

# Decision-Making & Modelling Under Uncertainty (DMU)

Oliver Maclaren ([oliver.maclaren@auckland.ac.nz](mailto: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 { models of [5/10] risk & intervention }

- ↳ probability, graphical models, & independence
- ↳ causal interpretations of graphical models
- ↳ stochastic process models (esp. Markov)
- ↳ simulation & estimation tools
- ↳ Tutorial sheet

Lecture 3 : statistical decision theory  
→ Basic concepts.

Idea: between risk & ignorance

→ don't know probability distribution

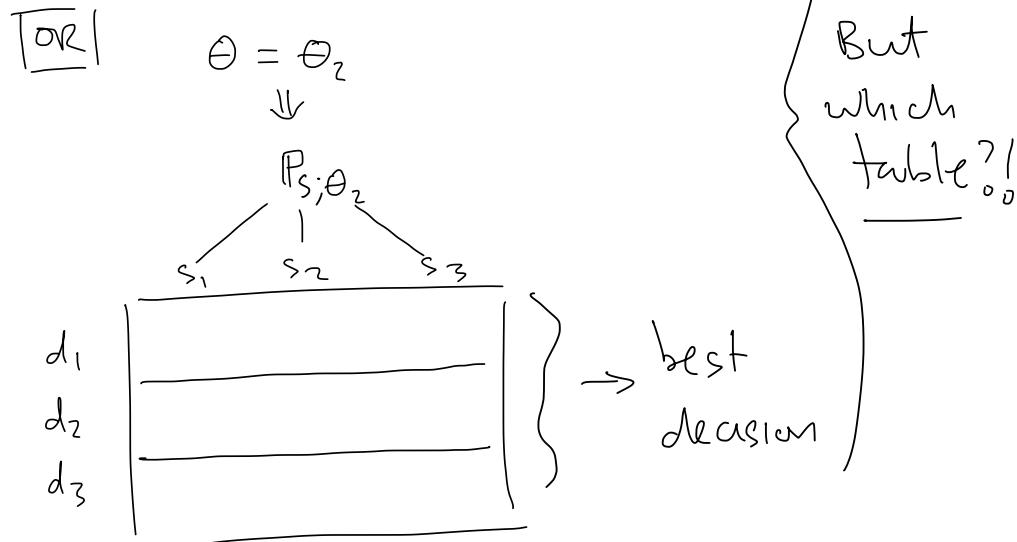
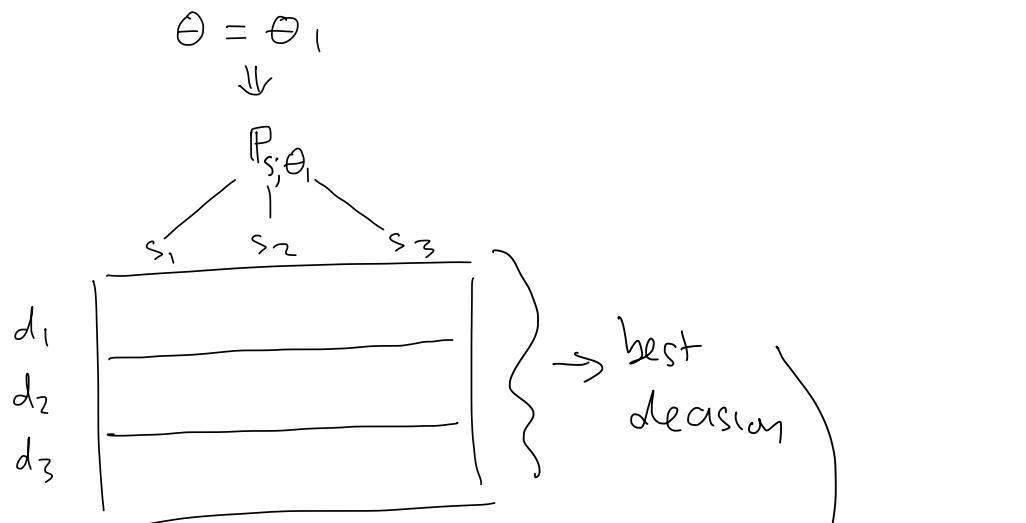
↳ might 'know' eg family of distributions though  
↳ ignorance?

→ have samples (realisations)  
from unknown distribution

$X \sim P_{X;\theta}$ ,  $\theta$  unknown  
 $x_0$  is realised value of  $X$ .

↳ risk?

First: No data  $\Rightarrow$  Multiple possible decision tables



CR . - - - ?

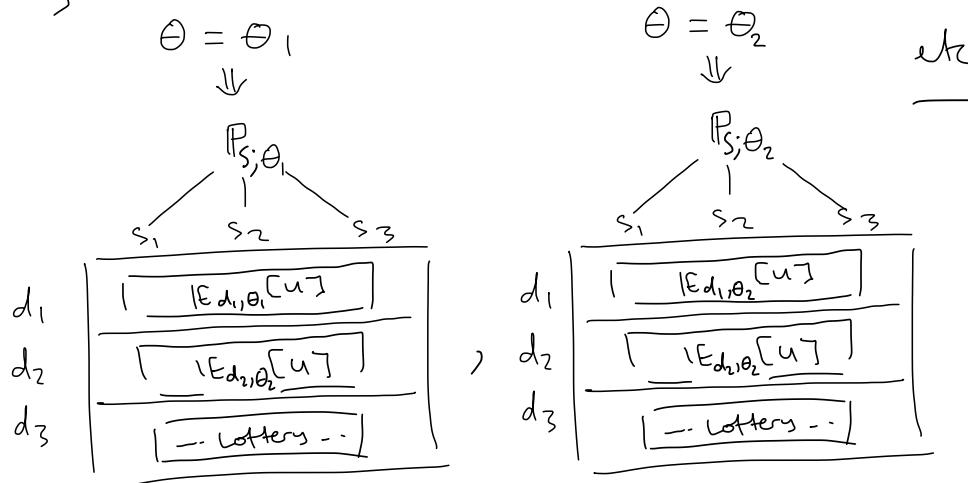
Don't know 'Nature's parameters'  $\theta$ !

$\rightarrow$  Note: in statistical decision theory,  
 $\theta$  is often called the  
'state of nature', though in  
terms of simple decision theory  
this is more like a distribution  
or lottery over states of nature

$\rightarrow$  Better term might be:

- 'Nature's distribution'
- 'Nature's lottery' etc.

However, from Lecture 2 we know  
we can value (assign utilities to)  
lotteries using its expected utility,  
effectively reducing an unknown  
distribution to an unknown  
state of nature in a new decision table



Risk or ignorance?

---

Now we have a table of the form:

---

state of nature = 'nature's dist.'

		$\Theta_1$	$\Theta_2$
decision	$d_1$	$u(d_1, \Theta_1)$	$u(d_1, \Theta_2)$
	$d_2$	$u(d_2, \Theta_1)$	$u(d_2, \Theta_2)$
	$d_3$	:	:

typical decision table

---

Eq value each decision (row) for each  $\Theta$ :

$$U(d_i, \Theta_j) = U(L(P_{S; d_i, \Theta_j}; S)) = E_{d_i, \Theta_j}[u(d_i, S)]$$


---

⇒ gives:

		$\Theta_1$	$\Theta_2$
d	$d_1$	$u(d_1, \Theta_1)$	$u(d_1, \Theta_2)$
	$d_2$	$u(d_2, \Theta_1)$	$u(d_2, \Theta_2)$
	$d_3$	:	:

typical decision table for stat.

→ Wald-style decision theory usually starts  
from tables like above (reduced)

↪ see later ... first →

However, unless we have a distribution over distributions we can't apply expected utility again!

Options:

- minimax (Wald)
- assume / come up with a 'distribution over distributions'  
→ a prior in Bayes
- change loss/decisions available

Note: no data used yet!

No data minimax

Suppose our decision options are to

guess the parameter, i.e.

$$d_i \in \{\theta_1, \theta_2, \dots\}$$

$$\text{& loss} = \begin{cases} 1 & \text{if } d_i \neq \theta_{\text{true}} \\ 0 & \text{if } d_i = \theta_{\text{true}} \end{cases}$$

	$\theta_1$	$\theta_2$	max
$d_1 = \theta_1$	0	1	1
$d_2 = \theta_2$	1	0	1
$d_3 = \theta_3$	1	1	1

min      | (all)

No good solution!

- ? Options
- change loss function, eg  $(d - \theta)^2$
  - change 'decision space', eg sets of decisions
  - randomised (mixed) strategy over  $d$
  - prior over states of nature (Bayes)
  - use data!

First: Dominance

consider:

	$\theta_1$	$\theta_2$	max
$d_1 = \theta_1$	0	1	1
$d_2 = \theta_2$	1	0	1
$d_3 = \theta_3$	1	1	1

min      | (all)

Note:  $d_3$  is always as bad or worse than  $d_1$  &  $d_2$

→ it is 'dominated' by them  
& can be ignored

Dominance applies under

- Ignorance
- risk if prob. are indep. of decisions (see Resnik lecture I)

## Randomised (mixed) strategy?

(note: terminology in stats &

game theory can conflict:

- mixed strategy  $\neq$  randomised estimator
- 'estimator' can include mixed or randomised strategy (in game theory language)

Consider two player game

nature (player 1)

$\theta_1 \quad \theta_2$

		$\theta_1$	$\theta_2$
you (player 2)	$d_1 = \theta_1$	0	1
	$d_2 = \theta_2$	1	0

- o Nature choose a  $\theta$  first, doesn't tell you
- o You have to guess  $\theta$   
→ what should you do?

## 'Zero sum' games ?

stat estimation

		$\theta_1$	$\theta_2$
$d_1$	0	1	
$d_2$	1	0	

~~~~~  
your loss  
form

game theory  
(zero sum)

|       |    | $\theta_1$ | $\theta_2$ |
|-------|----|------------|------------|
| $d_1$ | -1 | 1          |            |
| $d_2$ | 1  | -1         |            |

~~~~~  
prize form:  
player 1 gains  
 $1(d, \theta)$ , player  
2 loses  $1(d, \theta)$

Equivalent: positive linear transf.  
of utility, where...

positive linear trans.:  $U_{new} = a \cdot U_{old} + b$ ,  
 $a > 0$

## Deterministic vs randomised decision

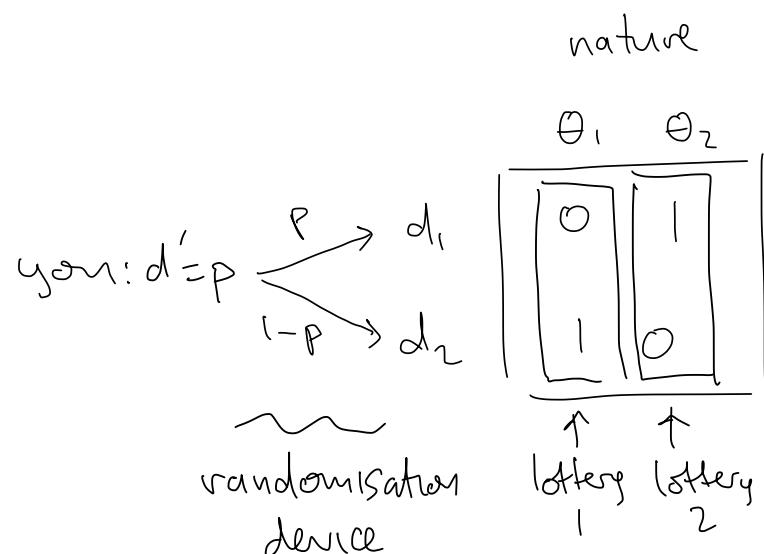
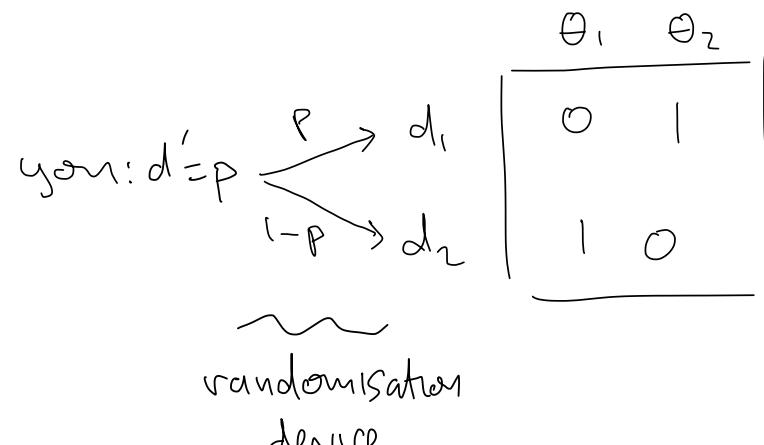
Deterministic ('pure') strategy:

		nature		max
you	$d_1$	0	1	1
	$d_2$	1	0	
		min		1 (not unique)

Randomised ('mixed') strategy:

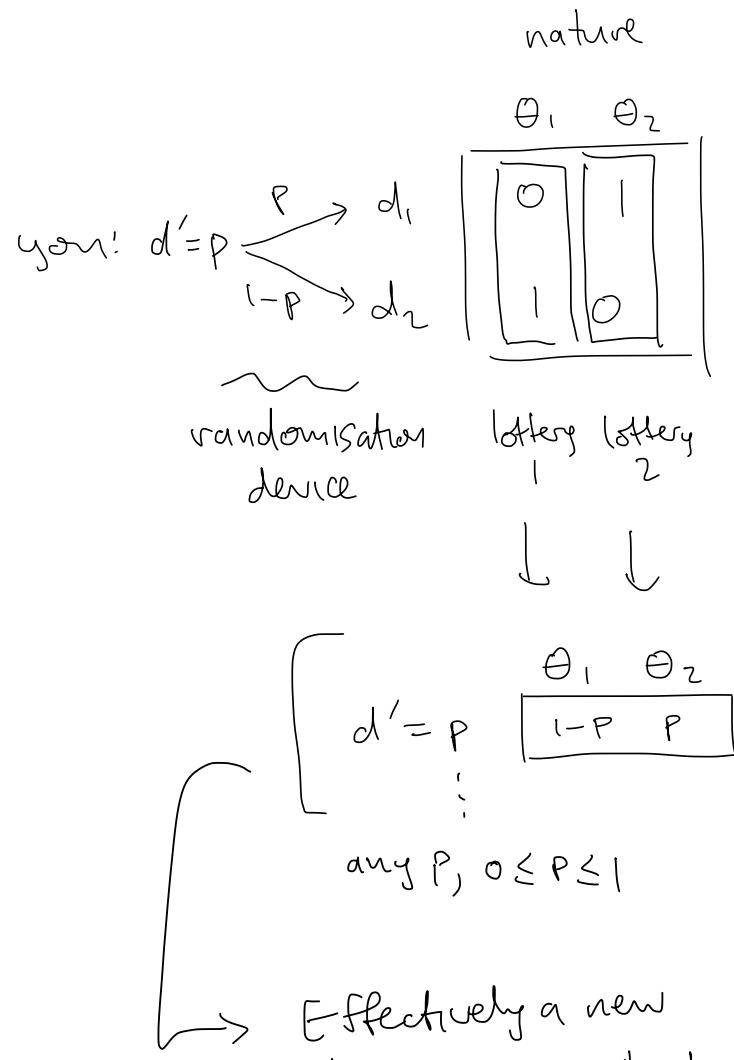
		nature		
you: $d' = p$	$d_1$	0	1	
	$d_2$	1	0	
		randomisation device (choices)		

Randomised:



For given  $p$ , get two lotteries,  
depending on  $\theta$  →

value lotteries via expected loss:



Back to Ignorance?

→ No more randomness in  $d'$  or  $\theta$

→ which P to use for fixed but unknown  $\theta$ ?

Try minimax!

$$d' = p \begin{array}{c|c} \theta_1 & \theta_2 \\ \hline 1-P & P \end{array} \quad \begin{array}{l} \max \\ \min \{ \min \{ \max \{ 1-P, P \} \} \} \end{array}$$

$$\Rightarrow \min \{ \max \{ 1-P, P \} \} \quad (0 \leq P \leq 1)$$

$$= 0.5$$

(if  $P > 1-P$  then  $P \geq 0.5$ )

$$\& \max \{ P, 1-P \} = P \\ \min \{ P \} = 0.5.$$

Similarly for  $1-P$ .

So

No data, two options/states of nature

→ flip a coin!

Note tho': gets bad as number of options gets large!

Further randomisation  
doesn't help

Options:

- change loss function
- change decision space
- use prior (Bayes)
- Add constraints
- use data!

What is the role of the sample data?

→ different approaches use differently

◦ Naive / plug-in / empirical / analog / sample approx

→ use empirical distribution

in place of true distribution

→ solve as if problem under known risk (empirical risk approx).

→ analyse (outside of decision theory?) when this is a good approximation of the 'true' problem

◦ Define decision functions mapping sample

to decision: like an informed random (under data distribution) or 'mixed' strategy!

Then analyse as decision problem using

↳ minimax (regret) loss

↳ Bayes loss

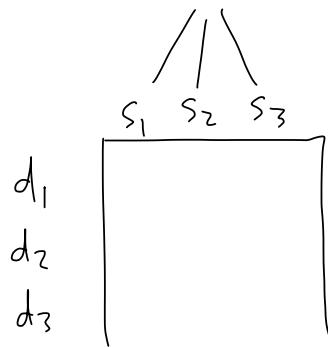
]

later.

Let's consider the 'naive' empirical approach to using data first

Go back to 'pre-reduction' form:

$\theta$  ? unknown



Suppose have a sample of iid data according to  $P_{S; \theta_{\text{true}}}$

i.e.  $X \sim P_{S; \theta_{\text{true}}}$

&  $x_n$  is a vector of n iid realisations of  $X$ .

(using  $X$  instead of  $S$  to avoid notation clashes).

We can consider the empirical distribution over states of nature  $S$

(ie it's a repeatable & repeated experiment like flipping a coin)

① replace  $P_{S; \theta}$  by " $P_n$ " the empirical distribution with prob. = observed frequencies

For continuous data will be  
prob. =  $\frac{1}{n}$  on each point that occurs, 0 otherwise  
( $P(\text{"no ties"}) = 1$ )

For discrete will be  $\frac{\# \text{times outcome}}{n}$

② Solve as 'empirical loss minimisation'

## Basis of / also known as :

- Analog estimation ('by analogy')
  - Plug-in estimation ('plug-in' empirical dist)
  - Maximum likelihood (when loss is -ve log likelihood)      Fisher, Efron, Manski etc (stats, econ)
  - M-estimation ('max likelihood-like' est.   
 → max a sample ave of loss)      Huber etc (robust stats)
  - Empirical risk minimisation      Vapnik etc. (machine learning)
  - Sample average approximation      Shapiro etc? (operations research)
  - Minimum score estimation      David etc (stat. decision theory)
- etc!

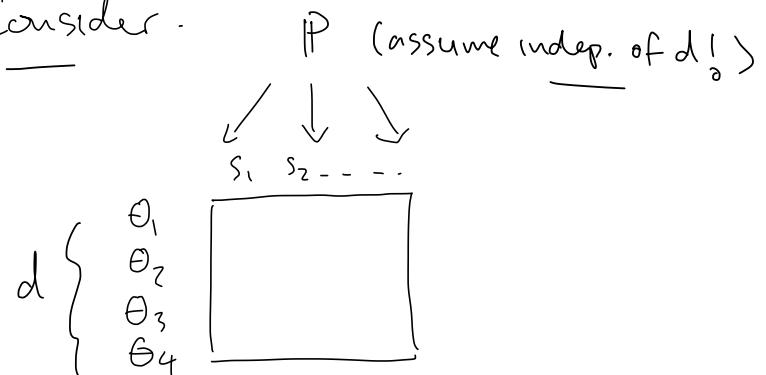
Example ? Statistics !

Goal : best summary of a distribution ?

## Defining target decision : population parameters

- Imagine we had 'infinite samples' ie knew the distribution over state outcomes ↳ Back to risk for now
- we haven't parameterised this, just a distribution  $P$  over  $s$
- We want to "decide" on a single number summary of this distribution
- This is a population ('summary/descriptive/parameter)

Consider:



Define loss :  $\lambda(\theta, s) = \frac{1}{2}(\theta - s)^2$  } need as input!

Best 'decision':

$$\min_{\theta} [E \left[ \frac{1}{2} (\theta - s)^2 \right]]$$

Then:

$$\begin{aligned} & \underset{\theta}{\min} \mathbb{E} \left[ \frac{1}{2} (\theta - s)^2 \right] \\ & \Rightarrow \frac{d}{d\theta} \overbrace{\mathbb{E} \left[ \frac{1}{2} (\theta - s)^2 \right]}^{f(\theta)} = 0 \quad \left. \begin{array}{l} \text{slightly} \\ \text{subtle} \\ \text{math!} \\ \rightarrow \end{array} \right. \\ & = \mathbb{E} \left[ \frac{\partial}{\partial \theta} \frac{1}{2} (\theta - s)^2 \right] \\ & \Rightarrow \mathbb{E} [(\theta - s)] = 0 \end{aligned}$$

$$(\text{linear}) \Rightarrow \mathbb{E} [\theta] = \mathbb{E} [s]$$

$$(\theta \text{ constant}) \Rightarrow \theta = \mathbb{E} [s] = \underline{\text{mean of } s!}$$

$$\text{ie } \boxed{\theta = \bar{s}}$$

→ best summary of distribution (under risk)  
using quadratic loss is the mean!

→ different loss → different summary!  
(eg median = L<sub>1</sub> norm)

→ constraints may change answer!

Alternative:

$$\begin{aligned} & \mathbb{E} \left[ \frac{1}{2} (\theta - s)^2 \right] \\ & = \frac{1}{2} \mathbb{E} [\theta^2 - 2\theta s + s^2] \\ & = \frac{1}{2} \theta^2 - \theta \cdot \mathbb{E}[s] + \frac{1}{2} \mathbb{E}[s^2] \\ & = f(\theta) \end{aligned} \quad \left. \begin{array}{l} \text{linearity} \\ \text{of} \\ \text{expect-} \\ \text{+} \\ \theta \text{ const.} \end{array} \right]$$

$$\Rightarrow \frac{d}{d\theta} f(\theta) = \theta - \mathbb{E}[s] = 0$$

$$\Rightarrow \theta = \underline{\mathbb{E}[s]} \quad \boxed{\quad}$$

Sample version { Don't know IP but can use data to form  $P_n$  (empirical dist) & solve 'min empirical loss'

---

Consider:

$$\min_{\theta} \mathbb{E} \left[ \frac{1}{2} (\theta - s)^2 \right] \xrightarrow{\text{empirical loss approx.}} \min_{\theta} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\theta - s_i)^2$$

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} (\theta - s_i)^2$$

$$\Rightarrow \frac{1}{n} \sum (\theta - s_i) = 0$$

$$\Rightarrow \frac{1}{n} \sum \theta = \frac{1}{n} \sum s_i$$

$$\Rightarrow \theta = \frac{1}{n} \sum s_i$$

= sample average!

i.e. the 'sample analog' of our 'true decision' target

Note: Need theorems for when empirical loss minimisation is a good approx. to actual loss minimisation

→ both the loss & the resulting decision!

Literature:

→ Vapnik etc } machine learning

→ Fisher, Huber, Wald, Le Cam etc } statistics

Short version: not too many options (decisions, states) + lots of data!

sample size  $\gg$  #parameters } 'asymptotic'

Machine learning & modern stat:

- Need 'regularisation' (since  $p \gg n$ ) → 721!
- Not in asymptotic regime without constraints
- See e.g. Vapnik (attached → structural risk min).

## Example problems

- A). Suppose you flip a potentially biased coin (unknown prob. of 'heads' =  $P$ ) 100 times & obtain 53 'heads' & 47 'tails'
- what is the minimum empirical quadratic loss estimate ('decision') of  $P$ ?
- would you trust the min. empirical loss rule if you only had data from one coin flip?
- B).
- What is the 'no data', non-randomised/non-mixed minimum estimate of  $P$  under quadratic loss? (Note:  $0 \leq P \leq 1$ )
- How does this differ from the '0/1' loss of  $\lambda(P_{est}, P) = \begin{cases} 0, & P_{est}=P \\ 1, & P_{est} \neq P \end{cases}$ ?
- C). Look up "maximum likelihood estimation of the mean under normal/Gaussian noise with known variance" → what is the soln?

## Answers

$$A) \circ P_{est} = \frac{53}{100} = 0.53$$

$$\circ \text{No! Always } P_{est} = \begin{cases} 0 \\ 1 \end{cases}$$

intuitively, not enough samples

B). Table is:

		$0 < P (\text{true}) \leq 1$	max
		$\lambda(P_{est}, P) = \frac{1}{2}(P_{est} - P)^2$	
$P_{est}$	0		$1/2$
	$1/2$		$1/4$
	1		$1/2$
			min   $1/4$ ( $P_{est}=1/2$ )

Since:

$$\frac{1}{2}(P_{est} - P)^2, \quad 0 < P \leq 1, \quad \text{max?} \quad \text{min?}$$

is maximised for fixed  $P_{est}$  by  $P$  as different to  $P_{est}$  as possible

(on boundary of constraint)

$$\rightarrow \text{Can't just do } \frac{\partial}{\partial P} \lambda = 0 \quad \dots$$

So:

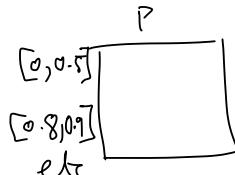
$P_{est} = 0.5$  is the  
"no data minimax est.  
under quadratic loss".

→ For '0/1' loss given, minimax loss

is 1 for all  $P_{est}$

⇒ no good/unique  
solution.

Comments: o solution depends on loss  
function!

→ try taking the decision  
space to be interval subsets →   
the parameter space &  
define loss as

$$l(d, \theta) = \begin{cases} 0 & \text{if } p \in d \text{ (an interval)} \\ 1 & \text{if } p \notin d \quad [a, b] \subseteq [0, 1] \end{cases}$$

→ What is the no data minimax  
solution now?

(c) Mean! Take loss =  $-\log(\text{likelihood})$  }  $\Leftrightarrow \max_{\text{likelihood}}$   
 $\Rightarrow$  quadratic loss!

Chernoff & Moses (1959)

'Elementary Decision Theory'

## CHAPTER 1

### Introduction

#### 1. INTRODUCTION

Beginning students are generally interested in what constitutes the subject matter of the theory of statistics. Years ago a statistician might have claimed that statistics deals with the processing of data. As a result of relatively recent formulations of statistical theory, today's statistician will be more likely to say that statistics is concerned with decision making in the face of uncertainty. Its applicability ranges from almost all inductive sciences to many situations that people face in everyday life when it is not perfectly obvious what they should do.

What constitutes uncertainty? There are two kinds of uncertainty. One is that due to randomness. When someone tosses an ordinary coin, the outcome is random and not at all certain. It is as likely to be heads as tails. This type of uncertainty is in principle relatively simple to treat. For example, if someone were offered two dollars if the coin falls heads, on the condition that he pay one dollar otherwise, he would be inclined to accept the offer since he "knows" that heads is as likely to fall as tails. His knowledge concerns the laws of randomness involved in this particular problem.

The other type of uncertainty arises when it is not known which laws of randomness apply. For example, suppose that the above offer were made in connection with a coin that was obviously bent. Then one could assume that heads and tails were not equally likely but that one face was probably favored. In statistical terminology we shall equate the laws of randomness which apply with the state of nature.

What can be done in the case where the state of nature is unknown? The statistician can perform relevant experiments and take observations. In the above problem, a statistician would (if he were permitted) toss the coin many times to estimate what is

the state of nature. The decision on whether or not to accept the offer would be based on his estimate of the state of nature.

One may ask what constitutes enough observations. That is, how many times should one toss the coin before deciding? A precise answer would be difficult to give at this point. For the time being it suffices to say that the answer would depend on (1) the cost of tossing the coin, and (2) the cost of making the wrong decision. For example, if one were charged a nickel per toss, one would be inclined to take very few observations compared with the case when one were charged one cent per toss. On the other hand, if the wager were changed to \$2000 against \$1000, then it would pay to take many observations so that one could be quite sure that the estimate of the state of nature were good enough to make it almost certain that the right action is taken.

It is important to realize that no matter how many times the coin is tossed, one may never know for sure what the state of nature is. For example, it is possible, although very unlikely, that an ordinary coin will give 100 heads in a row. It is also possible that a coin which in the long run favors heads will give more tails than heads in 100 tosses. To evaluate the chances of being led astray by such phenomena, the statistician must apply the theory of probability.

Originally we stated that statistics is the theory of decision making in the face of uncertainty. One may argue that, in the above example, the statistician merely estimated the state of nature and made his decision accordingly, and hence, decision making is an overly pretentious name for merely estimating the state of nature. But even in this example, the statistician does more than estimate the state of nature and act accordingly. In the \$2000 to \$1000 bet he should decide, among other things, whether his estimate is good enough to warrant accepting or rejecting the wager or whether he should take more observations to get a better estimate. An estimate which would be satisfactory for the \$2 to \$1 bet may be unsatisfactory for deciding the \$2000 to \$1000 bet.

## 2. AN EXAMPLE

To illustrate statistical theory and the main factors that enter into decision making, we shall treat a simplified problem in some

See also book: 'The nature of statistical learning theory' by Vapnik (1995)

## Principles of Risk Minimization for Learning Theory

V. Vapnik  
AT&T Bell Laboratories  
Holmdel, NJ 07733, USA

### Abstract

Learning is posed as a problem of function estimation, for which two principles of solution are considered: empirical risk minimization and structural risk minimization. These two principles are applied to two different statements of the function estimation problem: global and local. Systematic improvements in prediction power are illustrated in application to zip-code recognition.

### 1 INTRODUCTION

The structure of the theory of learning differs from that of most other theories for applied problems. The search for a solution to an applied problem usually requires the three following steps:

1. State the problem in mathematical terms.
2. Formulate a general principle to look for a solution to the problem.
3. Develop an algorithm based on such general principle.

The first two steps of this procedure offer in general no major difficulties; the third step requires most efforts, in developing computational algorithms to solve the problem at hand.

In the case of learning theory, however, many algorithms have been developed, but we still lack a clear understanding of the mathematical statement needed to describe the learning procedure, and of the general principle on which the search for solutions

should be based. This paper is devoted to these first two steps, the statement of the problem and the general principle of solution.

The paper is organized as follows. First, the problem of function estimation is stated, and two principles of solution are discussed: the principle of empirical risk minimization and the principle of structural risk minimization. A new statement is then given: that of local estimation of function, to which the same principles are applied. An application to zip-code recognition is used to illustrate these ideas.

## 2 FUNCTION ESTIMATION MODEL

The learning process is described through three components:

1. A generator of random vectors  $x$ , drawn independently from a fixed but unknown distribution  $P(x)$ .
2. A supervisor which returns an output vector  $y$  to every input vector  $x$ , according to a conditional distribution function  $P(y|x)$ , also fixed but unknown.
3. A learning machine capable of implementing a set of functions  $f(x, w)$ ,  $w \in W$ .

The problem of learning is that of choosing from the given set of functions the one which approximates best the supervisor's response. The selection is based on a training set of  $\ell$  independent observations:

$$(x_1, y_1), \dots, (x_\ell, y_\ell). \quad (1)$$

The formulation given above implies that learning corresponds to the problem of function approximation.

## 3 PROBLEM OF RISK MINIMIZATION

In order to choose the best available approximation to the supervisor's response, we measure the loss or discrepancy  $L(y, f(x, w))$  between the response  $y$  of the supervisor to a given input  $x$  and the response  $f(x, w)$  provided by the learning machine. Consider the expected value of the loss, given by the risk functional

$$R(w) = \int L(y, f(x, w)) dP(x, y). \quad (2)$$

The goal is to minimize the risk functional  $R(w)$  over the class of functions  $f(x, w)$ ,  $w \in W$ . But the joint probability distribution  $P(x, y) = P(y|x)P(x)$  is unknown and the only available information is contained in the training set (1).

## 4 EMPIRICAL RISK MINIMIZATION

In order to solve this problem, the following induction principle is proposed: the risk functional  $R(w)$  is replaced by the empirical risk functional

$$E(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i, w)) \quad (3)$$

constructed on the basis of the training set (1). The induction principle of empirical risk minimization (ERM) assumes that the function  $f(x, w_\ell^*)$ , which minimizes  $E(w)$  over the set  $w \in W$ , results in a risk  $R(w_\ell^*)$  which is close to its minimum.

This induction principle is quite general; many classical methods such as least square or maximum likelihood are realizations of the ERM principle.

The evaluation of the soundness of the ERM principle requires answers to the following two questions:

1. Is the principle consistent? (Does  $R(w_\ell^*)$  converge to its minimum value on the set  $w \in W$  when  $\ell \rightarrow \infty$ ?)
2. How fast is the convergence as  $\ell$  increases?

The answers to these two questions have been shown (Vapnik et al., 1989) to be equivalent to the answers to the following two questions:

1. Does the empirical risk  $E(w)$  converge uniformly to the actual risk  $R(w)$  over the full set  $f(x, w)$ ,  $w \in W$ ? Uniform convergence is defined as

$$\text{Prob}\left\{ \sup_{w \in W} |R(w) - E(w)| > \varepsilon \right\} \longrightarrow 0 \quad \text{as } \ell \rightarrow \infty. \quad (4)$$

2. What is the rate of convergence?

It is important to stress that uniform convergence (4) for the full set of functions is a *necessary* and *sufficient* condition for the consistency of the ERM principle.

## 5 VC-DIMENSION OF THE SET OF FUNCTIONS

The theory of uniform convergence of empirical risk to actual risk developed in the 70's and 80's, includes a description of necessary and sufficient conditions as well as bounds for the rate of convergence (Vapnik, 1982). These bounds, which are independent of the distribution function  $P(x, y)$ , are based on a quantitative measure of the capacity of the set of functions implemented by the learning machine: the VC-dimension of the set.

For simplicity, these bounds will be discussed here only for the case of binary pattern recognition, for which  $y \in \{0, 1\}$  and  $f(x, w)$ ,  $w \in W$  is the class of indicator functions. The loss function takes only two values  $L(y, f(x, w)) = 0$  if  $y = f(x, w)$  and  $L(y, f(x, w)) = 1$  otherwise. In this case, the risk functional (2) is the probability of error, denoted by  $P(w)$ . The empirical risk functional (3), denoted by  $\nu(w)$ , is the frequency of error in the training set.

The VC-dimension of a set of indicator functions is the maximum number  $h$  of vectors which can be shattered in all possible  $2^h$  ways using functions in the set. For instance,  $h = n + 1$  for linear decision rules in  $n$ -dimensional space, since they can shatter at most  $n + 1$  points.

## 6 RATES OF UNIFORM CONVERGENCE

The notion of VC-dimension provides a bound to the rate of uniform convergence. For a set of indicator functions with VC-dimension  $h$ , the following inequality holds:

$$\text{Prob}\{\sup_{w \in W} |P(w) - \nu(w)| > \varepsilon\} < \left(\frac{2\ell}{h}\right)^h \exp\{-\varepsilon^2 \ell\}. \quad (5)$$

It then follows that with probability  $1 - \eta$ , simultaneously for all  $w \in W$ ,

$$P(w) < \nu(w) + C_0(\ell/h, \eta), \quad (6)$$

with confidence interval

$$C_0(\ell/h, \eta) = \sqrt{\frac{h(\ln 2\ell/h + 1) - \ln \eta}{\ell}}. \quad (7)$$

This important result provides a bound to the actual risk  $P(w)$  for all  $w \in W$ , including the  $w^*$  which minimizes the empirical risk  $\nu(w)$ .

The deviation  $|P(w) - \nu(w)|$  in (5) is expected to be maximum for  $P(w)$  close to  $1/2$ , since it is this value of  $P(w)$  which maximizes the error variance  $\sigma(w) = \sqrt{P(w)(1 - P(w))}$ . The worst case bound for the confidence interval (7) is thus likely be controlled by the worst decision rule. The bound (6) is achieved for the worst case  $P(w) = 1/2$ , but not for small  $P(w)$ , which is the case of interest. A uniformly good approximation to  $P(w)$  follows from considering

$$\text{Prob}\{\sup_{w \in W} \frac{|P(w) - \nu(w)|}{\sigma(w)} > \varepsilon\}. \quad (8)$$

The variance of the relative deviation  $(P(w) - \nu(w))/\sigma(w)$  is now independent of  $w$ . A bound for the probability (8), if available, would yield a uniformly good bound for actual risks for all  $P(w)$ .

Such a bound has not yet been established. But for  $P(w) \ll 1$ , the approximation  $\sigma(w) \approx \sqrt{P(w)}$  is true, and the following inequality holds:

$$\text{Prob}\{\sup_{w \in W} \frac{|P(w) - \nu(w)|}{\sqrt{P(w)}} > \varepsilon\} < \left(\frac{2\ell}{h}\right)^h \exp\{-\frac{\varepsilon^2 \ell}{4}\}. \quad (9)$$

It then follows that with probability  $1 - \eta$ , simultaneously for all  $w \in W$ ,

$$P(w) < \nu(w) + C_1(\ell/h, \nu(w), \eta), \quad (10)$$

with confidence interval

$$C_1(\ell/h, \nu(w), \eta) = 2 \left( \frac{h(\ln 2\ell/h + 1) - \ln \eta}{\ell} \right) \left( 1 + \sqrt{1 + \frac{\nu(w)\ell}{h(\ln 2\ell/h + 1) - \ln \eta}} \right). \quad (11)$$

Note that the confidence interval now depends on  $\nu(w)$ , and that for  $\nu(w) = 0$  it reduces to

$$C_1(\ell/h, 0, \eta) = 2C_0^2(\ell/h, \eta),$$

which provides a more precise bound for real case learning.

## 7 STRUCTURAL RISK MINIMIZATION

The method of ERM can be theoretically justified by considering the inequalities (6) or (10). When  $\ell/h$  is large, the confidence intervals  $C_0$  or  $C_1$  become small, and

can be neglected. The actual risk is then bound by only the empirical risk, and the probability of error on the test set can be expected to be small when the frequency of error in the training set is small.

However, if  $\ell/h$  is small, the confidence interval cannot be neglected, and even  $\nu(w) = 0$  does not guarantee a small probability of error. In this case the minimization of  $P(w)$  requires a new principle, based on the simultaneous minimization of  $\nu(w)$  and the confidence interval. It is then necessary to control the VC-dimension of the learning machine.

To do this, we introduce a nested structure of subsets  $S_p = \{f(x, w), w \in W_p\}$ , such that

$$S_1 \subset S_2 \subset \dots \subset S_n.$$

The corresponding VC-dimensions of the subsets satisfy

$$h_1 < h_2 < \dots < h_n.$$

The principle of structure risk minimization (SRM) requires a two-step process: the empirical risk has to be minimized for each element of the structure. The optimal element  $S^*$  is then selected to minimize the guaranteed risk, defined as the sum of the empirical risk and the confidence interval. This process involves a trade-off: as  $h$  increases the minimum empirical risk decreases, but the confidence interval increases.

## 8 EXAMPLES OF STRUCTURES FOR NEURAL NETS

The general principle of SRM can be implemented in many different ways. Here we consider three different examples of structures built for the set of functions implemented by a neural network.

**1. Structure given by the architecture of the neural network.** Consider an ensemble of fully connected neural networks in which the number of units in one of the hidden layers is monotonically increased. The set of implementable functions makes a structure as the number of hidden units is increased.

**2. Structure given by the learning procedure.** Consider the set of functions  $S = \{f(x, w), w \in W\}$  implementable by a neural net of fixed architecture. The parameters  $\{w\}$  are the weights of the neural network. A structure is introduced through  $S_p = \{f(x, w), \|w\| \leq C_p\}$  and  $C_1 < C_2 < \dots < C_n$ . For a convex loss function, the minimization of the empirical risk within the element  $S_p$  of the structure is achieved through the minimization of

$$E(w, \gamma_p) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i, w)) + \gamma_p \|w\|^2$$

with appropriately chosen Lagrange multipliers  $\gamma_1 > \gamma_2 > \dots > \gamma_n$ . The well-known "weight decay" procedure refers to the minimization of this functional.

**3. Structure given by preprocessing.** Consider a neural net with fixed architecture. The input representation is modified by a transformation  $z = K(x, \beta)$ , where the parameter  $\beta$  controls the degree of the degeneracy introduced by this transformation (for instance  $\beta$  could be the width of a smoothing kernel).

A structure is introduced in the set of functions  $S = \{f(K(x, \beta), w), w \in W\}$  through  $\beta \geq C_p$ , and  $C_1 > C_2 > \dots > C_n$ .

## 9 PROBLEM OF LOCAL FUNCTION ESTIMATION

The problem of learning has been formulated as the problem of selecting from the class of functions  $f(x, w)$ ,  $w \in W$  that which provides the best available approximation to the response of the supervisor. Such a statement of the learning problem implies that a unique function  $f(x, w^*)$  will be used for prediction over the full input space  $X$ . This is not necessarily a good strategy: the set  $f(x, w)$ ,  $w \in W$  might not contain a good predictor for the full input space, but might contain functions capable of good prediction on specified regions of input space.

In order to formulate the learning problem as a problem of local function approximation, consider a kernel  $K(x - x_0, b) \geq 0$  which selects a region of input space of width  $b$ , centered at  $x_0$ . For example, consider the rectangular kernel,

$$K_r(x - x_0, b) = \begin{cases} 1 & \text{if } |x - x_0| \leq b \\ 0 & \text{otherwise} \end{cases}$$

and a more general continuous kernel, such as the gaussian

$$K_g(x - x_0, b) = \exp -\left\{\frac{(x - x_0)^2}{b^2}\right\}.$$

The goal is to minimize the local risk functional

$$R(w, b, x_0) = \int L(y, f(x, w)) \frac{K(x - x_0, b)}{K(x_0, b)} dP(x, y). \quad (12)$$

The normalization is defined by

$$K(x_0, b) = \int K(x - x_0, b) dP(x). \quad (13)$$

The local risk functional (12) is to be minimized over the class of functions  $f(x, w)$ ,  $w \in W$  and over all possible neighborhoods  $b \in (0, \infty)$  centered at  $x_0$ . As before, the joint probability distribution  $P(x, y)$  is unknown, and the only available information is contained in the training set (1).

## 10 EMPIRICAL RISK MINIMIZATION FOR LOCAL ESTIMATION

In order to solve this problem, the following induction principle is proposed: for fixed  $b$ , the local risk functional (12) is replaced by the empirical risk functional

$$E(w, b, x_0) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(y_i, f(x_i, w)) \frac{K(x_i - x_0, b)}{K(x_0, b)}, \quad (14)$$

constructed on the basis of the training set. The empirical risk functional (14) is to be minimized over  $w \in W$ . In the simplest case, the class of functions is that of constant functions,  $f(x, w) = C(w)$ . Consider the following examples:

**1. K-Nearest Neighbors Method:** For the case of binary pattern recognition, the class of constant indicator functions contains only two functions: either  $f(x, w) = 0$  for all  $x$ , or  $f(x, w) = 1$  for all  $x$ . The minimization of the empirical risk functional (14) with the rectangular kernel  $K_r(x - x_0, b)$  leads to the K-nearest neighbors algorithm.

**2. Watson-Nadaraya Method:** For the case  $y \in R$ , the class of constant functions contains an infinite number of elements,  $f(x, w) = C(w)$ ,  $C(w) \in R$ . The minimization of the empirical risk functional (14) for general kernel and a quadratic loss function  $L(y, f(x, w)) = (y - f(x, w))^2$  leads to the estimator

$$f(x_0) = \sum_{i=1}^{\ell} y_i \frac{K(x_i - x_0, b)}{\sum_{j=1}^{\ell} K(x_j - x_0, b)},$$

which defines the Watson-Nadaraya algorithm.

These classical methods minimize (14) with a fixed  $b$  over the class of constant functions. The supervisor's response in the vicinity of  $x_0$  is thus approximated by a constant, and the characteristic size  $b$  of the neighborhood is kept fixed, independent of  $x_0$ .

A truly local algorithm would adjust the parameter  $b$  to the characteristics of the region in input space centered at  $x_0$ . Further improvement is possible by allowing for a richer class of predictor functions  $f(x, w)$  within the selected neighborhood. The SRM principle for local estimation provides a tool for incorporating these two features.

## 11 STRUCTURAL RISK MINIMIZATION FOR LOCAL ESTIMATION

The arguments that lead to the inequality (6) for the risk functional (2) can be extended to the local risk functional (12), to obtain the following result: with probability  $1 - \eta$ , and simultaneously for all  $w \in W$  and all  $b \in (0, \infty)$

$$R(w, b, x_0) < E(w, b, x_0) + C_2(\ell/h, b, \eta). \quad (15)$$

The confidence interval  $C_2(\ell/h, b, \eta)$  reduces to  $C_0(\ell/h, \eta)$  in the  $b \rightarrow \infty$  limit.

As before, a nested structure is introduced in the class of functions, and the empirical risk (14) is minimized with respect to both  $w \in W$  and  $b \in (0, \infty)$  for each element of the structure. The optimal element is then selected to minimize the guaranteed risk, defined as the sum of the empirical risk and the confidence interval. For fixed  $b$  this process involves an already discussed trade-off: as  $h$  increases, the empirical risk decreases but the confidence interval increases. A new trade-off appears by varying  $b$  at fixed  $h$ : as  $b$  increases the empirical risk increases, but the confidence interval decreases. The use of  $b$  as an additional free parameter allows us to find deeper minima of the guaranteed risk.

## 12 APPLICATION TO ZIP-CODE RECOGNITION

We now discuss results for the recognition of the hand written and printed digits in the US Postal database, containing 9709 training examples and 2007 testing examples. Human recognition of this task results in an approximately 2.5% prediction error (Säckinger et al., 1991).

The learning machine considered here is a five-layer neural network with shared weights and limited receptive fields. When trained with a back-propagation algorithm for the minimization of the empirical risk, the network achieves 5.1% prediction error (Le Cun et al., 1990).

Further performance improvement with the same network architecture has required the introduction a new induction principle. Methods based on SRM have achieved prediction errors of 4.1% (training based on a double-back-propagation algorithm which incorporates a special form of weight decay (Drucker, 1991)) and 3.95% (using a smoothing transformation in input space (Simard, 1991)).

The best result achieved so far, of 3.3% prediction error, is based on the use of the SRM for local estimation of the predictor function (Bottou, 1991).

It is obvious from these results that dramatic gains cannot be achieved through minor algorithmic modifications, but require the introduction of new principles.

### Acknowledgements

I thank the members of the Neural Networks research group at Bell Labs, Holmdel, for supportive and useful discussions. Sara Solla, Leon Bottou, and Larry Jackel provided invaluable help to render my presentation more clear and accessible to the neural networks community.

### References

- V. N. Vapnik (1982), *Estimation of Dependencies Based on Empirical Data*, Springer-Verlag (New York).
- V. N. Vapnik and A. Ja. Chervonenkis (1989) 'Necessary and sufficient conditions for consistency of the method of empirical risk minimization' [in Russian], *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification, and Forecasting*, 2, 217-249, Nauka (Moscow) (English translation in preparation).
- E. Säckinger and J. Bromley (1991), private communication.
- Y. Le Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel (1990) 'Handwritten digit recognition with a back-propagation network', *Neural Information Processing Systems* 2, 396-404, ed. by D. S. Touretzky, Morgan Kaufmann (California).
- H. Drucker (1991), private communication.
- P. Simard (1991), private communication.
- L. Bottou (1991), private communication.