Decision-Making & Modelling Under Uncertainty (DMU)

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

[10 lectures/tutorials]

- Decision-making under uncertainty [5/10]
  - Basic concepts
  - Risk, probability, utility
  - Statistical: extended setup
    - formulation & empirical risk approx.
    - minimax & Bayes
  - Tutorial sheet
- Modelling under uncertainty [5/10] { models of risk & intervention
  - probability, graphical models, & independence
  - causal interpretations of graphical models
  - stochastic process models (esp. Markov)
  - simulation & estimation tools
  - Tutorial sheet

Lecture 4: Statistical decision theory after Wald

-> Statistical decision functions

-> Minimax & Bayes

So far we have looked at

- no data minimax
- lots of data (empirical loss)

Now we want to look at

- 'some data' case &
- Bayesian approach & relation to minimax

Setting so far

Recall that we reduced the no data problem down to eg

	θ_1	θ_2	θ_3
d_1	$l(d_1,\theta_1)$	...	-
d_2	⋮	⋱	
d_3	$l(d_3,\theta_1)$	..	$l(d_3,\theta_3)$

etc by considering the expected loss under each distribution indexed by θ:

note: updated in L3

$$l(d_i,\theta_j) = \mathbb{E}_{d_i,\theta_j}[l(d_i,s)]$$

Then we considered eg no data minimax with & without randomised (mixed) strategies ...

Statistical decision theory (Wald)

The Wald model begins from tables of this form, ie:

	θ_1	θ_2	θ_3
d_1	$l(d_1,\theta_1)$	--	-
d_2	⋮	⋱	
d_3	$l(d_3,\theta_1)$	--	$l(d_3,\theta_3)$

& typically (not always) assumes the losses are in 'regret form', ie

$$\min_d l(d,\theta) = 0 \quad \text{for all } \theta$$

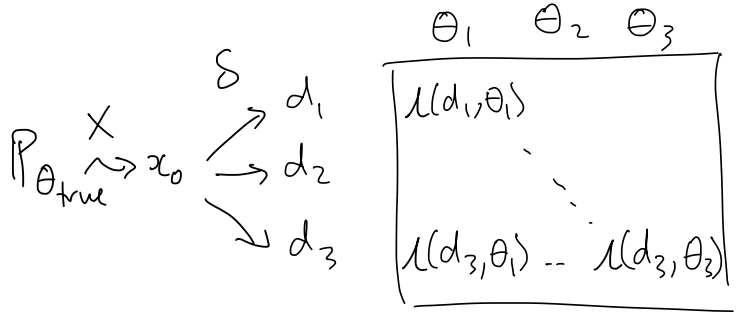
(eg $l(d=\theta,\theta) = 0$)

→ The question is how to use observed data eg x_0 assumed to be a realisation of $X \sim \mathbb{P}_{X;\theta}$

→ will use X & x here rather than s

Statistical decision functions: using data

Wald (~1950) had the idea to define 'statistical decision functions' $\delta(X)$ that map realisations of data to decisions:



ie:

$\delta : x \longmapsto$ decision

ie: data $\longmapsto$ decision.

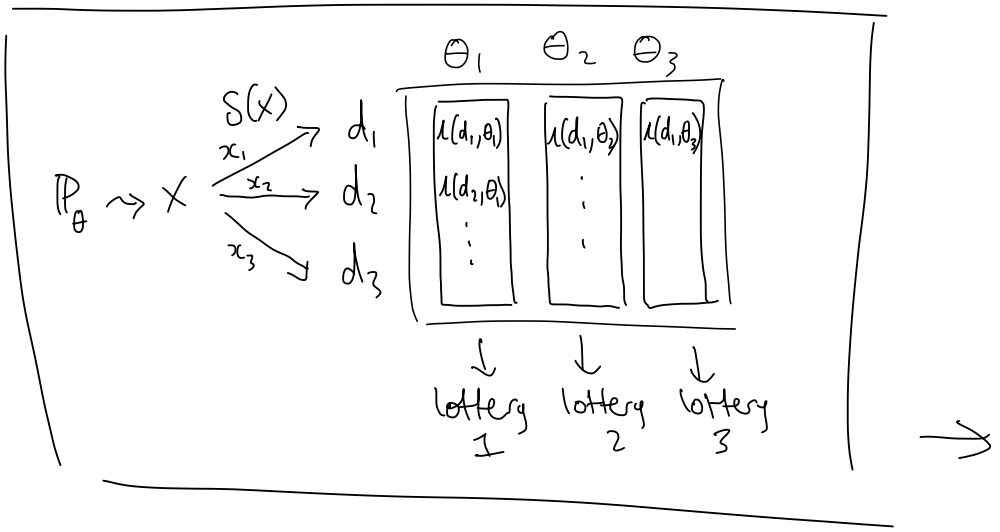
- These are also called 'decision rules' in the literature
- In statistics they are often called estimators

→

Statistical decision functions cont'd.

Treating the data x_0 as a realisation of a random variable X, these are like the randomised/mixed strategies considered previously, except now the random variable is related to the parameter:

→ We assume $X \sim P_{X;\theta}$, ie the data comes from a distribution indexed by θ. We again get lotteries:



→

From Barnett (1999),

'Comparative Statistical Inference'

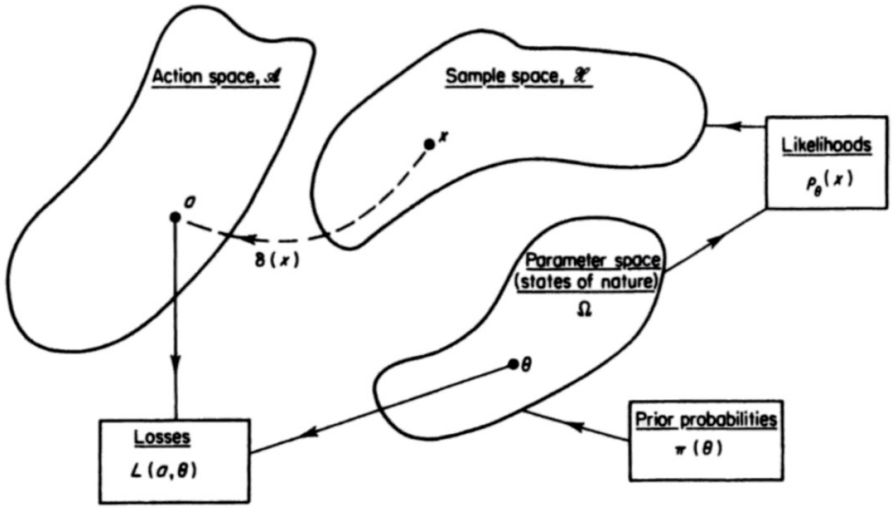


Figure 7.2.2 The superstructure of decision theory

Terminology:

- Note: in the statistical literature a decision rule/function is only called randomised if $\delta(x)$ maps to a further probability distribution over decisions (rather than to a single decision).

  → ie if additional randomness beyond the data distribution randomness is included.

  → Even non-randomised statistical decision functions are still like 'mixed' or 'randomised' strategies in ('no data') simple decision theory / game theory as they include randomness from the data distribution

  → The question in stat. is whether additional randomness is useful (sometimes, yes!)

Risk function

Lotteries ⇒ can value any choice of decision function δ (like how we valued coin flip randomised strategies):

$$l(\delta, \theta) = E_{X;\theta}[l(\delta(X), \theta)]$$

note: decision function

random decision resulting from decision function applied to random variable

The above is often called the 'risk function' of a decision rule δ in statistics: expected (regret) loss for decision function:

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

Risk function?

→ This can be justified in terms of expected utility theory for one-off decisions

→ You can also think of it as providing a measure of the 'expected performance' of the rule $\delta(X)$ if you applied it over & over again for repeated realisations of X

↳ '(frequentist) performance under repeated sampling'

Recall: expectation notation & calc.

⇒ given a distribution over a random variable X write:

$$\mathbb{E}[X] \equiv \mathbb{E}_{P(X)}[X] \equiv \mathbb{E}_X[X]$$

$$= \begin{cases} \sum x \cdot p(x), & \text{discrete} \\ \int x \cdot p(x)\,dx, & \text{continuous} \end{cases}$$

where $p(x) =$
- prob. mass function (pmf) on x (discrete)
- prob. density function (pdf) on x (continuous)

↳ '$p(x)dx$' is pmf for 'dx' at x.

(more later)

→ indexed by θ: $P_{X;\theta} \equiv P(X;\theta)$

$$\mathbb{E}[X] \equiv \mathbb{E}_{P(X;\theta)}[X] \equiv \mathbb{E}_{X;\theta}[X]$$

etc.

Using risk functions

A risk function further transforms our table to the form:

	θ_1	θ_2	θ_3
δ_1	$R(\delta_1,\theta_1)$ - - -		
δ_2	⋮		
δ_3			

$(R(\delta_1,\theta_1) = L(\delta_1,\theta_1)$ etc$)$

(like reducing coin flip problem to choice of p)

In general, the risk of one rule will not be uniformly better than another

Example →

Interlude: Properties of squared error loss

suppose $\boxed{\mathcal{L}(\delta(X), \theta) = (\delta(X) - \theta)^2}$

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

(Note θ not mean of $\delta(X)$ dist in general)

$$\Rightarrow R(\delta, \theta) = \mathbb{E}_{X;\theta}[\mathcal{L}(\delta(X), \theta)] = \mathbb{E}_{X;\theta}[\mathcal{L}(\delta, \theta)]$$

$$= \mathbb{E}_{X;\theta}[(\delta - \theta)^2] = \text{"mean squared error"}$$

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

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

Now: $2(\delta - \bar{\delta})(\bar{\delta} - \theta)$

const. ↑ const. (const. $\delta(X)$) constant

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

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

$$= 0 \text{ since } \mathbb{E}_{X;\theta}(\delta(X)) = \bar{\delta}$$

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

Interlude cont'd

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

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

$$\Rightarrow \boxed{R(\delta, \theta) = \text{Var}(\delta(X)) + [\text{bias}(\delta(X))]^2}$$

risk from summary of $\delta(X)$ dist. — risk from summary not matching θ

Implies:

"Bias-variance trade-off"

If bias ≈ 0 then minimise variance

But trading some bias increase for further variance reduction can be useful. →

Example (Wasserman 2004)

Suppose $X \sim N(\theta, 1)$ *, take one sample & we use squared error loss

Consider $\delta_1(X) = X$

$\delta_2(X) = 3$

which is better?

Can show

$R(\delta_1, \theta) = \text{var}(\delta_1) = 1$ (unbiased)

$R(\delta_2, \theta) = (\text{bias}(\delta_2))^2 = (3-\theta)^2$ (zero variance)

→

* X is normally-distributed with mean $= \theta$, variance $= 1$

Wasserman (2004) 'All of Statistics':

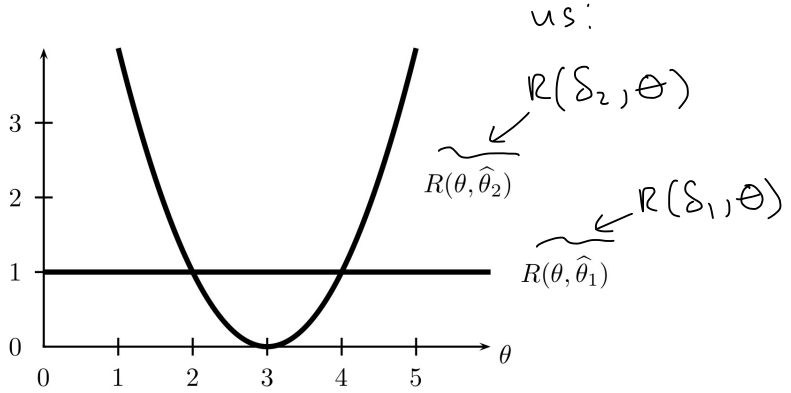


FIGURE 12.1. Comparing two risk functions. Neither risk function dominates the other at all values of θ.

'Silly' estimator δ_2 (here $\vec{\theta}_2$)

⇒ Can get lucky if $\theta = 3$ etc.

Minimax again!

- One approach is to apply minimax again to choose a decision rule (decision function)

Example:

$$\underbrace{\max R(\delta_1, \theta)}_{1} \leq \underbrace{\max R(\delta_2, \theta)}_{\text{eg } 4 \text{ if } \theta = 1}$$

$\Rightarrow$ minimax says choose

$$\delta_1(x) = x \quad \text{over} \quad \delta_2$$

(δ_1 is minimax wrt any other as well!)

Bayes

- Minimax is one way to reduce a risk function to a single number to compare rules

- Another way is to use a Bayesian approach.

- Here the idea is to assume a distribution over parameters (states of nature), ie a distribution over distributions (prior), & consider the associated expected loss

$\longrightarrow$

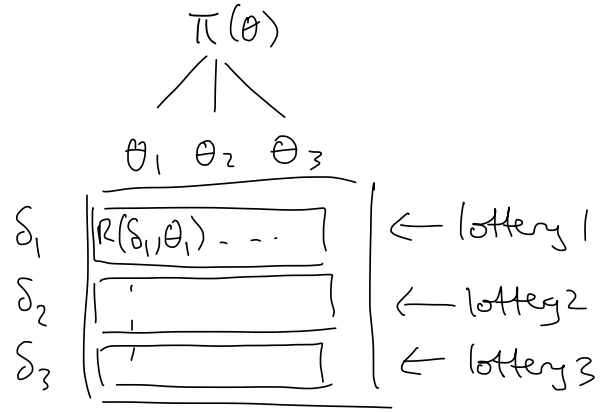
(Prior distribution: your belief about true parameter)

Bayes loss / Bayes risk

- Given a prior $\pi(\theta)$ over Θ we can compute the expected risk under that prior of any decision function (← distribution over θ)
- Common to use the symbol $r(\delta, \theta)$ for Bayes risk & $\pi(\theta)$ for prior prob. distribution

---

Ref:



(follow principles of expected utility given prior . . . .

Gives Bayes risk:

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

expectation under $\pi(\theta)$

"Bayes risk" of δ under $\pi(\theta)$

How to use?

Note:

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

Averages over both data & parameter distributions

also $\begin{cases} \mathbb{E}_{x;\theta} = \mathbb{E}_{x|\theta} \text{ since } \theta \text{ has prob. dist} \\ \mathbb{E}_{\theta|\pi} \equiv \mathbb{E}_{\pi(\theta)} \equiv \text{expectation under } \pi(\theta) \end{cases}$

→

Note: Conditional expectation notation:

"$\mathbb{E}[X|Y]$" another name for...

$$\mathbb{E}_X[X|Y] \equiv \mathbb{E}_{X|Y}[X] \equiv \mathbb{E}_{P(X|Y)}[X] \text{ etc}$$

$$= \begin{cases} \sum x \cdot p(x|y) \\ \int x \cdot p(x|y)\,dy \end{cases}$$

ie take expectation of X but with $X \sim P(X|Y)$ rather than $P(X)$.

Same for $f(X)$ in general.

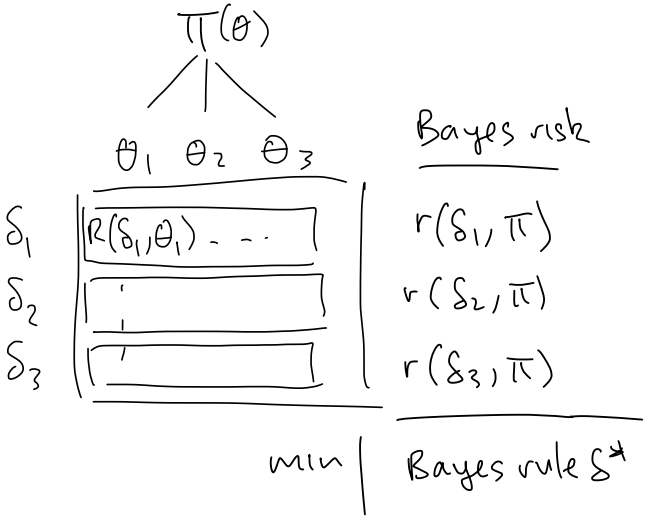
Should really write:

$$\mathbb{E}[X|Y=y] \equiv \mathbb{E}[X|y]$$ for fixed y. } will also use.

Bayes decision rule

- Given the Bayes risk for a chosen prior & set of decision rules, we can then find the decision rule that minimises the Bayes risk for this prior

→ This is called the Bayes rule with respect to the prior:



Posterior expected loss

So far the expectation has been with respect to the (pre-data!) prior (& data dist).

→ we then choose the Bayes rule as the best rule under this prior & data distribution

It turns out* that Bayes rules are equivalent to those that minimise the posterior loss:

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

expectation under posterior!**

Bayes' theorem:

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

(*proof: eg Wasserman (2004) & appendix →)

$E_{\pi(\theta|x)} \equiv E_{\theta|x,\pi}$ } diff. notation for same thing

Bayes rule: find posterior!

This is actually much simpler

→ for any given realisation x of X we find the posterior $\pi(\theta|x)$ use Bayes' rule (see later for more on computing)

→ we compute a posterior summary by taking the minimum expected value of $l(\delta,\theta)$ for given $\pi(\theta|x)$ & 'free choice' of δ

→ this is our decision rule for any x

ie $$\delta^*(x) = \min_{\delta} E_{\pi(\theta|x)}[l(\delta,\theta)]$$

$$E_{\pi(\theta|x)}[l(\delta,\theta)] = \int l(\delta(x),\theta)\,\pi(\theta|x)\,d\theta$$

Bayes summary:

- Given prior $\pi(\theta)$
- Given x (observed value)
- Find posterior $\pi(\theta|x)$
- Compute 'best summary' δ^* of posterior $\pi(\theta|x)$ under loss $l(\delta, \theta)$ & $\theta \sim \pi(\theta|x)$
- Defines $\delta(x) = \delta^*$ as Bayes rule value for given x
- Only need to consider x that occurs!

Example: squared error

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

$$\min_{\delta} \mathbb{E}_{\theta \sim \pi(\theta|x)} [(\delta - \theta)^2] \qquad \mathbb{E}_{\theta|\pi,x} =$$

ie best summary of distribution over θ in terms of squared error

$\Rightarrow$ **posterior mean!** (same as L3)

$$\delta^* = \mathbb{E}_{\theta \sim \pi(\theta|x)}[\theta] \approx \frac{1}{n} \sum \theta_i \text{ for } \theta_i \text{ samples from posterior}$$

[if $l(\delta, \theta) = |\delta - \theta|$

$\Rightarrow$ posterior median etc]

Note — we will consider computing posteriors etc in more detail in 'modelling under uncertainty' part.

Bayes & minimax

A limitation of Bayes cf minimax is that the prior is fairly arbitrary .... & a different (personal) interpretation of probability is required!

Compromise:

There are a number of theorems connecting Bayes rules using a least favourable prior to minimax

→ use Bayes as computational tool to find minimax solutions (Wald!)

Beyond scope to go into detail but:

Bayes & minimax

Basic idea: instead of 'belief' think of 'prior' as a 'randomised' or 'mixed' strategy of nature

→ Nature's best choice = your worst
=
'least favourable prior distribution'

Eg Key results

(1) If δ_0 is a Bayes rule wrt π_0 & $R(\delta_0, \theta) \leq r(\delta_0, \pi_0)$ then π_0 is a least favourable prior & δ_0 is minimax

(2) An 'equaliser' decision rule, ie $R(\delta_0, \theta) = \text{constant}$ in θ, that is also Bayes (or 'extended Bayes') for some prior is minimax & the prior is least favourable

Further reading (selected!)

Wasserman (2004) All of statistics

Barnett (1999) Comparative statistical inference

Parmigiani & Inoue (2009) Decision Theory

Chernoff & Moses (1959) Elementary Decision Theory

Ferguson (1967) Mathematical statistics

Wald (1950) Statistical decision functions

+ most 'standard' books on mathematical statistics

Wasserman (2004): good reference! (see library)

12

Statistical Decision Theory

12.1 Preliminaries

We have considered several point estimators such as the maximum likelihood estimator, the method of moments estimator, and the posterior mean. In fact, there are many other ways to generate estimators. How do we choose among them? The answer is found in **decision theory** which is a formal theory for comparing statistical procedures.

Consider a parameter θ which lives in a parameter space Θ. Let $\widehat{\theta}$ be an estimator of θ. In the language of decision theory, an estimator is sometimes called a **decision rule** and the possible values of the decision rule are called **actions**.

We shall measure the discrepancy between θ and $\widehat{\theta}$ using a **loss function** $L(\theta, \widehat{\theta})$. Formally, L maps $\Theta \times \Theta$ into $\mathbb{R}$. Here are some examples of loss functions:

$L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$	squared error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert$	absolute error loss,
$L(\theta, \widehat{\theta}) = \lvert\theta - \widehat{\theta}\rvert^p$	L_p loss,
$L(\theta, \widehat{\theta}) = 0$ if $\theta = \widehat{\theta}$ or 1 if $\theta \neq \widehat{\theta}$	zero–one loss,
$L(\theta, \widehat{\theta}) = \int \log\left(\frac{f(x;\theta)}{f(x;\widehat{\theta})}\right) f(x;\theta)dx$	Kullback–Leibler loss.

Appendix: 'proof' of

Bayes decision rule → min over (x, θ)

$\Rightarrow$

min posterior expected loss = min over $(\theta|x)$

Write: $\mathbb{E}_\theta \equiv \mathbb{E}_{\pi(\theta)}$ (expectation wrt $\pi(\theta)$)

Bayes risk:

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

$$= \int \left[\int l(\delta(x), \theta)\, p(x|\theta)\, dx \right] p(\theta)\, d\theta$$

$$= \int\int l(\delta(x), \theta)\, p(x|\theta)\, p(\theta)\, dx\, d\theta$$

$$= \int\int l(\delta(x), \theta)\, p(\theta|x)\, p(x)\, dx\, d\theta$$

$$= \int \left[\int l(\delta(x), \theta)\, p(\theta|x)\, d\theta \right] p(x)\, dx$$

$$= \mathbb{E}_x \left[\mathbb{E}_{\theta|x} \left[l(\delta(x), \theta) \right] \right] \Rightarrow$$

Appendix cont'd.

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

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

$\Rightarrow$ minimise this for each x ie solve via

$\Rightarrow$ minimise $\mathbb{E}_{\theta|x} \left[l(\delta, \theta) \right]$

for each fixed value of x.

$$\mathbb{E}_{\theta|x} \left[l(\delta, \theta) \right] = \int l(\delta, \theta)\, p(\theta|x)\, d\theta$$

= posterior expected loss.

Make best Bayes decision for each posterior induced by each realisation x of X

Uses: Rule of iterated expectations

General form:

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

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

Note:

if $f(X,Y) = Y$ then reduces to

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

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

ie $\mathbb{E}_X\left[\mathbb{E}_{Y|X}[Y]\right] = \mathbb{E}_Y[Y]$

also written $\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$

which is the standard but (imo) unnecessarily confusing form often used