mathbb{E}_X[X] = \text{mean of } X!$$

Sample version

Unfortunately, of course, we don't know the probability distribution of X and can't take the expectation. However, we can take the **empirical analog** of the expectation!

We go from

$$\min_\theta \mathbb{E}_X[\frac{1}{2}(\theta - X)^2]$$

to

$$\min_\theta \frac{1}{n}\sum_{i=1}^{n}[\frac{1}{2}(\theta - \mathrm{x}_i)^2]$$

where x_i is the value of the ith sample realisation.

Analogously (here using the fact that the derivative of a sum is the sum of derivatives, since the derivative is a linear operation), we get that the solution satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(\theta - \mathrm{x}_i) = 0$$

i.e. (as the sum operation is linear and the sum of n copies of θ is $n\theta$)

$$\frac{1}{n}n\theta = \frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i$$

i.e.

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i = \text{sample mean}.$$

This is simply the **sample analog** of our target, i.e. the **sample mean**!

> Thus our 'empirical decision' is to use the sample mean, while our 'target decision' is the true mean.

Literature

As indicated, our empirical decision is **not exactly the correct/target decision, but is hopefully 'close'**.

The literature mentioned previously extensively discusses the conditions for this 'closeness' to hold (e.g. Vapnik, Fisher, Huber, Wald, Le Cam etc in machine learning and statistics).

The short version is the most solid guarantees hold if the **'amount of data'** (e.g. sample size n) is **'big' relative to the 'space of allowed decisions'** (e.g. parameter space dimension p). This is the 'asymptotic', $n \gg p$ regime of **classical statistics**.

However, it is worth noting that **modern machine learning and statistics** typically address the **opposite setting** where the decision (parameter) space is very large compared to the sample size, i.e. $p \gg n$.

This leads to the need for 'regularisation' of some sort and other ideas. See e.g. **ENGSCI 721**! See also Vapnik handout.

In this course we will instead consider the classical, simpler 'lots of data, few decisions' setting.

Problems

A) Reconsider the statistical decision problem from the lecture with the loss function $l(\theta, X) = \frac{1}{2}(\theta - X)^2$. This can be considered the problem of finding the best 'constant summary' of the random variable X. Follow this alternative derivation approach that doesn't require any exchange of differentiation and expectation. Rather than differentiate this immediately, instead:

- First expand out the quadratic
- Then take the expectation to 'remove the randomness'
- Then differentiate the remaining deterministic expression and solve for the best θ.

B) What is the best summary of the distribution of a function $f(X)$ of X under squared loss $l(\theta, f(X))$. How does this relate to the result for the best summary of X? Do they give 'consistent' answers (hint: consider the cases of linear/nonlinear f)?

C) Given a sample of size n what are the best empirical summaries of X and $f(X)$ according to the 'empirical risk' approach?

Statistical decision theory following Wald: Decision functions, minimax

Overview

Recall that in **Wald-style** statistical decision theory we usually start from ('reduced') tables of the form:

	$\theta = \theta_1$	$\theta = \theta_2$
d_1	$u(d_1, \theta_1)$	$u(d_1, \theta_2)$
d_2	$u(d_2, \theta_1)$	$u(d_2, \theta_2)$
d_3	$u(d_3, \theta_1)$	$u(d_3, \theta_2)$

In contrast to the 'empirical risk' approach, here we are measuring the 'actual' quality of the decision relative to the truth. In fact, **we could use this approach to help assess the quality of the empirical risk decision approach**: empirical risk gives us a decision that is best under the empirical distribution approximation that we then aim to assess under the actual distribution...in principle!

Note: in Wald framework usually assume in 'regret form' i.e. that we have $\min_d l(d, \theta) = 0$.

The problem

However, as mentioned, unless we have a **distribution over distributions**, i.e. over the θ values, we can't apply expected utility again. Recall some options:

- Minimax (without data)
- Change the loss or the decisions available
- Use some data (in a new way?) if we have it!
- Assume/come up with a distribution over distributions (a *prior* in the Bayesian approach)

Let's (briefly) consider the first couple of options.

No data minimax

Consider the simple game of **guessing an unknown parameter that can take only one of two values**, $\{\theta_1, \theta_2\}$.

Because we are silly, **we will consider guesses (decisions) from the set of three values** $\{\theta_1, \theta_2, \theta_3\}$, even though θ_3 is known to be wrong. (Note that here we enforce a requirement that we must choose a single value or 'point estimate' – in principle though our decisions don't have to be point estimates and could be e.g. 'interval estimates' like confidence intervals).

For simplicity, will use the 0/1 loss

$$\text{loss}(d_i, \theta) = \begin{cases} 1 \text{ if } d_i \neq \theta \\ 0 \text{ if } d_i = \theta \end{cases} .$$

This leads to...

No-data Wald decision table

Our loss table with minimax calculations shown:

d	θ_1	θ_2	max
θ_1	0	1	1
θ_2	1	0	1
θ_3	1	1	1
		min	1 (all d)

There is **no solution that is any better than the others!** (Maybe that's fair enough!)

However, note that decision d_3 is always **at least as bad** and is also **sometimes worse** than d_1 and d_2. The **dominance principle** says that d_3 is **dominated** by d_1 and d_2 and hence can be discarded.

> **Dominance**: a decision d dominates another decision d' if d is **sometimes better** than d' and **never worse**.

The dominance principle applies under

- Ignorance
- Risk, **if the probabilities are independent of decisions** (see Resnik handout from Lecture 1).

Note: this principle goes (slightly) **beyond minimax: a decision can obtain the same minimax value as another decision that dominates it!** Pure minimax would treat these as no different, dominance prefers one to the other.

Reduced table

If we accept the dominance principle we obtain

d	θ_1	θ_2	max
θ_1	0	1	1
θ_2	1	0	1
		min	1 (all d)

However, this still doesn't lead to a definitive choice (makes sense without data!).

Two player game

Consider the previous table as a 'two player' game where nature chooses a θ (but doesn't tell you which!) and then you choose a θ that you hope matches:

player 1 (you)	player 2 (Nature) θ_1	player 2 (Nature) θ_2
θ_1	0	1
θ_2	1	0

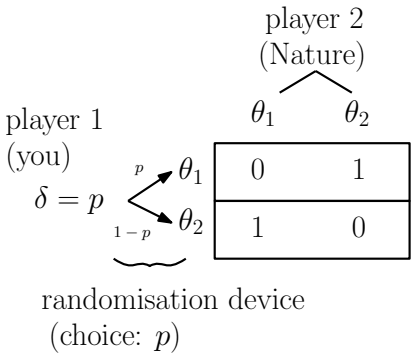
You play this game over and over. What should you do?

Note: this can be transformed via a positive linear transformation into a so-called 'zero sum two player game' in the sense of game theory, where your loss is Nature's gain and vice-versa.

Mixed/randomised strategies

In a mixed/randomised strategy, you deliberately introduce randomness into your choice of decision. Think: 'keep your opponent guessing!'

For example, you introduce a 'biased coin' with probability p of choosing θ_1 and probability $1-p$ of choosing θ_2:



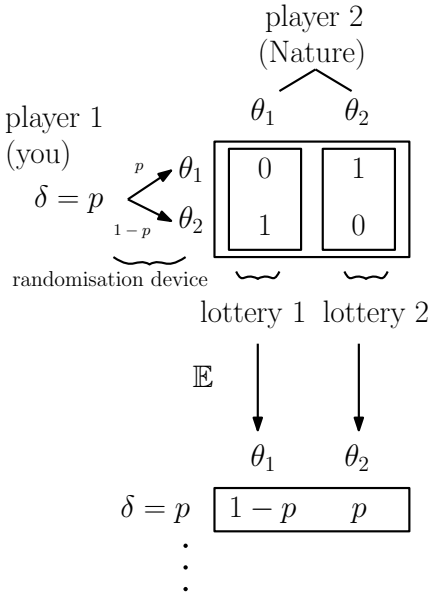
(We have labelled this new type of (randomised) decision as δ, rather than d).

The **new decision problem** is **which probability to use** for your coin (randomisation device).

Note: in statistical decision theory an 'estimator' can generally correspond to a mixed/randomised strategy in game theory. A 'randomised estimator' adds in additional randomness on top of this!

Lotteries again

Mixed/randomised strategies generate lotteries in the 'transpose' direction (now columns are lotteries):

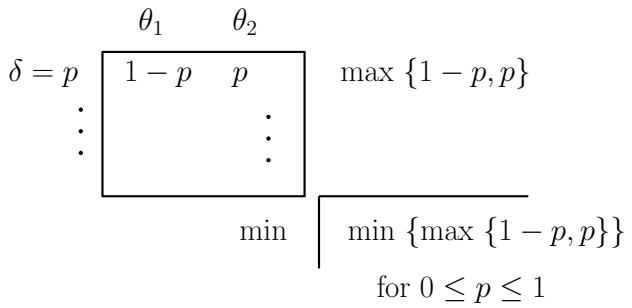


Above, for a given p, we get two lotteries, depending on the true value of θ. We have then valued these according to expected loss/utility!

The problem is to determine the 'best' p for $0 \leq p \leq 1$.

Back to ignorance

Now we have a decision problem with no more randomness. How do we choose p given θ is some fixed but unknown value? Try minimax again!



The solution to this is

$$\min_{\{p|0\leq p\leq 1\}}\{\max_{\{p|0\leq p\leq 1\}}\{1-p, p\}\} = 0.5$$

with $p = 0.5$.

Proof: if $p \geq 0.5$ say, then $1 - p \leq 0.5 \leq p$ and so $\max\{1-p, p\} = p$.

Then $\min\{p\} = 0.5$ (still assuming $p \geq 0.5$). The same idea works for $p \leq 0.5$ i.e. $1 - p \geq 0.5$, giving the same answer of $p = 0.5$ and hence covering both cases.

So…**have two options/states of nature, no data? Flip a fair coin!**

However

- This gets bad for 0/1 loss as the number of possible states of nature gets large
- Further randomisation doesn't help

Options for dealing with many states of nature

Options:

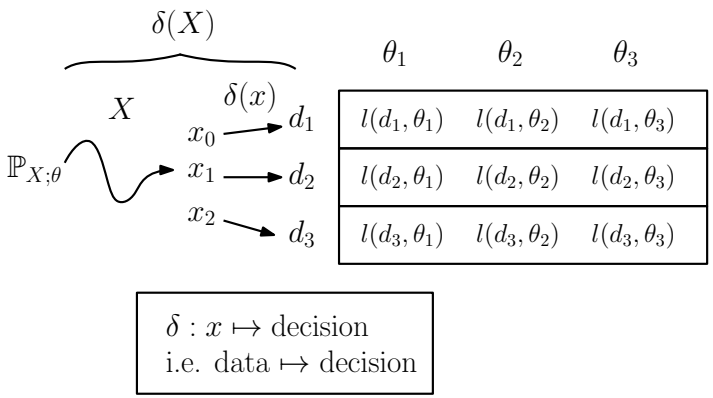
- Change loss function (e.g. quadratic loss)
- Change decision space (e.g. sets of values)
- Use a prior (Bayes)
- Add constraints
- Use data somehow

Consider using data in a new way…(c.f. empirical risk):

Using data in the Wald framework

The idea of the Wald framework is to consider **decision functions** which are like 'randomised/mixed strategies' in game theory but where **the randomness comes from the actual distribution via the data**.

This is done by considering **statistical decision functions** δ which are functions **mapping realisations of data to decisions**:



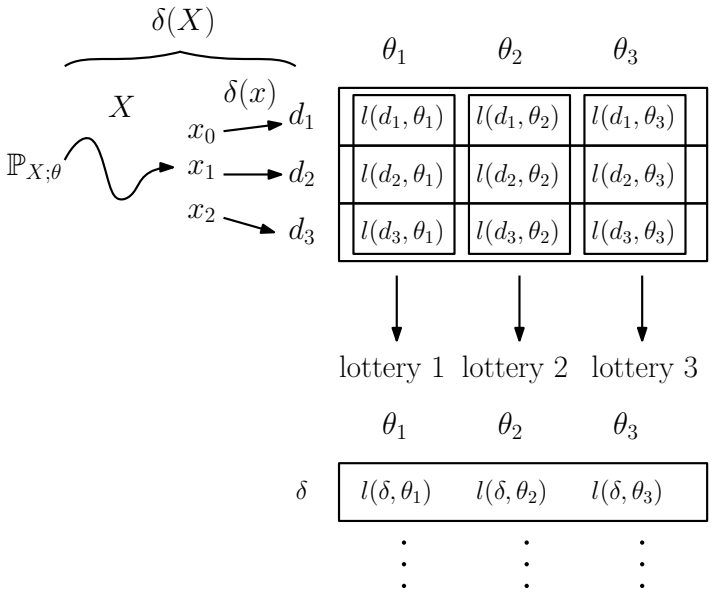
Notes:

- For a given data realisation x, $\delta(x)$ is a deterministic function. The randomness comes from x being the result of a realisation of the random variable X. The associated 'mixed/randomised' strategy (in the sense of game theory) is $\delta(X)$. We will call the **function itself, i.e. δ, the statistical decision function**.
- Statistical decision functions are also known as **decision rules** in the decision theory literature and **estimators** in the statistics etc literatue.

Statistical decision functions continued

As mentioned, these are effectively 'randomised/mixed strategies', except the randomness is based on the true distribution via having access to data realisations. The decision problem now is **which decision function** to choose.

For any candidate δ we can do the usual expected loss reduction to remove the randomness:



As indicated by the dots above, we can do this for any candidate δ and hence get a row for each candidate decision function. This leads to...

Risk function

We can hence assign a loss to the decision function (randomised/mixed strategy) itself:

$$l(\delta, \theta) = \mathbb{E}_{X;\theta}[l(\delta(X), \theta)]$$

This is often called the **risk function** of a **decision rule** δ in statistics: the expected loss (where the loss is usually in regret form) for the decision function.

Though it has a special name, and often denoted by R, this is just the usual loss that follows from the expected loss reduction above, i.e.

$$R(\delta, \theta) := l(\delta, \theta).$$

When we want to avoid confusion with other uses of 'risk', we will refer to this as the **Wald risk**. As with the no-data case, we get a 'table' (collection of all function values if continuous) like:

	θ_1	θ_2	θ_3
δ_1	$R(\delta_1, \theta_1)$	$R(\delta_1, \theta_2)$	$R(\delta_1, \theta_3)$
δ_2	$R(\delta_2, \theta_1)$	$R(\delta_2, \theta_2)$	$R(\delta_2, \theta_3)$
δ_3	$R(\delta_3, \theta_1)$	$R(\delta_3, \theta_2)$	$R(\delta_3, \theta_3)$
	$\vdots$	$\vdots$	$\vdots$

where $R(\delta_i, \theta_j) := l(\delta_i, \theta_j)$

In general there will be no one decision that dominates the others. Minimax can help pick out a solution but again will not in general lead to a unique solution.

Changing the loss function can sometimes help a little. For example, typically decision problems in e.g. statistics involve continuous variables for which 0/1 loss doesn't really make sense for a point estimate (though it does for an interval estimate!).

Squared error loss in the Wald setting

Consider the squared error loss between **the output of decision function** and a **fixed but unknown parameter/state of nature** θ:

$$l(\delta(x), \theta) = (\delta(x) - \theta)^2.$$

The associated risk (loss for decision function itself) is then

$$R(\delta, \theta) = l(\delta, \theta) = \mathbb{E}_{X;\theta}[(\delta(X) - \theta)^2].$$

Note: this is **subtly different** to the case where *we choose a constant* to summarise a *random outcome of Nature*. Here the **decision function** leads to a **random decision** for a fixed constant *chosen by Nature*.

This actually has important implications for determining the 'best decision function' for a fixed unknown state of nature, c.f. the 'best (constant) decision' for a randomly varying state of nature. Here we want to solve e.g.

$$\min_{\delta} \mathbb{E}_{X;\theta}[(\delta(X) - \theta)^2]$$

which is a minimisation over **functions**, vs e.g. what we had in the 'empirical risk case' where we approximated:

$$\min_{d} \mathbb{E}_{X;\theta}[(f(X) - d)^2].$$

The problem involving $\delta(X)$ is much harder in general. However, in the case of squared error loss (c.f. general loss), we can derive a useful **bias-variance decomposition** of the total risk.

Bias-variance decomposition for squared loss

Consider our decision function, $\delta(X)$, here as a random variable (as it is a function of the random variable X). As we have seen, this induces the random loss $l(\delta(X), \theta)$ and we take the expecation of this to get the (Wald) risk: $R(\delta, \theta) = \mathbb{E}_X[l(\delta(X), \theta)]$.

Now consider the expected value of the decision function defined as:

$$\bar{\delta} := \mathbb{E}_{X;\theta}[\delta(X)].$$

(Note that this expectation depends on θ so we could write this as e.g. $\bar{\delta}_\theta$ to be more explicit.) Now it shouldn't be too suprising that for an arbitrary decision function δ (which we can choose) and a θ 'chosen by Nature', $\bar{\delta} \neq \theta$. (E.g. choose $\delta(X) = 3$ independently of the data.)

More interestingly, **we can show that the best decision rule does not in general require** $\bar{\delta} = \theta$. To gain some intuition for this, we can remember that $\bar{\delta}$ can be considered the best 'single number summary' of the decision rule $\delta(X)$. Requiring $\bar{\delta} = \theta$ means we are requiring our 'best single number summary' of δ to match the true parameter. However, as we have seen e.g. in the coin flip example, introducing additional randomisation can often improve our performance – there is more to our decision function that its mean!

In particular, in our squared-error case we have…

Bias-variance decomposition continued

$$\begin{aligned}
R(\delta, \theta) &= \mathbb{E}_{X;\theta}[l(\delta(X), \theta)] \\
&= \mathbb{E}_{X;\theta}[(\delta(X) - \theta)^2] \\
&= \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta} + \bar{\delta} - \theta)^2] \\
&= \mathbb{E}_{X;\theta}[\big((\delta(X) - \bar{\delta}) + (\bar{\delta} - \theta)\big)^2] \\
&= \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})^2 + 2(\delta(X) - \bar{\delta})(\bar{\delta} - \theta) + (\bar{\delta} - \theta)^2] \\
&= \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})^2] + \mathbb{E}_{X;\theta}[2(\delta(X) - \bar{\delta})(\bar{\delta} - \theta)] + \mathbb{E}_{X;\theta}[(\bar{\delta} - \theta)^2]
\end{aligned}$$

Considering the middle term, we see that all the terms in

$$2(\delta(X) - \bar{\delta})(\bar{\delta} - \theta)$$

are constant except $\delta(X)$ and $\mathbb{E}_{X;\theta}[af(X)] = a\mathbb{E}_{X;\theta}[f(X)]$ for any constant a, we have

$$\mathbb{E}_{X;\theta}[2(\delta(X) - \bar{\delta})(\bar{\delta} - \theta)] = 2(\bar{\delta} - \theta)\mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})].$$

But

$$\mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})] = 0$$

as $\mathbb{E}_{X;\theta}[\delta(X)] = \mathbb{E}_{X;\theta}[\bar{\delta}] = \bar{\delta}$ by definition.

For the last term, we note that the components are all constant and so

$$\mathbb{E}_{X;\theta}[(\bar{\delta} - \theta)^2] = (\bar{\delta} - \theta)^2$$

So…all that algebra leads to…

Bias-variance decomposition continued

$$R(\delta, \theta) = \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})^2] + (\bar{\delta} - \theta)^2$$

Defining

$$\text{bias}_\theta(\delta) = \bar{\delta} - \theta = \mathbb{E}_{X;\theta}[\delta] - \theta$$

and

$$\text{variance}_\theta(\delta) = \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})^2]$$

We have, (for squared error loss)

$$R(\delta, \theta) = \text{variance}_\theta(\delta) + \text{bias}^2_\theta(\delta)$$

This implies the so-called **bias-variance tradeoff**: if we restrict to 'unbiased decision functions' (not a great name really!) where the mean of the decision function matches the true parameter, i.e. $\bar{\delta} = \theta$, then the total Wald risk becomes just the variance and we just minimise the variance of our estimator...

However, there are some cases where we can 'accept some bias' in the decision function in order to further reduce its variance such that the total risk goes down.

Note: This trade-off is particularly important for 'small sample sizes' – generally we want/expect the bias to go to zero faster than the variance as the sample size gets large (asymptotically unbiased), though in general we do want both to go to zero as sample size gets large (giving a 'consistent' decision function).

After all that, let's do an example!

Example

From Wasserman (2004), 'All of statistics'.

Suppose $X \sim N(\theta, 1)$ for some unknown θ we want to estiamte. Suppose further we take one sample and will use squared error loss.

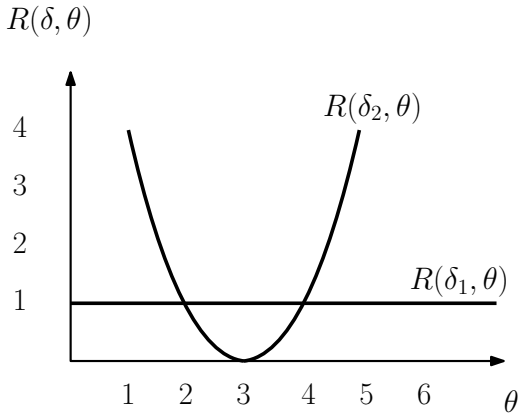
Consider the following two candidate decision functions:

$$\delta_1(X) = X$$
$$\delta_2(X) = 3$$

Which is better? We can show

$R(\delta_1, \theta) = \text{variance}_\theta(\delta_1) = 1$, i.e. δ_1 is unbiased
$R(\delta_2, \theta) = (\text{bias}_\theta(\delta_2))^2 = (3 - \theta)^2$, i.e. δ_2 has zero variance

The risk functions look something like:



Notes:
- Neither dominates the other, even though δ_2 is pretty silly (it can get lucky!)
- Minimax prefers δ_1 (why?)
- The risk function for δ_1 is constant. This is a common feature of minimax decision functions (but isn't always true)

Problems

A) Suppose you plan to flip a potentially biased coin (unknown probability of heads p, $0 \leq p \leq 1$) 100 times use this to estimate the probability. Before you collect any data, however, what is the 'no-data' minimax estimate of p under quadratic loss? How does this differ from the same estimate under 0/1 loss?

B) Show that under the squared error loss $l(\delta(X), \theta) = (\delta(X) - \theta)^2$ the Wald risk function $R(\delta, \theta) = \mathbb{E}_{X;\theta}[l(\delta(X), \theta)]$ can be written

$$R(\delta, \theta) = \mathrm{Var}_\theta(\delta) + (\mathrm{bias}_\theta(\delta))^2$$

where $\mathrm{bias}_\theta(\delta) = \bar{\delta} - \theta$ is a function of both the expected value of the estimator (here represented with an overbar, i.e. $\bar{\delta} = \mathbb{E}_{X;\theta}[\delta(X)]$ is the 'population' expectation) and the unknown parameter θ, while $\mathrm{Var}_\theta(\delta) = \mathbb{E}_{X;\theta}[(\delta(X) - \bar{\delta})^2]$ is an expectation of a function that only explicitly depends on the estimator, though the result depends on the parameter via the distribution associated with θ and hence we still write Var_θ.

C) Using the previous result, explain the idea of a 'bias-variance trade-off' in general terms.

Statistical decision theory: Bayes

Overview

Recall the typical 'Wald risk' form we obtained previously:

	θ_1	θ_2	θ_3
δ_1	$R(\delta_1, \theta_1)$	$R(\delta_1, \theta_2)$	$R(\delta_1, \theta_3)$
δ_2	$R(\delta_2, \theta_1)$	$R(\delta_2, \theta_2)$	$R(\delta_2, \theta_3)$
δ_3	$R(\delta_3, \theta_1)$	$R(\delta_3, \theta_2)$	$R(\delta_3, \theta_3)$
	⋮	⋮	⋮

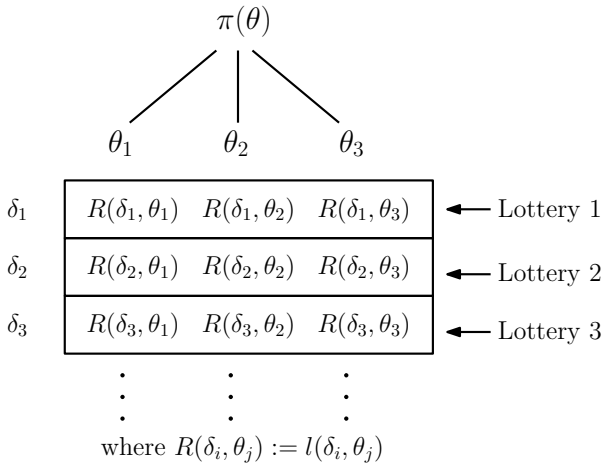
where $R(\delta_i, \theta_j) := l(\delta_i, \theta_j)$

We saw that **minimax** is one way to (try to) reduce a risk function $R(\delta, \theta)$ to a single number. Another way is to use a **Bayesian** approach. Here the idea is to assume a **distribution over the parameters** (states of nature), i.e. essentially a **'distribution over distributions'**. This is called a **prior**. We can then consider the expected loss!

Note: a key aspect of the Bayesian approach is that **even fixed unknown constants (like the true θ) are modelled with probability distributions**. In particular this now represents your **'belief'** (or 'state of information') about what the true parameter is (if it was 'truly random' it would just be normal risk). The **interpretation of probability has shifted subtly** here.

Bayes loss/Bayes risk

Given a prior probability distribution (here, density) over θ, often denoted by $\pi(\theta)$ rather than $p(\theta)$ (for who-knows-what reasons!) we can compute the **expected risk under that prior distribution over** θ for any decision function:



which gives the **Bayes risk of** δ **under** $\pi(\theta)$ and is typically denoted by a lowercase r:

$$r(\delta,\pi) = \mathbb{E}_{\pi(\theta)}[R(\delta,\theta)]$$

where $\mathbb{E}_{\pi(\theta)}$ denotes the expectation under the prior $\pi(\theta)$ over θ. This is also sometimes written $\mathbb{E}_{\theta;\pi}$ or $\mathbb{E}_{\theta|\pi}$, or even just $\mathbb{E}_\theta$, where the prior is implicit. So $r(\delta,\pi) = \mathbb{E}_{\theta;\pi}[R(\delta,\theta)] = \mathbb{E}_\theta[R(\delta,\theta)]$.

Note, the Bayes risk expands as:

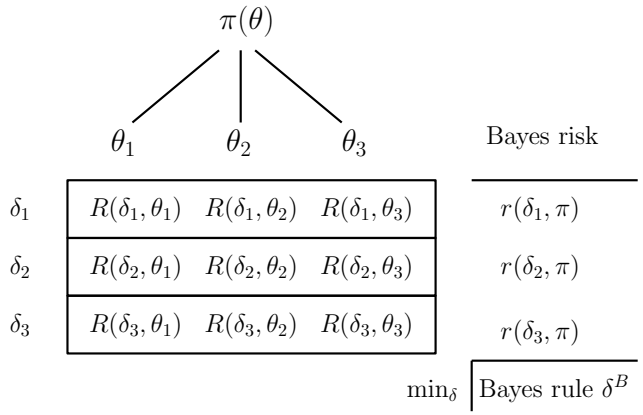
$$r(\delta,\pi) = \mathbb{E}_{\pi(\theta)}[\mathbb{E}_{X|\theta}[l(\delta(X),\theta)]] = \mathbb{E}_\theta[\mathbb{E}_{X|\theta}[l(\delta(X),\theta)]]$$

where the '|" in $\mathbb{E}_{X|\theta}$ indicates this is a proper conditional distribution for X as θ is now random, and the second expression leaves the prior implicit.

How do we use the Bayes risk?

Given the Bayes risk function for a chosen prior and set of decision rules, we can find the **decision rule** that **minimises the Bayes risk** for this prior.

This is called the **Bayes rule** δ^B with respect to the prior:



i.e. the Bayes rule δ^B minimises

$$r(\delta,\pi) = \mathbb{E}_{\pi(\theta)}[R(\delta,\theta)] = \mathbb{E}_{\pi(\theta)}[\mathbb{E}_{X|\theta}[l(\delta(X),\theta)]]$$

with respect to decision function δ for fixed prior π. As before, we are trying to solve a decision problem by optimising over **functions** rather than e.g. constants. This is hard in general! Luckily, we can use the …

Posterior expected loss

So far we have considered the **expectation with respect to the prior** (and the distribution of the data).

We then **defined the Bayes rule** as the **best** expected loss rule under this **prior and data distribution**.

It turns out, however (see attached proof), that Bayes rules as defined so far are equivalent to those that minimise the so-called **posterior** expected loss where the data is 'conditioned on' and hence fixed:

$$r(\delta, \pi(\theta|x)) = \mathbb{E}_{\pi(\theta|x)}[l(\delta, \theta)]$$

where $\mathbb{E}_{\pi(\theta|x)}[l(\delta, \theta)]$ indicates the expectation over θ is taken using the **conditional** distribution of the parameter given the observed data, $\pi(\theta|x)$. This distribution is called the **posterior**, and comes from **Bayes' theorem** for the conditional probability (density) of θ given x:

$$\pi(\theta|x) = \frac{\pi(x|\theta)\pi(\theta)}{\pi(x)}.$$

Again, we sometimes equivalently write $\mathbb{E}_{\pi(\theta|x)}$ as $\mathbb{E}_{\theta|x,\pi}$ or even just $\mathbb{E}_{\theta|x}$ where the prior etc distributional assumptions are implicit.

Note in the above that we are now finding a single value of δ for any fixed data realisation x. Thus...

Bayes vs Bayes' theorem

To use Bayes (decision) rule, find the posterior via Bayes' theorem!

This is actually a much simpler problem in general:

- For an given realisation x of X we find the **posterior** $\pi(\theta|x)$ using **Bayes' theorem** (see later, also 721 for how to find).
- We then compute a **best posterior summary** by taking the minimum expected value decision of $l(\delta, \theta)$ for given $\pi(\theta|x)$. Here x is fixed and we are just e.g. summarising a distribution by a constant.
- This defines our **decision rule** $\delta^B(x)$ for **any observation** x

i.e.

$$\delta^B(x) = \min_{\delta} \mathbb{E}_{\pi(\theta|x)}[l(\delta, \theta)]$$

Summary of Bayes approach

- Given prior $\pi(\theta)$
- Given observed data x
- Find posterior $\pi(\theta|x)$
- Compute a 'best summary' δ^* of the posterior $\pi(\theta|x)$ under loss $l(\delta, \theta)$ and $\theta \sim \pi(\theta|x)$ for x fixed
- This defines the Bayes rule $\delta^B(x) = \delta^*$ for that x and only need to consider the data x that actually occurs!

Example: squared-error loss

Consider a loss function of the form

$$l(\delta(X), \theta) = (\delta(X) - \theta)^2$$

Now the approach of minimising the expectation over the posterior means we solve

$$\min_{\delta} \mathbb{E}_{\pi(\theta|x)}(\delta(x) - \theta)^2$$

Note that we 'condition on' x and so it's fixed – we are simply solving for the best constant summary of a distribution (the posterior), which we know how to do! I.e. we are solving

$$\min_{\delta} \mathbb{E}_{\pi(\theta|x)}(\delta - \theta)^2$$

This means that, under squared error, the best decision for a given x is the **posterior mean**:

$$\delta^B(x) = \mathbb{E}_{\pi(\theta|x)}[\theta]$$

Notes:
- If we had e.g. $l(\delta, \theta) = |\delta - \theta|$ the bet solution would be the posterior median.
- If we had samples from the posterior, rather than the posterior itself, we could use the empirical mean over these samples as our estimate.
- Computing a posterior is not easy in general. We will talk a little about how to compute posteriors in the second part of this module. **Much more in EngSci 721!**

Bayes and minimax

A limitation of the Bayesian approach cf minimax is that the prior is fairly arbitrary...and a different (personal) interpretation of probability is required.

While minimax allows for a non-personal, 'frequentist' interpretation of probability, it is often hard to solve (and can be pessimistic...)

A compromise is to use one of the various theorems **connecting minimax and Bayes rules that use a so-called least favourable prior**.

This means we can use Bayes as a computational tool to find minimax solutions. This was Wald's point of view. Bayesians aim to find Bayesian solutions for their own sake!

It's beyond the scope here to go into detail but if you imagine that instead of representing your belief, a Bayes prior instead represents a 'mixed' or 'randomised' nature of Nature, then Nature's best choice of randomising distribution is your worst or 'least favourable' prior. Some key results include

- If δ^B is a Bayes rule with respect to some prior π and the Wald risk for that rule, i.e. $R(\delta^B, \theta)$, is less than or equal to the Bayes risk $r(\delta^B, \pi)$, then π is a least favourable prior and δ^B is minimax.
- An **'equaliser' decision rule**, i.e. one that means the Wald risk is constant, $R(\delta, \theta) = C$ for some C independent of θ, that is **also a Bayes rule for some prior**, or **a *limit* of Bayes rules for a sequence of priors** (extended Bayes), is **minimax** and the prior is **least favourable**.

Further reading: Wasserman (2004) 'All of statistics', Chapter 12.

Problems

A) Suppose we are trying to estimate the mean of a normal distribution 'data model', $N(\mu_d, \sigma_d^2)$, where μ_d is unknown. Suppose as a good Bayesian you make up a convenient prior of the form $N(\mu_p, \sigma_p^2)$ where you (of course) get to choose the prior mean and variance. Given a single observation of the data, i.e. arising as a realisation x_0 of the random variable $X \sim N(\mu_d, \sigma_d^2)$, it can be shown (see later in course or EngSci 721!) the posterior is a normal distribution with mean $\frac{\sigma_p^2}{\sigma_d^2+\sigma_p^2}x_0 + \frac{\sigma_d^2}{\sigma_d^2+\sigma_p^2}\mu_p$ and variance $\frac{\sigma_d^2\sigma_p^2}{\sigma_d^2+\sigma_p^2}$.

i) What is the Bayes estimator under a squared error loss function?

ii) What is the Bayes estimator under an absolute error loss function, i.e. $l(\delta(X), \theta) = |\delta(X) - \theta|$?

iii) What is the Bayes estimator as the prior variance goes to infinity? How does this relate to the estimator considered in Lecture 4? What would you guess this 'limiting prior' is (technical note: it is 'improper', i.e. converges to something that no longer integrates to 1, but this is 'allowed' as long as the posterior is proper and/or if the corresponding estimator is well-defined).

Appendix: conditional expectations

The **conditional expectation** of X given Y is written variously as $\mathbb{E}(X|Y)$, $\mathbb{E}_{X|Y}(X)$, $\mathbb{E}_{\mathbb{P}(X|Y)}(X)$ etc. These latter two make it somewhat more obvious that the idea is to take the expection of X under a new distribution, the conditional distribution $\mathbb{P}(X|Y)$:

$$\mathbb{E}_{X|Y=y} = \begin{cases} \sum_i x_i p(x_i|y), \text{ if discrete} \\ \int x p(x|y) dx, \text{ if continuous} \end{cases}$$

Here $\mathbb{E}_{X|Y=y}(X)$ gives a single number for the given y value. The expression $\mathbb{E}_{X|Y}(X)$ denotes a random variable in Y with value $\mathbb{E}_{X|Y=y}(X)$ when $Y = y$.

The same ideas apply for a general function $f(X)$: $\mathbb{E}_{f(X)|Y=y}(X)$ is given by the expectation ('average') of $f(X)$ as taken under the conditional distribution $\mathbb{P}(X|Y=y)$ with density $p(X|Y=y)$.

A key property of expectations and conditional expectations is the so called **rule of iterated expectations**:

$$\mathbb{E}_{(X,Y)}[f(X,Y)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[f(X,Y)]] = \mathbb{E}_X[\mathbb{E}_{Y|X}[f(X,Y)]]$$

which follows from the definition of joint and conditional probabilities and associated factorisations.

This law is usually given in the special case when $f(X,Y) = Y$, which means $\mathbb{E}_{(X,Y)}[Y] = \mathbb{E}_Y[Y]$ and so

$$\mathbb{E}_{(X,Y)}[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}[Y]] = \mathbb{E}_Y[Y]$$

The last equality is often written (obscurely imo!) as

$$\mathbb{E}[\mathbb{E}[Y]] = \mathbb{E}[Y],$$

with the subscripts suppressed. This is usually what is given the name **the rule of iterated expectations**.

Appendix: Proof sketch of 'Bayes rule via min expected posterior loss'

Consider the Bayes risk

$$r(\delta, \pi) = \mathbb{E}_\theta \mathbb{E}_{X|\Theta=\theta}[l(\delta(X), \Theta)] = \mathbb{E}_X \mathbb{E}_{\Theta|X=x}[l(\delta(X), \Theta)]$$

where the last equality follows from the law of iterated expectations (see previous Appendix) and we use Θ for the random variable associated with θ to explicitly distinguish between it and values of θ.

Now consider minimising this with respect to δ:

$$\min_\delta r(\delta, \pi) = \min_\delta \mathbb{E}_X \mathbb{E}_{\Theta|X=x}[l(\delta(X), \Theta)]$$

It can be shown that this expression is minimised when we minimise the inner expression for each fixed x value (if we didn't for some x we could do better by improving it for that x and keeping it the same for other values). That is, we solve

$$\min_\delta \mathbb{E}_{\Theta|X=x}[l(\delta(X), \Theta)] = \min_\delta \mathbb{E}_{\Theta|X=x}[l(\delta(x), \Theta)].$$

But this is just minimising the posterior expectation of the loss function, as desired!

So, the idea is we **make the best Bayes decision (summary of distribution) for each posterior induced by each realisation x of X.** Furthermore, we only need to solve the case of the actual x we obtain.

Part II

Modelling under uncertainty: risk and intervention

Probability models for risk: graphical models for modular representation

Overview

In the next part of this module, we will look in more detail at efficient and modular representation of **probability models**, i.e. **models for risk** (in the sense that we assume the uncertainty is of a known form).

Notes:

– These models may represent **'empirically-connected'** (frequentist or propensity-based) probabilities for 'naturally random' processes like repeatedly flipping a coin or repeating an experiment, or they may represent an agent's (Bayesian) **'degree of belief'**.

– It doesn't matter too much here what they represent, as **the mathematics is all the same**. However, a warning that the terminology can make this unclear. E.g., using Bayes' theorem (a theorem of mathematical probability theory that also applies to frequentist or propensity theories) is not the same as 'being a Bayesian' (interpreting probability as degree of belief). Similarly, we will look at what are sometimes called 'Bayesian networks', but these often represent 'frequentist' probability models. Hence we use more neutral but still common terms for these models such as 'directed acyclic graphical (DAG) models'.

– For a good discussion, see **Wasserman (2004) "All of statistics"** and his blog post.

Probability preamble: density and mass functions

For **discrete** random variables, each **outcome** has a **finite probability** and so e.g. $\mathbb{P}(X = 1)$ makes sense. The function giving the probabilities of each basic outcome, $\mathbb{P}(X = x)$, is called the **probability mass function (pmf)** and we will write

$$\mathbb{P}(X = x) = p(x)$$

for this, i.e. introducing the notation $p(x)$.

For **continuous** random variables, each outcome typically only has **'infinitesimal' probability**. To get finite probabilities we have to consider e.g. finite intervals (or sets) like $\mathbb{P}(1 \leq X \leq 2)$ etc. These can be obtained from an integration of the form

$$\mathbb{P}(x_0 \leq X \leq x_0 + \Delta x) = \int_{x_0}^{x_0+\Delta x} p(x)dx$$

where $p(x)$ here is the **probability density function (pdf)**, assuming it exists (there are technical conditions here that we won't discuss!).

If we think about the limit as Δx gets 'small', becoming an 'infinitesimal' dx, we can semi-formally define the 'probability mass of a continuous outcome' as

$$\mathbb{P}(X = x) = p(x)dx.$$

Importantly, in our context, most key relationships can be expresses in terms of '$p(x)$' for both continuous and discrete variables, i.e. **regardless** of whether it is the **pmf** of a discrete variable or the **pdf** of a continuous variable.

We will hence call both versions of $p(x)$ **probability functions** and let context indicate whether $p(x)$ is a pdf or pmf. Intuitively, the only difference is multiplying by some constant 'dx' etc factors that don't change anything important.

Probability preamble: joint and conditional probability functions

The probability function, $p(x, y, z)$, of the **joint** distribution of three random variables X, Y, Z, is defined via

$$\mathbb{P}(X = x \text{ and } Y = y \text{ and } Z = z) = \begin{cases} p(x, y, z), \text{ discrete case} \\ p(x, y, z)dx\, dy\, dz, \text{ continuous case.} \end{cases}$$

The joint distribution can be formed in any order, e.g.

$$\mathbb{P}(X = x \text{ and } Y = y \text{ and } Z = z) = \mathbb{P}(Y = y \text{ and } Z = z \text{ and } X = x),$$

etc. The above is written in terms of p as $p(x, y, z) = p(y, z, x)$ etc., where it is understood that the lower case variables refer to values of their upper case equivalents. If we needed to be more explicit, we would write $p_{XYZ}(x, y, z) = p_{YZX}(y, z, x)$, but we won't here.

The **marginal probability** function for X, $p(x)$, is given by

$$p(x) = \begin{cases} \sum_y \sum_z p(x, y, z), \text{ discrete case} \\ \int \int p(x, y, z)\, dz\, dy, \text{ continuous case,} \end{cases}$$

and similarly for Y or Z. That is, we 'sum/integrate out' the other variables. We also have e.g.

$$p(x, y) = \begin{cases} \sum_z p(x, y, z), \text{ discrete case} \\ \int p(x, y, z)dz \text{ continuous case,} \end{cases}$$

etc, giving the marginal probability function for X and Y.

To **condition on** a variable, we divide the joint probability function by the marginal we are conditioning on, giving the **conditional probability** function:

$$p(x, y \mid z) = \frac{p(x, y, z)}{p(z)}$$

for $p(z) > 0$ and where this formula the same for both discrete and continuous variables (in terms of the pmf or pdf, respectively).

These formulae can easily be extended to an arbitrary number of variables.

Chain rule of probability theory

For two random variables X, Y, the definition of conditional probability can be rearranged to the 'factorised' form

$$p(y, x) = p(y|x)p(x) = p(x|y)p(y),$$

where the second equality follows from the fact the joint probability function satisfies $p(x, y) = p(y, x)$ in the sense mentioned previously.

The **chain rule of probability theory** extends this to a **collection of N random variables**: given **any ordering** of these variables, written $X_1, X_2, \ldots, X_N$, say, we can always write the joint distribution as

$$p(x_n \mid x_{n-1}, \ldots, x_1)p(x_{n-1} \mid x_{n-2}, \ldots, x_1) \ldots p(x_2 \mid x_1)p(x_1).$$

E.g.

$$\begin{aligned} p(x, y, z) &= p(x|y, z)p(y|z)p(z) \\ &= p(y|x, z)p(x|z)p(z) \\ &= p(z|x, y)p(x|y)p(y) \\ &\vdots \end{aligned}$$

and so on.

Representation troubles

To help understand what sort of trouble happens for probability models for complex problems with lots of variables, consider the simple case of N *binary* random variables

$$X_i,\ i = 1, ..., N$$

where each X_i takes values in $\{0, 1\}$.

The joint probability function over these is given by

$$p(x_1, x_2, ..., x_N).$$

This requires specifying 2^N possible probabilities, one per input combination. This grows fast with N.

Thus, 'representing', or just 'writing down', a probability model is hard with there are many variables.

Independence for modularity?

Using the chain rule of probability theory we can write, for our previous problem,

$$p(x_1, x_2, ..., x_N) = p(x_1 \mid x_2, \dots, x_N)p(x_2 \mid x_3, \dots, x_N)p(x_{N-1}|x_N)p(x_N)$$

(remember we can use any order).

So far, **each of these terms is as difficult to specify as the joint!** We need to specify the values for each argument combination on both sides of the conditioning bar '|'. However, the ideas of **independence**, **mutual independence**, **conditional independence** etc of random variables can help!

Independence and mutual independence

Random variables X and Y are called **independent** if

$$p(x, y) = p(x)p(y),$$

equivalently, if

$$p(y|x) = p(y),$$

equivalently, if

$$p(x|y) = p(x).$$

We write this as

$$Y \perp X.$$

These definitions can also be used for **collections of random variables**: we write

$$X_1 \perp \{X_2, X_3, \dots, X_N\}$$

to express that X_1 **is independent of the other random variables**. In terms of the probability function this represents that:

$$p(x_1 \mid x_2, \dots, x_N) = p(x_1).$$

The **mutual independence** of a set of random variables $\{X_1, X_2, \dots, X_N\}$ is expressed as

$$X_i \perp \{X_j\}_{j \neq i},\ \text{for all } X_i,$$

where $\{X_j\}_{j \neq i}$ is the **set of all random variables except** X_i.

Thus, mutual independence means that **each variable is independent of all the rest**.

Independence for modularity!

Returning to our example, we had that:

$$p(x_1, x_2, ..., x_N) = p(x_1 \,|x_2, \dots, x_N)p(x_2 \,|x_3, \dots, x_N)p(x_{N-1}|x_N)p(x_N).$$

Now, in the extreme case that this set of binary random variables are **mutually independent**, we have, for all i, that

$$p(x_i|x_{i+1}, \dots, x_N) = p(x_i),$$

and so

$$\begin{aligned} p(x_1, x_2, ..., x_N) &= p(x_1)p(x_2) \dots p(x_N) \\ &= \prod_{i=1}^{N} p(x_i). \end{aligned}$$

In contrast to the general case, **mutual independence now means we only need to specify the values of each of the single-variable marginal probability functions** $p(x_i)$. For binary variables we only need to specify $2N$ values, instead of 2^N.

While mutual independence is rare, in practice various forms of **conditional independence** are not! This similarly leads to **computationally efficient** ways of specifying and working with probability models, while also giving **conceptual/modelling** advantages over specifying a full joint model directly – e.g., as we will discuss, conditional independence is important for **causal modelling**.

We will hence look at defining **conditional independence** as well as a convenient **graphical** (in the sense of graph theory!) **language for specifying and determining various types of conditional independences**.

In many cases, this allows us to work 'qualitatively' to a large extent, in terms of various conditional independences rather than detailed/concrete probability models. Furthermore, these languages come with **algorithms** for determining which variables are independent of which in complex models.

Conditional independence

Random variables X and Y are said to be **conditionally independent given** Z if

$$p(x|y, z) = p(x|z),$$

equivalently

$$p(y|x, z) = p(y|z),$$

equivalently

$$p(x, y|z) = p(x|z)p(y|z).$$

We write that X **and** Y **are conditionally independent given** Z symbolically as

$$X \perp Y \mid Z$$

which is equivalent to

$$Y \perp X \mid Z$$

We can similarly consider relationships like

$$X \perp Y \mid \{Z_1, \dots, Z_N\}$$

which says that X and Y are **conditionally independent given the set of random variables** $\{Z_1, \dots, Z_N\}$. This corresponds to e.g.

$$p(x|y, z_1, \dots, z_n) = p(x|z_1, \dots, z_n)$$

and

$$p(y|x, z_1, \dots, z_n) = p(y|z_1, \dots, z_n)$$

and

$$p(x, y|z_1, \dots, z_n) = p(x|z_1, \dots, z_n)p(y|z_1, \dots, z_n)$$

etc.

Representing conditional independence with graphical models

So far we have two 'languages' to express conditional independence: the probability function ('p') language and the symbolic ('$\perp$') language. Now we introduce another: **graphical models!**

These use different types of **graphs** – in the sense of graph theory, i.e. vertices/nodes and edges – to represent conditional independencies. Here:

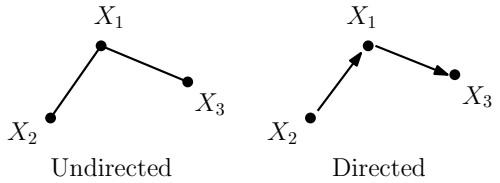
– **Nodes/vertices** represent variables

– **Edges** (or their absence!) represent (in)dependencies between variables.

There are **two main types of graphical model**:

- **Undirected graphical models**,
- **Directed graphical models**.

For the former the edges have no direction and for the latter the edgges are directed, i.e. arrows:



We write our graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the **set of nodes/vertices** in the graph and $\mathcal{E}$ the **set of edges**. A **path** between two vertices is a sequence of vertices starting at one and ending at the other such that there is an edge between each consecutive pair of vertices in the sequence. A **directed path** means all the arrows on the path point in the same direction.

We will mainly focus on directed graphs, but first briefly mention undirected graphs.

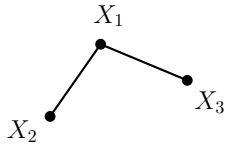
Undirected graphical models: independence

Undirected graphs represent purely probabilistic (c.f. causal – see later) independencies using the rule:

> Given a set of variables $X_1, \ldots, X_N$, we do **not** draw an edge between a **pair** X_i and X_j if $X_i \perp X_j \mid \{\text{the rest}\}$.

Intuitively, an edge represents a 'direct' (though not 'directed'!) dependence between two variables.

For example the graph:



represents (among other relationships) the symbolic independence relation:

$$X_2 \perp X_3 \mid X_1$$

and the probability function relationship:

$$p(x_2|x_3, x_1) = p(x_2|x_1)$$

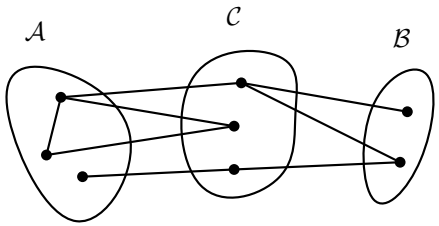
etc.

Undirected graphical models: separation

In a graphical model constructed as above, there is a **graphical property** called **u-separation** which relates sets of vertices/nodes in a graph as follows:

Given three **sets of vertices**, $\mathcal{A}, \mathcal{B}, \mathcal{C} \subseteq \mathcal{V}$, we say $\mathcal{C}$ **separates** $\mathcal{A}$ and $\mathcal{B}$ if **all paths** from any element of $\mathcal{A}$ to some element of $\mathcal{B}$ **pass through** $\mathcal{C}$.

For example:



We see that we can't get from $\mathcal{A}$ to $\mathcal{B}$ without passing through $\mathcal{C}$.

The **graphical** property of separation translates into the **probabilistic** property of **conditional indepedence** of sets of variables:

if:

$$\{Z_1, \dots, Z_p\} \text{ separates } \{X_1, \dots, X_n\} \text{ and } \{Y_1, \dots, Y_m\}$$

then:

$$\{X_1, \dots, X_n\} \perp \{Y_1, \dots, Y_m\} \mid \{Z_1, \dots, Z_p\}$$

This is called the **global Markov property** in graphical modelling.

What about directed graphical models?

Directed graphical models

In a directed graphical model **edges now have directions**. We will restrict attention to **acyclic** directed graphs or **DAGs** – directed acyclic graphs.

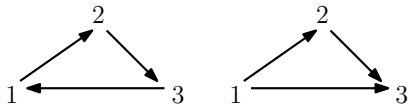
DAGs are designed to incorporate information about probabilistic (in)dependencies **as well as** 'causal' or 'directional' information. Recall: correlation $\neq$ causation!

In particular, a graph being a DAG means there are no **directed cycles** in the graph, where:

– A **cycle** is a path from a vertix/node back to itself

– A **directed cycle** is a cycle of directed edges that all point the same way.

Example:



Directed cycle No directed cycle

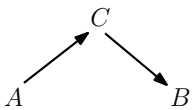
First we will look at how the encode **probabilistic** conditional independence relationships in a similar but different way to undirected graphs, later we will look at their **causal** interpreation.

Note: Directed graphical models are also sometimes called 'Bayesian networks', though as mentioned this isn't really a good name (nothing especially 'Bayesian' about them).

Directed graphical models: Predecessors, ancestors, parents, children

The idea of a DAG is to use the **ordering** indicated by the **arrow edges** to encode a type of conditional independence that follows from an **'ordered Markov factorisation'** of the probability model. To see how this works we need to define some graph theory terms for DAGS.

Firstly, a **predecessor** or **ancestor** of a vertex v is any other vertex occurring **'before'** v (earlier in the sequence of nodes) on a **directed path** connecting the two vertices. E.g. in



The node A has no predecessors/ancestors, C has A as an ancestor, and B has both A and C as ancestors.

A **parent** is an **immediate ancestor** (i.e. an ancestor on a single edge path).

Exercise: define **descendants** and **children** in the obvious way!

Directed graphical models: conditional independence

Given the above definitions, a DAG *directly* encodes **conditional independencies** of the form

$$X \perp \text{Pred}(X) \backslash \text{Pa}(X) \mid \text{Pa}(X)$$

where

$$\text{Pred}(X) = \text{Predecessors of X}$$
$$\text{Pa}(X) = \text{Parents of } X$$
$$\text{Pred}(X) \backslash \text{Pa}(X) = \text{"Non-parent ancestors" of X.}$$

That is, in a DAG

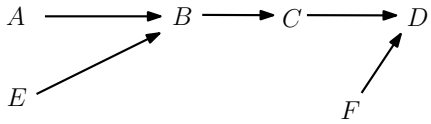
> X is independent of its non-parent predecessors, given its parents (only the parents matter...)

This translates in terms of the probability function to

$$p(x \mid \text{pred}(x)) = p(x \mid \text{pa}(x))$$

where $\text{pred}(x)$ and $\text{pa}(x)$ stand for the **values** of the predecessors and parents of X, respectively.

For example



encodes (among others!) independencies such as

$$C \perp \{A, E\} \mid B$$

i.e.

$$p(c|a, b, e) = p(c|b).$$

Product decomposition

Combined with the chain rule of probability, a DAG leads to a factorisation of the joint probability function of a set of N random variables, $X_1, ..., X_N$, of the form:

$$p(x_1, \dots, x_N) = \prod_{i=1}^{N} p(x_i \mid \text{pa}(x_i)).$$

Notes

– This does not require $X_1, \dots, X_N$ to be in any particular order. The DAG gives the ordering. Hence the use of '$\text{pa}(x_i)$".

– In general this allows us to specify the **joint** distribution as a **product of simpler distributions**, similar to (but more interesting than!) the case of mutual independence.

– In addition to computational convenience, it is often easier to **model** various phenomenon in terms of a combindation of **conditional distributions** rather than one big joint. This is related to intuitive ideas of 'what causes what'.

Other (in)dependencies in DAGs

A given DAG tells us 'directly' about the probabilistic dependences of a node on its ancestors ('past' nodes). It doesn't 'directly' tells us about the **probabilistic** dependencies of a node on its descendants...**even though these are there!**

E.g. if our DAG is $X \to Y$ then our product factorisation implies

$$\begin{aligned} p(x, y) &= p(y \mid \text{pa}(y))p(x \mid \text{pa}(x)) \\ &= p(y|x)p(x), \end{aligned}$$

but in general, from probability theory (which still applies here!), we also have

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \neq p(x),$$

where the last relation follows from $p(y|x) \neq p(y)$ as the DAG implies that Y is not independent of X (which we can write $Y \not\perp X$)

The above shows that in the DAG $X \to Y$ we also have that X is not independent of Y. The probabilistic dependence between X and Y can 'flow backwards'

Causal vs probabilistic dependencies

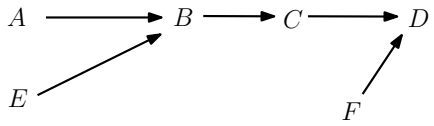
As we have seen, when we encode (in)dependencies that follow our natural 'causal' intuition, **we also indirectly encode other 'non-causal', i.e. probabilistic, (in)dependencies**. A key lesson is that **probabilistic (in)dependence is a different 'thing' to causal (in)dependence**.

It turns out **we can** add more components to the interpretation of DAGs **so that we can use DAGs to reason about cause and effect**. We will look at this next lecture, as well as how to understand all of the (non-causal) **probabilistic implications** of a DAG using something called **d-separation**.

Problems

A) Give reasonable definitions for descendants and children in a DAG and illustrate them for a DAG you choose.

B) Give some other independencies implied by the DAG



C) For a preview of d-separation, look up this blog post.

D) For more on probability, DAGs (and decision theory!) look up the book by Wasserman (2004) "All of Statistcs". You can get an electronic pdf copy via the uni library.

More DAGs: d-separation and intervention

Overview

Here we consider DAGs in more detail. We consider **two key questions** and how they interact

– How do we determine **all of the probabilistic (in)dependencies of a DAG?**

– How can we interpret and use DAGs to represent **causality** and **causal relationships**, not just probabilistic relationships?

To do this we introduce and study:

– **d-separation** for DAGs,

– **Interventions** in DAGs and associated causal effects.

We will briefly mention some details on how to **compute** answers to 'target queries' for probabilistic and causal models, and **'causal decision theory'** as an alternative to what we have seen previously, which is sometimes called 'evidential decision theory'.

Aside: DAG models were introduced in the 1920s/1930s (e.g. Wright), and were later developed by Pearl, Glymour, Scheines, Spirtes et al. from the 1980s on.

Conditional independencies implied by a DAG

To determine the other independencies implied by the **directed Markov factorisation** that the DAG implies and the **standard rules of probability**, it turns out we can use the **graphical** concept of **d-separation**.

This stands for **directed separation**, and was developed by Pearl and others in the 1980s. We start by giving a definition ... that uses terms we haven't yet defined!

d-separation and conditional independence

Given the definition of path blocking to follow next, **d-separation** (again, 'directed separation') is defined as:

> If **every** path between two nodes X and Y is **blocked given** $\mathcal{B}$, then X and Y are said to be **d-separated given** $\mathcal{B}$.

If two nodes are **not** d-separated, then we say they are **d-connected**.

We then have the **key result**, that follows from the standard 'directed Markov' interpretation of DAGs we saw in the previous lecture, along with standard probability theory, that

$$X \perp Y \mid \mathcal{B}$$

if and only if

$$X \ \& \ Y \text{ are } \textbf{d-separated given } \mathcal{B}.$$

Furthermore, **the d-separation criterion allows us to derive all conditional independencies implied by a DAG**.

Now we just need to know what it means for a path to be blocked!

Graphical concepts: paths, blocked paths

We first define **any** sequence of edges between two nodes, **regardless** of the edge (arrow) directions, a **dependence path** or just **path** for short.

E.g., in $A \rightarrow B \leftarrow C$, the sequence $(A \rightarrow B, B \leftarrow C)$ is a valid path.

We then define what it means for a dependence path to be **blocked by a set of nodes** $\mathcal{B}$:

> A (dependence) **path** between any two nodes X, Y in a DAG is **blocked given a set of nodes** $\mathcal{B}$ (which are not necessarily on the path) iff the **path** contains at least one **sub-path**, or **'junction'**, of the form:
>
> 1. $A \rightarrow B \rightarrow C$, where $B \in \mathcal{B}$ (a **chain**)
> 2. $A \leftarrow B \rightarrow C$ where $B \in \mathcal{B}$ (a **fork**)
> 3. $A \rightarrow D \leftarrow C$ where $D \notin \mathcal{B}$ & $\text{des}(D) \notin \mathcal{B}$ (a **collider**).

where $\text{des}(D)$ represents the descendants of D.

The last condition is perhaps slightly unintuitive but represents the idea that **a collider 'naturally blocks' information flow without conditioning on it**. In contrast, **conditioning on a collider or a descendant of a collider 'opens' ('unblocks') the information flow**.

Examples

Consider

1. $X \to U \to V \to W \to Y$.

We have that the sets $\{U\}$, $\{U, V\}$, $\{U, V, W\}$, $\{V, W\}$, $\{V\}$ etc are all sets that block $X \to ... Y$. So we can write $X \perp Y \mid \{U\}$ etc.

2. $X \leftarrow U \to Y$.

We have that $\{U\}$ d-separates X and Y, i.e. $X \perp Y \mid U$.

3. $X \to U \leftarrow Y$.

Here U does **not** d-separate X and Y: we have that $X \perp Y \mid \{\}$ i.e. $X \perp Y$ already, but $X \not\perp Y \mid U$. This is because a collider already blocks the flow of dependence (so we condition on 'nothing' = 'condition on nothing'), while conditioning on a collider 'opens' the flow of information. This is why you might hear phrases like 'don't condition on a collider'!

4. $Z \to W \leftarrow X \to Y$

What can we say about the (in)dependence of Z and Y?

Noting that the dependence path between Z and Y contains the sub-path $Z \to W \leftarrow X$, i.e. a collider, we see that we don't need to condition on anything to d-separate X and Z. Hence

$$Z \perp Y.$$

However,

$$Z \not\perp Y \mid W$$

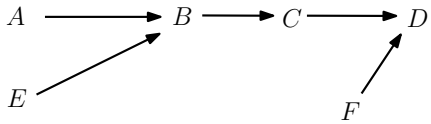
i.e. Z and Y are d-connected given W. Question: is $Z \perp Y \mid \{W, X\}$?

d-separation for sets of nodes

We have defined what it means for two **individual** nodes to be d-separated given a **set** of other nodes. We can extend this to d-separation fo sets of nodes by defining

> Given disjoint sets of nodes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$, we say that $\mathcal{A}$ and $\mathcal{B}$ are d-separated if **every** path between a node in $\mathcal{A}$ and a node in $\mathcal{B}$ is d-separated given $\mathcal{C}$.

E.g. in



we can say e.g. that $\{A, E\}$ and $\{C\}$ are d-separated given $\{B\}$. This translates to

$$C \perp \{A, E\} \mid B$$

and

$$p(c|a, b, e) = p(c|b).$$

These clearly also follow directly from the directed Markov interpretation of DAGs.

Problems

1. Consider $X \leftarrow Z \rightarrow Y$.

– Which of these is true: $X \perp Y$ or $X \perp Y \mid Z$

2. Suppose in (1) that we are sampling from a population and measuring the variables:

- Z is 'age'
- X is 'retirement savings'
- Y is 'number of wrinkles'

– What can you say about the relationship between retirement savings and wrinkles? Intuitively, how should you analyse this relationship to understand the 'causal effect' of wrinkles on retirement savings?

3. Consider $A \rightarrow B \rightarrow C$, $A \leftarrow B \rightarrow C$, and $A \leftarrow B \leftarrow C$.

– Derive the conditional independencies implied by the DAGs using d-separation.
– What do you notice?

4. Consider $A \rightarrow B \leftarrow C$.

– Is $A \perp C$? What about $A \perp C \mid B$?

5. In (4), suppose

- A is 'height'
- B is 'basketball success'
- C is 'basketball shooting skill'

– Interpret the associated conditional independencies.

Interventions and causality

Note that $X \rightarrow Y$ and $X \leftarrow Y$ encode the same conditional independencies – neither is independent of the other! Similarly

$$X \rightarrow Y \rightarrow Z$$
$$X \leftarrow Y \rightarrow Z,$$

encode the same independencies, but 'intuitively' we'd like (?) to think of them differently, e.g., the first involving the 'mediating of a cause' – X causes Y causes Z – and the second representing 'confounding' – Y causes both X and Z.

This led to people such as Pearl to distinguish 'seeing'-type relationships and 'doing'-type relationships.

Interventions: 'doing' and 'viewing' ('seeing')

To distinguish 'seeing' ('viewing') and 'doing', we first interpret

$$p(y|x)$$

as the probability that $Y = y$ given we (passively) **see** that $X = x$. We just **'view'** the probabilistic relationships implied by the DAG. This is **not necessarily a causal relationship** – remember the old saying 'correlation is not causation'!

In order to define (a type of) causal relationship, we want to try to define

$$p(y|\text{do}(x))$$

as the probability that $Y = y$ if we **intervene** on the system and **cause** $X = x$, i.e. the probability of the effect given the cause.

To precisely define what '$p(y|\text{do}(x))$' means, we first define what it means to 'intervene' on a DAG.

Intervention in a DAG

To define interventions on a variable X in a DAG $\mathcal{G}$, we define the following operation on the given DAG:

$\text{do}_X(\mathcal{G}) = \mathcal{G}_X$: remove all incoming arrows into X (i.e., remove all other causes of X!), giving a new DAG $\mathcal{G}_X$

For example,

$$\text{do}_X(X \to Y) = X \to Y$$

while

$$\text{do}_X(X \leftarrow Y) = X \quad Y$$

Thus the 'do' operation defined so far distinguishes $X \to Y$ and $X \leftarrow Y$!

Now we just need to combine this with 'viewing'...the idea is to get a causal relationship between say X as the cause and Y as the effect in a DAG (generally containing many other variables) we **'do then view'**:

1. Intervene on X to go from $\mathcal{G}$ to $\mathcal{G}_X$
2. Compute (in the standard way), the probability of Y given X under the new DAG $\mathcal{G}_X$

Because we have introduced a new DAG, we write $p_{\mathcal{G}}$ for the probability function under the original DAG and $p_{\mathcal{G}_X}$ for the probability function under the new DAG. We then define

$$p_{\mathcal{G}}(y|\text{do}(x)) = p_{\text{do}_{X(\mathcal{G})}}(y|x) = p_{\mathcal{G}_X}(y|x)$$

This definition operationalises the 'do then view' recipe. However, **we need to define the probabilities under the new DAG by relating them to those given by the old DAG!**

Stability of mechanisms and the invariance condition

We want to say that the intervention on our graph is 'minimalist' in the sense that most of the associated distributions are unchanged. Otherwise, an intervention could change things arbitrarily! Naturally for a DAG model, we specify the relationship between the terms of the directed Markov factorisation. If we label our random variables $X_1, X_2, \ldots, X_n$, then we express the **stability of mechanisms ('directed Markov factors') as**:

$$p_{\mathcal{G}_{X_j}}(x_i \mid \text{pa}_{\mathcal{G}_{X_j}}(x_i)) = \begin{cases} p_{\mathcal{G}}(x_i \mid \text{pa}_{\mathcal{G}}(x_i)), \text{ if } i \neq j \\ p_{\mathcal{G}}(x_i), \text{ if } i = j \end{cases}$$

That is, the key 'directed Markov factors' in the new graph are the same as in the original graph for any variables that aren't intervened on, while they become the simple marginals from the original graph when the parents are removed.

Another way to express this succinctly is to use the following **invariance condition** for the two probability functions, i.e. the probability functions of variables conditioned on their parents:

$$p_{\mathcal{G}_{X_j}}(x_i \mid \text{pa}_{\mathcal{G}}(x_i)) = p_{\mathcal{G}}(x_i \mid \text{pa}_{\mathcal{G}_{X_j}}(x_i))$$

for the values x_i of any random variable X_i in the DAGs (including possibly the variable we intervene on, X_j), and where $\text{pa}_{\mathcal{G}}(x_i)$ and $\text{pa}_{\mathcal{G}_{X_j}}(x_i)$ are the values of the parents of X_i in $\mathcal{G}$ and $\mathcal{G}_{X_j}$, respectively.

Notes:

- We write $p(z \mid \text{pa}(z)) = p(z)$ if Z has no parents.
- The above 'invariance condition', along with the directed Markov interpretation of any DAG and standard probability theory, enables us to derive the 'stability of mechanisms' rule above and vice-versa.

Causality defined??

Putting these together gives us a definition that gives us 'causal effect' probability functions (here in terms of X and Y):

Given a DAG $\mathcal{G}$ and associated probability function $p_{\mathcal{G}}(x, y, \dots)$ which factorises according to $\mathcal{G}$, we define the causal probability function of X_i given X_j as

$$p_{\mathcal{G}}(x_i \mid \mathrm{do}(x_j)) = p_{\mathcal{G}_{X_j}}(x_i|x_j)$$

where

$$p_{\mathcal{G}_{X_j}}(x_i \mid \mathrm{pa}_{\mathcal{G}_{X_j}}(x_i)) = \begin{cases} p_{\mathcal{G}}(x_i \mid \mathrm{pa}_{\mathcal{G}}(x_i)), \text{ if } i \neq j \\ p_{\mathcal{G}}(x_i), \text{ if } i = j \end{cases}$$

or, equivalently,

$$p_{\mathcal{G}_{X_j}}(x_i \mid \mathrm{pa}_{\mathcal{G}}(x_i)) = p_{\mathcal{G}}(x_i \mid \mathrm{pa}_{\mathcal{G}_{X_j}}(x_i))$$

Although the definition may seem a little complicated, in principle, this gives us a simple and automatic way to compute causal relationships! And we can say, e.g., if

$$p_{\mathcal{G}}(y \mid \mathrm{do}(x)) = p_{\mathcal{G}}(y)$$

using Y and X for simplicity, then X has **no causal effect on** Y, i.e. Y **is causally independent of** X. Note: Unlike probability relationships, this influence is **asymmetric**! For example, if X causes Y, then Y is causally dependent on X, but X is not causally dependent on Y.

Let's consider some examples!

Examples!

1. Consider $\mathcal{G} : X \to Y$.

Then $\mathcal{G}_X = X \to Y$, as X has no parents, and so

$$\begin{aligned} p_{\mathcal{G}}(y \mid \mathrm{do}(x)) &= p_{\mathcal{G}_X}(y \mid x) \\ &= p_{\mathcal{G}_X}(y \mid \mathrm{pa}_{\mathcal{G}_X}(y)) \\ &= p_{\mathcal{G}}(y \mid \mathrm{pa}_{\mathcal{G}}(y)) \\ &= p_{\mathcal{G}}(y \mid x), \end{aligned}$$

and so, **in this case, doing = seeing**, and our original conditional probability represents a causal relationship.

2. Now consider $\mathcal{G} : X \leftarrow Y$.

Then $\mathcal{G}_X = X \quad Y$, as Y is the parent of X and so is removed in the new graph, and so

$$\begin{aligned} p_{\mathcal{G}}(y \mid \mathrm{do}(x)) &= p_{\mathcal{G}_X}(y \mid x) \\ &= p_{\mathcal{G}_X}(y) \ \text{[by causal Markov condition in new graph]} \\ &= p_{\mathcal{G}_X}(y \mid \mathrm{pa}_{\mathcal{G}_X}(y)) \ [Y \text{ has no parents in new graph}] \\ &= p_{\mathcal{G}}(y \mid \mathrm{pa}_{\mathcal{G}}(y)) \ \text{[stability]} \\ &= p_{\mathcal{G}}(y) \ [Y \text{ has no parents in old graph}] \\ &\neq p_{\mathcal{G}}(y|x) \end{aligned}$$

Hence we see that now $p_{\mathcal{G}}(y \mid \mathrm{do}(x)) = p_{\mathcal{G}}(y)$ and so Y is causally independent of X i.e. X **has no causal effect on** Y.

The concept of 'intervening' or 'doing' has helped us define a difference in the **causal interpretation** of $X \to Y$ and $X \leftarrow Y$!

General recipe for automated answers to 'queries' of DAGs

We now have a general recipe for answering various questions given a DAG and associated probability function:

Given a joint distribution $p(x_1, \ldots, x_n)$ over random variables $X_1, \ldots, X_n$ that factorises according to a DAG $\mathcal{G}$, we can answer **'seeing'/'viewing'** (probabilistic dependence) questions using:

p(interest variables | observed variables)

and answer **'doing'** (causal dependence) questions using:

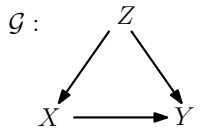
p(interest variables | do(observed variables)).

In the above:

- Any other variables are 'marginalised out', i.e., we use $p(x, y) = \int p(x, y, z)dz$ etc;
- To probabilistically condition, we use the standard definition, $p(y|x) = p(y, x)/p(x)$;
- $\text{do}_X(\mathcal{G})$ means we delete all arrows into X to obtain a new graph $\mathcal{G}_X$.
- We define $p(y \mid \text{do}(x))$ in terms of standard 'seeing' (probability) relationships in the new graph $\mathcal{G}_X$ obtained from $\text{do}_X(\mathcal{G})$.

Problems

Consider



Show:

$$p(y|x) = \int p(y|x, z)p(z|x)dz$$

$$p(y \mid \text{do}(x)) = \int p(y|x, z)p(z)dz \neq p(y|x)$$

where $p = p_{\mathcal{G}}$ is the probability function associated with $\mathcal{G}$.

Note: the expression for $p(y \mid \text{do}(x))$ above is called the 'adjustment formula'. It computes the probability function capturing the 'effect of X on Y, accounting for the confounder Z'.

Appendix: Naive computational implementation

Exact inference (computation of answers to 'queries' of a DAG) is 'NP-hard'. However, here is a simple, approximate approach (many more sophisticated algorithms exist!) to generating a sample $(x_1^i, x_2^i, \dots, x_n^i)$ from the joint, assuming we can sample from the (simpler!) directed Markov factors:

Algorithm 1 Direct sampling for a DAG.

Given a sorted list $X_1, \dots, X_n$ where all parents of X_i appear before X_i in the list [using e.g. a 'topological sort' algorithm on the DAG]
for $i = 1 : N$ **do**
  Sample x_i value according to $p(x_i \mid \text{pa}(x_i))$
end for

Next, given sample i from the joint, i.e. $(x_1^i, x_2^i, \dots, x_n^i)$, we can

Marginalise variables out by 'ignoring their values':

$$(x_1^i, x_2^i, \dots, x_n^i) \to (x_1^i, x_3^i)$$

gives a sample from $p(x_1, x_3)$.

View values of variables (for calculating joint and conditional probabilities) by 'restricting attention to' samples with required values:

$(x_1^i, x_2^i, \dots, x_n^i) \to$ ignore if $x_2^i \neq 3$ allows us to implement 'conditioning on $x_2^i = 3$'

Calculate probabilities by counting proportions of cases, e.g.

$$p(x_4 = 1 \mid x_2 = 3) \approx \frac{\#x_4^i = 1 \text{ and } \#x_2^i = 3}{\#x_2^i = 3}$$

Note: for continuous variables we need to implement e.g. $x_2^i = 3$ as $x_2^i \approx 3$, e.g., 'within dx' of 3, which can become very inefficient in high dimensions. See Kochenderfer (2015) 'Decision-making under uncertainty' for more.

Appendix: Causal decision theory

So far, we have considered decision theory based on standard probability theory. This is sometimes called 'Evidential decision theory', following Jeffrey (1965). This is based on expected utility:

$$\mathbb{E}_{S;d_i}[u(d_i, S)] = \int u(d_i, s) p(s|d_i) ds$$

However, what if we considered the **causal expected utility** instead:

$$\mathbb{E}_{S;\text{do}(d_i)}[u(d_i, S)] = \int u(d_i, s) p(s|\text{do}(d_i)) ds$$

This leads to so-called **causal decision theory** (Gibbard and Harper, 1976), and it can give different answers to 'evidential decision theory'!

A classic example is **Newcomb's problem**. Here is a simplified version, from Titelbaum (2022) 'Fundamentals of Bayesian Epistemology, vol. 2':

> I'm standing at the bar, trying to decide whether to order a third appletini. Drinking a third appletini is the kind of act much more typical of people with addictive personalities. People with addictive personalities also tend to become smokers. I'd kind of like to have another drink, but I really don't want to become a smoker.

Given a reasonable utility table reflecting my preferences, the answer reached by evidential decision theory is that I **shouldn't** order another appletini if it is sufficiently **correlated** with becoming a smoker. However, this doesn't make much sense if we assume having an addictive personality is not **caused** by having another appletini. Instead, if we imagine an addictive personality causes both a tendency to have another appletini and to become a smoker, we have a causal diagram

smoker $\leftarrow$ addictive personality $\rightarrow$ third appletini

This induces a **correlation** between smoking and appletini drinking, but there is no causation. A causal decision theory analysis instead indicates (again under an appropriate utility function) that I **should** order the third appletini (exercise!). We get different answers! Which is more reasonable to you?

Markov models in time

Overview

Here we consider probabilistic models that have explicit **time dynamics**. We might think of these as defined by the seemingly simple DAG:

$$X_0 \to X_1 \to \cdots \to X_n \to ...$$

for a sequences of (discrete) times $0, 1, ..., n, ...$ While this DAG appears simple, the complexity is that this this is a potentially **infinite** sequence of random variables. This requires new ideas!

In the graphical modelling literature (e.g. Koller and Friedman, 2009) this can be thought of as a **template** graphical model of the form:

$$X_n \to X_{n+1}$$

where now X_n and X_{n+1} are 'template' variables that can be instantiated for *any* n (or t, $t+1$ etc). The original DAG is sometimes said to be an **unrolling** of the template DAG.

One way to think of this is as a (stochastic) **dynamical system** (see EngSci 711). In terms of **probability and stochastic processes**, this is an example of a **stochastic process**, which generalises the idea of a finite collection of random variables. We next consider this perspective.

Stochastic processes

In general a **stochastic process** is an **indexed collection** (or sequence) of **random variables**:

$$(X_t \mid t \in T)$$

which is also written as e.g.

$$\{X_t \mid t \in T\}$$

or

$$(X_t)_{t \in T}$$

for **'time'** variable t in **index set** T.

The variables X_t each take values in the **state space** $\mathcal{X}$, i.e. $X_t \in \mathcal{X}$.

Examples include:

– IID trials
– Weather
– Stock prices
– Epidemic models

Etc. See tutorial!

Classification of stochastic processes

We can classify stochastic processes according to whether

– The **state space** $\mathcal{X}$ is **discrete** or **continuous**

– The **index (time)** set T can be **discrete** or **continuous**

Here we will just consider **discrete time**, **discrete state** stochastic processes for simplicity.

A **realisation** of our stochastic process will then be a **particular value** for the **whole sequence** e.g. $(X_1, X_2, ..., X_n, ...) = (1, 3, -1, ...)$ etc, or (sunny, rainy, cloudy, ...) etc.

Markov stochastic processes

A **Markov process** is a stochastic process for which the **future** only depends on the **current** state and **not on the rest of the past**. E.g., for a discrete-time process $(X_t)_{t\in T}$:

$$X_{t+1} \perp \{X_{t'}\}_{t'\leq t-1} \mid X_t$$

for all time t, where $\{X_{t'}\}_{t'\leq t-1}$ denotes all the random variables up to $t-1$. That is,

$$\text{future} \perp \text{past} \mid \text{present}$$

This is called the **Markov property**. This is just like what we saw for DAGs, but now we have a **template** representing an infinite number of nodes.

Markov chains

A **Markov chain** is stochastic process which is (1) **discrete time**, (2) **discrete state**, and (3) a **Markov process**.

The **Markov property** for **Markov chains** is, for all n (starting at $n = 0$):

$$\mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) = \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1})$$

i.e.

$$p(x_n \mid x_{n-1}, \dots, x_0) = p(x_n \mid x_{n-1})$$

This also implies, for all n, the joint distribution given x_0 has the form

$$p(x_n, x_{n-1}, \dots, x_1 \mid x_0) = \prod_{i=0}^{n-1} p(x_{i+1} \mid x_i)$$

i.e., $X_i \to X_{i+1}$, as expected. (X_0 is conditioned on here as it has no parents, though we can further multiply by its marginal to get the full joint.)

Homogeneous Markov chains

In principle, $p(x_{n+1} \mid x_n)$ could be different for each n ... a common further simplification is to assume that this is **not** the case, i.e.

$$p(x_{n+1} \mid x_n) = p(x_1 \mid x_0)$$

for all n. These are called **homogeneous Markov chains**.

We will restrict attention to homogeneous Markov chains in this course.

Transition probabilities

Given a homogeneous Markov chain and using positive integers like $i, j \in \{1, 2, 3, \dots, \}$ to label possible state values (c.f. e.g. 'sunny', 'rainy' etc), we can then define the **transition probabilities**

$$p_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i) = p_{X_{n+1}|X_n}(j \mid i)$$

and the **transition matrix** $\underline{P}$ as the **matrix** with ijth element $(\underline{P})_{ij}$:

$$(\underline{P})_{ij} = p_{ij}$$

Note the order of the 'from' and 'to' indices:

$$\text{from (i)} \left\{ \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \right. \overbrace{\begin{array}{c} \text{to (j)} \\ 1 \quad 2 \quad 3 \end{array}}^{} \begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.2 & 0.7 & 0.1 \\ 0.3 & 0.5 & 0.2 \end{pmatrix}$$

We will hence sometimes write p_{ij} as $p_{i\to j}$

Transition diagrams vs DAGs

We often use a **different** type of **graphical** diagram to represent **Markov chains** called **transition diagrams**.

While DAGs represent conditional independence relations between nodes representing random variables, transition diagrams use **nodes** to represent possible **values of the state** random variable and **arrows** for allowable **transitions** as time advances, $n \to n+1$. These diagrams are usually **cyclic**! Typically, we also include **transition probabilities** on arrows.

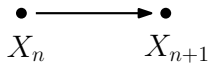
We can illustrate these with an example.

Example transition diagram

Given a homogeneous Markov chain:

$$\underline{P} = \begin{pmatrix} 0.5 & 0.0 & 0.5 \\ 0.8 & 0.0 & 0.2 \\ 0.0 & 0.75 & 0.25 \end{pmatrix}$$

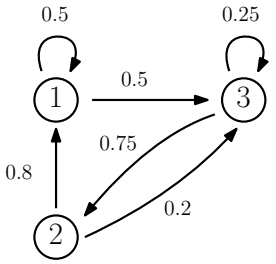
We have, as usual, the simple (template) DAG is just



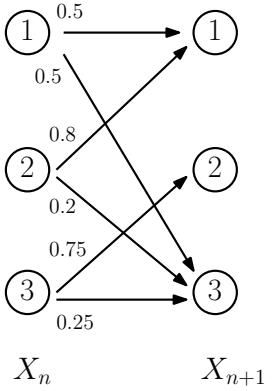
where X_n take values in $\mathcal{X} = \{1, 2, 3\}$, while nodes are random variables. The graph is acyclic.

In contrast, the **transition diagram** can be drawn as ...

Example transition diagram cont'd



Here nodes are **possible values** of $X_n \in \mathcal{X} = \{1, 2, 3\}$. This graph is cyclic, though we can also 'unrol' them and draw transition diagrams in the form:



Again, though, the nodes are the **values** of the random variables, rather than the random variables themselves (you might call these 'internal' diagrams in this form).

Simulating a Markov chain

Suppose we want to draw a sample of a full 'trajectory' from 0 to n from the joint probability function

$$p(x_n, x_{n-1}, \dots, x_0).$$

This factorises according to the Markovian assumption, whether homogeneous or not, as

$$p(x_n \mid x_{n-1})p(x_{n-1} \mid x_{n-2}) \dots p(x_1 \mid x_0)p(x_0).$$

We can simulate from this using the same **direct sampling** method mentioned in the Appendix of the previous section on DAGs.

Labelling the possible state values with positive integers as above, and using subscripts on the probability functions to be explicit, the algorithm can be written here as

Algorithm 2 Direct sampling for a Markov chain.

Sample i value from $p_{X_0}(i)$, i.e. $\mathbb{P}(X_0 = i)$, and record i.
for $n = 1 : N$ **do**
  Sample j value according to $p_{X_n|X_{n-1}}(j \mid i)$, i.e. $\mathbb{P}(X_n = j \mid X_{n-1} = i)$
  Record j
  Set $i \leftarrow j$
end for

For homogeneous Markov chains we have that the transition probability function doesn't depend on n and so

$$p_{X_n|X_{n-1}}(j \mid i) = \mathbb{P}(X_n = j \mid X_{n-1} = i) = p_{i \to j},$$

which simplifies the 'sample j' step above to:

Sample j value according to $p_{i \to j}$.

n-step transitions

The previous algorithm gives a sample of a **full trajectory**.

Sometimes, we **just want to know the probability we will be in some state after n steps**, not worrying about (= averaging over!) any intermediate steps. This leads, assuming a homogeneous Markov chain, to n-step transition probabilities:

$$p_{ij}(n) = p_{i \to j}(n) = \mathbb{P}(X_{m+n} = j \mid X_m = i).$$

For homogeneous Markov chains this gives an n-step transition matrix $\underline{P}_n$ with elements

$$(\underline{P}_n)_{ij} = p_{ij}(n)$$

Chapman-Kolmogorov equations

The n-step transition probabilities satisfy

$$p_{ij}(n+m) = \sum_k p_{ik}(m)p_{kj}(n),$$

which is just

$$\underline{P}_{m+n} = \underline{P}_m \underline{P}_n$$

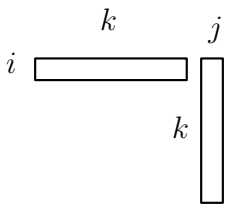
i.e. matrix multiplication! This can also be thought of as 'summing over the intermediate states', going from $i \to k$ in m steps and then $k \to j$ in an additional n steps.

Note that because of the 'from-to' layout of the matrices (a common convention for Markov chains), we think of the matrices operating in order from left to right rather than the more usual right to left (see later for more).

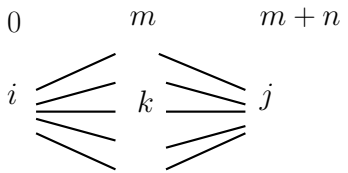
Chapman-Kolmogorov equations continued

i.e.

matrix multiplication



sum over intermediate states



Note:

$$\begin{aligned}\underline{P}_1 &= \underline{P},\\ \underline{P}_2 &= \underline{P}_1\underline{P}_1 = \underline{P}^2\\ &\cdots\\ \underline{P}_n &= \underline{P}^n\end{aligned}$$

i.e. to go n steps in one go, we take n single steps at a time.

Proof of Chapman-Kolmogorov

Here's a proof of the Chapman-Kolmogorov equation, assuming standard probability theory definitions (conditional, marginal etc. relationships):

$$\begin{aligned}p_{ij}(m+n) &= \mathbb{P}(X_{m+n} = j \mid X_0 = i) \text{ [definition of transition probabilities]}\\ &= \sum_k \mathbb{P}(X_{m+n} = j, X_m = k \mid X_0 = i) \text{ [definition of marginal probabili}\\ &= \sum_k \mathbb{P}(X_{m+n} = j \mid X_m = k, X_0 = i)\mathbb{P}(X_m = k \mid X_0 = i)\end{aligned}$$

where the last line folllows from the definition of conditional probability. Now, by the Markov property, we have

$$\begin{aligned}&\sum_k \mathbb{P}(X_{m+n} = j \mid X_m = k, X_0 = i)\mathbb{P}(X_m = k \mid X_0 = i)\\ &\qquad = \sum_k \mathbb{P}(X_{m+n} = j \mid X_m = k)\mathbb{P}(X_m = k \mid X_0 = i)\\ &\qquad = \sum_k \mathbb{P}(X_n = j \mid X_0 = k)\mathbb{P}(X_m = k \mid X_0 = i)\end{aligned}$$

by homogeneity. But this is just

$$\sum_k p_{kj}(n)p_{ik}(m) = \sum_k p_{ik}(m)p_{kj}(n)$$

since the terms are just numbers. Hence we have

$$p_{ij}(m+n) = \sum_k p_{ik}(m)p_{kj}(n)$$

i.e.

$$\underline{P}_{m+n} = \underline{P}_m\underline{P}_n$$

as required.

Example

Given

$$\underline{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$

we have

$$\underline{P}_1 = \underline{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$

and

$$\begin{aligned} \underline{P}_2 = \underline{P}^2 &= \underline{P}\,\underline{P} \\ &= \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix} \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix} \\ &= \begin{pmatrix} 3/8 & 5/8 \\ 5/16 & 11/16 \end{pmatrix} \end{aligned}$$

and

$$\underline{P}_3 = \underline{P}^3 = \begin{pmatrix} 22/64 & 42/64 \\ 21/64 & 43/64 \end{pmatrix}$$

$$\vdots$$

$$\underline{P}_n = \underline{P}^n \underset{n\to\infty}{\to} \begin{pmatrix} 1/3 & 2/3 \\ 1/3 & 2/3 \end{pmatrix}$$

Initial and marginal distributions

Recall our simulation started with a draw from our **initial distribution** $\mathbb{P}(X_0)$. Define the **row vector** (due to 'from-to' convention) μ_0 by

$$\mu_0(i) = \mathbb{P}(X_0 = i),$$

representing the **initial distribution**, i.e. the marginal distribution at $n = 0$. Then

$$\mu_1 = \mu_0 \underline{P}$$

is a **row vector** giving the **marginal distribution** vector at $n = 1$, where

$$\mu_1(i) = \mathbb{P}(X_1 = i).$$

In general we can find the **marginal distribution at any time** n using

$$\mu_n = \mu_0 \underline{P}^n$$

where $\mu_n = \mathbb{P}(X_n = i)$.

Example

Given

$$\underline{P} = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$

if

$$\mu_0 = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

then

$$\mu_1 = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix},$$

while if

$$\mu_0 = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}$$

then

$$\mu_1 = \begin{pmatrix} 3/8 & 5/8 \end{pmatrix}$$

etc.

Exercise: Find μ_2 for each of the μ_0 above.

Problems

A) Given

$$\underline{P} = \begin{pmatrix} 1/3 & 2/3 \\ 1/4 & 3/4 \end{pmatrix}$$

- Calculate the 2- and 3-step transition matrices.

B) Given

$$\underline{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

- Calculate $\underline{P}_2$.

C) Given the transition matrix from (B) and

$$\mu_0 = (1/2 \;\; 1/4 \;\; 1/4)$$

- Find μ_1
- Find μ_2 via **both** $\mu_2 = \mu_0 \underline{P}_2$ and $\mu_2 = \mu_1 \underline{P}_1$.
- Are the results the same? What justifies this in general? (E.g. a property or equation etc).

D) Consider

$$\underline{P} = \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.25 & 0.75 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.0 & 0.25 & ? \end{pmatrix}$$

- Fill in the missing entry
- Draw a state transition diagram with the states labelled from 1 to 4.

Markov models in time continued

Overview

Here we continue to look at Markov models in time, in particular **Markov chains**.

We will look at some aspects of the **long-term dynamics** by first **partitioning** the states into 'communication classes' according to the allowable **state transitions**, and then looking at **invariant and limiting distributions** over states. We will then consider calculating **expected values** under these distributions and end with simple ways of **estimating** transition matrices from data.

In the **Appendices**, we will briefly mention the (non-examinable) topics of filtering, smoothing, and prediction, Markov chain Monte Carlo (MCMC), and Markov decision processes (MDPs).

Communication properties of states

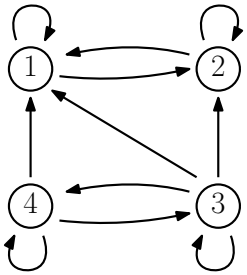
We say that a state i **reaches** a state j if $p_{ij}(n) > 0$ for some n (including $n = 0$, so all states reach themselves). Eequivalently, we also say e.g. j is **'accessible' from** i or even i **'communicates to'** j. We write this as

$$i \rightsquigarrow j$$

If $i \rightsquigarrow j$ and $j \rightsquigarrow i$ then we say **'i and j communicate'** and write

$$i \leftrightsquigarrow j$$

The 'reaches' relation is transitive: if $i \rightsquigarrow j$ and $j \rightsquigarrow k$ then $i \rightsquigarrow k$ (it is also reflexive, i.e. $i \rightsquigarrow i$, as we allow $n = 0$ steps above). Hence, **we can determine if $i \rightsquigarrow k$ by finding a directed path from i to k** in the state transition diagram:



'Reaches' relationships

$1 \rightsquigarrow 2$

$1 \not\rightsquigarrow 3$

$4 \rightsquigarrow 2$

etc.

Other concepts

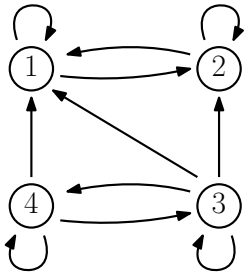
Other concepts for state classifications include

– Recurrent (persistent) and transient states
– Periodic and aperiodic states
– Ergodic states

We won't consider these here (though we will use the word 'ergodic' again…), but see Wasserman (2004) 'All of statistics'...of course! We will mainly focus on the communication relation.

Communication classes

The communication relation $\leftrightsquigarrow$ is reflexive, symmetric, and transitive, and hence defines an **equivalence relation**. In particular, this means it defines a **partitioning** of the states into **communication classes**, which are **disjoint sets** of states such that i **and** j **are in the same class if and only if** $i \leftrightsquigarrow j$:



Communicating classes

$\{1, 2\}$ i.e. $1 \leftrightsquigarrow 2$

$\{3, 4\}$ i.e. $3 \leftrightsquigarrow 4$

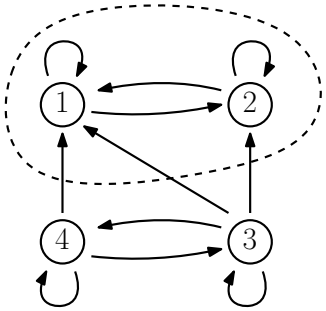
If all states communicate with all others, then we call the Markov chain **irreducible** (can't divide into subclasses).

Trapping sets

A set of states is called a **trapping set** (or closed set) if 'once you enter you never leave'. This means we **only have inward arrows** at the 'border' of the set or, more technically, **no outward arrows at the 'border' of the set**.

These sets can be nested, e.g. both $\{1, 2\}$ and $\{1, 2, 3, 4\}$ might be trapping sets for the same Markov chain.

Example:



Trapping sets
$\{1, 2\}$
$\{1, 2, 3, 4\}$

(whole state space is a trapping set!)

These are the key relationships governing long-term dynamics at the class level. Now we consider similar ideas at the (marginal) **distribution** level.

Invariant marginal distributions

Recall that the transition matrix $\underline{P}$ updates (maps) marginal distribution vectors (or just 'marginals') μ_n to marginals μ_{n+1} via

$$\mu_{n+1} = \mu_n \underline{P}$$

There are some 'special' marginal distributions that are **unchanged** when updating by $\underline{P}$, like the 'fixed points' in dynamical systems, typically denoted by π. These solve

$$\pi = \pi \underline{P}, \text{ such that } \sum_i \pi(i) = 1$$

and are called **invariant distributions** (or stationary/steady state etc distributions).

Eigenvalue/vector problem

This is an eigenvalue/vector problem! The solution π to

$$\pi = \pi \underline{P}$$

is a *left* eigenvector (row vector) for eigenvalue $\lambda = 1$.

We can solve this in the usual way (see below). It says that if we start at a distribution then we stay at that distribution for all time. It is defined by a 'single-step' equation.

Solving eigenproblem?

We can solve in the usual way, i.e., solve the vector equation

$$\pi(\underline{I} - \underline{P}) = 0$$

for π subject to $\sum_i \pi(i) = 1$ (as π is a probability distribution). Let's see an example!

Example

Let

$$\underline{P} = \begin{pmatrix} 0.5 & 0.5 \\ 0.25 & 0.75 \end{pmatrix}$$

then

$$\underline{I} - \underline{P} = \begin{pmatrix} 0.5 & -0.5 \\ -0.25 & 0.25 \end{pmatrix}$$

Let $\pi = [\pi_1, \pi_2]$ such that $\pi_1 + \pi_2 = 1$.

Then

$$\begin{aligned} \pi(\underline{I} - \underline{P}) &= [0.5\pi_1 - 0.25\pi_2, -0.5\pi_1 + 0.25\pi_2] \\ &= [0, 0] \end{aligned}$$

i.e.

$$\begin{aligned} &(1)\ \ 0.5\pi_1 - 0.25\pi_2 = 0 \\ &(2)\ \ -0.5\pi_1 + 0.25\pi_2 = 0. \end{aligned}$$

Next, we note, as in typical eigenvalue/eigenvector problems, that these are not linearly independent – really only one equation. However, we also have $\pi_1 + \pi_2 = 1$ giving the system of two equations in two variables:

$$\begin{aligned} &(1)\ \ 0.5\pi_1 - 0.25\pi_2 = 0 \\ &(2)'\ \pi_1 + \pi_2 = 1. \end{aligned}$$

Solving, e.g. by eliminating π by considering $(3) : \frac{1}{2}(2)' - (1)$, solving for π_2, then substituting back into (2)', gives $\pi_1 = 1/3,\ \pi_2 = 2/3$, i.e.

$$\pi = [1/3, 2/3].$$

Exercise: verify that the solution above satisfies $\pi\underline{P} = \pi$ by explicit matrix multiplication.

Limiting distributions

If the n-step transition matrix has a limit that looks like

$$\underline{P}^n \underset{n\to\infty}{\to} \begin{pmatrix} — & \pi & — \\ — & \pi & — \\ & \cdots & \\ — & \pi & — \end{pmatrix}$$

i.e. the **rows** of $\underline{P}^n$ tend to the **same vector** π, we call the limit defined by

$$\underline{P}^n \underset{n\to\infty}{\to} \underline{P}^\infty$$

the **limiting** n-step **transition matrix**

and

$$\pi$$

the **limiting distribution**.

Limiting?

Note: For any initial distribution μ, we have

$$\begin{aligned}(\mu \underline{P}^{\infty})(j) &= \sum_i \mu(i) \underline{P}^{\infty}_{ij} \\ &= \sum_i \mu(i)\pi(j) \\ &= \pi(j) \left(\sum_i \mu(i) \right) \\ &= \pi(j)\end{aligned}$$

since $\underline{P}^{\infty}_{ij} = \pi(j)$ for all rows i and since $\sum_i \mu(i) = 1$. Hence

$$\mu \underline{P}^{\infty} = \pi,$$

regardless of initial distribution μ.

Hence, the idea of a limiting distribution π and limiting transition matrix $\underline{P}^{\infty}$ is that

– $\underline{P}^{\infty}$ maps all initial distributions to π, i.e. all reach π eventually, regardless of initial distribution

– Once at π we 'stay' at π, i.e. $\pi \underline{P}^{\infty} = \pi$ (invariant)

Limiting vs invariant

Note that

– if a distribution **is limiting then it is invariant**, but

– if a distribution **is invariant then it is not necessarily limiting**

However, there are further condnitions that make these equivalent and in this course/in practice we will **just assume limiting = invariant**. E.g. find invariant via eigenvector then use as limiting.

Expected values

Suppose we have a 'value' (i.e. utility!) function defined on the state space of the Markov chain:

$$u(i) = \text{value/utility of state } i.$$

For a random state at time/stage n, i.e. X_n, the **value** is also a **random variable** at each n, $U_n = u(X_n)$. This defines another stochastic process $(U)_{n \in T}$. Suppose our Markov chain 'settles down' to a unique invariant/limiting distribution (we won't distinguish here) π. We can then compute the **expected value/utility** under the Markov chain in **two** ways.

– The long-term **time/trajectory average** based on the limit of the sequence of random variables:

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} u(X_i)$$

– The **'ensemble' average** based on the distribution:

$$\mathbb{E}_{\pi}[U] = \sum_j u(j)\pi(j) = \pi u^T = u \pi^T.$$

'Ergodic' Markov chains and expectations

When are the above expectations the same? We won't go into the formal definitions (see e.g. Wasserman, 2004), but briefly:

> An irreducible, **'ergodic'** (again, see Wasserman) Markov chain has unique invariant distribution π, also equal to the limiting distribution, which satisfies
>
> $$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} u(X_i) = \mathbb{E}_{\pi}[U] = \sum_j u(j)\pi(j)$$
>
> for all u

i.e. 'ergodic' means, in short, the **time average equals the ensemble average**.

Empirical computation

For us, the key idea is we can compute with either a **sample-based approximation** to the **sequence of random variables over time** (i.e., using a sequence of realisations), or as a **weighted average with respect to weightings given by** π, depending on what information we have available. This is a type of 'sample-based' Monte carlo (Markov chain Monte Carlo!).

Example

Q) Consider a Markov chain on state space $\mathcal{X} = \{0, 1, 2, 3, 4\}$.

– Suppose $u(j) = 2j + j^2$

– Suppose you find invariant distribution π as $\pi = \frac{1}{12}(1\ 2\ 3\ 5\ 1)$

Calculate the expected value under this distribution.

A) Consider the table (for convenience):

$j =$	0	1	2	3	4
$u =$	0	3	8	15	24
$\pi =$	1/12	2/12	3/12	5/12	1/12

Then

$$\begin{aligned}\mathbb{E}_\pi[U] &= \pi u^T \\ &= 0 \times \frac{1}{12} + 3 \times \frac{2}{12} + \cdots + 24 \times \frac{1}{12} \\ &= 10.75\end{aligned}$$

Note: if instead we had a long sequence of realisations from the Markov chain, we could approximate this (assuming ergodicity) by the **sample average** of utility value realisations (see tutorial/problem sets).

Estimating transition matrices

Recall that a realisation of a Markov chain is simply a sequence of state values. Define a matrix of **transition counts** $\underline{C}$ by

$$(\underline{C})_{ij} = \ \#\text{ times } j \text{ follows } i \text{ in a given sequence}$$

We can then estimate:

$$(\underline{P})_{ij} \approx \frac{(\underline{C})_{ij}}{\sum_j (\underline{C})_{ij}}$$

where the denominator normalises by the total count 'from i to anywhere' to ensure the rows are probability distributions. (This is the 'maximum likelihood' estimate of $\underline{P}$).

Example

Consider the following sequence from a Markov chain on $\{1, 2, 3\}$:

$$(1, 2, 1, 1, 1, 2, 3, 1, 1, 2, 1, 3, 1, 1, 3, 2)$$

We can estimate the transition matrix by

$$(\underline{P})_{ij} \approx \frac{(\underline{C})_{ij}}{\sum_j (\underline{C})_{ij}}$$

where

$$\underline{C} = \begin{pmatrix} 4 & 3 & 2 \\ 2 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$$

and the row sums are

$$\sum_j (\underline{C})_{:j} = \begin{pmatrix} 9 \\ 3 \\ 3 \end{pmatrix}$$

and so

$$(\underline{P})_{ij} \approx \begin{pmatrix} 4/9 & 3/9 & 2/9 \\ 2/3 & 0 & 1/3 \\ 2/3 & 1/3 & 0 \end{pmatrix}$$

Appendix: Further topics

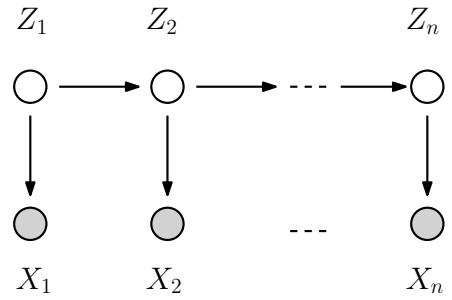
Here we just briefly mention the further topics:

- Filtering, smoothing, prediction
  - Time series etc.
- Markov chain Monte Carlo
  - Bayesian statistical inference (sampling posteriors)
- Markov decision processes
  - Probabilistic dynamic programming
  - Reinforcement learning

Appendix: Filtering, smoothing, prediction

Many discrete or continuous processes can be usefully be formulated as **state-space** models, also known as **state-observation** models, **hidden Markov models (HMMs)**, **hierarchical (process-observation) models** etc.

Idea: hidden states Z_n (white nodes), observations states X_n (dark nodes):



State evolution:

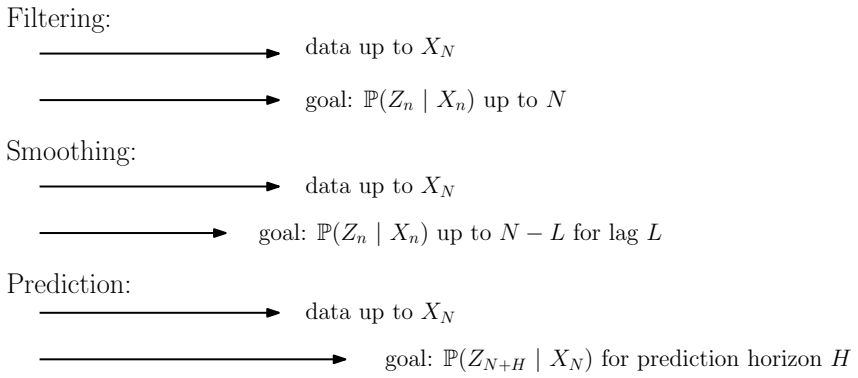
$$Z_{n+1} \perp (Z_{n'})_{n' \leq n-1} \mid Z_n$$

Observation model:

$$X_n \perp (Z_{n'})_{n' \leq n-1} \mid Z_n$$

Appendix: The filtering, smoothing, and prediction tasks

Conceptually, we can represent the tasks as:



In principle, these are just 'standard queries' that we can compute in standard, naive ways. However, **in practice**, we can use the structure of these models to define efficient, recursive computation algorithms, e.g.:

– **Forwards-backwards algorithm** for discrete state
– **Kalman filter** for continuous state space

Appendix: Markov chain Monte Carlo (MCMC)

Often in e.g. Bayesian statistics (but also in other cases), we want **samples** from a distribution or **expectations** with respect to a distribution, but we can't get a closed form solution.

The idea of **MCMC** is to construct an **ergodic Markov chain** with **limiting** distribution equal to the **target** (though still not known explicitly).

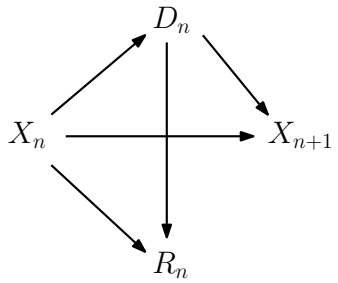
Time simulations will then 'settle down' into **draws from the distribution**, and, e.g., the **time averages will approximate ensemble expectations**.

Appendix: Markov decision processes (MDPs)

Markov decision processes, or **MDPs**, (also partially-observeed Markov decision processes or POMDPs) combine **sequential decision-making** (e.g. dynamic programming) and **Markov stochastic processes**. Get

– Influence diagrams (see figure)
– Stochastic dynamic programming

Influence diagram:



– Transitions are also influenced by actions/decisions e.g. consider $\mathbb{P}(X_{n+1} \mid X_n, D_n)$ for D_n at time/stage n in state X_n.

– Get 'rewards' R_n at each stage

– Given 'policy' (decisions for all states), reduces to Markov chain

Can evaluate and improve policies iteratively ('policy iteration'), among other strategies.

Importantly, MDPs form the foundation for **'reinforcement learning'** in machine learning (RL $\approx$ approximate dynamic programming)

For more, **see next module on dynamic programming**