Lecture Slides - Part 1

Bengt Holmstrom

MIT

February 2, 2016.

- Going to raise the level a little because 14.281 is now taught by Juuso and so it is also higher level
- Books: MWG (main book), BDT specifically for contract theory, others. MWG's mechanism design section is outdated

Comparison of Distributions

- First order stochastic dominance (FOSD):
- Definition: Take two distributions F, G. Then we say that $F \succ_1 G$ (F first order stochastically dominates G) iff
  1. $\forall u$ non-decreasing, $\int u(x)dF(x) \geq \int u(x)dG(x)$
  2. $F(x) \leq G(x)\ \forall x$
  3. There are $\tilde{x}, \tilde{z}$ random variables s.t. $\tilde{z} \geq 0$, $\tilde{x} \sim G$, $\tilde{x} + \tilde{z} \sim F$, and $\tilde{z} \sim H(z|x)$ (z's distribution could be conditional on x).
- All these definitions are equivalent.

- Second order stochastic dominance (SOSD):
- Take two distributions F, G with the same mean.
- Definition: We say that $F \succ_2 G$ (F SOSDs G) iff
  1. $\forall u$ concave and nondecreasing, $\int u(x)dF(x) \geq \int u(x)dG(x)$. ($F$ has less risk, thus is worth more to a risk-averse agent)
  2. $\int_0^x G(t)dt \geq \int_0^x F(t)dt \ \forall x$.
  3. There are $\tilde{x}$, $\tilde{z}$ random variables such that $\tilde{x} \sim F$, $\tilde{x} + \tilde{z} \sim G$ and $E(z|x) = 0$. ($\tilde{x} + \tilde{z}$ is a mean-preserving spread of $\tilde{x}$).
- All these definitions are equivalent.

- Monotone likelihood ratio property (MLRP):
- Let F, G be distributions given by densities f, g respectively. Let $l(x) = \frac{f(x)}{g(x)}$.
- Intuitively, the statistician observes a draw x from a random variable that may have distribution F or G and asks: given the realization, is it more likely to come from F or from G? $l(x)$ turns out to be the ratio by which we multiply the prior odds to get the posterior odds.

- Definition: The pair (f, g) has the MLRP property if $l(x)$ is non-decreasing.
- Intuitively, the higher the realized value x, the more likely that it was drawn from the high distribution, F.
- MLRP implies FOSD, but it is a stronger condition. You could have FOSD and still there might be some high signal values that likely come from G.
- For example: suppose $f(0) = f(2) = 0.5$ and $g(1) = g(3) = 0.5$. Then g FOSDs f but the MLRP property fails (1 is likely to come from g, 2 is likely to come from f).

- This is often used in models of moral hazard, adverse selection, etc., like so:
- Let $F(x|a)$ be a family of distributions parameterized/indexed by a. Here a is an action (e.g. effort) or type (e.g. ability) of an agent, and x is the outcome (e.g. the amount produced).
- MLRP tells us that if $x_2 > x_1$ and $a_2 > a_1$ then $\frac{f(x_2,a_2)}{f(x_2,a_1)} \geq \frac{f(x_1,a_2)}{f(x_1,a_1)}$. In other words, if the principal observes a higher x, it will guess a higher likelihood that it came about due to a higher a.

Decision making under uncertainty

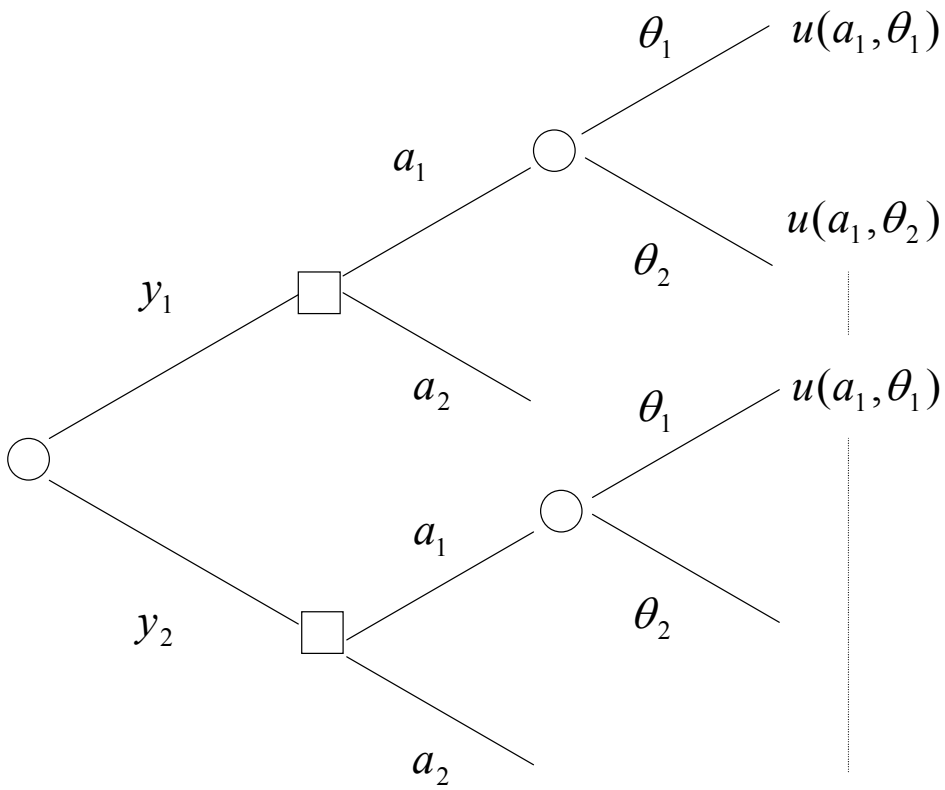
- Premise: you see a signal and then need to take an action. How should we react to the information?
- Goals:
  - Look for an optimal decision rule.
  - Calculate the value of the information we get. (How much more utility do we get vs. choosing under ignorance?)
  - Can information systems (experiments) be preference-ordered? (So you can say experiment A is “more useful” to me than B)

Basic structure:

- θ state of the world, e.g., market demand
- y is the information/signal/experimental outcome, e.g., sales forecast
- a (final) action, e.g., amount produced
- $u(a, \theta)$ payoff from choice a under state θ, e.g., profits

This may be money based: e.g., $x(a, \theta)$ is the money generated and $u(a, \theta) = \tilde{u}(x(a, \theta))$ where $\tilde{u}(x)$ is utility created by having x money.

Figure: A decision problem



- A *strategy* is a function $a : Y \to A$ where Y is the codomain of the signal, and $a(y)$ defines the chosen action after observing y.
- $\theta : \Omega \to \Theta$ is a random variable and $y : \Theta \to Y$ is the signal. Ω gives the entire probability space, Θ is the set of payoff-relevant states of the world, but the agent does not observe θ directly so must condition on y instead.

- How does the agent do this? He knows the joint distribution $p(y, \theta)$ of y and θ. In particular he has a prior belief about the state of the world, $p(\theta) = \int_y p(y, \theta)$. And he can calculate likelihoods $p(y|\theta)$ by Bayes' rule, $p(y, \theta) = p(\theta)p(y|\theta)$.
- As stated, the random variables with their joint distribution are the primitives and we back out the likelihoods. But since the experiment is fully described by these likelihoods, it can be cleaner to take them as the primitives.

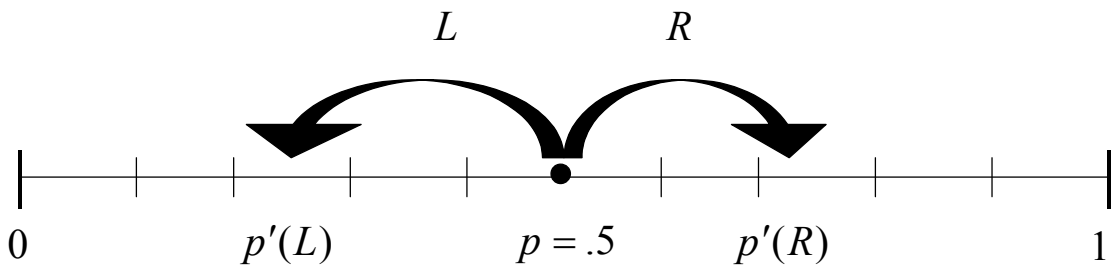
- In deciding what action to take, the agent will need the reverse likelihoods $p(\theta|y) = \frac{p(y,\theta)}{p(y)}$. These are the posterior beliefs, which tell the agent what states θ are more likely given the realization of the experiment y.
- IMPORTANT: every experiment induces a distribution over posteriors.

- By the Law of Total Probability, $p(\theta) = \sum_y p(y)p(\theta|y)$: the weighted average of the posterior must equal the prior. In other words, $p(\theta|\cdot)$, viewed as a random vector, is a martingale.
- Can also take posteriors as primitives!
- Every collection of posteriors $\{p(\theta|y)\}_{y \in Y}$ that is consistent with the priors and signal probabilities (i.e., $p_0(\theta) = \sum_y p(\theta|y)p(y)$) corresponds to an experiment.

- An example: coin toss
- A coin may be biased towards heads (θ_1) or tails (θ_2)
- $p(\theta_1) = p(\theta_2) = 0.5$
- $p(H|\theta_1) = 0.8$, $p(T|\theta_1) = 0.2$
- $p(H|\theta_2) = 0.4$, $p(T|\theta_2) = 0.6$

- We can then find:
- $p(H) = 0.8 * 0.5 + 0.4 * 0.5 = 0.6$, $p(T) = 0.4$
- $p(\theta_1|H) = \frac{0.8*0.5}{p(H)} = \frac{2}{3}$
- $p(\theta_1|T) = \frac{0.2*0.5}{p(T)} = \frac{1}{4}$

Figure: Updating after coin toss ($p'(R) = p(\theta_1|H)$, $p'(L) = p(\theta_1|T)$)

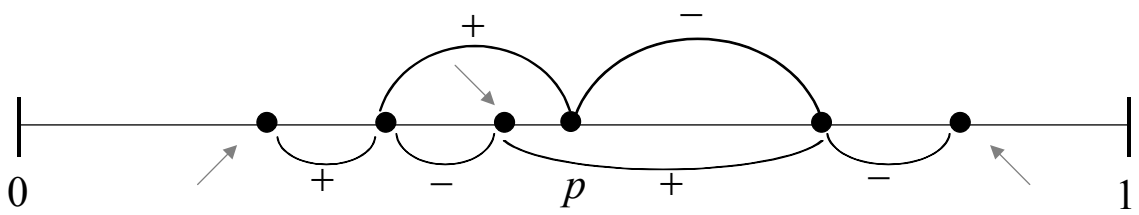


Sequential Updating

- Suppose we have signals y_1 and y_2 coming from two experiments (which may be correlated)
- It does not matter if you update based on experiment A first, then update on B or vice-versa; or even if you take the joint results (y_1, y_2) as a single experiment and update on that
- (However, if the first experiment conditions how or whether you do the second one, then of course this is no longer true)

- E.g., suppose that θ is the health of a patient, $\theta_1 =$ healthy, $\theta_2 =$ sick, and y_1, $y_2 = +$ or $-$ (positive or negative) are the results of two experiments (e.g. doctor's exam and blood test)

Figure: Sequential Updating



Lecture 2

- Note: experiments can be defined independently of prior beliefs about θ
- If we take an experiment as a set of posteriors $p(y|\theta)$, these can be used regardless of $p_0(\theta)$
- (But, of course, they will generate a different set of posteriors $p(\theta|y)$, depending on the priors)
- If you have a blood test for a disease, you can run it regardless of the fraction of sick people in the population, and its probability of type 1 and type 2 errors will be the same, but you will get different beliefs about probability of sickness after a positive (or negative) test

- One type of experiment is where $y = \theta + \epsilon$
- In particular, when $\theta \sim N(\mu, \sigma_\theta^2)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$, this is very tractable because the distribution of y, the distribution of $y|\theta$, and the distribution of $\theta|y$ are all normal
- Useful to define precision of a random variable: $\Sigma_\theta = \frac{1}{\sigma_\theta^2}$
- The lower the variance, the higher the precision
- Precision shows up in calculations of posteriors with normal distributions: in this example
  $\theta|y \sim N\left(\frac{\Sigma_\theta}{\Sigma_\theta+\Sigma_\epsilon}\mu + \frac{\Sigma_\epsilon}{\Sigma_\theta+\Sigma_\epsilon}y, \Sigma_\theta + \Sigma_\epsilon\right)$.

Statistics and Sufficient Statistics

- A statistic is any (vector-valued) function T mapping y to $T(y)$. Statistics are meant to aggregate information contained in y.
- Now suppose that $p(y|\theta) = p(y|T(y))p(T(y)|\theta)$.
- Then $T(y)$ contains all the relevant information that y gives me to figure out θ. In that case, we say T is a sufficient statistic.
- Formally, $p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|T(y))p(T(y)|\theta)p(\theta)}{p(y|T(y))p(T(y))} = \frac{p(T(y)|\theta)p(\theta)}{p(T(y))} = p(\theta|T(y))$.
- Here, we use that $p(y) = p(y, T(y))$ because $T(y)$ is a function of y.

- This is useful if, e.g., the mean of a vector of estimates is a sufficient statistic for θ, then I can forget about the vector and simplify the calculations.
- A minimal sufficient statistic is intuitively the simplest/coarsest possible. Formally, $T(y)$ is a minimal sufficient statistic if it is sufficient and, for any $S(y)$ that is sufficient, there is a function f such that $T(y) = f(S(y))$.

Decision Analysis

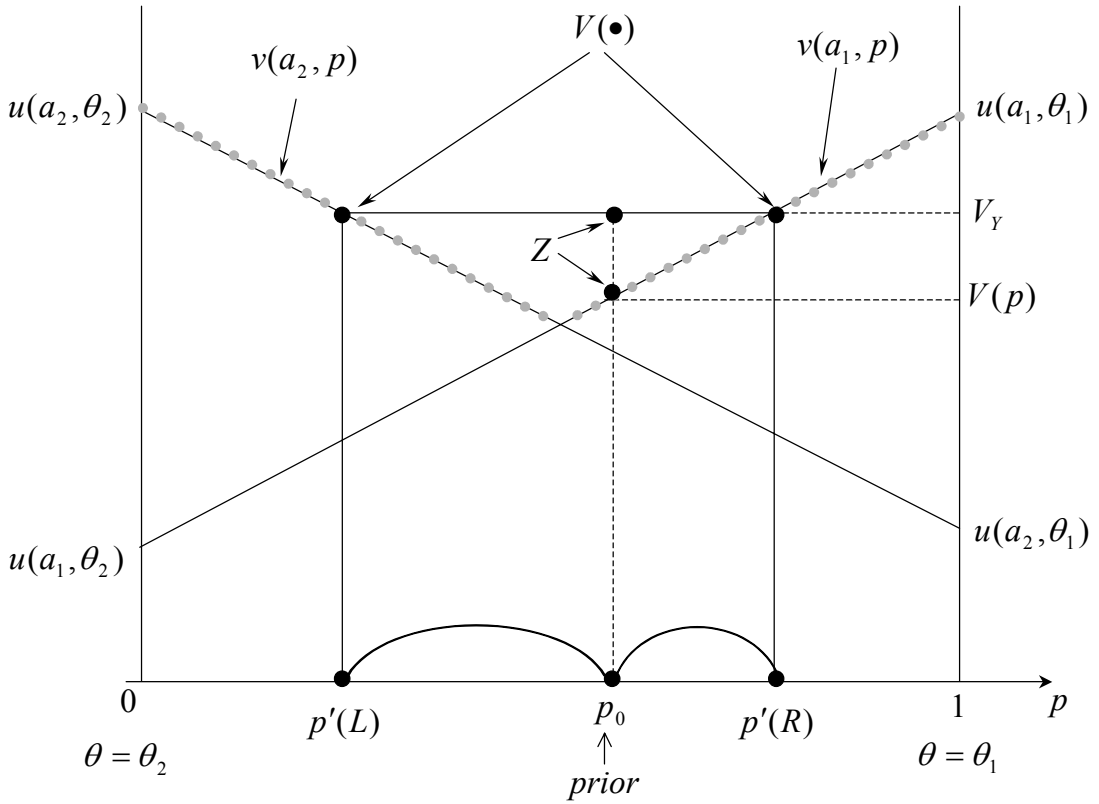
- Remember our framework: we want to take an action a to maximize $u(a, \theta)$, dependent on the state of the world
- We will condition on our information y, given by posteriors $p(y|\theta)$
- Three questions:
  - How to find the optimal decision rule?
  - What is the value of information? For a particular problem, this is given by how much your utility increases from getting the information
  - Can we say anything about experiments in general? E.g., experiment A will always give you weakly higher utility than B, *regardless of your decision problem*

- We can solve for the optimal decision ex post or ex ante
  - Either observe y and then calculate $a^*(y) = \max_a \int_\theta u(a,\theta)p(\theta|y)d\theta$ (ex post)
  - Or build a strategy $a^*(\cdot) = \max_{a(\cdot)} \int_y \int_\theta u(a(y),\theta)p(y,\theta)d\theta dy$ before seeing y (ex ante)
- In decision theory problems with only one agent, both are equivalent (as long as all y's have positive probability)

- Note: given a y, we can think of the ex post problem as a generic problem of the form $V(\tilde{p}) = \max_a v(a, \tilde{p})$ where $v(a, \tilde{p}) = \int_\theta u(a, \theta)\tilde{p}(\theta)d\theta$, and $\tilde{p}(\theta) = p(\theta|y)$.
- Note: $v(a, p)$ is linear in p.
- Exercise: show that $V(p)$ is convex in p.
- Definition: $V_Y = \int_y V(p_y)p(y)dy$ is the maximal utility I can get from information system Y (by taking optimal actions).
- Definition: $Z_Y = V_Y - V(p_0)$ is the value of information system Y (over just knowing the prior).

Consider an example with states θ_1, θ_2; actions a_1, a_2; and signal outcomes L, R

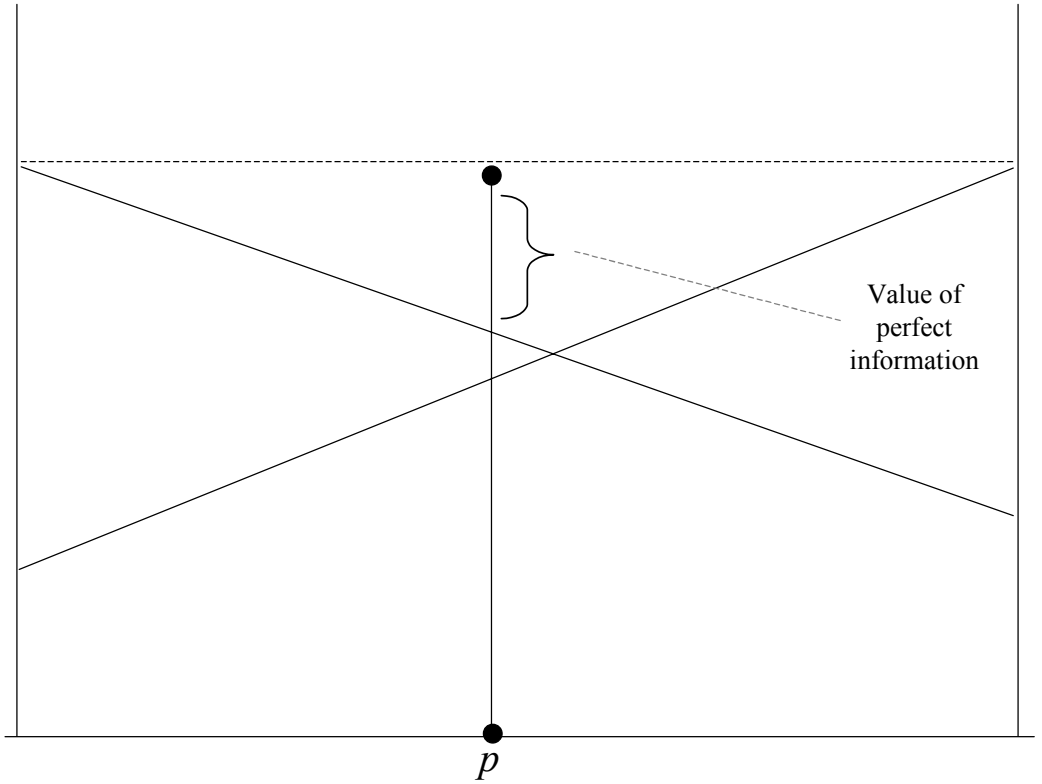
Figure: Value of Information



- In the graph $p = p_0(\theta_1)$
- For generic a and p, $v(a, p) = pu(a, \theta_1) + (1 - p)u(a, \theta_2)$
- In the graph $u(a_1, \theta_1) > u(a_2, \theta_1)$ but $u(a_2, \theta_2) > u(a_1, \theta_2)$ (want to match action to state)
- Under the prior, a_1 is the better action
- With information, we want to choose $a(L) = a_2$, $a(R) = a_1$
- V_Y is a weighted average of the resulting payoffs

This illustrates the maximal gap (between deciding with just the prior vs. exactly knowing the state):

Figure: Value of Perfect Information



Lecture 3

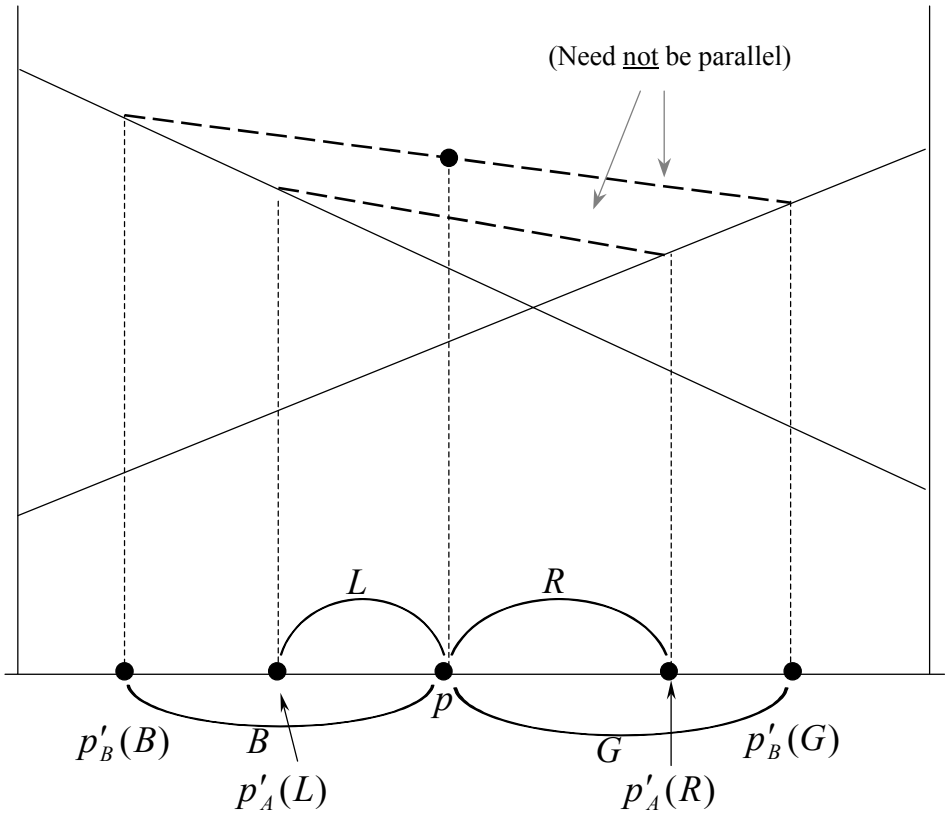
- Life lesson: even if two ways of writing a model are mathematically equivalent, it may make a huge difference how you think about it

Comparison of Experiments

- Question: given two experiments A, B, when is $Y_A \succ Y_B$ regardless of your decision problem?
- Answer 1: Iff the distribution of posteriors from Y_A is a MPS (mean-preserving spread) of the distribution of posteriors from Y_B.

In this example with two states and two two-outcome experiments, the one with more extreme posteriors gives higher utility

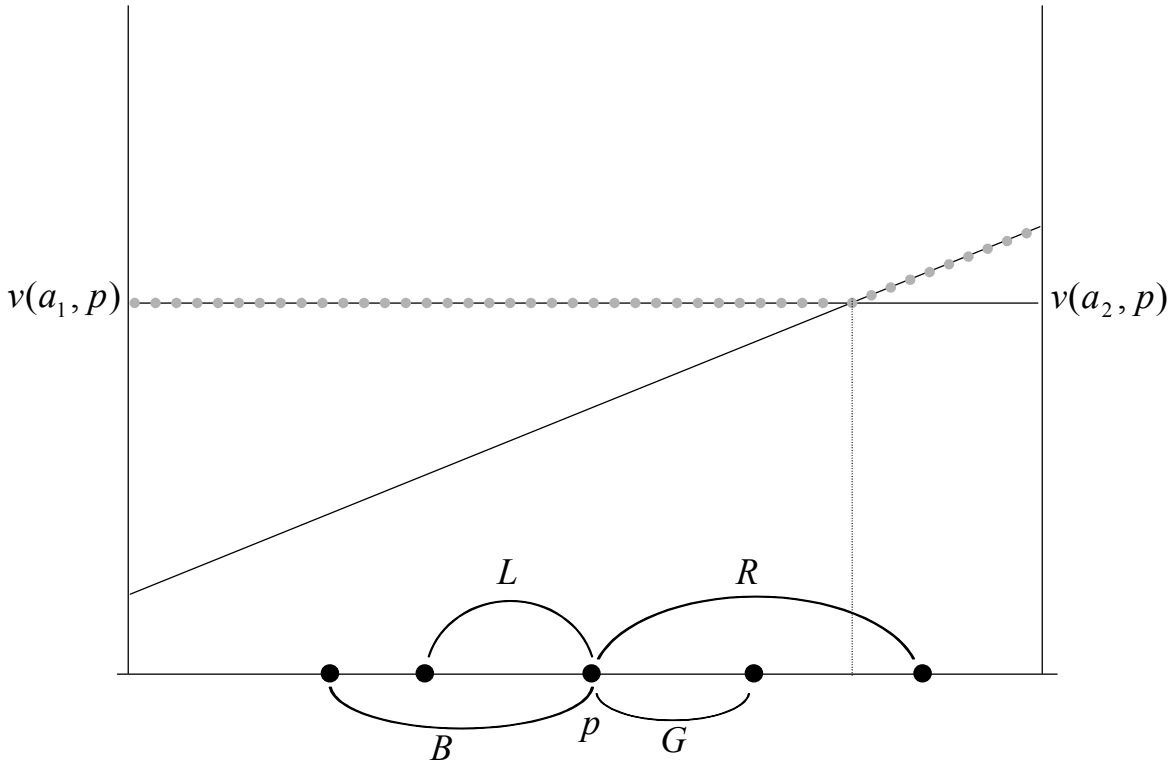
Figure: Mean-preserving spread of posteriors



- Idea: if Y_A's posterior distribution is a MPS of Y_B's, it is as though knowing the outcome of experiment A amounts to knowing the outcome of B, then being given some extra info (which generates the MPS)
- Note: MPS of the *signals* means less information, but MPS of the resulting *posteriors* means more information!
- Formally, since V is convex, averaging V over a more dispersed set gives a higher result (by Jensen's inequality).

- Given two experiments which do not bracket each other (neither's posteriors are a MPS of the other's), we can find decision problems for which either one is better
- In the graph, the experiment with outcomes B, G is uninformative for our purposes, hence worse than the one with outcomes L, R
- But conclusion is reversed if we change the payoff structure

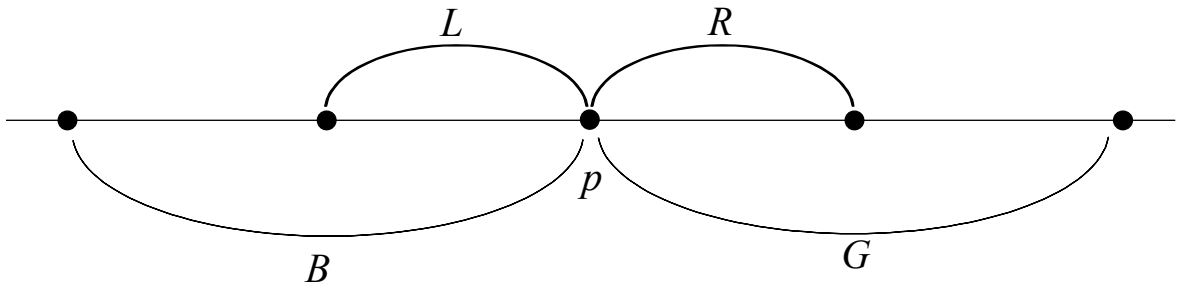
Figure: Unordered experiments



- Second attempt: use the concept of Blackwell garbling.
- Definition: Y_B is a garbling of Y_A if $P_B = MP_A^T$, where
  - $P_B = [p_{ij}]^B$, where $p_{ij}^B = P[y_B = i|\theta = j]$,
  - $P_A = [p_{kl}]^A$, where $p_{kl}^A = P[y_A = k|\theta = l]$,
  - $M = [m_{ik}]$, a Markov matrix (its columns add up to 1).
- The idea: B can be construed as an experiment that takes the outcomes of A and then mixes them up probabilistically. This makes B less informative (even if it e.g. had more outcomes than A).
- Answer 2: $Y_A \succ Y_B$ iff B is a Blackwell garbling of A.
- Corollary: B is a Blackwell garbling of A iff the posterior distribution of A is a MPS of B.

For two-outcome experiments, easy to show that a garbling has posteriors bracketed by those of the the "original" experiment

Figure: Garbled posteriors



14.124 Microeconomic Theory IV

Spring 2017