Decision Making under Uncertainty – Experiments and Value of Information

Bengt Holmstrom

14.124, Spring 2003

- So far, we've studied how individuals choose among a given set of lotteries.
- Here we are concerned with *choices contingent on information* that comes from an "experiment".
- Basic questions:
  - What's the optimal contingent decision rule?
  - What's the value of information?
  - Can information systems/experiments be ranked regardless of utility function?

1 Basic Structure

θ - state of nature
a - (final) action
$u(a, \theta)$ - utility payoff if a chosen and state is θ.

Could come from composition of monetary payoff $x(a, \theta)$ and utility function over money $\widetilde{u}(x)$:

$$u(a, \theta) = \widetilde{u}(x(a, \theta))$$

y - information signal/experimental outcome.

Decision Tree:

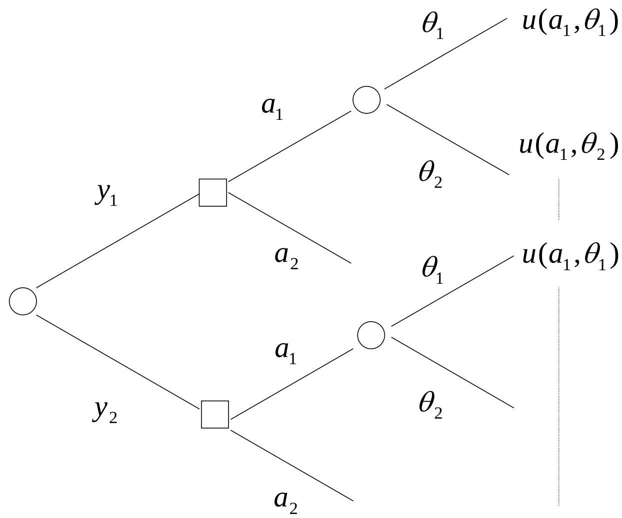

Figure 1

Example:

y – sales forecast
a – production decision
θ – realized demand
$x(a, \theta)$ – monetary payoff

Note: θ needs to include all payoff relevant events (but nothing more).

One and only one state of nature should "happen" in the end.

One and only one signal y should occur.

Strategy/decision rule: $\{a(y)\}$
What you decide if you observe signal y.

2 Priors and Posteriors

We take a Bayesian view: θ and y are random variables with a *joint distribution.*

$$p(y, \theta)$$

Often, we think of this joint distribution as stemming from two distributions:

i a *prior* distribution $p(\theta)$ (= marginal distribution of θ)

ii a set of *likelihoods* $\{p(y \mid \theta)\} \Rightarrow p(y, \theta) = p(\theta)p(y \mid \theta)$

The likelihoods $\{p(y \mid \theta)\}$ describe the full statistical characteristics of the *experiment*. For purposes of decision making, an experiment is identified with the set of likelihood functions $\{p(y \mid \theta)\}$.

The *Law of Total Probability* states:

$$p(y) = \int_{\theta} p(y \mid \theta) p(\theta) d\theta$$

Example: A medical test with signals {positive, negative} for identifying conditions {healthy, sick}.

Such tests are described by two numbers, e.g.,

$$\begin{gathered} p(\text{negative} \mid \text{healthy}) \\ p(\text{positive} \mid \text{sick}) \end{gathered}$$

From these we get the other two likelihoods:

$$\begin{aligned} p(\text{positive} &\mid \text{healthy}) = 1 - p(\text{negative} \mid \text{healthy}) \\ p(\text{negative} &\mid \text{sick}) = 1 - p(\text{positive} \mid \text{sick}) \end{aligned}$$

Bayes rule states:

$$p(\theta \mid y) = \frac{p(y \mid \theta) p(\theta)}{\int_{\theta} p(y \mid \theta) p(\theta) d\theta} \left[= \frac{p(y, \theta)}{p(y)} \right]$$

Probabilities $p(\theta \mid y)$ are called *posteriors*. Given y, $p(\theta \mid y)$ updates beliefs from the initial prior.

Every experiment induces a distribution over posteriors!

For any fixed θ, $p(\theta \mid \widetilde{y})$ is a random variable driven by the distribution of y's.

The function of $p(\cdot \mid \widetilde{y})$ is a random vector if there are a finite number of $\theta-$ outcomes.

This will be conceptually important.

Example:

$$\theta = \{\theta_1, \theta_2\}$$

Priors (and posteriors) are single numbers.

Prior: $p = \Pr(\theta = \theta_1) \Rightarrow 1 - p = \Pr(\theta = \theta_2)$

Posterior: $p'(y) = \Pr(\theta = \theta_1 \mid y)$

Suppose $\quad y = L$ or R

$$\begin{aligned} p &= .5 \\ p(R &\mid \theta_1) = .8 \Rightarrow p(L \mid \theta_1) = .2 \\ p(L &\mid \theta_2) = .6 \Rightarrow p(R \mid \theta_2) = .4 \end{aligned}$$

$$\Rightarrow p(R) = (.5)(.8 + .4) = .6$$

$$p(L) = (.5)(.2 + .6) = .4$$

$$\begin{aligned} p'(R) &= \frac{(.8)(.5)}{(.6)} = 2/3 \\ p'(L) &= \frac{(.2)(.5)}{(.4)} = 1/4 \end{aligned}$$

Note: $\quad E(p') = 2/3 \cdot (.6) + 1/4 \cdot (.4) = .5 = p$

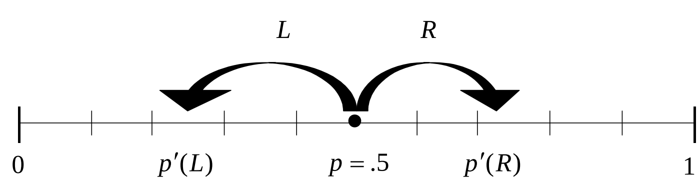

Figure 2

By the *Law of Total Probability*:

$$E_y[p(\theta \mid y)] = p(\theta)$$

$\Rightarrow p(\theta \mid \cdot)$, viewed as a random vector is a martingale.

Very important feature of the stochastic process taking priors into posteriors.

Sequential Updating.

Suppose y_1 and y_2 are outcomes from two separate experiments. We can view $y = (y_1, y_2)$ as the outcome of a single experiment and update beliefs about θ based on likelihoods $p(y \mid \theta)$. Or we can update beliefs sequentially: first incorporate the evidence from y_1 to go from $p(\theta)$ to $p(\theta \mid y_1)$ and then use the evidence from y_2 to go from $p(\theta \mid y_1)$ to $p(\theta \mid y_1, y_2)$.

Both procedures result in same final posterior.

Example:

$$\begin{aligned} \theta_1 &= \text{healthy} \qquad y_i = + \text{ or } - \qquad i = 1, 2 \\ \theta_2 &= \text{sick} \\ p &= \text{prob}(\theta = \theta_2) \end{aligned}$$

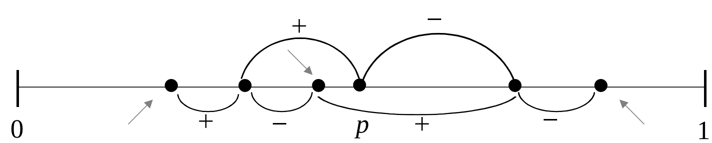

Figure 3

On Sufficient Statistics

In General, y is multi-dimensional. For instance, it may be a collection of facts or a large sample from an experiment (e.g., to test the effectiveness of a drug).

A *statistic* is any (vector-valued) function $T(y)$. For instance, the mean or average is a statistic. So is variance of a sample, median, etc.

Suppose

$$p(y \mid \theta) = p(y \mid T(y)) p(T(y) \mid \theta) \tag{1}$$

where the operational assumption is that the conditional probability $p(y \mid T(y))$ does *not* depend on θ (we can always write $p(y \mid \theta) = p(y \mid T(y), \theta)) P(T(y) \mid \theta))$. When (1) holds we call $T(y)$ a sufficient statistic.

The reason is this. Bayes rule gives

$$p(\theta \mid y) = \frac{p(y \mid T(y))p(T(y) \mid \theta)p(\theta)}{\int\limits_\theta p(y \mid T(y))p(T(y) \mid \theta)p(\theta)d\theta} = \frac{p(T(y) \mid \theta)p(\theta)}{\int\limits_\theta p(T(y) \mid \theta)p(\theta)d\theta}$$

$\Rightarrow$ posterior *only depends on y through $T(y)$.*

That is, for purposes of forming posteriors, it is enough to learn $T(y)$ (rather than all of y). Very often, sample averages are sufficient statistics for the mean of a distribution.

Note: The posterior $\{p(\cdot \mid y)\}$ is a sufficient statistic. Actually, it is a *minimal sufficient statistic* (the least one needs to know to form posteriors).

The reason sufficient statistics are of interest is that optimal decisions will only depend on posteriors.

3 Decision Analysis

A person can find an *optimal decision rule or strategy $a(y)$* in one of two ways:

Ex Post:

$$\max_a \int\limits_\theta u(a,\theta)p(\theta \mid y)d\theta \rightarrow a^*(y)$$

Ex Ante:

$$\max_{a(\cdot)} \int\limits_y \int\limits_\theta u(a(y),\theta)p(y,\theta)d\theta dy \rightarrow a^*(\cdot)$$

Both give the same answer, because *ex ante* optimality holds if and only if decision $a^*(y)$ is optimal *ex post* for every y.

Note: *Ex post* program can be written

$$\max_a \int\limits_\theta u(a,\theta)\frac{p(y \mid \theta)p(\theta)}{\int\limits_\theta p(y \mid \theta)p(\theta)}d\theta$$

$$\sim \max_a \int\limits_\theta u(a,\theta)p(y \mid \theta)p(\theta)d\theta$$

$$v(a,p) \equiv \int u(a,\theta)p(\theta)d\theta$$

v is *linear in probabilities* regardless of shape of $u(a,\theta)$.

$$\begin{aligned} a(y) &= \arg\max_a \quad v(a, p(\cdot \mid y)) \\ V(p) &\equiv \max_a \quad v(a,p) \end{aligned}$$

V is *convex*, because it is the upper envelope of linear functions.

$$V_I \equiv \int_y V(p(\cdot \mid y))p(y)dy$$

This is the maximal expected utility that a person can achieve with information system $Y = \{p(y \mid \theta)\}$.

Value of information system Y:

$$Z_Y \equiv V_Y - V(p_0) \qquad \text{where } p_0 \text{ is prior.}$$

Value of Y is the difference between maximal payoff with Y and payoff without Y (i.e., payoff achieved by choosing best action given prior $p(\cdot)$).

Example.

Two states: $\theta_1 \;\; \theta_2$

Two signal outcomes: $y = L$ or R

Two actions: a_1 or a_2

$$p = \Pr(\theta_1) \qquad 1 - p = \Pr(\theta_2)$$

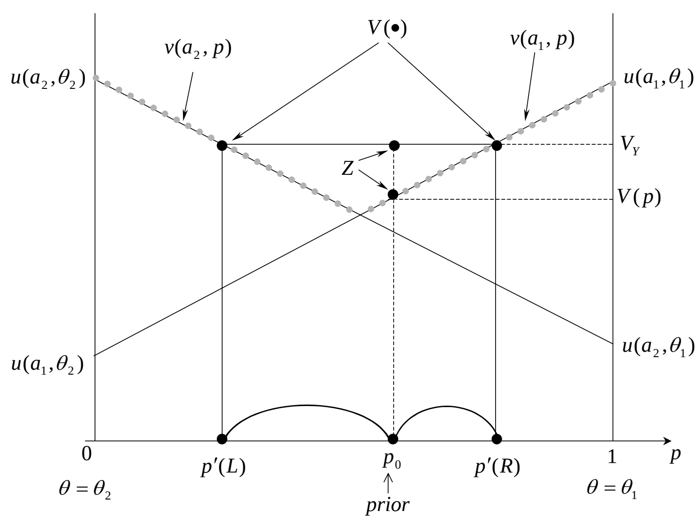

Figure 4

$$\begin{aligned} v(a,p) &= pu(a,\theta_1) + (1-p)u(a,\theta_2) \\ u(a_1,\theta_1) &> u(a_2,\theta_1) \\ u(a_2,\theta_2) &> u(a_1,\theta_2) \end{aligned}$$

Based on graph, the best decision without Y is:

$$a_1 = \arg\max_a \quad v(a, p_0)$$

$V(\cdot)$ is the squiggly line that identifies upper envelope.

According to the graph, if L is observed, a_2 will be optimal decision. If R occurs, a_1 will be optimal:

$$\begin{aligned} a(L) &= a_2 \\ a(R) &= a_1 \end{aligned}$$

Given this rule and considering the probability of L and R, which can be calculated from *Law of Total Probability*:

$$p_L \cdot p'(L) + (1 - p_L) \cdot p'(R) = p_0 \Rightarrow p_L = \Pr(L)$$

we get V_Y as the average of the value of $V(\cdot)$ at $p'(L)$ and $p'(R)$.

Z then is the distance between this average and $V(p)$ evaluated at p_0.

Perfect information system:

$$p'(L) = 0, \qquad p'(R) = 1$$

Totally uninformative information system:

$$p'(L) = p'(R) = p_0 \qquad \text{(prior)}$$

Value of perfect information is graphically:

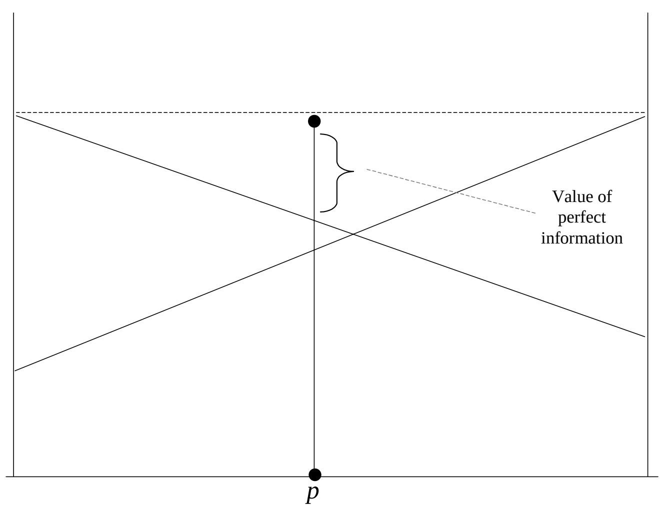

Figure 5

4 Comparison of Information Systems

We will consider only the case with two experimental outcomes:

$$\begin{aligned} y_A &= L \text{ or } R \\ y_B &= B \text{ or } G \end{aligned}$$

Immediate from the graph is that if posteriors from Y_B "*brackets*" posteriors from Y_A, then Y_B is at least as valuable as Y_A.

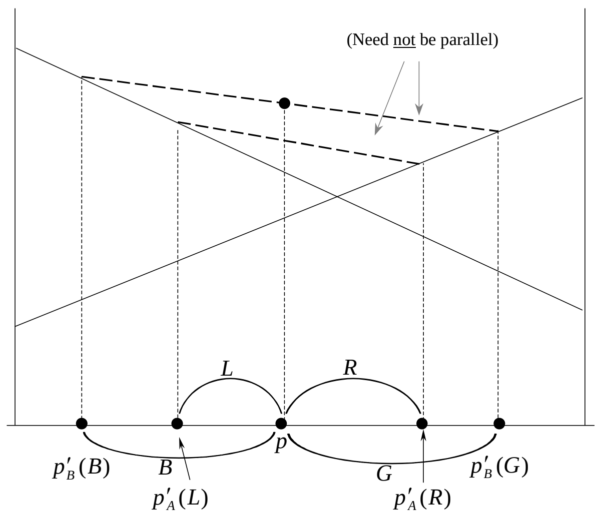

Figure 6

Note: The distribution of posteriors from Y_B is a mean-preserving spread of distribution of posteriors from Y_A.

Given convexity of $V(\cdot)$, this explains (Jensens' inequality) why Y_B is more valuable than Y_A (as can be seen from the graph).

More generally, the information system Y_B is (weakly) preferred to Y_A by all decision-makers (i.e., all utility functions $u(a, \theta)$) if and only if posteriors $p(\theta \mid y_B)$ form mean-preserving spread of posteriors $p(\theta \mid y_A)$ for all θ. (Note: This allows both multi-dimensional θ, a and y.) Mean-preserving spread is better because of Jensen and convexity of $V(\cdot)$.

Going the other way, find utility functions such that in one case Y_A is better, in the other case Y_B is better.

Illustration:

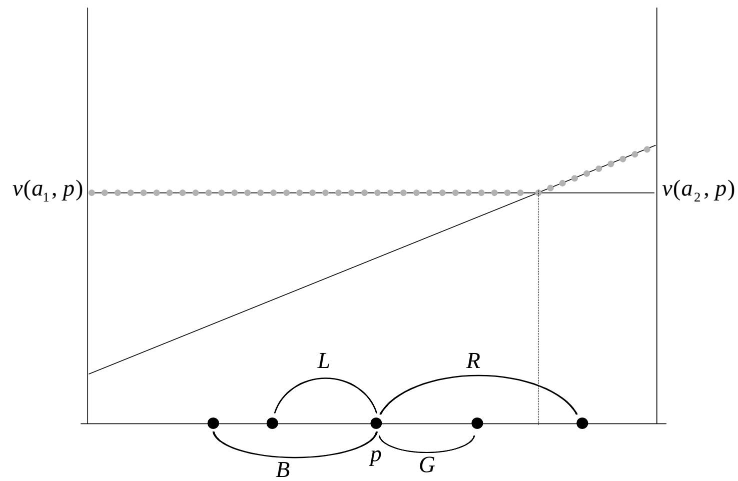

Figure 7

Here Y_A is better than Y_B. Flipping payoff functions around gives the opposite conclusion $\Rightarrow$ we cannot universally compare Y_A and Y_B, except when one distribution of posteriors is a mean-preserving spread of the other.

Garbling

Alternative characterization of information order can be obtained using the notion of *garbling*.

Y_A is a garbling of Y_B if

$$P_A = MP_B^T$$

where

$$\begin{array}{ll} P_A = [p_{ij}^A] & p_{ij}^A = \Pr[y_A = i \mid \theta = j] \\ P_B = [p_{kl}^B] & p_{kl}^B = \Pr[y_B = k \mid \theta = k] \\ M = [m_{ik}] & m_{ik} = \text{“}\Pr[y_A = i \mid y_B = k]\text{”} \end{array}$$

M is a *Markovian matrix*, that is, its columns add up to 1. (The conditional probability interpretation of m_{ik} is natural, but the garbling definition does not *per se* rest on that.)

Blackwell: Y_B is *more informative than* (i.e., every decision-maker prefers Y_B to Y_A (weakly)) if and only if Y_A is a garbling of Y_B.

Intuitively easy in one direction: Signals y_A can be construed as arising out of a two stage process: First, y_B signal observed, then *independently of* θ, but conditional on y_B, the signal y_A is generated (so y_A, given y_B, is pure noise).

Garbling $\Longleftrightarrow$ *MPS (mean-preserving spread) of Posteriors*

Easy to see in two-outcome systems Y_A, Y_B.

Garbling $\Rightarrow p(L \mid B), p(L \mid G)$ are *independent of* θ.

$$p(\theta_1 \mid L) = \frac{p(L \mid \theta_1)p(\theta_1)}{p(L)}$$

$$= \frac{[p(L \mid B)p(B \mid \theta_1) + p(L \mid G)p(G \mid \theta_1)]p(\theta)}{p(L)}$$

$$= \frac{p(L \mid B)p(B)p(\theta_1 \mid B)}{p(L)} + \frac{p(L \mid G)p(G)p(\theta_1 \mid G)}{p(L)}$$

$$= \alpha p(\theta_1 \mid B) + (1-\alpha)p(\theta_1 \mid G)$$

$\Rightarrow p(\theta_1 \mid L)$ is convex combination of posteriors from Y_B.
Similarly true for $p(\theta_1 \mid R)$.
$\Rightarrow$ Posteriors of Y_B bracket posteriors of Y_A when garbling condition holds.

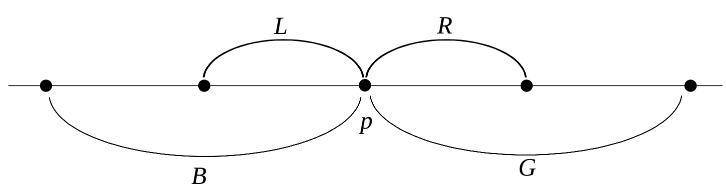

Figure 8

To prove result in other direction, note that given posteriors, we find $p(L)$, $p(R)$, $p(B)$, $p(G)$ from Law of Total Probability, (i.e., jump-probabilities in previous graph fixed by the location of the end points/posteriors).

Can then run argument in reverse to get Markov matrix. (Note again there is no presumption that $\widetilde{y}_A$ is the result of a draw conditional on observing y_B outcome.)

One implication of Blackwell's Theorem: Randomization is sub-optimal.