NBER WORKING PAPER SERIES

BEHAVIORAL IMPLICATIONS OF RATIONAL INATTENTION WITH SHANNON ENTROPY

Andrew Caplin
Mark Dean

Working Paper 19318
http://www.nber.org/papers/w19318

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2013

We thank Dirk Bergemann, Xavier Gabaix, Sen Geng, Andrei Gomberg, John Leahy, Daniel Martin, Filip Matejka, Alisdair McKay, Stephen Morris, and Michael Woodford for their constructive contributions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Behavioral Implications of Rational Inattention with Shannon Entropy
Andrew Caplin and Mark Dean
NBER Working Paper No. 19318
August 2013
JEL No. D83

ABSTRACT

The model of rational inattention with Shannon mutual information costs is increasingly ubiquitous. We introduce a new solution method that lays bare the general behavioral properties of this model and liberates development of alternative models. We experimentally test a key behavioral property characterizing the elasticity of choice mistakes with respect to attentional incentives. We find that subjects are less responsive to such changes than the model implies. We introduce generalized entropy cost functions that better match this feature of the data and that retain key simplifying features of the Shannon model.

Andrew Caplin
Department of Economics
New York University
19 W. 4th Street, 6th Floor
New York, NY 10012
and NBER
andrew.caplin@nyu.edu

Mark Dean
Department of Economics
New York University
19 W. 4th Street, 6th Floor
New York, NY 10012
mark.dean@nyu.edu

1 Introduction

The model of rational inattention with Shannon mutual information costs (henceforth the Shannon model) is increasingly ubiquitous. While the pioneering contributions of Sims [1998, 2003] considered its implications for macroeconomic dynamics, the ensuing period has seen applications to such diverse subjects as stochastic choice (Matejka and McKay [2011, 2013]), investment decisions (e.g van Nieuwerburgh and Veldkamp [2008]), global games (Yang [2011]), and pricing decisions (Mackowiak and Wiederholt [2009], Matejka [2010] Martin [2013]). Yet despite its growing importance, little of a general nature is known about the behaviors the Shannon model produces. Even solving the model can be challenging absent additional restrictions on the signal space and/or the utility function (Sims [2006]).

We introduce a new approach to solving the Shannon model that lays bare its behavioral properties and liberates development of alternative models. Our "posterior-based" approach explicitly models the decision maker's choice of signals, as characterized by the resulting posterior distribution over states of the world. The resulting necessary and sufficient conditions for rationality identify behavioral patterns that characterize the Shannon model.[1]

We establish two defining behavioral properties of the Shannon model. The first property relates to changes in incentives. We show that the ratio of the difference in utilities across chosen acts to the log difference in posterior beliefs is constant across states and decision problems. This "invariant likelihood ratio" (ILR) property pins down the rate at which choice mistakes respond to changes in the cost of these mistakes. The second property relates to changes in prior beliefs. We show that posterior beliefs are invariant to local changes in prior beliefs - the "locally invariant posteriors" (LIP) property.

[1] In general, rational inattention problems have been approached by considering the problem of directly choosing the optimal joint distribution of state and chosen act. One exception is Matejka and McKay [2011], who consider the choice of posteriors but do not use the net utility approach. The net utility approach has been used in a related setting by Kamenica and Gentzkow [2011].

We experimentally test the ILR property characterizing the elasticity of choice mistakes with respect to attentional incentives. We find that subjects are less responsive to such changes than the model implies. We introduce generalized entropy cost functions that better match this feature of the data while retaining key simplifying features of the Shannon model.

In addition to the two key invariants, the posterior-based approach enables us to analyze many other properties of the Shannon model. We show that an optimal strategy exists involving no more acts being taken than there are states of the world, that collections of chosen acts relate uniquely to posterior beliefs, and that an envelope condition characterizes dependence of the value of the optimal strategy on model parameters. We identify also conditions under which the model has a unique solution.

We adapt the design of Caplin and Dean [2013] (henceforth CD13) to conduct our experimental test of the ILR property. Subjects are shown a display with a number of red and blue balls, with the true state determined by the number of red balls. They are then asked to choose between acts with state dependent payoffs. We observe how subjects' patterns of choice vary with the value of choosing the correct act. We compare this experimental variation to that predicted by the Shannon model, as embodied in the ILR condition. We find that subjects are generally less responsive to changes in incentives than the Shannon model implies. To accommodate our findings, we generalize the model to a broader class of "posterior-separable" attention cost functions. We show that this class allows a better fit to the experimental data, while retaining key simplifying features of the Shannon model.

Section 2 introduces the posterior-based approach and uses it to solve the Shannon model. Section 3 derives the two behavioral invariants, with other theoretical results in section 4. Section 5 introduces our experimental design and results. Section 6 introduces posterior-separable models. Section 7 concludes. Our approach is of particular value in dynamic settings in which beliefs evolve due to the interaction between attentional effort and exogenous shocks. It is equally of value in strategic settings (e.g. Martin [2013]).

2 The Posterior-Based Approach

2.1 The Decision Problem

A decision making environment comprises possible states of the world $\Omega = \{1, , m, , M\}$ and a set F of acts with state dependent payoffs U_m^f for $f \in F$ and $m \in \Omega$. We define $\Gamma = \Delta(\Omega)$ as the set of probability distributions over states, with γ_m indicating the probability of state m given $\gamma \in \Gamma$. The state of the world is assumed to be knowable in principle, but obtaining (or processing) information is taken to be costly.

A decision problem consists of a prior distribution over these states of the world and the subset of acts from which the decision maker (DM) must choose.[2] We reserve the special notation $\mu \in \Gamma$ for prior beliefs, with Ω^μ the corresponding set of possible states of the world. To ensure that maximization problems have a solution, we assume that the set of feasible utility vectors is closed and bounded above.

Definition 1 *A* ***decision problem*** *comprises a pair* $(\mu, A) \in \Gamma \times \mathcal{F}$*, where* $\mathcal{F} \subset 2^F/\emptyset$ *comprises all non-empty sets* $A \subset F$ *such that* $\{U^f \in \mathbb{R}^M | f \in A\}$ *is closed and bounded above.*

2.2 Attention Strategies and Shannon Costs

Prior to choosing an act, the DM must choose an *attention strategy* detailing their method of learning about the state of the world. As is standard in the rational inattention literature, we model the DM as choosing a fixed attention strategy for each decision problem that (stochastically) maps states of the world to a set of signals. Since we will be characterizing an expected utility maximizing agent, we identify these signals with the posteriors $\gamma \in \Gamma$ which they produce. For simplicity, we consider only attention strategies with finitely

[2] Both of which are assumed known to the DM.

many possible posteriors. Since we are assuming a finite state space, this restriction is immaterial in the current context, as we discuss in section 4.2.

An attention strategy is identified with a function $\pi : \Omega^{\mu} \rightarrow \Delta(\Gamma)$ that specifies the probability of receiving each signal in each state of the world.[3] Feasible attention strategies are constrained to satisfy Bayes' law. Given $\mu \in \Gamma$, any corresponding attention strategy $\pi : \Omega^{\mu} \rightarrow \Delta(\Gamma)$ has (finite) image $\Gamma(\pi) \subset \Gamma$ and must be such that, for all $m \in \Omega^{\mu}$ and $\gamma \in \Gamma(\pi)$,

$$\gamma_m = \frac{\mu_m \pi_m(\gamma)}{\sum\limits_{j=1}^{M} \mu_j \pi_j(\gamma)},$$

where $\pi_m(\gamma) \equiv \pi(m)(\{\gamma\})$ is the probability of receiving the signal that gives posterior beliefs γ in objective state m.

We assume that information is costly, with costs measured in the same expected utility units in which the prizes are measured. Furthermore, we assume that the cost of an attentional strategy is linearly related to Shannon's measure of the mutual information between $\mu \in \Gamma$ and π,

$$I(\mu, \pi) = \sum_{\gamma \in \Gamma(\pi)} \sum_{m=1}^{M} \mu_m \pi_m(\gamma) \ln \left(\frac{\mu_m \pi_m(\gamma)}{\mu_m \left[\sum_{m=1}^{M} \mu_m \pi_m(\gamma) \right]} \right).$$

The Shannon mutual information cost function has been heavily used in the rational inattention literature since it was popularized by Sims [1998]. Such use has been justified both axiomatically and through links to optimal coding in information theory (see Sims [2010] and Matejka and McKay [2013] for discussions).

[3] We use $\Delta(\Gamma)$ to denote the set of all simple probability distributions on Γ.

2.3 The Posterior-Based Approach and Net Utility

The Shannon model assumes that a rationally inattentive DM selects an attention strategy which maximizes expected utility from the subsequent choice of acts net of mutual information attention costs. The most commonly used approach to solving such models is to focus directly on the DM's probability of choosing each act $f \in A$ in each state $m \in \Omega$ (see for example Sims [2006], Matejka and McKay [2013]). The attention strategy can then be mechanically computed from the joint distribution of states and acts. In our approach this procedure is reversed. We solve the model by focusing directly on choice of posterior beliefs, deriving the state dependent choice probabilities from this solution.

To develop the posterior-based method, we first reformulate the optimization problem as in Matejka and McKay [2011]. A feature of the Shannon cost function is that strictly more informative strategies are strictly more attentionally expensive than less informative such strategies. Hence optimization is inconsistent with choice of the same act at two distinct posteriors. This implies that, for purposes of optimization, an attention strategy can be specified as a subset of available acts $B \subset A$ chosen with strictly positive unconditional probabilities $P^f > 0$, and corresponding act-specific posteriors, $\gamma^f \in \Gamma$. The posteriors and probabilities must satisfy Bayes' law with respect to the prior beliefs.

Definition 2 *Given $(\mu, A) \in \Gamma \times \mathcal{F}$, a (posterior-based)* ***attention strategy*** *$(B, P, \gamma) \in \Lambda^{(\mu,A)}$ comprises a finite set $B \subset A$, probability weights $P : B \to \mathbb{R}_{++}$ with $\sum_{f \in B} P^f = 1$, and posteriors $\gamma : B \to \Gamma$ such that,*

$$\mu = \sum_{f \in B} P^f \gamma^f. \tag{1}$$

A standard result is that the mutual information of two random variables X and Y can be written as the difference between the Shannon entropy of X, and the expected Shannon entropy of X conditional on Y.[4] Hence the Shannon attention cost function with parameter $\kappa > 0$ is defined on $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ by,

$$S^{\kappa}(B, P, \gamma) = \kappa \left[H(\mu) - \sum_{f \in B} P^f H(\gamma^f) \right], \tag{2}$$

where $H(\mu) = -\sum_{j=1}^{M} \mu_j \ln \mu_j$ is the Shannon entropy function extended to boundary points using the limit condition $\lim_{\gamma \searrow 0} \gamma \ln \gamma = 0$.

Specifying information costs this way allows us to write the DM's optimization problem in a particularly convenient way. As the gross benefit associated with an attention strategy (B, P, γ) is given by $\sum_{f \in B} P^f \sum_{j=1}^{M} \gamma_m^f U_m^f$, the objective function of the DM facing decision problem (μ, A) can be defined in terms of the "net utility" associated with each posterior belief/act pair,

$$N^{(\mu,A)}(B, P, \gamma) \equiv \sum_{f \in B} P^f \sum_{j=1}^{M} \gamma_m^f U_m^f - S^k(B, P, \gamma) = \sum_{f \in B} P^f N^f(\gamma^f) - \kappa H(\mu), \tag{3}$$

where $N^f : \Gamma \to \mathbb{R}$ is the net utility of act f,

$$N^f(\gamma^f) \equiv \sum_{j=1}^{M} \gamma_m^f U_m^f + \kappa H(\gamma^f).$$

Each posterior belief/act pair has a net utility associated with it, which is equal to the expected utility of using that act at that posterior minus the information cost associated with that posterior. Since the term $H(\mu)$ is independent of the attention strategy, we can ignore it, and characterize the DM as maximizing the weighted average of act-specific net utilities.

It is convenient to treat the posterior probability of state M as the residual, and consider $X = \{(\mu_1, ..\mu_{M-1}) \in \mathbb{R}_+^{M-1} | \sum_{m=1}^{M-1} \mu_m \leq 1\}$ as the domain of the

[4] See for example Cover and Thomas [2006] p. 20.

net utility functions.[5] Having done so, a simple geometric construction illustrates computation of net utilities. Figure 1 illustrates for a decision problem with $M = 2$, $A = \{f, g\}$, $U_1^f = U_2^g = \ln(1 + e)$, $U_2^f = U_1^g = 0$, and $\kappa = 1$. The horizontal axis represents the probability $\gamma_1 \in [0, 1]$ of state 1.

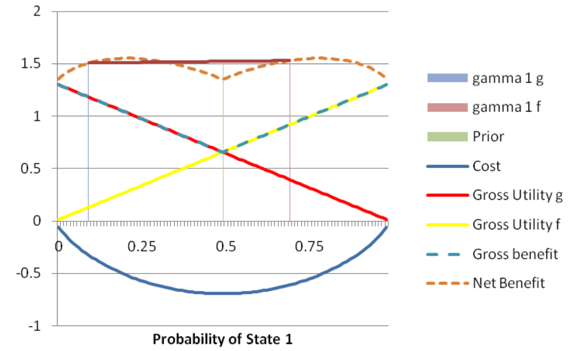

Figure 1: Net Utility Functions

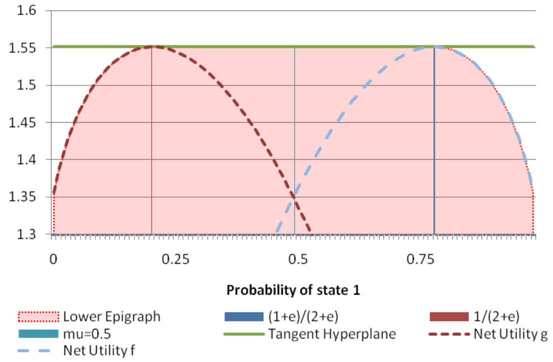

Figure 2: Optimal Posteriors

Figure 1 enables the value of all attention strategies to be visualized. For any given prior $\mu_1 \in [0, 1]$, a feasible attention strategy in which both acts are taken with positive probability corresponds to posteriors γ_1^f, γ_1^g that contain the prior interior to their convex hull. Assuming that acts are chosen optimally given posterior beliefs this requires $\gamma_1^f > \mu_1$ and $\gamma_1^g < \mu_1$. Given two such posteriors, the utility of the corresponding attention strategy is the weighted average of the net utility associated with each posterior, with the expectation taken using the act-specific probabilities. Given the choice of posteriors, the act specific probabilities are pinned down by condition 1. Geometrically, this means that the value of an attention strategy consisting of posteriors γ_1^f, γ_1^g

[5] Hence,

$$N^f(\gamma_1^f, ..\gamma_{M-1}^f) = \sum_{j=1}^{M-1} \gamma_m^f U_m^f + \left(1 - \sum_{m=1}^{M-1} \gamma_m\right) U_M^f + \kappa H \left(\gamma_1^f, ..\gamma_{M-1}^f, 1 - \sum_{m=1}^{M-1} \gamma_m\right).$$

is equal to the height of the chord connecting $\left(N^f(\gamma_1^f), \gamma_1^f\right)$ and $\left(N^f(\gamma_1^g), \gamma_1^g\right)$ as it goes over the prior, as demonstrated in Figure 1.

2.4 Solving the Shannon Model

In this section we characterize rationally inattentive behavior for the Shannon model.

Definition 3 *Given $(\mu, A) \in \Gamma \times \mathcal{F}$, strategy $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ is* ***rationally inattentive*** *if,*

$$N^{(\mu,A)}(B, P, \gamma) \geq N^{(\mu,A)}(B', P', \gamma'),$$

all $(B', P', \gamma') \in \Lambda^{(\mu,A)}$, with $\hat{\Lambda}^{(\mu,A)}$ the corresponding set of rationally inattentive strategies.

The geometric approach to computing the payoff to attention strategies suggests that rationally inattentive strategies are defined by the posteriors and acts whose associated net utility functions support the highest chord above the prior. In order to identify such posteriors, one can concavify the upper envelope of the net utility functions by finding the minimal concave function that majorizes them all, a construction familiar from the work of Kamenica and Gentzkow [2011] and others (see section 2.4.1 for further details). The optimal attention strategy for any prior is given by the posteriors that support this concavified net utility function above that prior.

Figure 2 illustrates this concavification operation for our running example in the case where $\mu_1 = 0.5$. The shaded region is the lower epigraph of the concavified version of the maximum net utility function. This epigraph is a closed, convex set that is bounded above in the net value coordinate. This implies that a rationally inattentive strategy has associated with it a hyperplane in $\mathbb{R}^M$ which supports it (also illustrated in figure 2). The net utility functions associated with all acts lie weakly below this hyperplane, while those of chosen acts touch the hyperplane at their associated posterior beliefs. Lemmas

1 and 2 in the appendix demonstrate that both of these results generalize to arbitrary decision problems.

The Shannon mutual information cost function has many features which make the hyperplane characterization particularly powerful. First, it is differentiable. Second, it effectively rules out corner solutions (as the marginal cost of information goes to infinity as a posterior belief approaches 0). These two conditions enable us to provide a derivative-based characterization of the supporting hyperplane (lemma 3 in the appendix). Finally, its functional form provides easily tractable tangency conditions. Theorem 1 uses these features to characterize rationally inattentive behavior in terms of suitably transformed utility parameters,

$$\nu_m^f \equiv \exp[\frac{U_m^f}{\kappa}].$$

Theorem 1 *Given* $(\mu, A) \in \Gamma \times \mathcal{F}$ *and* $\kappa > 0$, $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ *is rationally inattentive if and only if:*

1. ***Invariant Likelihood Ratio (ILR) Equations for Chosen Acts:*** *given* $f, g \in B$, *and* $1 \leq m \leq M$,

$$\frac{\gamma_m^f}{\nu_m^f} = \frac{\gamma_m^g}{\nu_m^g}.$$

2. ***Likelihood Ratio Inequalities for Unchosen Acts:*** *given* $f \in B$ *and* $g \in A \backslash B$,

$$\sum_{m=1}^{M} \left[\frac{\gamma_m^f}{\nu_m^f} \right] \nu_m^g \leq 1.$$

Proof. *Appendix.* ■

The first of these properties derives from the fact that the net utility functions of the chosen acts must support the same hyperplane at their associated posteriors. The second derives from the fact that the net utility functions of all unchosen acts must lie weakly below this hyperplane.

These equations can, in many cases, be used directly to solve for the optimal attention strategy. For example, the main behavioral equation in Yang [2011] (equation 10) corresponds to the analogous condition for continuous state spaces. The equations also highlight a number of important general properties of such solutions, as we discuss in sections 3 and 4.

2.4.1 Comparison to Other Approaches

The most commonly used approach to solving models of rational attention is to directly focus on the decision maker's state dependent stochastic choice function (see for example Sims [2003], [2006]). This method takes as the control variable the DM's probability of choosing each act $f \in A$ in state $m \in \Omega$, with information costs based on the mutual information between the distribution of chosen acts and prior distribution over states of the world. While this formulation is equivalent to ours, there are insights into behavior that are easier to identify based on choice of posterior. We discuss these insights in the next two sections.

Another approach, taken by Matejka and McKay [2011] is to focus on choice of posterior beliefs, but to attack the problem directly using the Karush-Kuhn-Tucker (KKT) conditions, rather than using concavified net utility and the separating hyperplane theorem. While this approach results in somewhat similar conditions, it turns out that, because the problem is not convex, the KKT conditions are necessary, but not sufficient for optimality (we provide a demonstration of the lack of sufficiency in the online appendix).

The approach we take shares many technical similarities with that taken by Kamenica and Gentkow [2011] (henceforth KG) to solve their model of optimal persuasion. KG consider a situation in which a sender chooses a signal structure with which to convey information about the true state of the world to a receiver, who then chooses an action to take. In essence, the sender chooses a set of posterior beliefs in order to maximize their expected "net utility", just as they do in our formulation of the rational inattention model. The difference is that, while in the rational inattention model the DM is constrained by the cost of information, in the KG model they are constrained by the fact that the receiver will choose an action at each posterior in order to maximize their own utility rather than that of the sender. The key formal connection is that the same concavification operation is used in both cases to identify optimal strategies. This suggests that certain of the insights and techniques we develop may be of value in models of optimal persuasion.

3 Two Behavioral Invariants

Theorem 1 highlights two important behavioral regularities associated with the Shannon model. These go far beyond the conditions that characterize the general class of rationally inattentive models (CD13).[6] They involve invariance properties of the model solution to changes in the underlying decision problem, in particular the incentives for attention and prior beliefs.

3.1 Invariant Likelihood Ratio and the Cost Elasticity of Mistakes

The ILR condition of theorem 1 embodies an invariance condition with respect to changes in the state dependent utilities associated with chosen acts. In any decision problem, for any two acts that are chosen with positive probability in some state, the ratio of the relative utility of those two acts in that state to the log ratio of the posterior probability of that state when each act is chosen is fixed, and equal to the cost of information. Thus, as the difference in utilities between two acts in some state changes, so the relative probabilities of choosing those acts must also change in a constrained way.

[6] A "No Improving Action Switch (NIAS)" condition which ensures that chosen acts are optimal at each posterior and a "No Improving Attention Cycle (NIAC)" condition which ensures that choice of attention strategy can be rationalized by some cost function.

We illustrate the power of this property by demonstrating how it pins down the response of choice "mistakes" to changes in incentives. From the point of view of a fully informed observer, a rationally inattentive DM may make mistakes: they will sometimes choose an action which is suboptimal given the true state of the world (though is optimal given their posterior beliefs). The ILR condition specifies how the probability of such mistakes responds to changes in incentives. To illustrate, consider symmetric decision problems with two states, two acts $\{f, g\}$ and $\mu = (0.5, 0.5)$ such that,

$$U_1^f = U_2^g = U_2^f + c = U_1^g + c,$$

with $c > 0$. The superior choice is therefore f in state 1 and g in state 2, while c parameterizes the cost of making the wrong choice. Direct application of the ILR equation for chosen acts shows that the probability of choosing the correct action in each state is related to the cost of a mistake in that state in a simple manner,

$$\pi_1(f) = \gamma_1^f = \frac{\exp\frac{U_1^f}{\kappa}}{\exp\frac{U_1^f}{\kappa} + \exp\frac{U_1^g}{\kappa}} = \frac{1}{1 + \exp(\frac{-c}{\kappa})} = \gamma_2^g = \pi_2(g).$$

This enables us to compute the elasticity of errors with respect to the cost of mistakes as,

$$\frac{\partial \pi_1(f)}{\partial c}\frac{c}{\pi_1(f)} = \frac{\frac{c}{\kappa}\exp\left(\frac{-c}{\kappa}\right)}{1 + \exp(\frac{-c}{\kappa})}.$$

We test this property explicitly in section 5

3.2 Locally Invariant Posteriors

We now consider how optimal attention strategies respond to changes in prior beliefs - an issue of interest in dynamic and in strategic applications of the Shannon model [see for example Martin [2013]]. Figure 2 suggests a powerful invariance condition - that optimal posterior distributions are locally invariant to changes in prior beliefs. As per lemma 2, we identify the optimal behavior for prior $\mu_1 = 0.5$ by identifying the posterior beliefs that support that tangent hyperplane to the lower epigraph of the concavified net utility function above that prior. However figure 2 makes clear these posteriors also support the tangent hyperplane above all priors in the range $[\gamma_1^1, \gamma_1^2]$. It follows that the optimal strategy for priors in this range uses the same posterior beliefs as for $\mu_1 = 0.5$.

This highlights a general result: if a set of acts B and act specific posteriors $\left\{\gamma^f\right\}_{f\in B}$ form the basis for an optimal strategy for some decision problem (μ, A), then they also form the basis for an optimal strategy for any decision problem $(\rho\ A)$ from which it is feasible to use these posteriors - that is for every prior belief that is in the convex hull of the posterior beliefs $\left\{\gamma^f\right\}_{f\in B}$. This and all other corollaries are proved in the online appendix.

Corollary 1 (Locally Invariant Posteriors - LIP): If $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $(C, Q, \eta) \in \Lambda^{(\rho,A)}$ with $C \subset B$ satisfying $\eta^f = \gamma^f$ all $f \in C$, then $(C, Q, \eta) \in \hat{\Lambda}^{(\rho,A)}$.

Corollary 1 has important implications for solving rational inattention models: the solution to one decision problem identifies a solution to many related problems. It also has comparative static implications for how the unconditional probability of choosing a given act must change with local changes in the prior belief. With posterior beliefs unchanged, the unconditional probabilities of choosing each act must change mechanically in order to ensure that Bayes' rule is obeyed. Consider for example the two act, two state, case in figure 2 with prior $\mu_1 \in (\gamma_1^f, \gamma_1^g)$. In this case we can use Bayes' rule to explicitly solve for $P^f(\mu_1)$ as a function of $\mu_{1,}$ γ_1^f, and γ_1^g, and establish that the local response of P^f to changes in prior belief depends only on the difference $\left(\gamma_1^f - \gamma_1^g\right)$,

$$P^f(\mu_1) = \frac{\mu_1 - \gamma_1^g}{\gamma_1^f - \gamma_1^g} \Longrightarrow \frac{\partial P^f}{\partial \mu_1} = \frac{1}{\gamma_1^f - \gamma_1^g}.$$

Thus, once posterior beliefs have been observed, the response of act choice probabilities to local changes in the prior can be calculated without recourse to any further details of the model.

4 Further Implications

We highlight additional features of the Shannon model that are revealed by the posterior-based approach which are of use for understanding the Shannon model.

4.1 The Envelope Condition

Our approach allows us to use the envelope condition to characterize the effect of changing prior beliefs on expected utility. Given $A \in \mathcal{F}$, define $V^A : \Gamma \longrightarrow \mathbb{R}$ to be the maximal value obtainable for decision problem (μ, A),

$$V^A(\mu) = \max_{(B,P,\gamma)\in\Lambda^{(\mu,A)}} N^{(\mu,A)}(B, P, \gamma).$$

It is a direct corollary of theorem 1 that the value function is differentiable and that a version of the standard envelope theorem characterizes local changes in value with respect to changes in prior beliefs.

Corollary 2 (Envelope Condition): Given $(\mu, A) \in \Gamma \times \mathcal{F}$ such that $\mu_m > 0$, the value function $V^A : \Gamma \longrightarrow \mathbb{R}$ is differentiable at μ and has continuous partial derivatives,

$$\frac{\partial V^A(\mu)}{\partial \mu_m} = \frac{\partial N^f}{\partial \gamma_m}(\hat{\gamma}^f),$$

where $f \in \hat{B}$ some $(\hat{B}, \hat{P}, \hat{\gamma}) \in \hat{\Lambda}^{(\mu,A)}$.

4.2 States Bound Acts

Theorem 1 implies that an optimal attention strategy exists that uses no more acts than there are states of the world. Intuitively, the optimal strategy for a given decision problem can be found by identifying acts whose net utility functions touch the hyperplane that supports the concavified net utility function above the prior. With M states of the world, Charateodory's theorem implies that any such hyperplane is defined by any M points it contains. For example, in the two dimensional case of figure 2, the supporting hyperplane is a line, as defined by any pair of its points. This in turn implies that the hyperplane can be supported by and M act/posterior pairs, which in turn form an optimal strategy.

Corollary 3 (States Bound Acts - SBA): Given $(\mu, A) \in \Gamma \times \mathcal{F}$, there exists a rationally inattentive strategy with $|B| \leq M$.

4.3 Unique Posteriors

Consider decision problems which share the same prior, have different available acts, and yet for which the same acts are optimally chosen. In this case we show in corollary 4 that the posteriors associated with all chosen acts will be identical. In two decision problems with the same prior in which the subjects make use of the same acts, they will have the same posteriors.

Corollary 4 (Unique Posteriors): If $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $(B, Q, \rho) \in \hat{\Lambda}^{(\mu,C)}$, then $\gamma(f) = \rho(f)$ all $f \in B$.

The Unique Posteriors property tells us that there can be at most one possible set of posteriors that satisfies the ILR conditions for a given prior and a given set of chosen acts.

4.4 Uniqueness

One natural question is whether the Shannon model always makes unique behavioral predictions. The answer is no, yet we can provide conditions under which uniqueness is guaranteed. Consider our running example with $A = \{f, g\}$, $\kappa = 1$, $U_1^f = U_2^g = \ln(1+e)$, and $U_2^f = U_1^g = 0$. Note that,

$$\nu_1^f = \nu_2^g = 1 + e; \nu_2^f = \nu_1^g = 1.$$

By the necessity aspect of theorem 1, both acts can be chosen only if the ILR equation is satisfied, which in this case uniquely pins down the posteriors:

$$\begin{aligned} \gamma^f &= (\gamma_1^f, \gamma_2^f) = (\frac{1+e}{2+e}, \frac{1}{2+e}); \\ \gamma^g &= (\gamma_1^g, \gamma_2^g) = (\frac{1}{2+e}, \frac{1+e}{2+e}); \end{aligned}$$

as already indicated in figure 2.

Now consider adding a third act $h \in F$ with $U_1^h = U_2^h = \frac{\ln(1+e)}{2}$, so that $\nu_1^h = \nu_2^h = \frac{2+e}{2}$. Note that the ILR equations are satisfied for acts f, g, h, at the corresponding posteriors γ^f, γ^g, and $\gamma^h = (0.5, 0.5)$. By the sufficiency aspect of theorem 1, choosing all three acts with equal probability at these posteriors identifies an optimal policy for prior $(\mu_1, \mu_2) = (0.5, 0.5)$, as does choosing act h for sure.

An independence condition rules examples of this kind.

Axiom 1 (Affine Independence) $\{\nu^f \in \mathbb{R}^m | f \in A\}$ *is affinely independent.*

By definition $\{\nu^f \in \mathbb{R}^m | f \in A\}$ is affinely independent if one cannot find scalars α^f, not all zero, such that $\sum_{f \in A} \alpha^f = 0$ and $\sum_{f \in A} \alpha^f \nu_m^f$=0. Affine independence rules out having $M+1$ transformed utility vectors in any hyperplane and ensures that there is one and only one optimal attention strategy.

Theorem 2: With affine independence, $\left|\hat{\Lambda}^{(\mu,A)}\right| = 1$ all $\mu \in \Gamma$.

Theorem 2 has strong implications for data derived from more than one decision problem. By theorem 1, observation of behavior in a single decision problem enables the cost parameter κ to be identified, provided two or more distinct acts are chosen. Theorem 2 implies that this uniquely pins down observed behavior in all other decision problems provided the independence condition is satisfied.

4.5 Global Comparative Statics

One of the major difficulties in solving rational inattention models is identifying which acts will be chosen with positive probability as part of the optimal strategy. As theorem 1 demonstrates, once this is known, it is relatively easy to solve for the associated optimal posteriors and unconditional probabilities associated with each act.

The posterior-based approach offers help in this regard by identifying, for any set $A \in \mathcal{F}$, all subsets of acts that can possibly be chosen together as part of an optimal strategy for some prior beliefs. Because optimal strategies can be characterized by a supporting hyperplane that touches the lower epigraph of the concavified net utility function at the net utility functions of the used acts, for any collection of acts to be used together there must be a tangent hyperplane that is supported by the net utility functions of those acts.

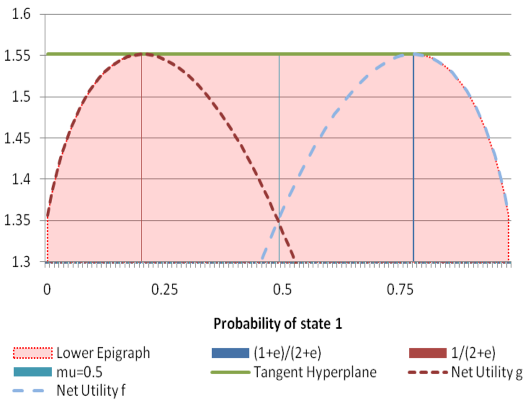

Figure 3: An example with 3 acts

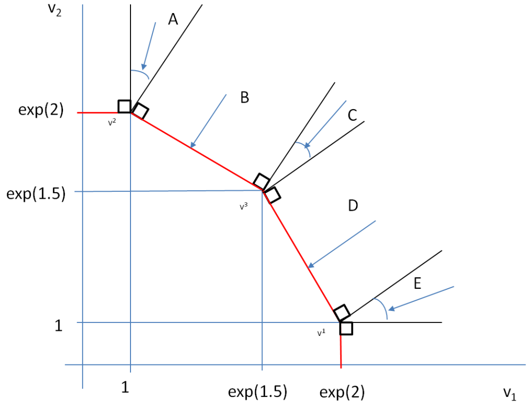

Figure 4: The Dual Cone

Figure 3 illustrates this concept in the case of two states of the world and three acts h_1 (which pays 2 in state 1 and 0 otherwise), h_2 (which pays 2 in state 2 and 0 otherwise) and h_3 (which pays 1.5 in both states). This makes it clear that there are five classes of solution to this problem, depending on prior beliefs.

Table 1

Prior μ_1	Acts Chosen	Slope of hyperplane	Region
$[0, \gamma_1^2]$	$\{h_2\}$	$(\infty, \frac{\partial N^2(\gamma^2)}{\partial \gamma}]$	A
$[\gamma_1^2, \gamma_1^{3a}]$	$\{h_2, h_3\}$	$\frac{\partial N^2(\gamma^2)}{\partial \gamma} = \frac{\partial N^3(\gamma^{3a})}{\partial \gamma}$	B
$[\gamma_1^{3a}, \gamma_1^{3b}]$	$\{h_3\}$	$[\frac{\partial N^3(\gamma^{3a})}{\partial \gamma}, \frac{\partial N^3(\gamma^{3b})}{\partial \gamma}]$	C
$[\gamma_1^{3b}, \gamma_1^1]$	$\{h_1, h_3\}$	$\frac{\partial N^3(\gamma^{3b})}{\partial \gamma} = \frac{\partial N^1(\gamma^1)}{\partial \gamma}$	D
$[\gamma_1^1, 1]$	$\{h_1\}$	$[\frac{\partial N^1(\gamma^1)}{\partial \gamma}, -\infty)$	E

Theorem 1 provides a general method for identifying these regions. and associated acts. Consider an arbitrary strictly positive vector $\pi \in \mathbb{R}_+^M$, and define $B(\pi)$ as all acts that maximize the corresponding dot product,

$$B(\pi) = \{f \in A | \pi.\upsilon^f \geq \pi.\upsilon^g \text{ all } g \in A\}.$$

Set $B(\pi)$ identifies acts whose net utility functions would support a hyperplane with slope π. With $\pi \in \mathbb{R}^M_+$ and $B(\pi)$ identified, the Unique Posterior property implies that there are unique corresponding posteriors $\left\{\gamma^f\right\}_{f \in B(\pi)}$ that satisfy the ILR property. By the sufficiency condition of theorem 1, these posteriors form the basis for an optimal attention strategy for any decision problem (μ, A) for which these posteriors are feasible. The related unconditional choice probabilities P^f are then determined by Bayes' law.

By identifying optimal acts for all vectors π, one can in this manner characterize rationally inattentive policies for all priors $\mu \in \Gamma$. Identification of the mapping from normal vectors to maximizers of the corresponding dot product at extreme points of a convex set is a well-studied problem. It is equivalent to finding the dual cone associated with extreme points of this convex set defined by $\{\upsilon^g | g \in A\}$(Rockafellar [1972]). Figure 4 illustrates this for the three act case. Table 1 indicates the link between the regions of the dual cone and the set of priors for which the supporting acts form an optimal solution.

5 An Experimental Test of ILR

In this section we present an experiment that allows us to observe subjects' attentional responses to changing incentives, which we compare to the predictions of the Shannon model. To perform these tests we generate "state dependent stochastic choice data" (see CD13). For a given decision problem $(\mu, A) \in \Gamma \times \mathcal{F}$, we estimate the probability of choosing each act in each state of the world - i.e. a mapping $q : \Omega^\mu \longrightarrow \Delta(A)$. This identifies not only all chosen acts, but also how unconditionally likely each such act is to be chosen, and the corresponding revealed posteriors. Together, these constitute the revealed posterior-based attention strategy.

5.1 Experimental Design

We use the method of CD13 to generate state dependent stochastic choice data. In a typical round of the experiment, a subject is shown a screen on which there are displayed 100 balls. Some of the balls are red with the remainder blue. The state of the world is identified by the precise number of balls that are red as opposed to blue. Prior to observing the screen, subjects are informed of the probability distribution over such states. Having seen the screen they choose from a number of different acts whose payoffs are state dependent. A decision problem is defined by this prior information and the set of available acts, as in section 2.1. Each subject faces each specific decision problem 50 times, allowing us to approximate their state dependent stochastic choice function. In a given session, each subject faced 4 distinct decision problems. All occurrences of the same problem were grouped, with the order of the problem block-randomized. At the end of the experiment, one question was selected at random for payment, which reward was added to the show up fee of $10.

5.2 Description of Experiments and Theoretical Predictions

Table 2: Experimental Design

Decision Problem	Payoffs
1	2
2	10
3	20
4	30

Our experiment involves 2 treatments. In both treatments there are two equiprobable states and two acts $f(x)$ and $g(x)$, with $f(x)$ paying off x in state 1 (and zero otherwise) and $g(x)$ paying x in state 2 (and zero otherwise). The value x varies between decision problems, as in table 2. The difference between the two treatments rests in the difficulty of the underlying perceptual task. In one case (treatment 1), it is relatively easy to discriminate, with states 1 and 2 involving 47 and 53 red balls respectively, while in the other (treatment 2) it is harder, with 49 and 51 red balls respectively. Overall, the experiment allows us to estimate how mistakes (i.e. the probability of choosing the lower payoff act) vary with rewards. The Shannon model makes strong predictions in this regard, summarized by the ILR condition. Rearranging the first condition of theorem 1 gives,

$$\frac{U(x)}{\ln(\gamma_1^{f(x)} - \gamma_1^{g(x)})} = \frac{U(x)}{\ln(\gamma_2^{g(x)} - \gamma_2^{f(x)})} = \kappa$$

where $U(x)$ is the expected utility of monetary prize x. Assuming that the cost of attention κ does not vary within a treatment, then neither should the ratio of the difference in utilities between prizes to the log differences in posterior beliefs. Further assuming that costs are higher in treatment 2 than treatment 1, this ratio should be higher in decision problems in the former than the latter. This leads to the following hypothesis.

Hypothesis: (ILR) Given any $x, y \in \{2, 10, 20, 30\}$ in the same treatment,

$$\frac{U(x)}{\ln(\gamma_1^{f(x)} - \gamma_1^{g(x)})} = \frac{U(y)}{\ln(\gamma_2^{g(y)} - \gamma_2^{f(y)})}.$$

Given $x \in \{2, 10, 20, 30\}$, $\left\{\gamma^{f(x)}, \gamma^{g(x)}\right\}$ and $\left\{\bar{\gamma}^{f(x)}, \bar{\gamma}^{g(x)}\right\}$ observed in treatments 1 and 2 respectively,

$$\frac{U(x)}{\ln(\gamma_1^{f(x)} - \gamma_1^{g(x)})} < \frac{U(x)}{\ln(\bar{\gamma}_1^{f(x)} - \bar{\gamma}_1^{fgx)})}.$$

In order to test this hypothesis it is necessary to observe the utilities associated with each prize x. We assume initially that utility is linear in money, before controlling for subject-specific utility curvature.

5.3 Results

41 subjects took part in treatment 1, while 46 took part in treatment 2. Figure 5 shows the aggregate probability of correct choice for each decision problem in both treatments. As expected, the probability of choosing correctly is higher in the easier experiment 1, and is increasing in the rewards for making the correct choice in both experiments. In order to compare these expansion paths to those predicted by the Shannon model, we can calculate the ratio $\frac{U(x)}{\ln \gamma_1^{f(x)} - \ln \gamma_1^{g(x)}}$ and $\frac{U(x)}{\ln(\gamma_2^{g(x)} - \gamma_2^{f(x)})}$ for each value of x in treatments 1 and 2, using the aggregate data and assuming U is the identity function. If the aggregate data can be explained by the Shannon model, then this ratio should equal to the cost parameter κ and be invariant within each treatment. Assuming that treatment 2 is harder than 1, we would expect estimated costs to be higher in the former than the latter. Figure 6 plots this ratio for aggregate data, and associated standard errors.[7]

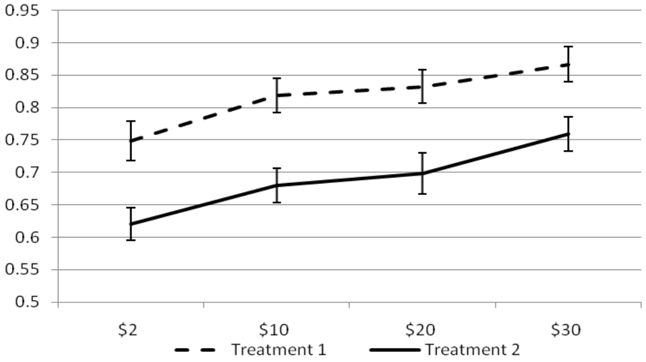

Figure 5: Probability of choosing the correct act as a function of x

[7] For all analysis in this paper, standard errors are calculated taking into account clustering at the subject level.

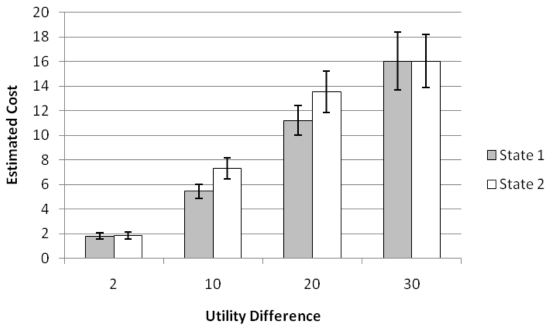

Figure 6a: Treatment 1

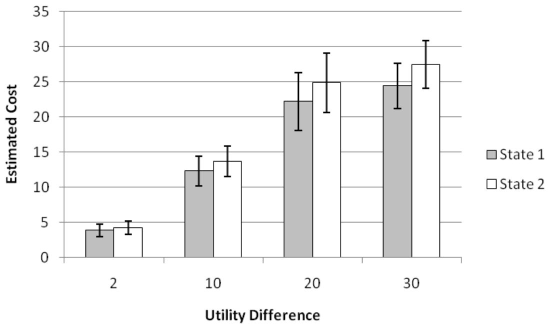

Figure 6b: Treatment 2

The key observation is that estimated costs appear significantly higher for higher reward levels, meaning that subjects do not increase attention in response to increasing rewards as much as the Shannon model predicts. For treatment 1, the estimated cost in state 1 is significantly different between all reward levels.[8] For treatment 2, estimated cost is significantly different between all reward levels apart from between \$20 and \$30.[9]

One possible explanation for deviations between our measured behavior and the predictions of the Shannon model is curvature of the utility function. In order to control for this, subjects also made choices between lotteries in the manner of the multiple price list task of Holt and Laury [2002].[10] This allows us to estimate subject-specific utility functions which can be used to

[8] At the 0.1% level between the \$2 and \$10 rewards and \$10 and \$20 rewards, and at the 5% level between the \$20 and \$30 rewards.

[9] In many cases there are also significant differences (at the 5% level) in estimated costs between states at the same reward amount. This is true for the \$10 and \$20 rewards in treatment 1 and the \$2, \$10 and \$20 rewards in treatment 2. This finding violates the predictions of the Shannon model, but is driven by the fact that subjects who are completely inattentive tend to choose option f rather than option g, meaning they are less accurate in state 2 than in state 1.

[10] Subjects had to make 10 choices between $p\$6.00+(1-p)\$4.80$ and $p\$11.55+(1-p)\$0.30$ for $p = 0.1$ to $p = 1$. Their responses are then used to estimate a CRRA utility function

convert prizes from monetary amounts to utility amounts. We use these data to estimate subject-specific utility functions which can be used to convert prizes from monetary amounts to utility amounts.[11]

In order to control for the effect of risk aversion, we use subject-specific data. We focus on subjects who (a) exhibited consistency in the risk aversion questions (i.e. obeyed stochastic dominance) (b) were never inattentive and (c) choose the correct act given their posterior beliefs (i.e. never violate the No Improving Action Switches condition from CD13). The first condition means that we have a well defined estimate of the subject's utility function, while the latter two mean that we can obtain a precise estimate of their costs from each decision problem. These conditions leave us with 28 subjects in treatment 1 and 22 in treatment 2. For each of these subjects, we estimate their costs using data from state 1 in each decision problem, and then test for statistically different costs across decision problems.[12] We do this both assuming linear utility, and using our subject-specific estimated utility function. In the former case, 34%, (or 17 of 50) of subjects across the two experiments exhibit significantly (at the 10% level) increased cost estimates between the \$2 and \$30 reward case. Controlling for risk aversion this number drops to 26% (13 of 50), with 2 subjects (4%) exhibiting a significant decrease in costs. We conclude that a significant fraction of subjects are less reactive to changes in rewards than the Shannon model would predict, even controlling for risk aversion.

[11] Note that, because only 1 in 200 questions is rewarded, the true value of choosing the correct act for any x is,

$$\frac{1}{200}U(x)+\frac{1}{199}U(0).$$

Normalizing $u(0)$ to 0, note that hypothesis ILR still holds.

[12] In some cases, our subjects were perfectly accurate at some reward level. The Shannon model predicts that subjects will never choose to be perfectly discriminatory for positive costs. In such cases, we test the probability of observing perfect accuracy in our sample in decision problem x given the costs estimated in decision problem y, and reject the Shannon model if this probability is less that 10%

6 Separable Models of Rational Inattention

Our experimental results suggests that there may be value to considering rational inattention models that allow for different responses to incentives than does the Shannon model. To that end, we introduce a family of attention cost functions that maintain much of the structure of the Shannon model, but allows for different response elasticities. These cost functions maintain the form of equation 2, which enables us to apply many of our same posterior-based methods. However, we allow the cost of a posterior distribution to differ from the negative of its entropy. We call this the posterior-separable class of attention cost functions.

Definition 4 *A Strictly convex function* $G : \mathbb{R}_+^M \to \mathbb{R} \in \mathcal{G}$ *generates a* ***posterior-separable*** *attention cost function* $K_G : \Lambda^{(\mu,A)} \to \mathbb{R}$ *if, for all* $(\mu, A) \in \Gamma \times \mathcal{F}$*,*

$$K_G(B, P, \gamma) = -G(\mu) + \sum_{\gamma \in B} P(\gamma) G(\gamma).$$

Note that the Shannon mutual information function fits into this class with $G(\mu) = -\kappa H(\mu)$. Note also that requiring strict convexity ensures that strictly more Blackwell informative signals always involve strictly higher costs. We use the notation $N_G^f(\gamma^f)$ to denote the net utility function associated with act f when the cost function is G. We let $\hat{\Lambda}_G^{(\mu,A)}$ denote the corresponding rationally inattentive strategies.

Because posterior-separable cost functions are structurally similar to Shannon mutual information costs, many of the results of the paper apply equally to this class. Importantly, lemmas 1 and 2 hold, so that rationally inattentive behavior can be characterized by a supporting hyperplane to the lower epigraph of the concavified net utility function. For differentiable cases, lemma 3 also holds, which carries with it the ability to solve the model using derivative conditions. This in turn implies that corollaries 1, 3, and 4 hold for all posterior-separable cost functions, while corollary 2 holds for all such functions which are differentiable.

We illustrate how the posterior-separable family allow for different elasticities of attention with respect to incentives by introducing a parametrized class of cost functions $G_{\{\rho,\kappa\}} \in \mathcal{G}$:

$$G_{\{\rho,\kappa\}}(\gamma) = \begin{cases} -\kappa\left(\sum_{m=1}^{M}\gamma_m\left[\frac{\gamma_m^{1-\rho}}{(\rho-1)(\rho-2)}\right]\right) & \text{if } \rho \neq 1 \text{ and } \rho \neq 2; \\ -\kappa\left(\sum_{m=1}^{M}\gamma_m \ln\gamma_m\right) & \text{if } \rho = 1. \\ -\kappa\left(\sum_{m=1}^{M}\gamma_m \frac{\ln\gamma_m}{\gamma_m}\right) & \text{if } \rho = 2. \end{cases},$$

In the two state case, derivatives with respect to γ_1 obey,

$$\frac{\partial G_{\{\rho,\kappa\}}(\gamma)}{\partial\gamma_1} = \begin{cases} \kappa\left(\frac{\gamma_1^{1-\rho}-(1-\gamma_1)^{1-\rho}}{(\rho-1)}\right) & \text{if } \rho \neq 1; \\ \kappa\left(\ln\gamma_1 - \ln(1-\gamma_1)\right) & \text{if } \rho = 1. \end{cases}$$

$$\frac{\partial^2 G_{\{\rho,\kappa\}}(\gamma)}{\partial\left(\gamma_1\right)^2} = \kappa\left(\gamma_1^{-\rho} + (1-\gamma_1)^{-\rho}\right) \text{ if } \rho \neq 1;$$

Note that the second derivative of these costs functions is continuous in ρ, with the Shannon entropy cost function fitting smoothly into the parametric class at $\rho = 1$.

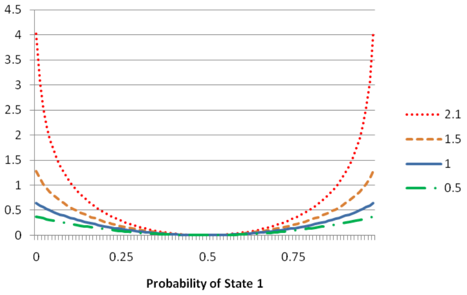

Figure 7a: Cost functions

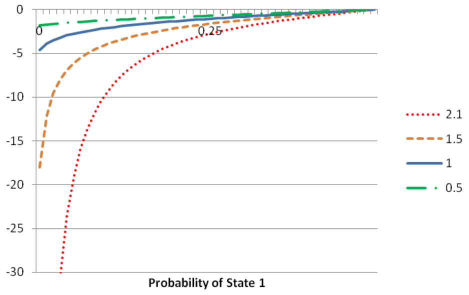

Figure 7b: Marginal Costs

Figures 7a and 7b illustrate the shape of these cost functions (normalized to equal 0 at $\gamma_1 = 0.5$) and their first derivatives in the symmetric two state case and with $\kappa = 1$. Functions with ρ less than 1 have a first derivative that does not tend to infinity as γ_1 tends to zero. This means that the marginal cost of information does not go to infinity, so that subjects may choose to be fully informed.

We solve the model computationally using the derivative characterization of lemma 3. Figure 8 plots the probability of correct choice against the cost of mistake for different values of ρ. For $\rho = 0.5$, the DM would choose to become fully informed when the utility difference hits about 1.1. For the other cost functions, subjects never become fully informed. Note that larger values of ρ imply that, for any given cost of mistakes, subjects will choose to be less informed, and will be less responsive to a change in the cost of mistakes.

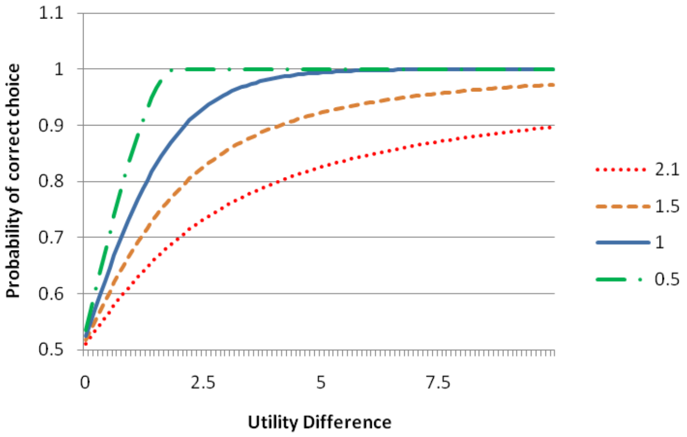

Figure 8: Attentional Response to Incentives

Figures 9a and 9b show how the extra degree of freedom introduced by the parameter ρ can help to match our experimental data. The two bars in figure 9a show the posterior probability of state 1 given the choice of f and the posterior probability of state 2 given choice of g for each decision problem in treatment 1 according to our aggregate data. Figure 9b shows the same data for treatment 2.

On both graphs, the dashed line shows the predictions of the Shannon model, using a treatment specific cost parameter chosen to minimize the mean squared difference between predicted and actual posterior beliefs (note that, due to symmetry, the Shannon model predicts that $\gamma_1^{f(x)} = \gamma_2^{g(x)}$ for every x). The best fitting cost parameters are 8.7 in treatment 1 and 22.2 in treatment 2, which leads to a mean squared difference between observed and predicted data of 0.020.

The solid line shows the predictions of the best fitting model in the class $G_{\{\rho,\kappa\}}$, with ρ fixed across treatments but costs allowed to vary.[13] The parameters that best fit the model are $\rho = 7.01$, with costs in the first treatment of 0.002 and in the second treatment of 0.046. Clearly this model provides a much better fit of the data, with a mean squared difference of 0.002.

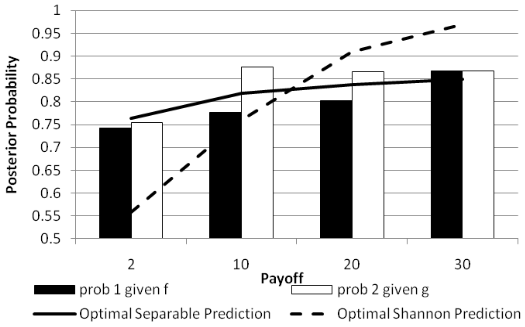

Figure 9a: Treatment 1

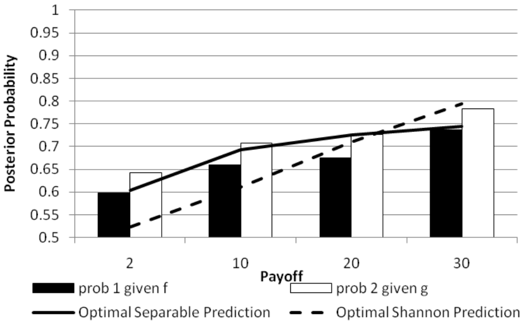

Figure 9b: Treatment 2

Since the Shannon model is nested within the class of $G_{\{\rho,\kappa\}}$ functions, this broader class must weakly provided a better fit of the data. However, criteria that punish models for having additional parameters suggest rejecting

[13] Thus the shape of the cost function is constrained to be the same in the two treatments, but the level of costs varies. This is equivalent to the way we treated the Shannon model.

Shannon in favor of the broader parametrized class. For example the Akaike information criterion is lower for the model that allows for variable ρ than for the Shannon model.[14]

7 Concluding Remarks

Rational inattention theory is of rapidly growing importance. Yet general behavioral implications can be hard to identify even with Shannon mutual information costs. We develop posterior-based methods that identify key behavioral properties of this model. We experimentally test a key implication of the Shannon model regarding changes in incentives. We find our subjects to be less responsive along this dimension than is implied by the Shannon model. We introduce a class of generalized entropy cost functions that allow for a more flexible such response, and identify the improved fit that results.

The posterior-based method is currently being applied in a variety of settings. It is of particular value in dynamic settings in which the beliefs evolve as a result of the interaction between attentional effort and exogenous shocks. It is equally of value in strategic settings (e.g. Martin [2013]).

[14] The AIC is defined by the equation,

$$AIC = 2k - 2\ln(L),$$

where k is the number of parameters of the model and L its likelihood. The AIC for the Shannon model (allowing for different costs in the two treatments) is 10028, while for the extended model (constant ρ but different costs in the two treatments) the AIC is 9521. The favored model is the one with the lowest AIC. The flexibility that our generalized entropy model adds is therefore of significant value in fitting our experimental data.

8 Bibliography

References

Andrew Caplin and Mark Dean. Rational inattention and state dependent stochastic choice. Mimeo, New York University, 2013.

T. Cover and J. Thomas. *Elements of Information Theory 2nd Edition.* John Wiley and Sons, Inc., New York, 2006.

C.A. Holt and S.K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655, 2002.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, September 2011.

Bartosz Mackowiak and Mirko Wiederholt. Optimal sticky prices under rational inattention. *American Economic Review*, 99(3):769–803, June 2009.

Daniel Martin. Strategic pricing and rational inattention to quality. Mimeo, New York University, 2013.

Filip Matejka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. CERGE-EI Working Papers wp442, The Center for Economic Research and Graduate Education - Economic Institute, Prague, June 2011.

Filip Matejka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. Mimeo, CERGE, 2013.

Filip Matejka. Rationally inattentive seller: Sales and discrete pricing. CERGE-EI Working Papers wp408, The Center for Economic Research and Graduate Education - Economic Institute, Prague, March 2010.

R. Tyrrell Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1972.

Christopher A. Sims. Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49(1):317–356, December 1998.

Christopher Sims. Implications of Rational Inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.

Christopher A. Sims. Rational inattention: Beyond the linear-quadratic case. *American Economic Review*, 96(2):158–163, May 2006.

Christopher A. Sims. Rational inattention and monetary economics. In Benjamin M. Friedman and Michael Woodford, editors, *Handbook of Monetary Economics*, volume 3 of *Handbook of Monetary Economics*, chapter 4, pages 155–181. Elsevier, 2010.

Stijn van Nieuwerburgh and Laura Veldkamp. Information Immobility and the Home Bias Puzzle. *Journal of Finance (forthcoming)*, 2008.

Ming Yang. Coordination with rational inattention. Technical report, 2011.

9 Appendix

9.1 Lemmas

Our proof of theorem 1 relies on three lemmas that hold more broadly than for the Shannon cost function. They apply to the general posterior-separable cost functions $K_G : \Lambda^{(\mu,A)} \to \mathbb{R}$ based on arbitrary strictly convex function $G : \mathbb{R}^M_+ \to \mathbb{R} \in \mathcal{G}$ as introduced in section 6 and the corresponding net utility functions,

$$N_G^{(\mu,A)}(B,P,\gamma) \equiv \sum_{f\in B} P^f \sum_{j=1}^{M} \gamma_m^f U_m^f - K_G(B,P,\gamma) = \sum_{f\in B} P^f N_G^f(\gamma^f) + G(\mu).$$

As for the Shannon model, the definition of rational inattention for $G \in \mathcal{G}$ involves maximization of this net utility function. Proofs of these Lemmas are in the online appendix.

Lemma 1 Given $A \in \mathcal{F}$ and $G \in \mathcal{G}$, the set $\mathcal{E}_G(A) \subset \mathbb{R}^M$ defined by,

$$\mathcal{E}_G(A) \equiv \left\{ \begin{array}{c} (y, \mu_1, .., \mu_{M-1}) \in \mathbb{R} \times X \\ \text{s.t. } \exists A \in \mathcal{F} \text{ and } (B,P,\gamma) \in \Lambda^{(\mu,A)} \text{ s.t. } y \leq \sum_{f\in B} P^f N_G^f(\gamma^f) \end{array} \right\},$$

is closed, convex, and bounded above in its first coordinate.

Lemma 2 Strategy $(B,P,\gamma) \in \Lambda^{(\mu,A)}$ is rationally inattentive for $G \in \mathcal{G}$ if and only if it there exists λ_m for $1 \leq m \leq M-1$ such that **property SH** holds:

$$N_G^g(\gamma) - \sum_{m=1}^{M-1} \lambda_m \gamma_m \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f;$$

all $f \in B$, $g \in A$ and $\gamma \in \Gamma$.

Lemma 3 Given $G \in \mathcal{G}$ that is differentiable on Γ^I, the interior of Γ, strategy $(B,P,\gamma) \in \Lambda^{(\mu,A)}$ with $\gamma_m^f \in (0,1)$ satisfies $(B,P,\gamma) \in \hat{\Lambda}_G^{(\mu,A)}$ if and only if it satisfies CT, ED, and UB:

A. **Common Tangent for Chosen Acts (CT)**: Given $f, g \in B$,

$$N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^f = N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^g(\gamma^g)}{\partial \gamma_m} \right] \gamma_m^g.$$

B. **Equal Derivative for Chosen Acts (ED)** : Given $f, g \in B$,

$$\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} = \frac{\partial N_G^g(\gamma^g)}{\partial \gamma_m}.$$

C. **Unchosen Act Bound (UB) :** Given $f \in B$ and $g \in B \backslash A$,

$$N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^g \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^f,$$

where $\gamma^g \in \Gamma$ maximizes on $N_G^g(\gamma) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m$ on $\gamma \in \Gamma$.

9.2 Theorem 1

Theorem 1 Given $(\mu, A) \in \Gamma \times \mathcal{F}$ and $\kappa > 0$, $(B, P, \gamma) \in \Lambda^{(\mu, A)}$ is rationally inattentive if and only if:

A. **ILR Equations for Chosen Acts**: given $f, g \in B$, and $1 \leq m \leq M$,

$$\frac{\gamma_m^f}{\nu_m^f} = \frac{\gamma_m^g}{\nu_m^g}.$$

B. I**LR Inequalities for Unchosen Acts**: given $f \in B$ and $g \in A \backslash B$,

$$\sum_{m=1}^{M} \left[\frac{\gamma_m^f}{\nu_m^f} \right] \nu_m^g \leq 1.$$

Proof. The Shannon model satisfies the differentiability condition of lemma 3, and also has unbounded derivatives at the boundary points of the domain,

$$\lim_{\gamma_m \searrow 0} \frac{\partial G(\gamma)}{\partial \gamma_m} = -\infty \text{ and } \lim_{\gamma_m \nearrow 1} \frac{\partial G(\gamma)}{\partial \gamma_m} = \infty.$$

so that all rationally inattentive strategies necessarily involve $\gamma_m^f > 0$. Hence condition UB reduces to the "bounded tangent" (BT) condition (as γ^g occurs at the tangent point),

$$\frac{\partial N_G^g(\gamma^g)}{\partial \gamma_m} = \lambda_m \Longrightarrow N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^g \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f.$$

Given the unbounded derivatives at corners of $G = -\kappa H$, there can only be interior rationally inattentive posteriors with $\gamma_m^f > 0$ all $f \in B$ as in lemma 3. Our proof of theorem 1 is therefore based on conditions CT, ED, and BT.

Necessity: Note that Shannon net utility functions have particularly simple form,

$$N^f(\gamma) \equiv \sum_{m=1}^{M-1} \gamma_m (U_m^f - \kappa \ln \gamma_m) + \left(1 - \sum_{m=1}^{M-1} \gamma_m\right) \left(U_M^f - \kappa \ln \left(1 - \sum_{m=1}^{M-1} \gamma_m\right)\right),$$

Define

$$\rho_m \equiv \max_{f \in B} \frac{\partial N^f(\gamma^f)}{\partial \gamma_m},$$

for $1 \leq m \leq M-1$. for $\kappa > 0$. Hence ED implies that,

$$U_m^f - \kappa \ln \gamma_m^f - \left[U_M^f - \kappa \ln \gamma_M^f\right] = \rho_m.$$

Substitution in CT yields,

$$\begin{aligned} N^f(\gamma^f) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f &= \sum_{m=1}^{M-1} \gamma_m^f \left(U_M^f - \kappa \ln \gamma_M^f + \rho_m \right) + \left(1 - \sum_{m=1}^{M-1} \gamma_m^f \right) \left(U_M^f - \kappa \ln \gamma_M^f \right) \\ &= U_M^f - \kappa \ln \gamma_M^f = U_M^g - \kappa \ln \gamma_M^g = N^g(\gamma^g) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f. \end{aligned}$$

Overall, given $f, g \in B$, we conclude therefore that,

$$\frac{U_m^f}{\kappa} - \ln \gamma_m^f = \frac{U_m^g}{\kappa} - \kappa \ln \gamma_m^f,$$

all $1 \leq m \leq M$. This establishes necessity of the ILR equations for chosen acts upon exponentiation.

To prove necessity of the ILR inequalities consider $g \in A \backslash B$ and $\gamma^g \in \Gamma$ satisfying,

$$\frac{\partial N^g(\gamma^g)}{\partial \gamma_m} = U_m^g - \kappa \ln \gamma_m^g - [U_M^g - \kappa \ln \gamma_M^g] = \rho_m,$$

all $1 \leq m \leq M - 1$. By property BT,

$$N^g(\gamma^g) - \sum_{m=1}^{M-1} \rho_m \gamma_m^g = U_M^g - \kappa \ln \gamma_M^g \leq U_M^f - \kappa \ln \gamma_M^f = N^f(\gamma^f) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f,$$

for $f \in B$. Hence

$$\frac{U_m^g}{\kappa} - \ln \gamma_m^g \leq \frac{U_m^f}{\kappa} - \ln \gamma_m^f \Longrightarrow \ln \gamma_m^f + \frac{U_m^g}{\kappa} - \frac{U_m^f}{\kappa} \leq \ln \gamma_m^g,$$

for $1 \leq m \leq M$. Exponentiating and taking the summation we arrive at,

$$\sum_{m=1}^{M} \left[\frac{\gamma_m^f}{\nu_m^f} \right] \nu_m^g \leq \sum_{m=1}^{M} \gamma_m^g = 1,$$

establishing necessity.

Sufficiency: Given $(\mu, A) \in \Gamma \times \mathcal{F}$, consider $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ such that the ILR conditions are met. We first confirm that the equal derivative condition is therefore met. The ILR equations imply that, given $f, g \in B$, and $1 \leq m \leq M$,

$$\frac{\gamma_m^f}{\nu_m^f} = \frac{\gamma_m^g}{\nu_m^g}.$$

Hence, upon substituting the logarithmic versions of the ILR equations for arguments m and M, we conclude that indeed derivatives are equal,

$$\frac{\partial N^f(\gamma^f)}{\partial \gamma_m} = U_m^f - \kappa \ln \gamma_m^f - \left[U_M^f - \kappa \ln \gamma_M^f\right] = U_m^g - \kappa \ln \gamma_m^g - [U_M^g - \kappa \ln \gamma_M^g] = \frac{\partial N^g(\gamma^g)}{\partial \gamma_m} \equiv \rho_m.$$

To establish the CT, note as before that for any $f \in B$,

$$N^f(\gamma^f) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f = U_M^f - \kappa \ln \gamma_M^f.$$

Applying again the ILR equation for chosen acts with $m = M$ we confirm that indeed, for all $f.g \in B$,

$$N^f(\gamma^f) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f = U_M^f - \kappa \ln \gamma_M^f = U_M^g - \kappa \ln \gamma_M^g = N^g(\gamma^g) - \sum_{m=1}^{M-1} \rho_m \gamma_m^f.$$

To establish BT, consider $g \in A \backslash B$ and $\gamma^g \in \Gamma$ with,

$$\frac{\partial N^g(\gamma^g)}{\partial \gamma_m} = U_m^g - \kappa \ln \gamma_m^g - U_M^g - \kappa \ln \gamma_M^g = \rho_m,$$

for all $1 \leq m \leq M - 1$. As before, this enables us to compute the value of the relevant tangent as,

$$N^g(\gamma^g) - \sum_{m=1}^{M-1} \rho_m \gamma_m^g = U_M^g - \kappa \ln \gamma_M^g.$$

Hence the bounded tangent property applies provided that,

$$\frac{U_M^g}{\kappa} - \ln \gamma_M^g \leq \frac{U_M^f}{\kappa} - \ln \gamma_M^f.$$

To the contrary, suppose that,

$$\frac{U_M^g}{\kappa} - \ln \gamma_M^g > \frac{U_M^f}{\kappa} - \ln \gamma_M^f.$$

Given their defining derivative conditions, this implies that the same inequality holds for all $1 \leq m \leq M-1$,

$$\frac{U_m^g}{\kappa} - \ln \gamma_m^g = [U_M^g - \kappa \ln \gamma_M^g] + \rho_m > U_M^f - \kappa \ln \gamma_M^f + \rho_m = \frac{U_m^f}{\kappa} - \ln \gamma_m^f.$$

Rearrangement yields,

$$\ln \gamma_m^f + \frac{U_m^g}{\kappa} - \frac{U_m^f}{\kappa} > \ln \gamma_m^g.$$

Exponentiating and taking the summation we arrive at,

$$\sum_{m=1}^{M} \left[\frac{\gamma_m^f}{\nu_m^f} \right] \nu_m^g > \sum_{m=1}^{M} \gamma_m^g = 1,$$

contradicting the ILR inequalities for unchosen acts, and with it establishing validity of the bounded tangent property. This rounds out the proof that the sufficient conditions of corollary 3 hold, completing the proof of theorem 1. ■

9.3 Theorem 2

Theorem 2 (Unique Optimal Strategy): With affine independence, $\left|\hat{\Lambda}^{(\mu,A)}\right| = 1$ all $\mu \in \Gamma$.

Proof. Given the differentiability of the value function $V : X \longrightarrow \mathbb{R}$ at all points with $\mu_m > 0$, we know that for such points there is a single tangent plane at the corresponding boundary of set $\mathcal{E}(A)$. Let $(1, \lambda_1, .., \lambda_{M-1})$ be a normal vector defining this supporting hyperplane at boundary point $(y, \mu_1, .., \mu_{M-1})$, and define corresponding posteriors γ^f on $f \in A$ to solve the first order conditions,

$$\frac{\partial N^f(\gamma^f)}{\partial \gamma_m} + \lambda_m = 0, \text{ for } 1 \leq m \leq M-1.$$

Now consider the set $C \subset A$ comprising all acts $f \in A$ that maximize the corresponding dot product,

$$C = \left\{ f \in A | \exists \gamma^f \in \Gamma \text{ such that } N^f(\gamma^f) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^f \geq N^g(\gamma^g) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^g \text{ all } g \in A \right\}.$$

By lemma 2 all optimal strategies $(B, P, \gamma) \in \hat{\Lambda}^{(\mu, A)}$ must satisfy $B \subset C$, since, given $f, g \in C$ and $g \in A \backslash C$ we know that,

$$\begin{aligned} N^f(\gamma^f) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^f &= N^g(\gamma^g) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^g; \\ N^f(\gamma^f) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^f &> N^g(\gamma^g) + \sum_{m=1}^{M-1} \lambda_m \gamma_m^g. \end{aligned}$$

By the logic of lemma 2, satisfaction of the upper equation implies satisfaction of the ILR equations, while satisfaction of the lower strict inequality produces a strict form of the ILR inequality, establishing indeed that $B \subset C$ is necessary for optimality.

Now suppose that there are two distinct sets of probability weights P^f, Q^f with $\sum_{f \in C} Q^f = \sum_{f \in C} P^f = 1$ (these need not all be strictly positive) such that,

$$\sum_{f \in C} P^f \gamma^f = \sum_{f \in C} Q^f \gamma^f = \mu.$$

Subtraction produces a non-zero set of weights $\delta^f \equiv P^f - Q^f$ with $\sum\limits_{f \in C} \delta^f = 0$ such that,

$$\sum_{f \in C} \delta^f \gamma_m^f = 0,$$

for all $1 \leq m \leq M$. Substitution of the ILR condition yields,

$$\sum_{f \in C} \delta^f \alpha_m \nu_m^f = 0,$$

whereupon division by any non-zero term $\delta^g \alpha_m$ yields,

$$\sum_{f \in C} \left[\frac{\delta^f}{\delta^g} \right] \nu_m^f = 0.$$

Given that $\sum\limits_{f \in C} \left[\frac{\delta^f}{\delta^g} \right] = 0$, this directly contradicts affine independence. This completes the proof for all cases with $\mu_m > 0$ for some $1 \leq m \leq M - 1$. This leaves only the case $\mu_M = 1$, for which case identical logic establishes that there can only be one optimizing act that must be chosen for sure. ■

10 Online Appendix

10.1 Proofs of Lemmas

Lemma 1 Given $A \in \mathcal{F}$ and $G \in \mathcal{G}$, the set $\mathcal{E}_G(A) \subset \mathbb{R}^M$ defined by,

$$\mathcal{E}_G(A) \equiv \left\{ (y, \mu_1, .., \mu_{M-1}) \in \mathbb{R} \times X | \exists A \in \mathcal{F} \text{ and } (B, P, \gamma) \in \Lambda^{(\mu, A)} \text{ s.t. } y \leq \sum_{f \in B} P^f N_G^f(\gamma^f) \right\},$$

is closed, convex, and bounded above in its first coordinate.

Proof. Given $A \in \mathcal{F}$ and $G \in \mathcal{G}$, consider $(y, \mu_1, .., \mu_{M-1})$, $(\tilde{y}, \tilde{\mu}_1, .., \tilde{\mu}_{M-1}) \in \mathcal{E}_G(A)$ together with finite sets $B, \tilde{B} \subset A$, probabilities on actions and associated posteriors, P^f, γ_m^f for $f \in B$ and $\tilde{P}^f, \tilde{\gamma}_m^f$ for $f \in \tilde{B}$, all $1 \leq m \leq M-1$ such that,

$$\begin{aligned} \mu_m &= \sum_{f \in B} P^f \gamma_m^f \text{ and } y \leq \sum_{f \in B} P^f N_G^f(\gamma^f); \\ \tilde{\mu}_m &= \sum_{f \in \tilde{B}}^{f} \tilde{P}^f \tilde{\gamma}_m^f \text{ and } \tilde{y} \leq \sum_{f \in \tilde{B}} \tilde{P}^f N_G^f(\tilde{\gamma}^f). \end{aligned}$$

Define $C = B \cup \tilde{B}$ and extend $P^f, \tilde{P}^f$ to this domain by setting them to zero on the unchosen acts.

Given $\lambda \in (0, 1)$, define $R^f = \lambda P^f + (1-\lambda)\tilde{P}^f$ and

$$\eta_m^f = \frac{\lambda P^f \gamma_m^f + (1-\lambda)\tilde{P}^f \tilde{\gamma}_m^f}{\lambda P^f + (1-\lambda)\tilde{P}^f}.$$

It is immediate that $\eta^f \in \Gamma$ all $f \in C$ and that $\sum\limits_{f\in C} R^f \eta^f_m = \lambda\mu_m + (1-\lambda)\tilde{\mu}_m$ so that $(C, \eta, R) \in \Lambda^{\frac{\mu+\tilde{\mu}}{2}}$ Note that, for each $f \in C$

$$\begin{aligned}
& N^f_G\left(\eta^f\right) \\
= & \sum_{j=1}^{M} \eta^f_m U^f_m - G(\eta^f) \\
= & \frac{\lambda P^f}{\lambda P^f + (1-\lambda)\tilde{P}^f} \sum_{j=1}^{M} \gamma^f_m U^f_m + \frac{(1-\lambda)\tilde{P}^f}{\lambda P^f + (1-\lambda)\tilde{P}^f} \sum_{j=1}^{M} \bar{\gamma}^f_m U^f_m - G(\eta^f) \\
\geq & \frac{\lambda P^f}{\lambda P^f + (1-\lambda)\tilde{P}^f} N^f_G(\gamma^f) + \frac{(1-\lambda)\tilde{P}^f}{\lambda P^f + (1-\lambda)\tilde{P}^f} N^f_G(\tilde{\gamma}^f),
\end{aligned}$$

By the convexity of G

Thus we have that

$$\begin{aligned}
\sum_{f\in C} R^f N^f_G\left(\eta^f\right) &= \sum_{f\in C} \left(\lambda P^f + (1-\lambda)\tilde{P}^f\right) N^f_G\left(\eta^f\right) \\
&\geq \lambda \sum_{f\in B} P^f N^f_G(\gamma^f) + (1-\lambda) \sum_{f\in\tilde{B}} \tilde{P}^f N^f_G(\tilde{\gamma}^f) = \lambda y + (1-\lambda)\bar{y},
\end{aligned}$$

confirming that $\lambda\left(y, \mu_1, .., \mu_{M-1}\right) + (1-\lambda)(\tilde{y}, \tilde{\mu}_1, .., \tilde{\mu}_{M-1}) \in \mathcal{E}_G(A)$

To establish closedness, consider a sequence $(y(n), \mu(n)) \in \mathcal{E}_G(A)$ converging to (y^L, μ^L) (to simplify notation we use the full prior as the second argument since μ_M is anyway implied) and corresponding triples $(B(n), P(n), \gamma(n)) \in \Lambda^{(\mu(n),A)}$, so that $\mu(n) = \sum_{f\in B(n)} P^f(n)\gamma^f(n)$ and $y(n) \leq \sum_{f\in B(n)} P^f(n) N^f_G(\gamma^f(n))$. We show now that there is no loss of generality in assuming $|B(n)| \leq M + 1$. Suppose initially that $|B(n)| > M$. By Charateodory's theorem, since $\{\gamma^f(n) \in \Gamma | f \in B(n)\}$ contain $\mu(n)$ in its convex hull, there exists $B_1(n) \subset B(n)$ with $|B_1(n)| \leq M + 1$ for which there exists a strictly positive probability weights $P^f_1(n) > 0$ on $f \in B_1(n)$ such that $\mu = \sum\limits_{f\in B^1(n)} P^f_1(n)\gamma^f(n)$. If expected net utility is no lower,

$$y(n) \leq \sum_{f \in B(n)} P^f(n) N_G^f(\gamma^f(n)) \leq \sum_{f \in B_1(n)} P_1^f(n) N_G^f(\gamma^f(n)),$$

we are done. If not, identify the smallest scalar $\alpha_1 \in (0,1)$ such that,

$$\alpha_1 P_1^f(n) = P^f(n),$$

some $f \in B_1(n)$. That such a scalar exists follows from the fact that

$$\sum_{f \in B_1(n)} P_1^f(n) = \sum_{f \in B(n)} P^f(n) = 1,$$

with all components in both sums strictly positive and with $|B(n)| > |B_1(n)|$.

We now define a second set of probability weights $P_2^f(n)$,

$$P_2^f(n) = \frac{P^f(n) - \alpha_1 P_1^f(n)}{1 - \alpha^1}.$$

for $f \in B_1(n)$. Correspondingly, we define,

$$B_2(n) = \{f \in B(n) | P_2^f(n) > 0\},$$

noting that $|B_2(n)| \leq |B(n)| - 1$. By construction $\mu = \sum\limits_{f \in B(n)} P_2^f(n) \gamma^f(n)$. Moreover,

$$\sum_{f \in B_2(n)} P_2^f(n) N_G^f(\gamma^f(n)) = \sum_{f \in B(n)} \left[\frac{P^f(n) - \alpha_1 P_1^f(n)}{1 - \alpha^1} \right] N_G^f(\gamma^f(n)) > \sum_{f \in B(n)} P^f(n) N_G^f(\gamma^f(n)).$$

Iteration from this point establishes that indeed we can identify a set $\tilde{B}(n) \subset B(n)$ with $\left|\tilde{B}(n)\right| \leq M + 1$ and $\tilde{P}(n) > 0$ on $f \in \tilde{B}(n)$ such that $\mu =$

$\sum_{f \in \tilde{B}(n)} P_1^f(n)\gamma^f(n)$ and,

$$\sum_{f \in \tilde{B}(n)} P_1^f(n) N_G^f(\gamma^f(n)) \geq y(n).$$

Given this, there is no loss of generality in assuming that $|B(n)| \leq M + 1$ in our original sequence.

With this, we can focus on a subsequence (we continue to index by n for notational simplicity) with all sets $B(n)$ of the same cardinality $K \leq M$. In each set $B(n)$ we index the acts in (arbitrary) order by $f(k, n) \in$ for $1 \leq k \leq K$, and correspondingly label that associated posteriors and act probabilities as $\gamma^k(n), P^k(n)$. Given the compactness of Γ, we can further select subsequences to ensure that there is a full set of limit posteriors and limit probabilities $\bar{\gamma}^k$ and $\bar{P}^k$, for $1 \leq k \leq K$,

$$\lim_{n \to \infty} \gamma^k(n) = \bar{\gamma}^k; \lim_{n \to \infty} P^k(n) = \bar{P}^k$$

For all acts $f \in A$, we can compute the net utility at all limit posteriors,

$$N_G^f(\bar{\gamma}^k) = \sum_{m=1}^{M} U_m^f \bar{\gamma}_m^k - G(\bar{\gamma}^k)$$

Since $\{U_m^f \in \mathbb{R}^M | f \in A\}$ is bounded above then so is $N_G^f(\bar{\gamma}^k)$ (with respect to $f \in A$). . Since $\{U_m^f \in \mathbb{R}^M | f \in A\}$ is closed, the upper bound is achieved. Hence we can find acts $\bar{f}(k) \in A$ that maximize the above net utilities,

$$N(\bar{f}(k), k) \geq N(f, k),$$

all $f \in A$.

We now define $\bar{B} = \cup_{k=1}^{K} \bar{f}(k)$. Note that, by construction

$$\sum_{k=1}^{K} \bar{P}^k \bar{\gamma}^k = \mu^L,$$

so that $(\bar{B}, \bar{\gamma}, \bar{P}) \in \Lambda^{(\mu^L, A)}$. Note also that, for each for all n,

$$\sum_{k=1}^{K} \bar{P}^k N^{\bar{f}(k)}(\bar{\gamma}^k) \geq \sum_{k=1}^{K} \bar{P}^k N_G^{f(k,n)}(\bar{\gamma}^k).$$

In light of continuity of all functions N_G^f, taking the limit on the RHS as $n \to \infty$ yields,

$$\sum_{k=1}^{K} \bar{P}^k N^{\bar{f}(k)}(\bar{\gamma}^k) \geq \lim_{n \to \infty} \sum_{k=1}^{K} P^k(n) N_G^f(\gamma^f(n)) \geq y^L, \tag{4}$$

This completes the proof that $(y^L, \mu^L) \in \mathcal{E}_G(A)$, hence that $\mathcal{E}_G(A)$ is closed. Boundedness above of the first coordinate follows from the fact that $\{U_m^f \in \mathbb{R}^M | f \in A\}$ is bounded above for all $A \in \mathcal{F}$. ■

Lemma 2 Strategy $(B, P, \gamma) \in \Lambda^{(\mu, A)}$ is rationally inattentive for $G \in \mathcal{G}$ if and only if it there exists λ_m for $1 \leq m \leq M - 1$ such that **property SH** holds:

$$N_G^g(\gamma) - \sum_{m=1}^{M-1} \lambda_m \gamma_m \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f;$$

all $f \in B$, $g \in A$ and $\gamma \in \Gamma$.

Proof. Necessity: Given $(B, P, \gamma) \in \hat{\Lambda}_G^{(\mu, A)}$, $\left(\sum_{f \in B} P^f N_G^f(\gamma^f), \mu_1, .., \mu_{M-1} \right)$ is an upper boundary point boundary of $\mathcal{E}_G(A)$. Lemma 1 establishes that such sets are always closed, convex, and bounded above in the first coordinate. This implies existence of a supporting hyperplane defined by normal vector

$(1, -\lambda_1, ..., -\lambda_{M-1})$ such that, for all $(y_0, y_1, .., y_{M-1}) \in \mathcal{E}_G(A)$,

$$y_0 - \sum_{m=1}^{M-1} \lambda_m y_m \leq \sum_{f \in B} P^f N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \mu_m = \sum_{f \in B} P^f [N_G^f(\gamma^f) - \lambda_m \gamma_m^f]. \quad (5)$$

We show now property SH is satisfied for such a normal vector. Substitution of $(N_G^f(\gamma^f), \gamma_1^f, .., \gamma_{M-1}^f) \in \mathcal{E}_G(A)$ on the LHS for $f \in B$ yields,

$$N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f \leq \sum_{f \in B} P^f [N_G^f(\gamma^f) - \lambda_m \gamma_m^f].$$

This implies that these inequalities are in fact equations for all $f \in B$, since this is the only way to prevent one of the sums on the RHS from being strictly higher than their weighted average on the LHS. This implies that $N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f$ can be plugged in to the right hand side of equation 5, which in turn establishes that, given $f, g \in B$, and $\gamma \in \Gamma$

$$N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^g = N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f,$$

as necessary for property SH. Again, equation 5 tells us that all $f \in B$, γ^f solves,

$$\max_{\gamma \in \Gamma} N_G^f(\gamma) - \sum_{m=1}^{M-1} \lambda_m \gamma_m,$$

as again required for SH. The final aspect of condition SH to confirm is that, given $f \in B$, $g \in A \backslash B$ and $\gamma \in \Gamma$

$$N_G^g(\gamma) - \sum_{m=1}^{M-1} \lambda_m \gamma_m \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \lambda_m \gamma_m^f,$$

This is again immediate from 5 since $(N^g(\gamma), \gamma_1, .., \gamma_{M-1}) \in \mathcal{E}_G(A)$.

Sufficiency: If property SH holds, it directly implies existence of a normal vector $(1, -\lambda_1, ..., -\lambda_{M-1})$ such that, given $(y_0, y_1, .., y_{M-1}) \in \mathcal{E}_G(A)$,

$$y_0 - \sum_{m=1}^{M-1} \lambda_m y_m \leq N_G^f(\gamma^f) - \lambda_m \gamma_m^f,$$

any $f \in B$. Applying lemma 1, this implies that all points $(N_G^f(\gamma^f), \gamma_1^f, .., \gamma_{M-1}^f) \in \mathcal{E}_G(A)$ are in the upper boundary of $\mathcal{E}_G(A)$. Hence this applies also to any convex combination of them such as that defined by $(\sum_{f \in B} P^f N_G^f(\gamma^f), \mu_1, .., \mu_{M-1}) \in \mathcal{E}_G(A)$, completing the proof. ■

Lemma 3 Given $G \in \mathcal{G}$ that is differentiable on Γ^I, the interior of Γ, strategy $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ with $\gamma_m^f \in (0,1)$ satisfies $(B, P, \gamma) \in \hat{\Lambda}_G^{(\mu,A)}$ if and only if it satisfies CT, ED, and UB:

A. **Common Tangent for Chosen Acts (CT)**: Given $f, g \in B$,

$$N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^f = N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^g(\gamma^g)}{\partial \gamma_m} \right] \gamma_m^g.$$

B. **Equal Derivative for Chosen Acts (ED)** : Given $f, g \in B$,

$$\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} = \frac{\partial N_G^g(\gamma^g)}{\partial \gamma_m}.$$

C. **Unchosen Act Bound (UB) :** Given $f \in B$ and $g \in B \backslash A$,

$$N_G^g(\gamma^g) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^g \leq N_G^f(\gamma^f) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m^f,$$

where $\gamma^g \in \Gamma$ maximizes on $N_G^g(\gamma) - \sum_{m=1}^{M-1} \left[\frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m} \right] \gamma_m$ on $\gamma \in \Gamma$.

Proof. In light of lemma 2, the first part requires us to show that, when $G \in \mathcal{G}$ is differentiable, property SH is satisfied for $(B, P, \gamma) \in \Lambda^{(\mu,A)}$ with $\gamma_m^f \in (0,1)$

if and only if (B, P, γ) satisfies conditions ED, CT, and UB. That these three conditions are sufficient for property SP is immediate using $\lambda_m = \frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m}$ for any $f \in B$ and applying UB. That they are necessary for SP to be satisfied in cases with $\gamma_m^f \in (0, 1)$ and with G differentiable derives from the fact that SP certainly requires that, for each $f \in B$, γ^f solves,

$$\max_{\gamma \in \Gamma} N_G^f(\gamma) - \sum_{m=1}^{M-1} \lambda_m \gamma_m.$$

Given that $\gamma_m^f \in (0, 1)$ and that $G \in \mathcal{G}$ is differentiable, solving this problem requires $\lambda_m = \frac{\partial N_G^f(\gamma^f)}{\partial \gamma_m}$. Given this, SP directly implies CT, ED, and UB as illustrated in the proof of lemma 1. ■

10.2 Proofs of Corollaries

All corollaries apply more generally than to the Shannon model. However for consistency with the text they are stated only for this case. The appropriate generalization to separable cost functions is in each case clear.

Corollary 1 (Locally Invariant Posteriors - LIP): If $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $(C, Q, \eta) \in \Lambda^{(\rho,A)}$ with $C \subset B$ satisfies $\eta^f = \gamma^f$ all $f \in C$, then $(C, Q, \eta) \in \hat{\Lambda}^{(\rho,A)}$.

Proof. Note by the necessity aspect of lemma 2 that if $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ then condition SH is satisfied. Neither the prior $\mu \in \Gamma$ nor the probability map $P : B \to \mathbb{R}$ feature in condition SP, while deletion of acts can only weaken the check. Hence if $(C, Q, \eta) \in \Lambda^{(\rho,A)}$ with $C \subset B$ satisfies $\eta^f = \gamma^f$ all $f \in C$, condition SH remains valid and the sufficiency aspect of lemma 2 implies it is optimal. ■

Corollary 2 (Envelope Condition): Given $(\mu, A) \in \Gamma \times \mathcal{F}$ such that $\mu_m > 0$, the value function $V^A : \Gamma \longrightarrow \mathbb{R}$ is differentiable at μ and has continuous partial derivatives,

$$\frac{\partial V^A(\mu)}{\partial \mu_m} = \frac{\partial N^f}{\partial \gamma_m}(\hat{\gamma}^f),$$

where $f \in \hat{B}$ some $(\hat{B}, \hat{P}, \hat{\gamma}) \in \hat{\Lambda}^{(\mu,A)}$.

Proof. Given $(\mu, A) \in \Gamma \times \mathcal{F}$, note that $\big(\mu_1, .., \mu_{M-1}\big)$ is in the interior of $X = \{(\mu_1, .., \mu_{M-1}) \in \mathbb{R}_+^{M-1} | \sum_{m=1}^{M-1} \mu_m \leq 1\}$. By lemma 1 an optimal policy exists. Consider a corresponding optimal strategy $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ that therefore achieves the value,

$$V^A(\mu) = \sum_{f \in B} P^f N^f(\gamma^f).$$

Define a composite act $\hat{h} \in F$ with state dependent payoffs,

$$U_m^{\hat{h}} = \sum P^f U_m^f.$$

Define the net payoff function to $N^{\hat{h}} : \Gamma \to \mathbb{R}$ in standard fashion, and apply the envelope theorem of Benveniste and Scheinkman [1979] to functions $V^A, N^{\hat{h}} : \Gamma \to \mathbb{R}$, noting that both are concave, that $V(\mu) \geq N^{\hat{h}}(\mu)$ on the interior of X, and that $V(\mu) = N^{\hat{h}}(\mu)$, and that $N^{\hat{h}}(\mu)$ is differentiable on the interior of X. With this we conclude that V^A is differentiable at μ and that,

$$\frac{\partial V^A}{\partial \mu_m}(\mu) = \frac{\partial N^{\hat{h}}}{\partial \gamma_m}(\mu) = \sum_{f \in B} P^f \frac{\partial N^f}{\partial \gamma_m}(\gamma^f).$$

By lemma 3, the ED conditions are satisfied,

$$f, g \in B \Longrightarrow \frac{\partial N^f}{\partial \gamma_m}(\gamma^f) = \frac{\partial N^g}{\partial \gamma_m}(\gamma^g),$$

completing the proof in light of $\sum_{f \in B} P^f = 1$. ■

Corollary 3 (States Bound Acts - SBA): Given $(\mu, A) \in \Gamma \times \mathcal{F}$, there exists a rationally inattentive strategy with $|B| \leq M$.

Proof. Consider $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ such that $|B| > M$. By the necessity aspect of lemma 2, condition SH is satisfied. This condition remains valid for any subset of acts $\tilde{B} \subset B$ with $\tilde{\gamma}^f = \gamma^f$ on $f \in \tilde{B}$. By the sufficiency aspect of lemma 2, $\left(\tilde{B}, \tilde{P}, \tilde{\gamma}\right) \in \hat{\Lambda}^{(\mu,A)}$ provided only that μ in the convex hull of the family of vectors $\left\{\gamma_m^f\right\}_{f \in \tilde{B}}$. Charateodory's theorem implies that we can reduce the cardinality of B to M while retaining μ in this convex hull, completing the proof. ■

Corollary 4 (Unique Posteriors): If $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $(B, Q, \rho) \in \hat{\Lambda}^{(\mu,C)}$, then $\gamma(f) = \rho(f)$ all $f \in B$.

Proof. Note first that if $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $B \subset C \subset A$, then $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,C)}$. To see this, note from the necessity aspect of lemma 2 that if $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$, then condition SH is satisfied. Since $B \subset C \subset A$, $(B, P, \gamma) \in \Lambda^{(\mu,C)}$ and condition SH is still satisfied. Hence $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,C)}$ follows in light of the sufficiency aspect of lemma 2. We conclude that since $(B, P, \gamma) \in \hat{\Lambda}^{(\mu,A)}$ and $(B, Q, \rho) \in \hat{\Lambda}^{(\mu,C)}$, then $(B, P, \gamma), (B, Q, \rho) \in \hat{\Lambda}^{(\mu,B)}$.

Given $f \in B$ define $R^f = \frac{P^f + Q^f}{2}$ and $\eta^f \in \Gamma$ by,

$$\eta_m^f = \frac{P^f \gamma_m^f + Q^f \rho_m^f}{P^f + Q^f}.$$

By construction, $(B, \eta, R) \in \Lambda^{(\mu,B)}$. If $\gamma(f) \neq \rho(f)$ some $f \in B$, we can apply the strict version of Jensen's inequality as in lemma 1 to establish the contradiction that net utility must be strictly higher at (B, η, R) than at either (B, P, γ) and (B, Q, ρ),

$$\sum_{f \in B} R^f N^f(\eta^f) > \lambda \sum_{f \in B} P^f N^f(\gamma^f) + (1 - \lambda) \sum_{f \in B} Q^f N^f(\rho^f).$$

■

11 Comparison with KKT Conditions

Following Matejka and McKay [2011], consider the constrained optimization problem of maximizing expected prize utility less Shannon attention costs, subject to constraints associated with rational expectations, with act-specific posteriors adding to unity, and with probabilities being non-negative. Let $\pi \in \mathbb{R}^M$ be the multipliers on the rational expectations constraints and $\eta : A \to \mathbb{R}$ the multipliers on posteriors. With act set A countable, the associated Lagrangean is,

$$\mathcal{L} = \sum_{f \in A} P^f \sum_{m=1}^{M} \gamma_m^f (U_m^f - \kappa \ln \gamma_m^f) - \sum_{m=1}^{M} \pi_m (\sum_{f \in A} P^f \gamma_m^f - \mu_m) - \eta^f \ (\sum_{m=1}^{M} \gamma_m^f - 1).$$

Treating this using standard KKTcondition, a necessary condition for (B, P, γ) to be rationally inattentive is that there exists $\hat{\pi} \in \mathbb{R}^M$, $\hat{\eta} : A \to \mathbb{R}$, and posteriors $\gamma_m^f \in \Gamma$ for $f \in A/B$ such that conditions KKT 1, KKT 2, and KKT 3 are satisfied:

KKT1: For $f \in B$,

$$P^f \left[U_m^f - \kappa \ln \gamma_m^f - \kappa - \hat{\pi}_m \right] - \hat{\eta}^f \ = 0 \text{ for } 1 \le m \le M.$$

KKT2: For $f \in B$, if

$$\begin{aligned} P^f &\in (0,1) \Longrightarrow \sum_{m=1}^{M} \gamma_m^f \left(U_m^f - \hat{\pi}_m - \kappa \ln \gamma_m^f \right) = 0; \\ P^f &= 1 \Longrightarrow \sum_{m=1}^{M} \gamma_m^f \left(U_m^f - \hat{\pi}_m - \kappa \ln \gamma_m^f \right) \ge 0. \end{aligned}$$

KKT3: For $f \in A/B$,

$$\sum_{m=1}^{M} \gamma_m^f \left(U_m^f - \hat{\pi}_m - \kappa \ln \gamma_m^f \right) \leq 0.$$

The reason that these KKT conditions do not characterize rationally inattentive policies is that the objective function is not concave in the choice variables, involving as it does product terms as between beliefs and posteriors. As a result, one can find non-optimal solutions. To illustrate, consider the case in the text with two acts, f and g, with $\kappa = 1$, and with $U_1^f = U_2^g = \ln(1+e)$ and $U_2^f = U_1^g = 0$. Note that an attention strategy can be fully specified by P^f and $\gamma_1^{f,g} \in [0,1] \geq 0$. Now consider the equal prior $\mu = 0.5$ and note that the following strategy is feasible and, together with the specified multipliers, satisfies all KKT necessary conditions:

$$(P^f, \gamma_1^f, \gamma_1^g) = (1, 0.5, 0.5); \hat{\pi}_1 = \ln(1+e) + \ln 2; \hat{\pi}_2 = \ln 2, \hat{\eta}^1 = \hat{\eta}^2 = 0.$$

Yet $(P^f, \gamma_1^1, \gamma_1^2)$ is not optimal, since net utility to the feasible triple $(0.5, \frac{1+e}{2+e}, \frac{1}{2+e})$ is strictly higher,

$$N(0.5, \frac{1+e}{2+e}, \frac{1}{2+e}) = \left(\frac{1+e}{2+e}\right) \ln(1+e) - \ln 0.5 > \frac{\ln(2+e)}{2} - \ln 0.5 = N(1, 0.5, 0.5).$$