Rational Inattention with Shannon Mutual Information Costs

Mark Dean

ECON 2090 Spring 2015

Shannon Information Costs

- We have so far considered what we can say when we are agnostic about information costs
- We now move consider behavior under a specific assumed cost for information
- Based on the concept of Shannon Entropy
- Popular in the applied literature
- Consider this the ‘Cobb Douglas’ case to last week’s ‘revealed preference’ treatment
- Read Cover and Thomas for more information

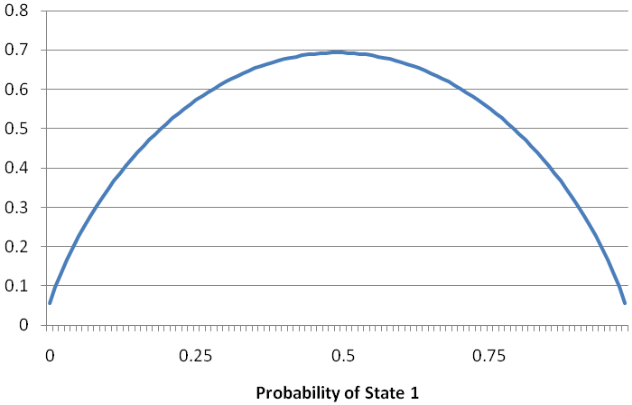
Outline

Shannon Entropy

- Shannon Entropy is a measure of how much 'missing information' there is in a probability distribution
- In other words - how much we do not know, or how much we would learn from resolving the uncertainty
- For a random variable X that takes the value x_i with probability $p(x_i)$ for $i = 1...n$, defined as

$$\begin{aligned} H(X) &= E(-\ln(p(x_i))) \\ &= -\sum_i p(x_i)\ln(p_i) \end{aligned}$$

Shannon Entropy



- Can think of it as how much we learn from result of experiment

Justification for Shannon Entropy

- Say we want our measure of entropy to have the following features
- Depends only on the probability distribution
  - $H(X) = H(p)$

Justification for Shannon Entropy

- Say we want our measure of entropy to have the following features
- Depends only on the probability distribution
- Maximized at a uniform probability distribution
  - $\max_{p \in \Delta^M} H(p) = H\left(\left\{\frac{1}{M}, \frac{1}{M}, ..., \frac{1}{M}\right\}\right)$

Justification for Shannon Entropy

- Say we want our measure of entropy to have the following features
- Depends only on the probability distribution
- Maximized at a uniform probability distribution
- Unaffected by adding zero probability state
    - $H(\{p_1....p_M\}) = H(\{p_1....p_M, 0\})$

Justification for Shannon Entropy

- Say we want our measure of entropy to have the following features
- Depends only on the probability distribution
- Maximized at a uniform probability distribution
- Unaffected by adding zero probability state
- Additive
    - $H(X, Y) = H(X) + \sum_x p(x)H(Y|x)$
    - (Most 'controversial' - other entropies relax this assumption)

Justification for Shannon Entropy

- Say we want our measure of entropy to have the following features
- Depends only on the probability distribution
- Maximized at a uniform probability distribution
- Unaffected by adding zero probability state
- Additive
- Then Entropy must be of the form (Khinchin 1957)

$$H(X) = -k \sum_i p(x_i) \ln(p_i)$$

- Related to the notion of entropy is the notion of Mutual Information

$$I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- Measure of how much information one variable tells you about another
- Note that $I(X,Y) = 0$ if X and Y are independent

- Note also that mutual information can be rewritten in the following way

$$
\begin{aligned}
I(X,Y) &= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= \sum_y \sum_x p(x,y) \ln P(x|y) - \sum_x \sum_y p(x,y) \ln p(x) \\
&= \sum_y p(y) \sum_x p(x|y) \ln P(x|y) - \sum_y p(x) \ln p(x) \\
&= H(X) - E(H(X|Y))
\end{aligned}
$$

- Difference between entropy of X and the expected entropy of X once Y is known

Mutual Information and Information Costs

- Mutual Information between prior and posteriors can be used to model information costs

$$
\begin{aligned}
K(\mu, \pi) &= \lambda(H(\mu) - E(H(\gamma)) \\
&= \lambda \begin{pmatrix} \sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) \sum_{\Omega} \gamma(\omega) \ln \gamma(\omega) \\ - \sum_{\Omega} \mu(\omega) \ln \mu(\omega) \end{pmatrix}
\end{aligned}
$$

- Can be justified by information theory
  - Homework

- Key feature: Entropy is strictly *concave*
- So negative of entropy is strictly convex
- Say we choose a signal structure with two posteriors γ and γ'
- It must be that

$$P(\gamma)\gamma + P(\gamma')\gamma' = \mu$$

- so

$$\begin{aligned} P(\gamma)H(\gamma) + P(\gamma')H(\gamma') &< H(P(\gamma)\gamma + p(\gamma')\gamma') \\ &= H(\mu) \end{aligned}$$

- So the cost of 'learning something' is always positive

Solving Rational Inattention Models

- Solving Rational Attention models can be difficult analytically
- General approach - ignore choice of information structure, instead focus on joint distribution of choice variable and state
  - i.e. choose state dependent stochastic choice directly
- Example (Matejka and McKay 2015) - continuous state space, finite action space

Solving Rational Inattention Models

- $\mathcal{P}$ set of all state contingent stochastic choice functions for some state space Ω and set of acts A
- Remember $P(a|\omega)$ is the probability of choosing a in state ω
- Remember that, for $P \in \mathcal{P}$, the mutual information between choices a and objective state ω is given by

$$I(A, \Omega) = H(A) - H(A|\Omega)$$

- Decision problem of agent is to choose $P \in \mathcal{P}$ to maximize

$$\sum_{a \in A} \int_{\omega} u(a(\omega)) P(a|\omega) \mu(d\omega)$$
$$-\lambda \left[\sum_{a \in A} \int_{\omega} P(a|\omega) \ln P(a|\omega) \mu(d\omega) + \sum_{a \in A} P(a) \ln P(a) \right]$$

- Subject to

$$\sum_{a \in A} P(a|\omega) = 1 \text{ Almost surely}$$

- Where $P(a)$ is the unconditional probability of choosing a

$$\sum_{a\in A}\int_{\omega} u(a(\omega))P(a|\omega)\mu(d\omega)$$
$$-\lambda\left[\sum_{a\in A}\int_{\omega} P(a|\omega)\ln P(a|\omega)\mu(d\omega)+\sum_{a\in A}P(a)\ln P(a)\right]$$
$$-\int_{\omega}\rho(\omega)\left[\sum_{a\in A}P(a|\omega)-1\right]\mu(d\omega)$$

- $\rho(\omega)$ Legrangian multiplier on the condition that $\sum_{a\in A}P(a|\omega)=1$
- FOC WRT $P(a|\omega)$ (assuming >0)

$$u(a(\omega))-\rho(\omega)+\lambda[\ln P(a)+1-\ln P(a|\omega)-1]=0$$

- Note that this is a convex problem

Solution

- FOC WRT $P(a|\omega)$ (assuming >0)

$$u(a(\omega)) - \rho(\omega) + \lambda[\ln P(a) + 1 - \ln P(a|\omega) - 1] = 0$$

- Which gives

$$P(a|\omega) = P(a) \exp^{\frac{u(a(\omega)) - \rho(\omega)}{\lambda}}$$

- Plug this into

$$\begin{aligned} \sum_{a \in A} P(a|\omega) &= 1 \\ \Rightarrow \quad \exp^{\frac{\rho(\omega)}{\lambda}} &= \sum_{a \in A} P(a) \exp^{\frac{u(a(\omega))}{\lambda}} \end{aligned}$$

- Which in turn gives...

Comments

$$P(a|\omega) = \frac{P(a)\exp^{\frac{u(a(\omega))}{\lambda}}}{\sum_{a\in A} P(a)\exp^{\frac{u(a(\omega))}{\lambda}}}$$

- Similar in form to logistic random choice
- If alternatives are ex ante identical, this *is* logistic choice
- Otherwise choice probabilities are 'warped' by $P(a)$ - which contains information on the prior value of each option
- As costs go to zero, deterministically pick best option in that state
- As costs go to infinity, deterministically pick the best option ex ante

Comments

- The above is not a complete solution
- Does not solve for $P(a)$
- One can completely characterize solution in closed form if one knows what acts are chosen with positive probability
- In general, not all acts will be chosen (see Matejka and Sims 2010)
- Also, they are only **necessary** not **sufficient** conditions
    - Always satisfied by assuming that only one act will be chosen

- Caplin, Dean and Leahy [2015]
- Let $z(a(\omega))$ be 'normalized utilities'

$$z(a(\omega)) = \exp\left\{\frac{U(a(\omega))}{\lambda}\right\}$$

- $Z_\omega(P)$ be 'unconditional expected utility' in state ω generated by P

$$Z_\omega(P) = \sum_{b \in A} P(b) z(b(\omega))$$

- P is consistent with rational inattention with mutual information costs **if and only if**

$$\sum_{\omega}\left[\frac{\mu(\omega)z(a(\omega))}{Z_{\omega}(P)}\right] \leq 1 \text{ all } a \in A$$

$$\sum_{\omega}\left[\frac{\mu(\omega)z(a(\omega))}{Z_{\omega}(P)}\right] = 1 \text{ all } a \text{ s.t. } P(a) > 0$$

and

$$P(a|\omega) = \frac{P(a)z(a(\omega))}{Z_{\omega}(P)}$$

Necessary and Sufficient Conditions

- P is consistent with rational inattention with mutual information costs **if and only if**

$$\sum_{\omega}\left[\frac{\mu(\omega)z(a(\omega))}{Z_{\omega}(P)}\right] \leq 1 \text{ all } a \in A$$

$$\sum_{\omega}\left[\frac{\mu(\omega)z(a(\omega))}{Z_{\omega}(P)}\right] = 1 \text{ all } a \text{ s.t. } P(a) > 0$$

and

$$P(a|\omega) = \frac{P(a)z(a(\omega))}{Z_{\omega}(P)}$$

1. Identify correct **unconditional** choice probabilities
   - Equality condition for chosen actions
   - Check inequality condition for unchosen actions
     - Those not good enough at prior beliefs
     - Big advantage of necessary and sufficient conditions
2. Read off **conditional** choice probabilities

The Linear Quadratic Gaussian Case

- One case in which this problem becomes more tractable is if the input and output signal are both normal
- The entropy of a normal variable $X \sim N(\mu, \sigma_x^2)$ is given by

$$H(Y) = \frac{1}{2}\ln(2\pi e \sigma_x^2)$$

- If Y and X are both normal, then

$$H(Y|X) = \int_x f(x) \int_y f(y|x) \ln(y|x) d(y) d(x)$$

- As $y|x$ is distributed normally with variance $(1-\rho^2)\sigma_y^2$, this becomes

$$\begin{aligned} H(Y|X) &= \int_x f(x) \frac{1}{2} \ln(2\pi e \sigma_{y|x}^2) d(x) \\ &= \frac{1}{2}\ln(2\pi e(1-\rho^2)\sigma_y^2) \end{aligned}$$

The Linear Quadratic Gaussian Case

- As mutual information is given by

$$
\begin{aligned}
& H(Y) - H(Y|X) \\
= \quad & \frac{1}{2}\ln(2\pi e\sigma_y^2) - \frac{1}{2}\ln(2\pi e(1-\rho^2)\sigma_y^2)
\end{aligned}
$$

- In this case, the mutual information is given by

$$\frac{1}{2}\ln(1-\rho^2)$$

- So information costs depend only on the covariance of the two signals!
- It turns out that joint normality is optimal if the utility function is quadratic in the relationship between the objective and subjective state
  - Choice of variance on some normally distributed error term
- However, note that some papers *assume* normality (this is bad)

A Posterior Based Approach

- Can write the objective function as

$$\sum_{\gamma \in \Gamma(\pi)} P(\gamma) \left(W(\gamma) - \lambda H(\gamma) \right) + \lambda H(\mu)$$

- Where
  - $P(\gamma)$ is the unconditional probability of posterior γ
  - $W(\gamma) = \sum_{\omega \in \Omega} \gamma(\omega) u(a^*(\omega))$ be the expected utility of a^*, optimal choice at posterior γ
  - $H(\gamma)$ is the entropty associated with γ

Implications

- For each posterior we can define the net utility

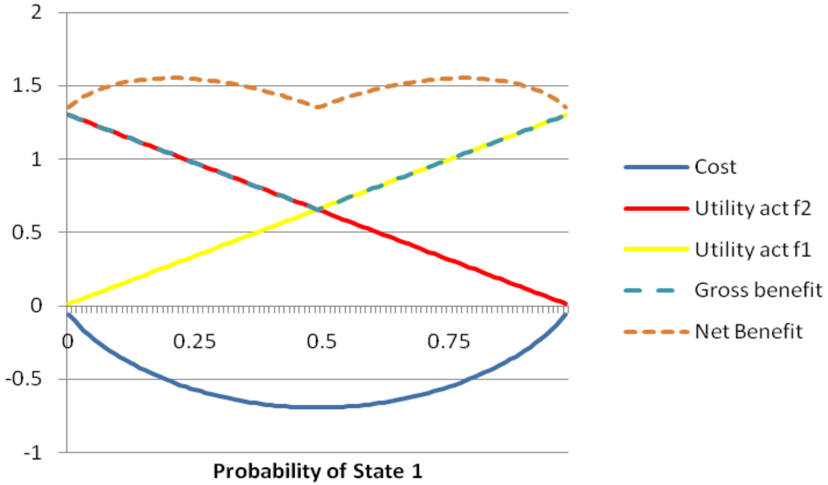
$$N(\gamma) = W(\gamma) - \lambda H(\gamma)$$

- Optimal strategy: Choose posteriors to maximize the weighted average of $N(\gamma)$, subject to

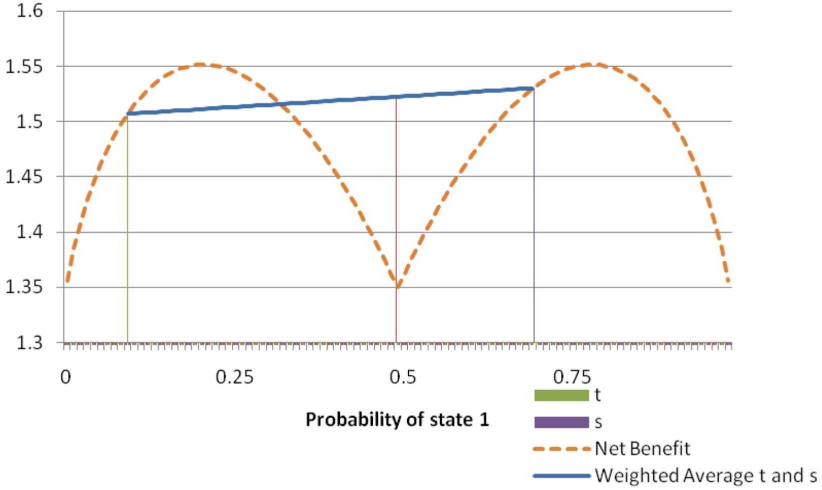
$$\sum_{\gamma \in \Gamma(\pi)} P(\gamma)\gamma = \mu$$

- If same number of posteriors as states this pins down $P(\gamma)$ once posteriors have been chosen

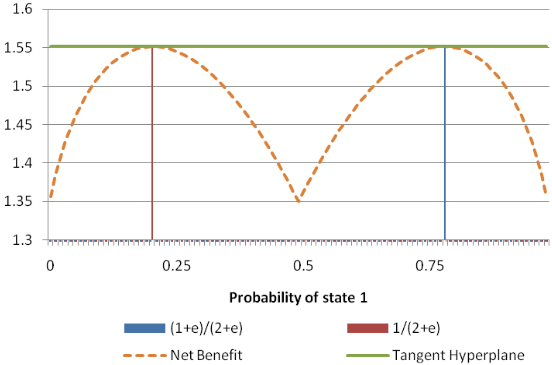
Constructing the Net Utility Function



Value as a Weighted Average of Net Utility



Finding the Optimal Strategy



- Optimal posteriors identified by hyperplane that supports the set of feasible net utilities.

Theorem

Given decision problem $(\mu, A) \in \Gamma \times \mathcal{F}$ a set of posteriors are rationally inattentive if and only if:

1. **Invariant Likelihood Ratio (ILR) Equations for Chosen Acts**: *given $a, b \in B$, and $\omega \in \Omega$,*

$$\frac{\gamma^a(\omega)}{z(a(\omega))} = \frac{\gamma^b(\omega)}{z(b(\omega))}$$

2. **Likelihood Ratio Inequalities for Unchosen Acts**: *given act a chosen with positive probability and $b \in A$,*

$$\sum_{\omega \in \Omega} \left[\frac{\gamma^a(\omega)}{z(a(\omega))} \right] z(b(\omega)) \leq 1.$$

Behavioral Properties

- Locally Invariant Posteriors
- Invariant Likelihood Ratio and Response to Incentives
- Symmetry

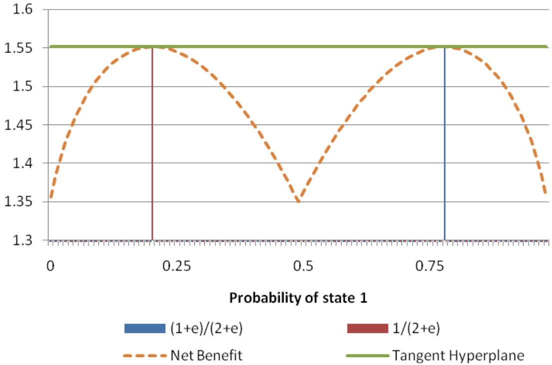
Behavioral Properties

- Locally Invariant Posteriors
- Invariant Likelihood Ratio and Response to Incentives
- Symmetry

- Example: 2 states, 2 actions

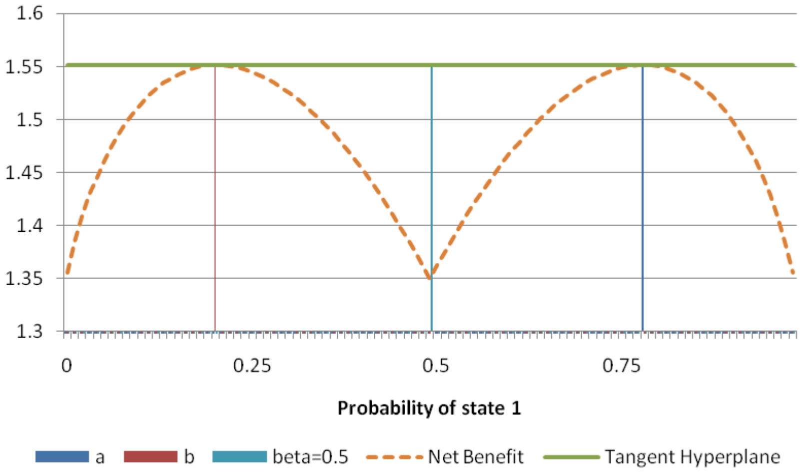
Action	**Payoff in state 1**	**Payoff in state 2**
$\mathbf{f}^1$	x	0
f^2	0	x

Finding the Optimal Strategy

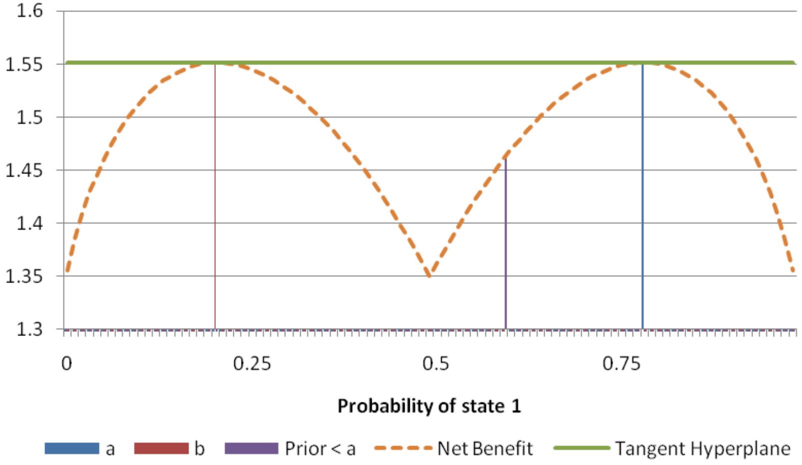


- Optimal posteriors identified by hyperplane that supports the set of feasible net utilities.
- What happens when priors change?

Behavior at 0.5 Prior



Behavior for prior>0.5



Theorem (Locally Invariant Posteriors)

If a set of posteriors $\{\gamma^a\}_{a \in A}$ *are optimal for decision problem* $\{\mu, A\}$ *and are also feasible for* $\{\mu', A\}$ *then they are also optimal for that decision problem*

- Choice probabilities move 'mechanically' with prior to maintain posteriors
- Useful in, for example, models in which consumers are rationally inattentive to quality
  - As the prior distribution of quality changes, posterior beliefs do not
  - See Martin [2014]

- Locally Invariant Posteriors
- Invariant Likelihood Ratio and Response to Incentives
- Symmetry

Invariant Likelihood Ratio and Responses to Incentives

- For chosen actions our condition implies

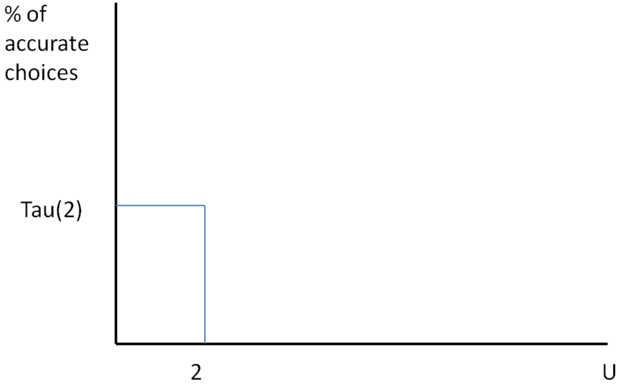
$$\frac{u(a(\omega)) - u(b(\omega))}{\ln \bar{\gamma}^a(\omega) - \ln \bar{\gamma}^b(\omega)} = \lambda$$

- Constrains how DM responds to changes in incentives

Table 1: Experiment 1				
Decision Problem	Payoffs			
	$u(a(1))$	$u(a(2))$	$u(b(1))$	$u(b(2))$
1	2	0	0	2
2	10	0	0	10
3	20	0	0	20
4	30	0	0	30

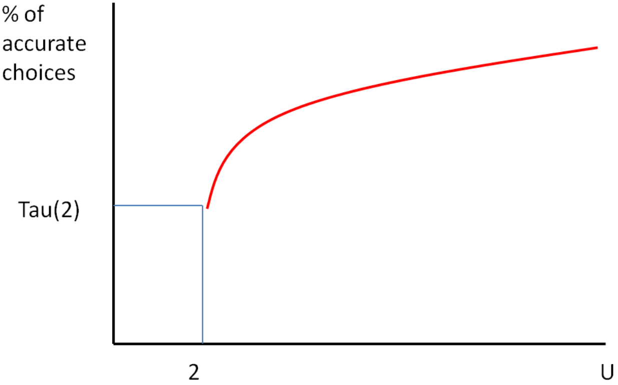
$$\frac{2}{\ln \bar{\gamma}^a(2) - \ln \bar{\gamma}^b(2)} = \frac{10}{\ln \bar{\gamma}^a(10) - \ln \bar{\gamma}^b(10)} = ... = \lambda$$

- One observation pins down λ
- Determines behavior in all other treatments



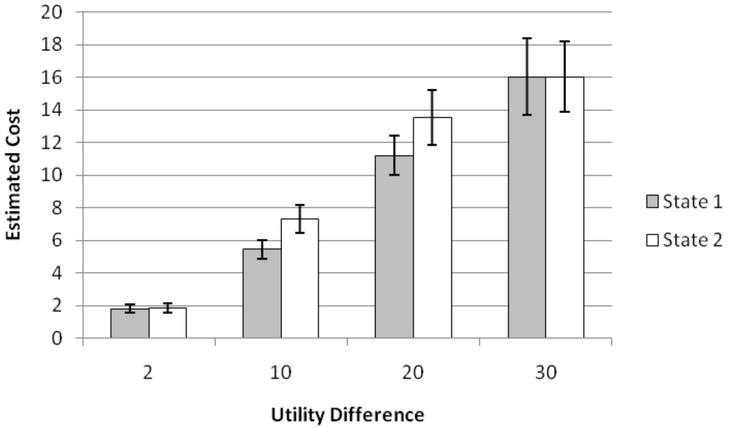
- Observation of choice accuracy for $x = 2$ pins down λ

Invariant Likelihood Ratio - Example



- Implies expansion path for all other values of x
- This does not hold in our experimental data

Invariant Likelihood Ratio - An Experimental Test



Posterior Separable Cost Functions

- Subjects do not respond enough to changes in incentives
- This is not due to curvature of the utility function
- In the paper we introduce a set of cost functions that
  - Maintain structure of Shannon Costs
  - Allow for different response to incentives

Posterior Separable Cost Functions

- Shannon Cost function:

$$K(\pi, \mu) = \lambda \left[-H(\mu) + \sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) H(\gamma) \right].$$

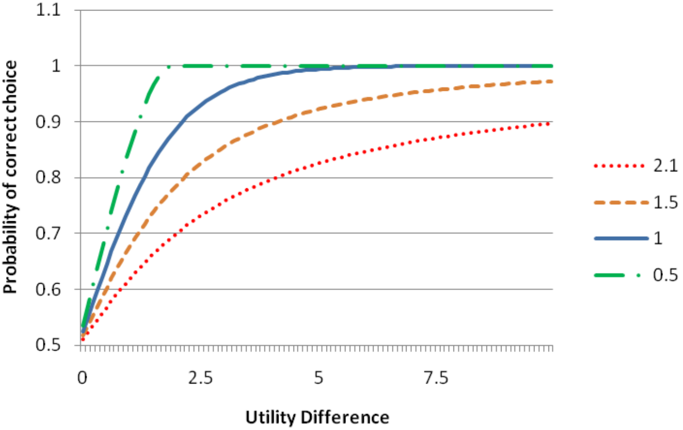
- Posterior- Separable cost functions:

$$K(\pi, \mu) = \lambda \left[-L(\mu) + \sum_{\gamma \in \Gamma(\pi)} \pi(\gamma) L(\gamma) \right].$$

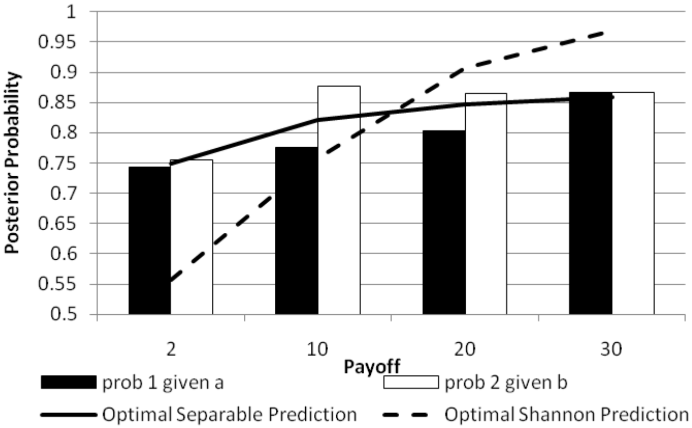
- where

$$L_{\{\rho,\lambda\}}(\gamma) = \begin{cases} -\lambda \left(\sum_{\Omega} \gamma(\omega) \left[\frac{\gamma(\omega)^{1-\rho}}{(\rho-1)(\rho-2)} \right] \right) & \text{if } \rho \neq 1 \text{ and } \rho \neq 2; \\ -\lambda \left(\sum_{\Omega} \gamma(\omega) \ln \gamma(\omega) \right) & \text{if } \rho = 1. \\ -\lambda \left(\sum_{\Omega} \gamma(\omega) \frac{\ln \gamma(\omega)}{\gamma(\omega)} \right) & \text{if } \rho = 2. \end{cases},$$

Response to Incentives: Posterior Separable Cost Functions



Fitting the Data



Behavioral Properties

- Locally Invariant Posteriors
- Invariant Likelihood Ratio and Response to Incentives
- Symmetry

- Shannon Mutual Information has the property of **symmetry**
- Behavior invariant to the labelling of states

$$\frac{u(a(\omega)) - u(b(\omega))}{\ln \bar{\gamma}^a(\omega) - \ln \bar{\gamma}^b(\omega)} = \lambda$$

- Optimal beliefs depend **only** on the relative value of actions in that state
- Implies that there is no concept of 'perceptual distance'

A Simple Example

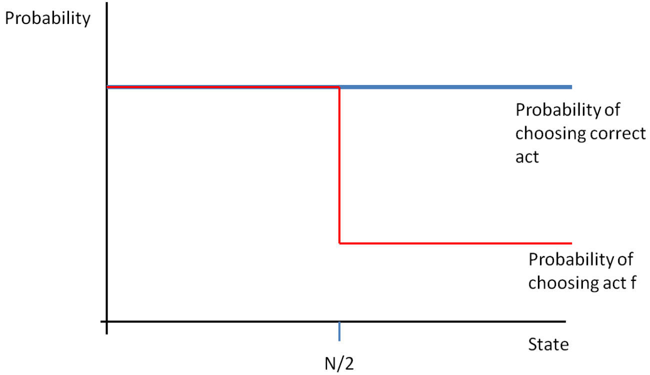
- N equally likely **states of the world** $\{1, 2....., N\}$
- Two **actions**

	Payoffs	
States	$1, \ldots \frac{N}{2}$	$\frac{N}{2} + 1, .., N$
action f	10	0
action g	0	10

- Mutual Information predicts a *quantized* information structure
  - Optimal information structure has **2 signals**
  - Probability of making correct choice is **independent of state**

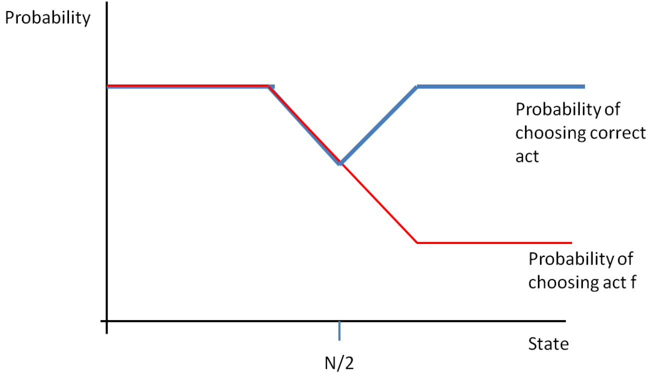
$$\frac{\exp\left(\frac{u(10)}{\lambda}\right)}{1 + \exp\left(\frac{u(10)}{\lambda}\right)}$$

Predictions for the Simple Problem - Shannon



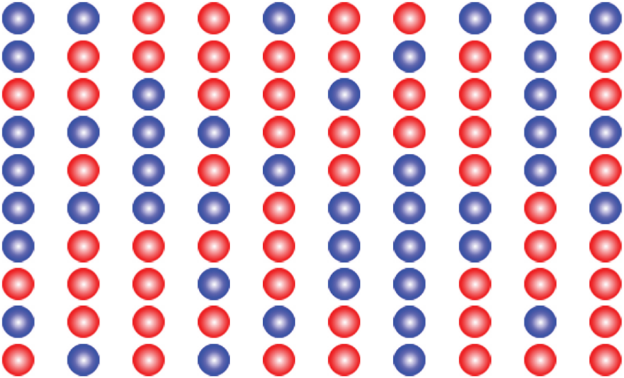
- Probability of correct choice does not go down near threshold

Predictions for the Simple Problem - Shannon



- Not true of other information structures (e.g. uniform signals)

- Shannon Model makes strong predictions for the simple problem
  - Accuracy not affected by closeness to threshold
  - In contrast to (e.g.) uniform signals
- Which model is correct?
  - It may depend on the **perceptual environment**
- Test prediction in two different environments



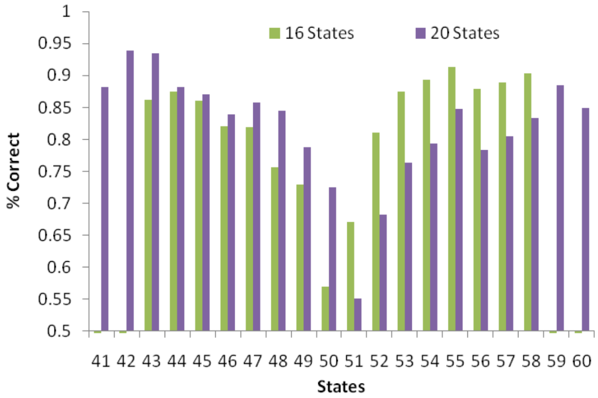
Action	Payoff $\leq$ 50 Red	Payoff $>$ 50 Red
f	10	0
g	0	10

Environment 2 (Letters)

J	P	P	J	J	L
P	N	K	N	K	M
J	Q	M	O	L	O
O	M	L	N	Q	J
Q	K	J			

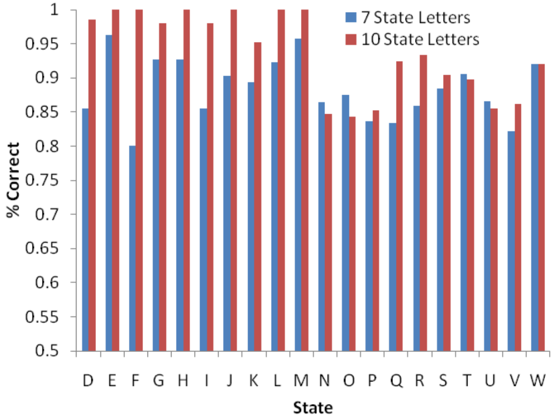
Action	Payoff state letter $<$ N	Payoff state letter $\geq$ N
f	10	0
g	0	10

- 2 treatments
- ‘Balls’ Experiment
  - 23 subjects
  - Vary the number of states
- ‘Letters’ Experiment
  - 24 subjects
  - Vary the relative frequency of the state letter
- Test whether probability of correct choice is lower nearer the threshold



- Probability of correct choice significantly correlated with distance from threshold ($p<0.001$)

Letters Experiment



- Probability of correct choice does vary between states
- But is not correlated with distance from threshold ($p=0.694$)