Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy*

Andrew Caplin†, Mark Dean‡, and John Leahy§

February 2019

Abstract

We provide a full behavioral characterization of the standard Shannon model of rational inattention. The key axiom is "Invariance under Compression", which identifies this model as capturing an ideal form of attention-constrained choice. We introduce tractable generalizations that allow for many of the known behavioral violations from this ideal, including asymmetries and complementarities in learning, context effects, and low responsiveness to incentives. We provide an even more general method of recovering attention costs from behavioral data. The data set in which we characterize all behavioral patterns is "state dependent" stochastic choice data.

1 Introduction

Understanding limits on private information has been central to economic analysis since the pioneering work of Hayek [1937, 1945]. While there are many models of information acquisition in use (see Hellwig *et al.* [2012]), a major new route to such understanding was initiated by Sims [1998, 2003], who introduced the theory of rational inattention. He considered the implications of attention costs based on Shannon mutual information for macroeconomic dynamics. The ensuing period has seen applications of the Shannon cost function to such diverse subjects as stochastic choice (Matejka and McKay [2015]), investment decisions (Mondria [2010]), global games (Yang [2015]), pricing decisions (Woodford [2009], Mackowiak and Wiederholt [2009] and Matějka [2015]), dynamic learning (Steiner *et al.* [2015]) and social learning (Caplin *et al.* [2015]).

One reason for the appeal of the Shannon cost function is analytic tractability (Matejka and McKay [2015], Caplin *et al.* [2018]). This makes it the go-to model not only in studies of individ-

*We thank Sandro Ambuehl, Dirk Bergemann, Daniel Csaba, Tommaso Denti, Henrique de Oliveira, Xavier Gabaix, Sen Geng, Andrei Gomberg, Michael Magill, Daniel Martin, Filip Matejka, Alisdair McKay, Stephen Morris, Efe Ok, Larry Samuelson, and Michael Woodford for their constructive contributions. This paper builds on the material contained in the working paper "The Behavioral Implications of Rational Inattention with Shannon Entropy" by Andrew Caplin and Mark Dean [2013], and subsumes all common parts.

†Center for Experimental Social Science and Department of Economics, New York University. Email: andrew.caplin@nyu.edu

‡Department of Economics, Columbia University. Email: mark.dean@columbia.edu

§Department of Economics and the Gerald R. Ford School of Public Policy, University of Michigan and NBER Email: jvleahy@umich.edu

ual decision making and error patterns, but also for applications of limited attention to market settings.[1] A second reason for its appeal is that, while it gives rise to highly sophisticated behaviors[2] the resulting model appears to capture key qualitative aspects of costly yet flexible attention. A third is that Shannon costs have a justification in information theory: Mutual information is related to the rate of information flow needed to generate a given conditional distribution of signals given a distribution of states, assuming optimal coding (see for example Cover and Thomas [2012] chapter 10). Pioneering work by Shannon [1948] and Khinchin [1957] provides direct axiomatizations of Shannon entropy as a measure of disorder, characterizing its 'ideal' nature. This provides grounds for having special interest in the Shannon model as representing a well calibrated attentional machine (see for example Sims [2003] and Matejka and McKay [2015]).

Despite all of these positives, the experimental literature in economics and psychology establishes that there are important behavioral reasons to look beyond that model. In fact it is now known that behavior often violates key features of the Shannon model. These include: its implication that all states are equally easy to identify and to discriminate among (see Dewan and Neligh [2017] for behavioral counterexamples); the implied flexibility of response to payoff incentives (see Caplin and Dean [2013] for behavioral counterexamples); and essential independence of behavior from event likelihoods (see Woodford [2012] for behavioral counter examples).

In this paper we address two central questions the above highlights. First, there is clearly need for more flexible models that capture features that the Shannon model does not. How can one allow for rich behavioral features while retaining at least some of the tractability that makes the Shannon model so useful? Second, is there a behavioral sense in which the Shannon model is 'ideal', as suggested by the information theory literature? If there is, can this property guide us on the situations in which the model is most likely to capture key properties of behavior?

We give positive answers to both questions by providing a complete characterization of the behavior consistent with rational inattention for a nested class of cost functions which includes Shannon as the most restrictive case. This answers the first question by identifying the broadest class of models consistent with the "concavification" operation that is in heavy use in areas of rational inattention and Bayesian persuasion (see Caplin and Dean [2013], Gentzkow and Kamenica [2014], Steiner *et al.* [2015], Clark [2016], Morris and Strack [2017], and Hébert and Woodford [2017]). These 'Posterior Separable' (PS) models all allow standard tools of convex analysis to be used for purposes of solution. Moreover there is good reason to believe that these cost functions may lend themselves to efficient algorithmic computation along the lines of the Blahut-Arimoto algorithm.[3] While tractable, PS models allow for essentially all behaviors uncovered to date that contradict the Shannon model such as: asymmetric costs of learning about distinct states; differing perceptual distance between distinct states; complementarities in learning about distinct states; and complex responses to payoff incentives. They also allow the cost of obtaining a given posterior to depend on prior beliefs and hence vary from context to context.[4] In this sense, they represent a reasonable compromise between tractability and behavioral flexibility. Finally, both Morris and Strack [2017], and Hébert and Woodford [2017] show that cost functions in this class are consistent with models of optimal sequential learning (see section 9.3).

[1] Recent examples include Caplin *et al.* [2015], Martin [2017] and Ravid [2017].

[2] Such as the incomplete consideration of options (Caplin *et al.* [2018]) or attentional discrimination (Bartoš *et al.* [2016]).

[3] We owe this point to Daniel Csaba who is actively researching the conditions under which this holds.

[4] Note that models in which the cost of posteriors is invariant to the prior have the feature that costs of a particular experiment - i.e. a distribution of signals in each state - depends on the prior.

With regard to the second question, our characterization of the Shannon model introduces a new behavioral axiom that pinpoints one respect in which the Shannon cost function is 'ideal' within the separable class. This axiom, Invariance Under Compression (IUC), constrains behavior to be efficient in a particular sense: it asserts that behavior is invariant to changes in the decision problem that leave the probabilistic structure of payoffs unchanged. The nature of the state space, per se, is behaviorally irrelevant. We show not only that the Shannon model implies IUC, but that this axiom is enough to identify this model within the separable class: adding IUC to the axioms which characterize separability is both necessary and sufficient for Shannon. In this sense, the Shannon model alone produces an idealized form of attention-constrained behavior in which only payoff relevant information matters. While the necessity of IUC follows immediately from prior work, its sufficiency is non-trivial and, to us, surprising. It means that IUC cleanly identifies the Shannon model, and so implies the myriad other features that have been identified in the theoretical and experimental literatures.

In addition to these two contributions, our 'soup-to-nuts' behavioral characterization provides a constructive method of recovering attention costs from behavioral data for a much broader class of cost functions. This method is both general and intuitively reasonable, resting as it does on standard balancing of marginal costs and marginal utility. The cost function is fully pinned down as long as three axioms are satisfied. The first two are necessary for existence of a rationalizing cost function of any form: no improving action switches (NIAS: see Caplin and Martin [2015]) and no improving attention cycles (NIAC; see Caplin and Dean [2015]). The final axiom requires "completeness" of the behavioral data: broadly speaking, all possible information structures must be observed in some decision problem.

There is other work that characterizes the Shannon cost function, albeit in very different manners. First, our axioms are related in a spiritual sense to the classic characterizations of Shannon entropy summarized in Csiszár [2008]. However this literature is concerned with placing axioms directly on measures of disorder and these do not have clear behavioral counterparts. For example, IUC might be viewed as bearing a superficial resemblance to 'recursivity', which has been used as a pivotal axiom in the characterization of Shannon costs. Yet this resemblance is only skin deep. Recursivity places conditions directly on measures of disorder (or information costs in our framework): it states that the difference in costs between two different distributions of posteriors must be a particular multiple of the cost of a distinct posterior distribution. The interpretation is that it is as costly to convey information directly as it is to convey it indirectly. The fact that this does not translate directly to behavior is not a criticism of the axiom, but rather a comment on the vast gulf between our behavioral characterization and classical cost based axiomatizations. Interestingly, as further detailed in section 9, the direct implication of our IUC axiom for the cost function bears no relation to the logic of recursivity. It is therefore possible that our characterization of behavior could be used to build a new axiomatic characterization of the Shannon cost function. It is hard to see how one could travel in the reverse direction and use any of the existing cost-based axiomatizations to develop a behavioral characterization.

A second related strand of the literature uses first order conditions to solve the Shannon model (for example Stevens [2014], Matejka and McKay [2015], Steiner *et al.* [2015], Caplin *et al.* [2018]), showing that the resulting behavior is similar to the classic logit model. While complementary to our work, these papers do not address the two fundamental issues we raise above. First, because only the Shannon model is characterized, they say nothing about relaxations that could better capture observed behavior. Second, these first order conditions do not make clear the sense in which the Shannon model captures ideal information acquisition.

Central to our approach is a particular specification of the choice data available to an ideal observer, such as an econometrician or economic theorist. The data set that we study is "state-dependent" stochastic choice (SDSC) data, as introduced in Caplin and Martin [2015] and Caplin and Dean [2015]. This treats both the payoff determining states of the world and the behavioral choice as observable. It rests on the idea that attentional constraints do not apply to an ideal observer. While consumers may have difficulty assessing whether or not sales tax is included in the price paid at the register, the econometrician knows (Chetty *et al.* [2009]). The resulting data strongly reflects the match between perception and reality. In fact our results show that SDSC data can capture the full behavioral footprint of attention costs, in stark contrast with standard stochastic choice data. de Oliveira [2014] considers the behavioral implications of the Shannon model, but for a data set which consists of observed choices over different menus of alternatives.

Our work is also related to a growing literature aimed at understanding the behavioral implications of models with limited attention. Notable recent contributions include Masatlioglu *et al.* [2012], Manzini and Mariotti [2014], Oliveira *et al.* [2017] and Steiner and Stewart [2016]. Relying as it does on stochastic choice data, our work also links in with the recent renewed interest in modelling random choice in general (e.g. Agranov and Ortoleva [2017], Manzini and Mariotti [2016], Apesteguia *et al.* [2017]), and it's relationship to information acquisition in particular (e.g. Krajbich and Rangel [2011]).

Section 2 defines attention strategies in analytically appropriate form, and introduces the various classes of attention cost functions. Section 3 establishes general applicability of Lagrangian methods of identifying optimal strategies. Section 4 introduces SDSC data and links it to attention strategies. Section 5 introduces our IUC axiom and the associated characterization theorem. Section 6 introduces the recoverability result. Our characterizations of the PS and UPS models are in Section 7. Section 8 provides additional analyses concerning alternative formulations of the representation theorems and the properties of our axioms. Section 9 relates our work to the existing literature on attention. Section 10 concludes. Throughout the paper we present the main Theorems and discuss informally why they are true. Formal proofs are in the Appendix.

2 Attention Strategies and Costs

2.1 A Note on Notation

Before proceeding to the model, it is useful to preview a few of the notational conventions that will be in force throughout the paper. Our focus is on the conditions under which observed behavior is consistent with the predictions of a theoretical model. Many objects therefore appear twice: once as objects implied by theory and once as objects implied by the data. We will use a subscript P to denote data objects. For example, if theory implies that an agent chooses a posterior beliefs γ, we will use γ_P to denote the posterior beliefs implied by the observed choices.

We will also need to move back and forth between strategies, which are the natural theoretical objects, and observed choices, which are the natural data objects. We use bold letters to denote operators that transform strategies into data and data into strategies. $\mathbf{P}_\lambda$ is then the data generated by the strategy λ and $\boldsymbol{\lambda}_P$ is the strategy implicit in the data P.

We let Γ denote the set of all distributions over the set of states of the world Ω. $\Gamma(\cdot)$ is then used throughout to restrict Γ in a way implied by its argument. For example, given μ, a distribution

over states of the world, $\Gamma(\mu)$ is the restriction to distributions that are absolutely continuous with respect to μ. Given Q, a distribution over distributions, $\Gamma(Q)$ is the support of Q.

Hats will be used throughout to determine sets of optimal choices. For example, if Λ is the set of feasible policies, then $\hat{\Lambda}$ will be the set of optimal policies. Tildes will be used to denote interiors of sets. For example, if $\Gamma(\mu)$ is the set of distributions that are absolutely continuous with respect to μ, $\tilde{\Gamma}(\mu)$ is the set of distributions that place positive weight on all elements in the support of μ. We use script letters to denote universal sets. For example, $\mathcal{A}$ is the set of all possible actions and $\mathcal{D}$ is the set of all possible decision problems.

Section 11 provides a complete list of notation including the point in the paper where that notation is defined.

2.2 Posterior-Based Attention Strategies

We consider a decision maker (DM) who faces a large class of decision problems related to an infinite (countable or uncountable) underlying set Ω of conceivable states of the world and an uncountably infinite set of potentially available actions, $\mathcal{A}$. In a given decision problem, the DM is endowed with a prior with finite support as well as a finite set of available actions. When taking action $a \in \mathcal{A}$ in state $\omega \in \Omega$, the DM receives a prize with a known, state independent utility. We denote the utility of the prize received when $a \in A$ is chosen and the state is $\omega \in \Omega$ as $u(a, \omega)$.[5]

Definition 1 *Given $\mu \in \Delta(\Omega) \equiv \Gamma$, $\Omega(\mu) \equiv \{\omega \in \Omega | \mu(\omega) > 0\}$ specifies possible states (where Δ denote simple distributions over the space); $\Gamma(\mu) = \{\gamma \in \Gamma | \Omega(\gamma) \subset \Omega(\mu)\}$ possible posteriors; and $\tilde{\Gamma}(\mu) = \{\gamma \in \Gamma(\mu) | \Omega(\gamma) = \Omega(\mu)\}$ interior posteriors with precisely the same support as μ.*

Definition 2 *A **decision problem** comprises a pair $(\mu, A) \in \Gamma \times \mathcal{A}$ with $A \subset \mathcal{A}$ finite. We assume that $\mathcal{A}$ is **rich**: For any function $f : \Omega \to \mathbb{R}$ there exists $a \in \mathcal{A}$ such that $u(a, \omega) = f(\omega)\ \forall\ \omega \in \Omega$. We define $\mathcal{D}$ as the set of decision problems.*

The key role of the richness assumption is that it means that, for the class of cost functions we study, a rich set of posterior beliefs will be form part of an optimal strategy in some decision problem. Indeed, with Shannon costs, all posteriors will be optimal in some decision problem. We make use of this feature at a number of points, most obviously in Theorem 2.

The central decision that we model concerns how much to learn. The DM decides this by comparing the incremental improvement in decision quality associated with improved information with the cost of incremental information. In formalizing the cost of learning, we will focus on the outcome of the learning process and assign costs directly to each Bayes-consistent distribution of posteriors, as in Caplin and Dean [2015]. To this end, we define an attention strategy in terms of the resulting posteriors and their implications for choice.

Definition 3 *Given $(\mu, A) \in \mathcal{D}$, the set of **posterior-based attention strategies** comprises all*

[5] While we assume that this utility function is observable, one could alternatively recover it from choice data using standard techniques: for example by assuming an Anscombe-Aumann type set up, assuming expected utility, and recovering utilities from choices over degenerate acts that supply the same lottery in each state.

simple probability distributions over posteriors and corresponding mixed action strategies,

$$\Lambda(\mu, A) \equiv \{\lambda = (Q_\lambda, q_\lambda) | Q_\lambda \in \mathcal{Q}(\mu),\ q_\lambda : \Gamma(Q_\lambda) \to \Delta(A)\},$$

with $\mathcal{A}(\lambda) \subset A$ *the chosen actions,* $\Gamma(Q_\lambda)$ *the support of* Q_λ *in* Γ *and* $\mathcal{Q}(\mu)$ *the Bayes-consistent distributions,*

$$\mathcal{Q}(\mu) = \{Q \in \Delta(\Gamma(\mu)) |\ \mu = \sum_{\gamma \in \Gamma(Q)} \gamma Q(\gamma)\}.$$

We also define $\Lambda^I(\mu) \subset \Lambda(\mu)$ *as the set of* ***inattentive strategies*** *such that* $\Gamma(Q_\lambda) = \mu$.

Here Q is a distribution over posteriors and q specifies a distribution over actions for each posterior in the support of Q. This posterior-based approach departs from the standard signal-based approach which specifies the cost of an available set of signals correlated with the true state of the world (see for example Caplin and Dean [2015]). There are two key advantages of the posterior-based formulation. First, our behavioral characterizations are more naturally stated in terms of posteriors. Second, this formulation allows for several interesting generalizations of the Shannon cost function. Of course, there is in general a mapping between signals and posteriors. We discuss in Section 8 why behavioral results are independent of how strategies are formulated.

Figure 1 illustrates the strategy λ^* which we use as a running example. The underlying decision problem consists of a prior μ with two states in its support, $\Omega(\mu) = \{\omega_1, \omega_2\}$, each of which is equally likely.[6] The support of the strategy comprises two posteriors, $\Gamma(Q_{\lambda^*}) = \{\gamma^a, \gamma^b\}$:

$$\gamma^a = \begin{pmatrix} 0.8 \\ 0.2 \end{pmatrix} \quad \gamma^b = \begin{pmatrix} 0.4 \\ 0.6 \end{pmatrix};$$

and specifies $Q_{\lambda^*}(\gamma^a) = 0.25$ and $Q_{\lambda^*}(\gamma^b) = 0.75$. Actions a and b are chosen deterministically from γ^a and γ^b respectively, $q_{\lambda^*}(a|\gamma^a) = q_{\lambda^*}(b|\gamma^b) = 1$.

[6] We use the notation

$$\gamma = \begin{pmatrix} \gamma(\omega_1) \\ \gamma(\omega_2) \end{pmatrix}$$

to describe probability distributions.

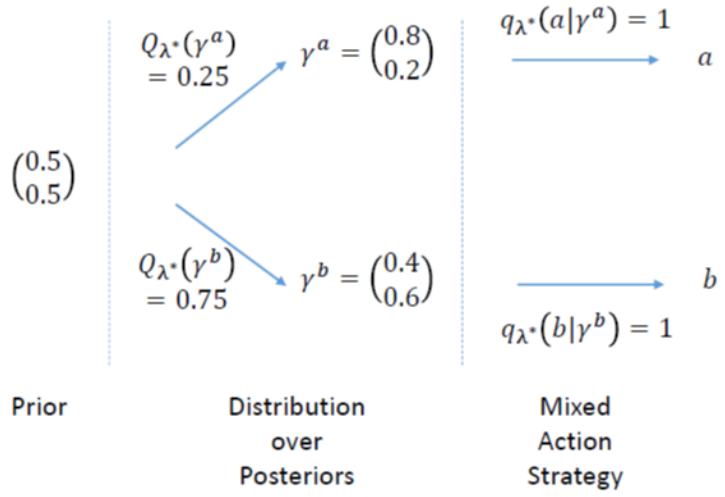

Figure 1: Strategy λ^*

2.3 Utility, Costs and Optimal Strategies

The goal of the DM is to maximize prize-based expected utility (EU) net of additively separable attention costs. Given $\lambda \in \Lambda(\mu, A)$, prize based EU is computed in the standard manner,

$$U(\lambda) \equiv \sum_{\gamma \in \Gamma(Q_\lambda)} \sum_{a \in A} Q_\lambda(\gamma) q_\lambda(a|\gamma) \bar{u}(\gamma, a),$$

where $\bar{u}(\gamma, a)$ is expected utility conditional on the posterior γ

$$\bar{u}(\gamma, a) \equiv \sum_{\omega \in \Omega(\mu)} \gamma(\omega) u(a, \omega). \tag{1}$$

Attention costs for strategy $\lambda \in \Lambda(\mu, A)$ depend only on the distribution of posteriors $Q_\lambda \in \mathcal{Q}(\mu)$. We assume that inattention is always possible, and normalize its cost to zero. We allow for the possibility that some distributions of posteriors are infeasible by setting their costs to infinity. For example, there are interesting cases in which it is prohibitively costly to entirely rule out ex ante possible states, so that it is infeasible to choose posteriors on the boundary of $\Gamma(\mu)$.

Definition 4 *We define $\mathcal{F}$ as the set of all priors and Bayes' consistent posterior distributions,*

$$\mathcal{F} = \{(\mu, Q) | \mu \in \Gamma, Q \in \mathcal{Q}(\mu)\}. \tag{2}$$

We define $\mathcal{K}$ as the set of all ***attention cost functions*** *$K : \mathcal{F} \to \bar{\mathbb{R}}$ such that $K(\mu, Q_\lambda) = 0$ for $\lambda \in \Lambda^I(\mu)$.*

The value of strategy $\lambda \in \Lambda(\mu, A)$ is computed based on additive separability of prize utility and attention costs.

$$V(\mu, \lambda|K) \equiv U(\lambda) - K(\mu, Q_\lambda).$$

The value function and corresponding optimal strategies are then defined as:

$$\begin{aligned} \hat{V}(\mu, A|K) &\equiv \sup_{\{\lambda \in \Lambda(\mu, A)\}} V(\mu, \lambda|K); \\ \hat{\Lambda}(\mu, A|K) &\equiv \left\{\lambda \in \Lambda(\mu, A) \,|V(\mu, \lambda|K) = \hat{V}(\mu, A|K)\right\}. \end{aligned}$$

2.4 The Shannon Cost Function

By far the best studied cost function that can be expressed directly in terms of priors and posteriors is the Shannon function, in which the costs are linear in the mutual information between prior and posteriors. It is standard that one can compute mutual information by comparing the Shannon entropy of the prior, $H(\mu) = -\sum_{\omega \in \Omega(\gamma)} \mu(\omega) \ln \mu(\omega)$, to the expected Shannon entropy of the posteriors. In translating this into an attention cost function, note that what is costly is increasing predictability, or **reducing** entropy. Given $(\mu, Q) \in \mathcal{F}$, the Shannon attention cost function K^S_κ with multiplicative factor $\kappa > 0$ is therefore specified as,

$$K^S_\kappa(\mu, Q) \equiv \kappa \left[\sum_{\gamma \in \Gamma(Q)} Q(\gamma)\left[-H(\gamma)\right] - \left[-H(\mu)\right]\right] = \kappa \left[\sum_{\gamma \in \Gamma(Q)} -Q(\gamma)H(\gamma) + H(\mu)\right]. \qquad (3)$$

By way of illustration, consider attention strategy λ^* from Figure 1. Figure 2 records the probability of state ω_1 on the horizontal axis. The Figure reflects the fact that Shannon entropy is strictly concave and symmetric around its maximized value at uniformity and that it is zero at the end-points of the interval (since $\lim_{x \downarrow 0} x \ln x = 0$), at which it has unbounded derivative. Following (3), we shift up the negative of the entropy function, which is strictly convex, to zero at the prior of 0.5. The cost of strategy λ^* is then found as the height of the chord joining the points on the function corresponding to the two possible posterior likelihoods of ω_1 (0.4 and 0.8) as it passes over the prior, as Figure 2 illustrates.

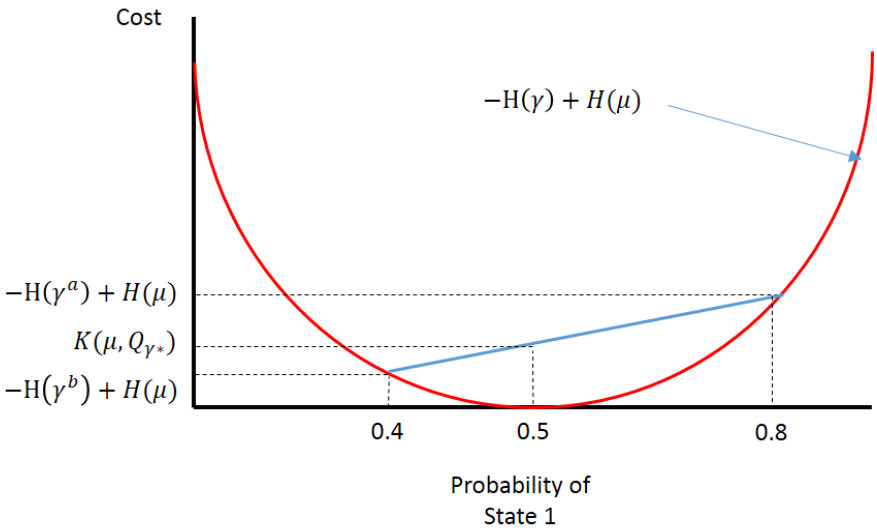

Figure 2: Cost of Strategy λ^*

Note that the Figure shows that all attentive strategies have strictly positive cost.

2.5 PS Cost Functions

The posterior-separable (PS) cost functions we study have the same form as (3), yet generalize the underlying measure of disorder, or "entropy", of the probability distribution over prior possible states of the world. The only properties that are retained relate to the strict convexity of this function and the specification of inattention as feasible and free.

Before introducing PS functions, we introduce the optimal posterior set. For later purposes we also introduce the optimal distributions of posteriors and the corresponding restricted domain for the cost function.

Definition 5 *Given $K \in \mathcal{K}$ we define the **optimal posterior set** $\hat{\Gamma}(\mu|K)$ for every $\mu \in \Gamma$ as*

$$\hat{\Gamma}(\mu|K) = \{\gamma \in \Gamma | \exists\, (\mu, A) \in \mathcal{D} \text{ and } \lambda \in \hat{\Lambda}(\mu, A|K) \text{ with } \gamma \in \Gamma(Q_\lambda)\}, \tag{4}$$

We define $\hat{\mathcal{Q}}(\mu|K) \equiv \mathcal{Q}(\mu) \cap \Delta(\hat{\Gamma}(\mu|K))$ and $\hat{\mathcal{F}}(\mu|K)$ as the subset of $\mathcal{F}(\mu)$ consistent with optimality,

$$\hat{\mathcal{F}}(\mu|K) = \left\{(\mu, Q) \in \mathcal{F}(\mu) | Q \in \hat{\mathcal{Q}}(\mu|K)\right\}.$$

It is the fact that all state dependent utility vectors are possible for any given utility function that makes the set $\hat{\Gamma}(\mu|K)$ independent of the utility function. This function determines only which actions give which payoffs in which states not the union over all decision problems.

Definition 6 *An attention cost function is **posterior-separable (PS)**, $K \in \mathcal{K}^{PS}$, if, given $\mu \in \Gamma$, there exists a strictly convex function $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$, real-valued on $\tilde{\Gamma}(\mu)$, such that, given*

$Q \in \mathcal{Q}(\mu)$,

$$K(\mu, Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma) T_\mu(\gamma) - T_\mu(\mu), \tag{5}$$

and such that the optimal posterior set $\hat{\Gamma}(\mu|K)$*,is convex.*

To clarify fine points in the definition, note that allowing T_μ to take infinite values for boundary posteriors both covers various interesting forms of entropy (see section 8.3) and simplifies our behavioral characterization. Strict convexity in this case means that, given distinct posteriors γ_1,γ_2 at which T_μ is real-valued (the set *dom* T_μ in the notation of Rockafellar [1970] p. 23),

$$T_\mu(\alpha\gamma_1 + (1-\alpha)\gamma_2) < \alpha T_\mu(\gamma_1) + (1-\alpha) T_\mu(\gamma_2),$$

for all $\alpha \in (0,1)$: hence *dom* T_μ itself is a convex set. Our insistence that $\hat{\Gamma}(\mu|K)$ is also convex avoids complications associated with possible non-existence of sub-differentials on the boundary.[7]

As noted in Section 9, functions of the PS form have featured in the literature on measures of the information content of experiments following Blackwell [1951]. A straight forward result with this functional form is that the strict convexity of T_μ ensures that the corresponding measure strictly respects the Blackwell partial ordering of information content (see Torgersen [1991]).

In addition to allowing for general convex cost functions, note that this definition allows costs to differ arbitrarily across priors, e.g. according to the cardinality of the state space. Subtraction of $T_\mu(\mu)$ is a normalization which ensures that inattentive strategies are free as per the general definition. Note that there are many different T functions that give rise to precisely the same cost function. In particular, we show in the Appendix that K is invariant to the addition of affine functions of γ to T (Lemma 4.3).

2.6 UPS Cost Functions

While the PS case allows for arbitrary dependence of the cost function on the prior, the Shannon model does not exploit this freedom. Given distinct priors $\mu, \mu' \in \Gamma$, the function $T_\mu(\gamma)$ and $T_{\mu'}(\gamma)$ can be written in a manner that is independent of the prior. A fine point relates to the possibly infinite costs of ruling out ex ante possible states. Note that even with Shannon cost functions, the incremental cost of fully ruling out any prior possible state is unbounded at the margin. This means that there is not full independence between the prior and the cost of the corresponding posterior. However this dependence is limited. We can cover all such cases by insisting on a common T function only for posteriors consistent with optimality.

Definition 7 *A PS cost function* $K \in \mathcal{K}^{PS}$ *is* ***uniformly posterior-separable (UPS)****,* $K \in \mathcal{K}^{UPS}$*, if there exists a strictly convex function* $T : \Gamma \to \mathbb{R}$ *such that,*

$$K(\mu, Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma) T(\gamma) - T(\mu). \tag{6}$$

[7] In general, the set of posteriors at which sub-differentials exist (*dom* ∂T_μ in the notation of Rockafellar [1970], p. 227) need not be convex in particular contrived cases. Our results are most straight forward with $\hat{\Gamma}(\mu|K)$ convex, which holds for all standard forms of entropy. While both are convex, note that $\hat{\Gamma}(\mu|K)$ may be a strict subset of *dom* T_μ. For example, the Shannon cost function is real-valued on the convex set $\Gamma(\mu)$, while $\hat{\Gamma}(\mu|K^S_\kappa)$ comprises only interior posteriors, $\hat{\Gamma}(\mu|K^S_\kappa) = \tilde{\Gamma}(\mu)$.

for all $(\mu, Q) \in \hat{\mathcal{F}}(\mu|K)$.

Examples of cost functions which fall into the UPS category are those based on alternative measures of entropy, such as that introduced by Tsallis [1988]. We discuss the relationship between Tsallis and Shannon costs in Section 8.3.

3 PS Models, Optimal Strategies, and Lagrangians

In this section we identify optimal strategies using Lagrangian methods. We develop the geometric intuition in the body of the text, with technical arguments in Appendix 1.

3.1 Net Utility

Given $K \in \mathcal{K}^{PS}$ we establish that optimal strategies always exist and that there are Lagrangian methods of characterizing all optimal strategies. Yet the fact that costs can depend on the prior in the PS model gives rise to certain notational complexities. Hence for expository purposes, we focus on the UPS case, noting at the end that the approach generalizes to the PS case.

The key geometric observation is that the value of any given strategy, modulo the normalizing factor $T(\mu)$, can be decomposed into action specific net utilities, $N^a(\gamma)$,

$$N^a(\gamma) \equiv \bar{u}(\gamma, a) - T(\gamma). \tag{7}$$

To confirm, note that since $\sum_{a \in A} q_\lambda(a|\gamma) = 1$ for all $\gamma \in \Gamma(Q_\lambda)$,

$$\begin{aligned} V(\mu, \lambda|K) + T(\mu) &= \sum_{\gamma \in \Gamma(Q_\lambda)} \sum_{a \in A} Q_\lambda(\gamma) q_\lambda(a|\gamma) \bar{u}(\gamma, a) - \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) \sum_{a \in A} q_\lambda(a|\gamma) T(\gamma) \\ &= \sum_{\gamma \in \Gamma(Q_\lambda)} \sum_{a \in A} Q_\lambda(\gamma) q_\lambda(a|\gamma) N^a(\gamma). \end{aligned}$$

Hence optimal strategies can be identified as those that maximize the weighted averages of net utilities.

The net utility approach has simple geometric content. In Figure 3 we illustrate action-specific net utilities in a simple two-state case with $\Omega(\mu) = \{\omega_1, \omega_2\}$ and $\mu(\omega_1) = 0.5$. The probability of state 1 is on the horizontal axis. The red, dashed line graphs $T(\gamma)$ as a function of $\gamma(\omega_1)$. The green line represents the prize-based expected utility of an action a in which we have assumed that: $u(a, \omega_1) = 1$ and $u(a, \omega_2) = 0$. To compute net utility we simply subtract the cost from the benefit (for clarity in the Figure we illustrate $N^a(\gamma) + T(\mu)$ which allows us to see the tangency of net costs with gross costs when $\gamma = \mu$). The result is the blue line in the Figure. Note that since net utility is the difference between a line and a strictly convex function, it is strictly concave.

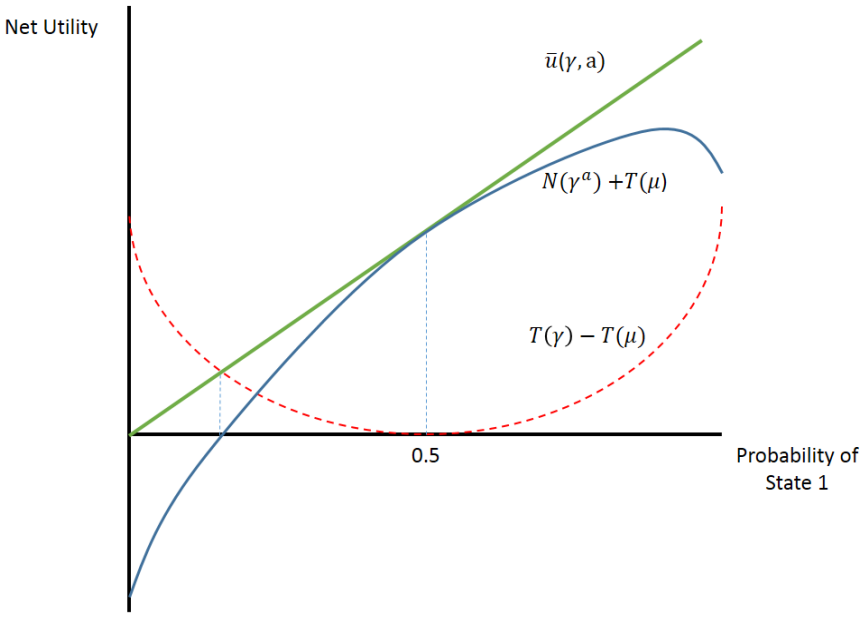

Figure 3: Net Utility of Action a.

Figure 4 illustrates net utilities for a decision problem (μ, A) with two equiprobable states and two actions, $A = \{a, b\}$. The second action is the mirror image of the first, with $u(b, \omega_1) = 0$ and $u(b, \omega_2) = 1$. We illustrate in the Figure computation of the net utility of strategy λ^*. Precisely as when computing the cost, the value is found by joining the points on the net utility function corresponding to possible posteriors with a chord, and finding the value of the chord as it passes over the prior. Thinking of all such chords identifies optimal strategies as defined by the posteriors that support the highest chord passing over the prior. In Figure 4 posteriors $\hat{\gamma}^a$ and $\hat{\gamma}^b$ have this property, and so form the support of an optimal strategy for this decision problem. Note that our example strategy λ^* is non-optimal, since the corresponding chord passes strictly below the top

chord.

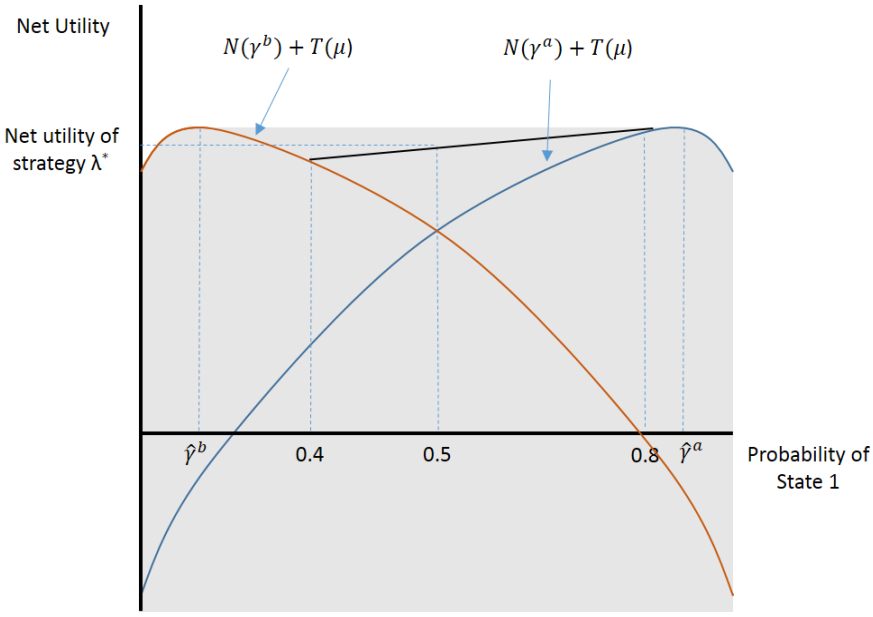

Figure 4: Net Utility of Strategy λ^*

3.2 Lagrange Multipliers and the PS model

The shaded area in Figure 4 is the lower epigraph of the concavified net utility function, defined as the minimal concave function that majorizes all net utilities (Rockafellar [1970]). The applicability of Lagrangian methods rests on the fact that the lower epigraph is always a convex set. This is geometrically clear in the simple case illustrated in Figure 4, and applies quite generally. Indeed, the same geometric approach works not only for UPS cost functions, but also for PS cost functions, in which net utilities are specific to the prior. For PS cost functions, we fix the prior μ and again define action-specific net utilities as $N^a_\mu(\gamma)$,

$$N^a_\mu(\gamma) \equiv \bar{u}(\gamma, a) - T_\mu(\gamma). \tag{8}$$

The key geometric observation is that one can still compute optimal strategies by appropriately averaging these action and prior specific net utilities. Hence identical convex analytic methods apply.

The geometric approach in Figure 4 is completely general. There is one important point to note in so generalizing, which derives from the adding up constraint on probabilities. Given this constraint, Figure 4 represents a two-dimensional state space in one dimension. This transformation is of great general value. Given $\mu \in \Gamma$ with $|\Omega(\mu)| = J$, we transform $\Omega(\mu)$ into the equivalent subspace of $\mathbb{R}^{J-1}$. To simplify, we give all states distinct integer labels $1 \leq j \leq J$, and let Γ^{J-1}

denote the corresponding space of probability distributions:

$$\Gamma^{J-1} = \left\{ \mu \in \mathbb{R}_+^{J-1} \left| \sum_{j=1}^{J-1} \mu(j) \leq 1 \right. \right\}; \tag{9}$$

with $\mu(J) = 1 - \sum_{j=1}^{J-1} \mu(j)$ left as implicit.

In Appendix 2 we establish a "Lagrangian" lemma that shows that there is always a supporting hyperplane to the lower epigraph of the concavified net utility function (Lemma 2.6). This is a formal statement of the concavification operation introduced geometrically above. The analytic translation of this geometrically clear result is that optimal attention strategies are characterized by Lagrange multipliers $\theta(j)$ conveying the change in net utility as each posterior $\gamma(j)$ for $1 \leq j \leq J-1$ is raised at the expense of reducing $\gamma(J)$. The Lagrange multipliers define the slope of the supporting hyperplane at the optimum. All chosen actions have net utilities that lie on this hyperplane at the corresponding chosen posterior, while no net utility function breaches the hyperplane for any posterior.

Lagrangean Lemma: Given $K \in \mathcal{K}^{PS}$ and $(\mu, A) \in \mathcal{D}$, $\lambda \in \hat{\Lambda}(\mu, A|K)$ if and only if $\exists \theta \in \mathbb{R}^{J-1}$ s.t.,

$$N_\mu^a(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq \sup_{a' \in A, \gamma' \in \Gamma(\bar{\mu})} N_\mu^{a'}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j),$$

for all $\gamma \in \Gamma(\mu)$ and $a \in A$, with equality if $\gamma \in \Gamma(Q_\lambda)$ and $q_\lambda(a|\gamma) > 0$.

Note that this lemma characterizes optimal strategies, and opens up standard methods of model solution. In addition, it conveys important qualitative features of the behavior implied by PS and UPS models. We return to this in later sections.

4 SDSC and Representations

In this section we introduce the data set and the sought after representations.

4.1 State Dependent Stochastic Choice Data

The key question in applied work on attention is the extent to which DMs internalize the actual decision making environment in which they find themselves. Do they notice whether or not a sales tax is included in the price paid at the register (Chetty *et al.* [2009])? Do they notice fluctuating prices of the same good in a supermarket (Matějka [2015])? Essentially all such situations can be captured using the general model above, by appropriately specifying available actions, the various factors (states of the world) that determine their payoffs, and prior beliefs about how likely is each such state.

Our goal is to specify observable patterns in choice data that narrow down the theories of inattentive choice. Before we begin, however, we must first specify exactly what sort of data is sufficient for this task. An important first point is that standard stochastic choice data, in which one

only observes the unconditional likelihood of each choice, is fundamentally inadequate for capturing attentional constraints. To see this, consider the two action decision problem illustrated in Figure 4. Note that the symmetry of the decision problem implies that the optimal strategy results in each action being chosen equally often. In the particular strategy chosen, this reflects partial information. Yet the same unconditional probabilities are also consistent with perfect information, with each action chosen precisely when it is optimal. These probabilities are also consistent with completely inattentive choice, with a fair coin flipped to decide which action is taken. Unconditional choice probabilities in no way reflect the extent to which behavioral patterns are impacted by reality. One must also know how well the action suited reality.

As first noted by Block and Marschak [1960] (p. 98-99), the way forward lies in realistically enriching the ideal behavioral data available to an ideal observer (IO), such as an econometrician or economic theorist, in which of costs of attention are to be identified. The key to our data enrichment is the observation that the information constraints that impact the DM do not apply to the IO. For example, while the DM may have difficulty assessing whether or not a sales tax is included in the purchase price or what the actual price of each good is in a supermarket, the IO with access to the underlying reality does not. In defining our data, we therefore specify that the IO observes both the state of the world as well as the action.

In formal terms, our behavioral data set is state dependent stochastic choice (SDSC) data, as in Caplin and Martin [2015] and Caplin and Dean [2015]. We specify both states and actions as being fully observed by the IO. We further specify our IO as able to watch this DM facing this same decision infinitely often, with precisely this strategy used each time.[8] For the IO to treat repeated observations of the DM as deriving from the same decision problem implies that the set of available actions, A, is the same. It requires also that the DM is seen as having the same prior μ over possible states of the world. We assume that the IO then observes the full distribution of actual state realizations and action choices. In terms of interpreting the data as revealing of patterns of attentional choice, we make the simplifying assumption that there are common probability assessments between IO and DM. We call this "rational expectations" with which it has spiritual commonalities.

A key observation is that rationality of expectations enables the IO to infer the DM's presumed prior as the actual proportion of times each state is realized. We therefore treat the prior itself as observable in specifying our behavioral data set in its most general form. Note this approach is standard in the rational inattention literature (Matejka and McKay [2015], Caplin and Dean [2015])

Definition 8 *Given $(\mu, A) \in \mathcal{D}$, we define* ***state dependent stochastic choice (SDSC)*** *data as mapping from possible states to action probabilities,*

$$\mathcal{P}(\mu, A) \equiv \{P : \Omega(\mu) \to \Delta(A)\},$$

with $P(a|\omega)$ the probability of action a in state ω. We define $\mathcal{P}$ as the union over all decision problems, $\mathcal{P} \equiv \cup_{(\mu,A)\in\mathcal{D}} \mathcal{P}(\mu, A)$.

Implicit in this definition is the assumption that the expected utility function of the DM is part

[8] In practice one might apply a model of this form to a population rather than an individual, as in the literature on discrete choice following McFadden [2005].

of the data. One could readily replace this assumption with an enrichment of the data set that allowed for utilities to be recovered from behavior, as discussed in Caplin and Dean [2015].[9]

While only recently introduced in economics, SDSC data has a long and storied history in psychometrics. The Weber-Fechner laws, which are based on corresponding data, identify regularities in how well humans perceive objective differences in the strength of various external stimuli.

There are two key differences between our approach and the standard psychometric approach. First, we follow classical economic logic, so that the stimuli are levels of utility, or reward. Second, we model perceptual effort as chosen in light of potential rewards. Given this, we will show that rich behavioral data has patterns in it that fully reveal costs of accurately recognizing external reward stimuli.

4.2 From Strategy to Data

We illustrate in Figure 5 how seeing data on states and actions captures the behavioral imprint of our running example, strategy λ^*. Given the assumed rationality of expectations, the subjective probabilities of the DM agree with the data frequencies as seen by the IO. What the IO will then see is a joint distribution of states and actions with precise probabilities determined by the prior, the posteriors, and the mixed action strategy.

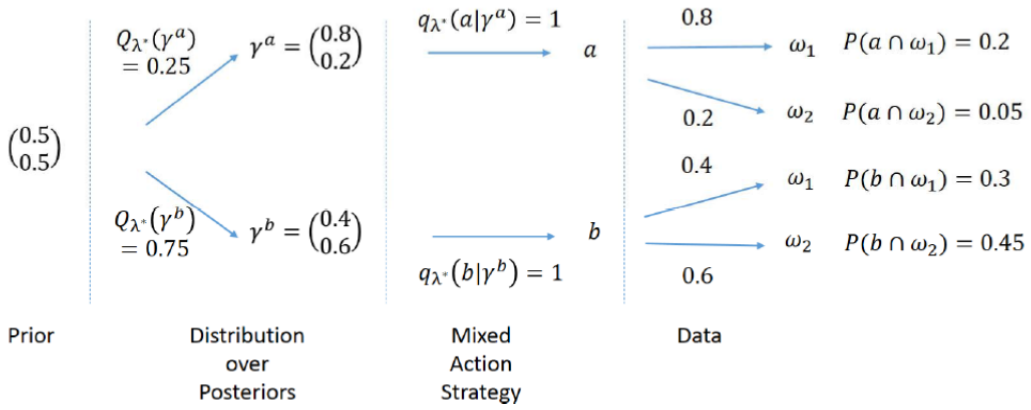

Figure 5: Data Generated by Strategy λ^*

The fact that action a is chosen if and only if the DM receives γ^a, and that $Q_{\lambda^*}(\gamma^a) = 0.25$ means that action a will be chosen 25% of the time (and b the remaining 75% of the time). Because γ^a is associated with an 80% probability of ω_1 (and a 20% probability of ω_2), the resulting joint probability of a and ω_1 is 20%. All other joint probabilities can be calculated in a similar way, as shown in Figure 5. These joint probabilities can be converted into conditional probabilities using

[9] One could replace the "Savage style" actions we use in this paper with "Anscombe-Aumann" acts that map states of the world to probability distributions over the prize space. Assuming the DM does maximize expected utility, u could then be recovered by observing choices over degenerate acts (i.e. acts whose payoffs are state independent).

Bayes' rule, giving the SDSC P^* associated with λ^*:

$$\begin{aligned} P^*(a|\omega_1) &= 0.4; \quad P^*(b|\omega_1) = 0.6; \\ P^*(a|\omega_2) &= 0.1; \quad P^*(b|\omega_2) = 0.9. \end{aligned}$$

This method for generating data from strategies is more general. Following the logic of Figure 5, we translate each strategy $\lambda \in \Lambda(\mu, A)$ into its observable counterpart in SDSC data, $\mathbf{P}_\lambda$, assuming rational expectations. With this notation, note that $P^* = \mathbf{P}_{\lambda^*}$.

Definition 9 *Given $\lambda \in \Lambda(\mu, A)$ we define the **generated** SDSC data $\mathbf{P}_\lambda : \Omega(\mu) \to \Delta(A)$ and the corresponding action choice probabilities $\mathbf{P}_\lambda(a)$ on $a \in \mathcal{A}(\lambda)$ by:*

$$\begin{aligned} \mathbf{P}_\lambda(a|\omega) &= \frac{\sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) q_\lambda(a|\gamma)\gamma(\omega)}{\mu(\omega)}; \\ \mathbf{P}_\lambda(a) &= \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) q_\lambda(a|\gamma). \end{aligned}$$

4.3 Choice Correspondence and Representations

In the idealized data set that we consider, SDSC data is available for all decision problems. As indicated, and as in Caplin and Dean [2015], we assume that the IO knows all details of the decision problem faced by the DM, which includes the prior and the payoffs to all available actions. For technical reasons, it simplifies the statement of our representation theorems to imagine that the IO sees a data set that is deep as well as broad. It specifies for each decision problem a corresponding set of qualifying SDSC functions - i.e. all such functions used by the DM in that decision problem. Following Richter [1966], this is in the spirit of standard choice analysis based on a correspondence mapping a choice set to a subset of suitable alternatives. $\mathcal{C}$ is the set of such data sets:

$$\mathcal{C} \equiv \left\{ C : \mathcal{D} \to 2^{\mathcal{P}} / \emptyset | C(\mu, A) \subset \mathcal{P}(\mu, A) \right\}.$$

This level of artificiality turns out to be substantively irrelevant. We discuss in Section 8 how our results extend to cases in which one sees only a selection from this data correspondence.

The relationship between $C(\mu, A)$ and $\mathcal{P}(\mu, A)$ is similar to the relationship between $\hat{\Lambda}(\mu, A|K)$ and $\Lambda(\mu, A)$. Just as $\Lambda(\mu, A)$ is the set of possible strategies and $\hat{\Lambda}(\mu, A|K)$ is the set of optimal strategies, $\mathcal{P}(\mu, A)$ is the set of possible data and $C(\mu, A)$ is the set of observed data.

We say that a data set C has a costly information representation based on a cost function K if the observed SDSC data $C(\mu, A)$ corresponding to each decision problem (μ, A) coincides with the SDSC data $\mathbf{P}_\lambda$ generated by optimal strategies $\lambda \in \hat{\Lambda}(\mu, A|K)$.

Definition 10 *Data set $C \in \mathcal{C}$ has a **costly information representation** (CIR) based on $K \in \mathcal{K}$ if, for all $(\mu, A) \in \mathcal{D}$,*

$$C(\mu, A) = \{\mathbf{P}_\lambda \in \mathcal{P} | \lambda \in \hat{\Lambda}(\mu, A|K)\} \equiv \hat{P}(\mu, A|K).$$

1. *It has a* ***posterior-separable*** *(PS) representation if it has a CIR* $K \in \mathcal{K}^{PS}$.

2. *It has a* ***uniformly posterior-separable*** *(PS) representation if it has a CIR* $K \in \mathcal{K}^{UPS}$.

3. *It has a* ***Shannon representation*** *if it has a CIR* $K = K^S_\kappa$ *for* $\kappa > 0$.

4.4 The Revealed Strategy

Caplin and Dean [2015] show that, while there is a multiplicity of strategies that could have generated any SDSC data, there is always a unique least Blackwell informative strategy consistent with the data. The first step in constructing this strategy is to identify with each chosen action a the corresponding "revealed posterior" γ^a_P. This treats the action as chosen at one and only one posterior which can be inferred from our behavioral data using Bayes' rule. Building on this, the "revealed posterior-based strategy" is the least Blackwell informative strategy consistent with the data. As such, it is the least costly for all our PS cost functions. It follows that for the class of models we consider in this paper, optimality implies that the revealed attention strategy is used by the DM in each decision problem.

Definition 11 *Given* $(\mu, A) \in \mathcal{D}$, $P \in \mathcal{P}(\mu, A)$, *and* $a \in A$, *we define* ***revealed action probability*** $P(a) = \sum_{\omega \in \Omega(\mu)} \mu(\omega) P(a|\omega)$. *We define* $\mathcal{A}(P)$ *as the actions chosen with positive probability. If* $a \in \mathcal{A}(P) \subset A$, *we define also* ***revealed posterior*** $\gamma^a_P \in \Gamma(\mu)$

$$\gamma^a_P(\omega) = \frac{\mu(\omega) P(a|\omega)}{P(a)};$$

with $\Gamma(P)$ *the union of* γ^a_P *across* $a \in \mathcal{A}(P)$. *The* ***revealed posterior-based attention strategy*** $\boldsymbol{\lambda}_P = (Q_P, q_P) \in \Lambda(\mu, A)$[10] *is defined by* $\Gamma(Q_P) = \cup_{a \in \mathcal{A}(P)} \gamma^a_P$ *and:*

$$\begin{aligned} Q_P(\gamma) &= \sum_{\{a \in \mathcal{A}(P) | \gamma^a_P = \gamma\}} P(a); \\ q_P(a|\gamma) &= \begin{cases} \frac{P(a)}{\mathbf{Q}_P(\gamma)} & \text{if } \gamma^a_P = \gamma; \\ 0 & \text{if } \gamma^a_P \neq \gamma. \end{cases} \end{aligned}$$

To illustrate construction of the revealed attention strategy, consider the data set $P^* = \mathbf{P}_{\lambda^*}$. The revealed posterior associated with the choice of action a is,

$$\gamma^a_{P^*}(\omega_1) = \frac{\mu(\omega_1) P^*(a|\omega_1)}{P^*(a)} = \frac{0.5 \times 0.4}{0.25} = 0.8.$$

Similarly

$$\gamma^a_{P^*}(\omega_2) = 0.2; \ \gamma^b_{P^*}(\omega_1) = 0.4; \text{ and } \gamma^b_{P^*}(\omega_2) = 0.6$$

We can then calculate the revealed strategy as involving

$$\begin{aligned} Q_{P^*}(\gamma^a_{P^*}) &= P^*(a) = \mu(\omega_1) P^*(a|\omega_1) + \mu(\omega_2) P^*(a|\omega_2) \\ &= 0.5 * (0.4 + 0.1) = 0.25. \end{aligned}$$

[10] See Appendix 1 for direct confirmation that $\boldsymbol{\lambda}(P) \in \Lambda(\mu, A)$.

Hence,

$$Q_{P^*}(\gamma^b_{P^*}) = P^*(b) = 0.75;$$

Furthermore,

$$q_{P^*}(a|\gamma^a_{P^*}) = 1 = q_{P^*}(b|\gamma^b_{P^*}).$$

Note in this case that λ^* is in fact the revealed strategy associated with data set $\mathbf{P}_{\lambda^*} = P^*$,

$$\lambda^* = \boldsymbol{\lambda}_{P^*} = \boldsymbol{\lambda}_{\mathbf{P}_{\lambda^*}}.$$

While this does not hold for arbitrary strategies, it is general for data observed in our representations, as discussed in Caplin and Dean [2015]. Appendix 2 contains this result together with other general results that link strategies revealed by data in costly information representations with the SDSC data generated by optimal strategies.

5 Compression and the Shannon Model

Having introduced the key model elements and data-related definitions, we turn now to the results themselves. In this section we start with a data set having a UPS representation and identify additional behavioral restrictions that make this a Shannon representation. As the definitions show, the UPS form allows for a general convex function $T(\gamma)$, while Shannon restricts $T(\gamma)$ to a particular one parameter family, $T(\gamma) = \kappa \ln(\gamma)$. This restriction on T implies many qualitative restrictions on behavior. For example, there are strong symmetry properties, so that behavior must indicate that all states individually are equally easy or difficult to perceive. There are also no complementarities, so that learning about one state makes it no easier (or more difficult) to learn about any separate state. There are also very strong smoothness properties and profound quantitative restrictions e.g. in terms of the response to payoff changes.

Our first theorem establishes that a single behavioral invariance axiom is enough to move us from a UPS representation to a Shannon representation, hence conveying all of these particular properties noted above and all others besides. This axiom insists that choices not change when payoff equivalent states are "compressed" into a single state. In the remainder of the section we first introduce this behavioral axiom intuitively. We then formalize it and state the main theorem. Finally, we sketch the proof. The proof itself, which is involved, is in Appendix 5.

5.1 Basic Decision Problems and Basic Forms

What precisely does it mean to say that payoffs alone matter? To specify, consider first decision problems in which all states are distinct in terms of payoffs, so that no two possible states have identical payoffs for all available actions. We call these "basic" decision problems.

Definition 12 *Decision problem (μ, A) is **basic,** $(\mu, A) \in \mathcal{B} \subset \mathcal{D}$ if, given $\omega \neq \omega' \in \Omega(\mu)$, there exists $a \in A$ such that $u(a, \omega) \neq u(a, \omega')$.*

Consider now a non-basic decision problem with three possible states: $\Omega(\mu) = \{\omega_1, \omega_2, \omega_3\}$ and two actions $A = \{a, b\}$. In this problem, states ω_1 and ω_2 are equivalent:

$$\begin{aligned} u(a, \omega_1) &= 1,\ u(b, \omega_1) = 0; \\ u(a, \omega_2) &= 1,\ u(b, \omega_2) = 0; \\ u(a, \omega_3) &= 0,\ u(b, \omega_3) = 1. \end{aligned}$$

There are two obvious ways to shift all probability from the two equivalent states to one or the other of them. One way is to set $\bar{\mu}(\omega_1) = \mu(\omega_1) + \mu(\omega_2)$ and $\bar{\mu}(\omega_2) = 0$, with $\bar{\mu}(\omega_3) = \mu(\omega_3)$. The alternative is to set $\hat{\mu}(\omega_2) = \mu(\omega_1) + \mu(\omega_2)$ and rule out state ω_1. These priors associated with these two "basic forms" of (μ, A) are illustrated in Figure 6.

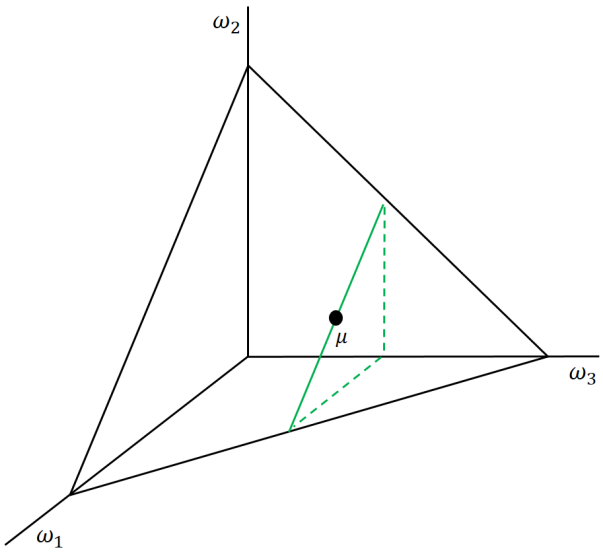

Figure 6: Basic Forms of Decision Problem (μ, A)

We now provide the general technical definitions.

Definition 13 *We associate $(\mu, A) \in \mathcal{D}$ with a set of **basic forms** $(\bar{\mu}, A) \in \mathcal{B}$ by:*

1. *Partitioning $\Omega(\mu)$ into L **basic sets** $\{\Omega^l(\mu)\}_{1 \leq l \leq L}$ comprising payoff equivalent states, so that, given $\omega \in \Omega^l(\mu)$ and $\omega' \in \Omega^m(\mu)$,*

$$l = m \text{ iff } u(a, \omega) = u(a, \omega') \text{ for all } a \in A.$$

2. *For $1 \leq l \leq L$, defining $I(l) = |\Omega^l(\mu)|$, and indexing by i the states $\omega_i^l \in \Omega^l(\mu)$, so that:*

$$\Omega^l(\mu) = \{\omega_i^l \in \Omega(\mu) | 1 \leq i \leq I(l)\}.$$

3. *Selecting $\bar{\imath}(l) \in \{1, .., I(l)\}$ for all l and defining $\bar{\Omega}(\mu) = \cup_{l=1}^{L} \omega_{\bar{\imath}(l)}^l$.*

4. Defining $\bar{\mu} \in \Gamma$ *by:*

$$\bar{\mu}(\omega_i^l) = \begin{cases} \sum_{j=1}^{I(l)} \mu(\omega_j^l) \text{ if } i = \bar{\imath}(l); \\ 0 \text{ if } i \neq \bar{\imath}(l). \end{cases}$$

We let $\mathcal{B}(\mu, A) \subset \mathcal{B}$ *be all basic forms corresponding to* $(\mu, A) \in \mathcal{D}$*. Given* $\bar{\imath}(l) \in \{1, .., I(l)\}$ *on* $1 \leq l \leq L$*, we write* $(\bar{\mu}, A) \in \mathcal{B}(\mu, A)$ *for* $\bar{\imath}$*.*

5.2 Invariance Under Compression

Note that there is no functional value for the DM in distinguishing between states that assign the same payoff to all actions. Hence an ideally designed machine for encoding states would not waste any of its scarce resources on this task. The stochastic structure of choice would not change if distinct yet payoff equivalent states were "compressed" into a single state.

Our Invariance under Compression axiom insists that patterns of choice are equivalent in all decision problems with a common basic form.

Axiom A1 Invariance under Compression (IUC): *Given* $(\mu, A), (\bar{\mu}, A) \in \mathcal{D}$ such that $(\bar{\mu}, A) \in \mathcal{B}(\mu, A)$ *for some* $\bar{\imath}$*:*

$$P \in C(\mu, A) \Longleftrightarrow \exists \bar{P} \in C(\bar{\mu}, A) \text{ s.t. } P(a|\omega_i^l) = \bar{P}(a|\omega_{\bar{\imath}(l)}^l),$$

for all $1 \leq i \leq I(l)$, $1 \leq l \leq L$ and $a \in A$.

We can illustrate the meaning of IUC using the example discussed above, in which the decision problem (μ, A) is such that $\Omega(\mu)$ has two basic sets: $\{\omega_1, \omega_2\}$ and $\{\omega_3\}$. Note first that IUC implies that, for any observed $P \in C(\mu, A)$ it must be the case that $P(a|\omega_1) = P(a|\omega_2)$ for all $a \in A$: the DM must behave identically in any states that belong to the same basic set. Moreover, behavior in (μ, A) must be similar to behavior in the basic version of the problem. For example, given $\bar{\mu}(\omega_1) = \mu(\omega_1) + \mu(\omega_2)$, $\bar{P} \in C(\bar{\mu}, A)$ if and only if $\bar{P}(a|\omega_1) = P(a|\omega_1) = P(a|\omega_2)$ for some $P \in C(\mu, A)$ and for all $a \in A$. The fact that this also holds for the basic version of the problem in which $\hat{\mu}(\omega_1) = \mu(\omega_1) + \mu(\omega_2)$ means furthermore that behavior in the two basic versions of the problem must be the same: $\bar{P} \in C(\bar{\mu}, A)$ if and only if $\bar{P}(a|\omega_1) = \hat{P}(a|\omega_2)$ for some $\hat{P} \in C(\hat{\mu}, A)$. An immediate corollary is that, for any prior μ^* such that $\mu^*(\omega_3) = \mu(\omega_3)$ and $\Omega(\mu^*) \subset \Omega(\mu)$ it must be the case that $C(\mu, A) = C(\mu^*, A)$.

The key result is that the Shannon cost function alone among UPS cost functions satisfies this invariance axiom.

Theorem 1: *Data set* $C \in \mathcal{C}$ *with a UPS representation has a Shannon representation if and only if it satisfies IUC.*

5.3 Necessity

That IUC is necessary for a Shannon representation follows directly from the posterior-based characterization of the solution to the Shannon model. Caplin and Dean [2013] provide an "invariant

likelihood ratio" condition for optimality. This states that $P \in C(\mu, A)$ is consistent with optimality for a cost function K_κ^S if and only if:

1. Given $a, b \in \mathcal{A}(P)$,

$$\frac{\gamma_P^a(\omega)}{\exp(u(a,\omega)/\kappa)} = \frac{\gamma_P^b(\omega)}{\exp(u(b,\omega)/\kappa)} \text{ for all } \omega \in \Omega(\mu). \tag{10}$$

2. Given $a \in \mathcal{A}(P)$ and $c \in A \backslash \mathcal{A}(P)$,

$$\sum_{\omega \in \Omega(\mu)} \left[\frac{\gamma_P^a(\omega)}{\exp(u(a,\omega)/\kappa)} \right] \exp(u(c,\omega)/\kappa) \leq 1.$$

It is the fact that these conditions are invariant under the compression operation that establishes IUC as necessary for a Shannon representation, as formalized in Appendix 5.

5.4 Guide to the Sufficiency Proof

While the necessity proof is straight forward, the sufficiency proof is not. Theorem 1 establishes that IUC is profoundly powerful. It implies that, starting with behavior generated by a general strictly convex function, IUC plus one attentive choice pins down behavior in all decision problems. This follows since the attentive choice pins down the single parameter $\kappa > 0$ in the Shannon function, leaving no more degrees of freedom.

Given the vast distance that the proof must travel to rule out all other forms of the cost function, it involves several stages that we elaborate on briefly here. The proof itself involves many corresponding lemmas that provide details.

One line of argument uses IUC to establish strong **symmetry** properties of the cost function: here the argument is direct. Two other key aspects of the proof take up issues of smoothness and functional form. In particular, there are strong **differentiability** and **additive separability** arguments. With these established, we identify a **second order PDE** that must be satisfied and that implies the Shannon form. The smoothness and separability arguments work in a fixed state space of cardinality 4 or higher. The final step in the proof involves using IUC to link cost functions across dimensions and to iterate down to dimensions below four. We briefly outline what is accomplished in each stage, leaving the full treatment to the Appendix.

5.4.1 Symmetry

The first step in the proof is to introduce and demonstrate the powerful symmetry implications of IUC. The definition of symmetry in beliefs is direct: $\gamma_1, \gamma_2 \in \Gamma$ are symmetric, $\gamma_1 \sim_\Gamma \gamma_2$, if there exists a bijection $\sigma : \Omega(\gamma_1) \to \Omega(\gamma_2)$ such that, for all $\omega \in \Omega(\gamma_1)$,

$$\gamma_1(\omega) = \gamma_2(\sigma(\omega)).$$

Correspondingly, the strictly convex function $T : \Gamma \longrightarrow \mathbb{R}$ is symmetric if,

$$\gamma_1 \sim_\Gamma \gamma_2 \Longrightarrow T(\gamma_1) = T(\gamma_2).$$

A sequence of results establishes that IUC implies symmetry of the T function in a UPS representation (Lemma 5.7).

Symmetric Cost Lemma: Given $C \in \mathcal{C}$ satisfying Axiom A1, any function $T : \Gamma \longrightarrow \mathbb{R}$ in a UPS representation $K(Q) = \sum_{\Gamma(Q)} Q(\gamma) T(\gamma)$ must be symmetric.

Intuitively, Axiom A1 implies that relabeling the states cannot affect choice. To see this consider two basic decision problems (μ_1, A_1) and (μ_2, A_2) that are symmetric in the sense that $\mu_1(\omega) = \mu_2(\sigma(\omega))$ for each $\omega \in \Omega(\gamma_1)$ and such that, for each $a \in A_1$, there exists $b \in A_2$ such that $u(a, \omega) = u(b, \sigma(\omega))$ where the implied mapping between A_1 and A_2 is bijective. Now consider a third problem (μ_3, A_3) which involves replicating (μ_1, A_1) and (μ_2, A_2) on a set of states $\Omega(\mu_3)$ disjoint from $\Omega(\mu_1) \cup \Omega(\mu_2)$, and then consider the problem $(\frac{\mu_1}{3} + \frac{\mu_2}{3} + \frac{\mu_3}{2}, A_1 \cup A_2 \cup A_3)$. (μ_1, A_1), (μ_2, A_2), and (μ_3, A_3) are all basic versions of this last problem and therefore the SDSC data generated by (μ_1, A_1) and (μ_2, A_2) is similar to the SDSC data generated by (μ_3, A_3) and hence each is similar to the other. It is a small step from this observation to the Symmetric Cost Lemma.

5.4.2 Differentiability

As noted above, much of the proof involves working within a fixed state space $\tilde{\Omega} \subset \Omega$ of cardinality $J \geq 4$, with the states indexed by j. Recall that $\tilde{\Gamma}$ comprise the interior posteriors with $\Omega(\gamma) = \tilde{\Omega}$ and we correspondingly let $\tilde{T}$ be the restriction of T to $\tilde{\Gamma}$. By symmetry, the form of this function depends only on the cardinality J.

Given $\gamma \in \tilde{\Gamma}$ and any pair of states $i \neq j$ we define the one-sided derivative in direction ji, $\tilde{T}_{\overrightarrow{ji}}(\gamma)$, as the directional derivative associated with increasing the ith coordinate and equally reducing the jth:

$$\tilde{T}_{\overrightarrow{ji}}(\gamma) = \lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon};$$

where $e_k \in \mathbb{R}^J$ is the corresponding unit vector.[11]

Since $\tilde{T}$ is convex, we know that $\tilde{T}_{\overrightarrow{ji}}(\gamma)$ exists. We define also the two-sided derivative in direction ji, $\tilde{T}_{(ji)}$, by:

$$\tilde{T}_{(ji)}(\gamma) = \lim_{\epsilon \to 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon}.$$

While in principle the two-sided derivative need not exist, we show that it always does (Lemma 5.32). The proof makes heavy use of results in Rockafellar [1970] and the profound structure that IUC conveys. The proof of differentiability comes fairly late in the proof of Theorem 1. For expositional clarity, we will assume in what follows that the two-sided derivatives exist. In the proof, most of the results are first proved for the one-sided derivatives and only apply to the two-sided derivatives once differentiability has been established.

[11] This is defined in Rockafellar [1970] as the directional derivative of $\tilde{T}$ at γ in direction $e_i - e_j$ direction, $\tilde{T}'(\gamma | e_i - e_j)$. Its existence is established in his Theorem 23.1.

With $\tilde{T}_{(ji)}(\gamma)$ existing always, we can define cross directional derivatives of $\tilde{T}$. Given $\gamma \in \tilde{\Gamma}$ and any two pairs of states $i \neq j$ and $k \neq l$, we define the corresponding cross derivative of $\tilde{T}_{(ji)}$ in direction lk as the corresponding (two-sided) directional derivative,

$$\tilde{T}_{(ji)(lk)}(\gamma) = \lim_{\epsilon \to 0} \frac{\tilde{T}_{(ji)}(\gamma + \epsilon(e_k - e_l)) - \tilde{T}_{(ji)}(\gamma)}{\epsilon}$$

Again, we show that these cross-derivatives exist everywhere in $\tilde{\Gamma}$ (Lemma 5.36).

5.4.3 Additive Separability

The proof of additive separability is staged and inter-leaved with the proof of differentiability. While we cannot in the text convey the full flavor of the additivity and differentiability results, it may be helpful to point out several key insights.

The first observation is that the Lagrangian Lemma implies that there is a common hyperplane tangent to each of the net utility functions at each chosen posterior. This links directional derivatives of the net utility function at distinct optimal posteriors (Lemma 5.11).

Equalization of Derivatives: Suppose $C \in \mathcal{C}$ has a UPS representation K, and consider $(\mu, A) \in \mathcal{D}$ and $P \in C(\mu, A)$ with $a, b \in \mathcal{A}(P)$ with $\{\gamma_P^a, \gamma_P^b\} \subset \tilde{\Gamma}$. Suppose that both $\tilde{T}^a_{(ji)}(\gamma_P^a)$ and $\tilde{T}^b_{(ji)}(\gamma_P^b)$ exist, then

$$N^a_{(ji)}(\gamma_P^a) = N^b_{(ji)}(\gamma_P^b).$$

A second observation is that IUC places structure on the sets of posteriors that can be linked by considering decision problems with equivalent states. In Figure 7, we illustrate this implication of IUC with three states, but the intuition applies generally. Consider a decision problem with three states $(\omega_1, \omega_2, \omega_3)$ and two actions $A = \{a, b\}$, in which states ω_1 and ω_2 are equivalent. Figure 7 displays the space of potential priors and posteriors. Suppose that $\bar{\mu}_1$ is the prior in the basic problem in which all of the combined probability of ω_1 and ω_2 is assigned to ω_1 and $\bar{\mu}_2$ is the prior in the case in which ω_2 receives all of the weight. Since $\bar{\mu}_1(\omega_1) = \bar{\mu}(\omega_2)$ the line segment connecting these two priors is parallel to the segment connecting $(1, 0, 0)$ and $(0, 1, 0)$. The line segment connecting $\bar{\mu}_1$ and $\bar{\mu}_2$ represents the set of potential priors for which,

$$\mu(\omega_1) + \mu(\omega_2) = \bar{\mu}_1(\omega_1) = \bar{\mu}_2(\omega_2),$$

so that $(\bar{\mu}_1, A)$ and $(\bar{\mu}_2, A)$ are basic versions of (μ, A).

The above shows that, letting μ to be an arbitrary prior in this set, IUC places restrictions on the relationship between the optimal posteriors for the problems (μ, A), $(\bar{\mu}_1, A)$ and $(\bar{\mu}_2, A)$ (Lemma 5.12). Consider γ^a. Bayes rule states that $\gamma^a(\omega) = P(a|\omega)\mu(\omega)/P(a)$. IUC implies that $P(a|\omega)$ and $P(a)$ are the same for all μ on the segment connecting $\bar{\mu}_1$ and $\bar{\mu}_2$, including $\bar{\mu}_1$ and $\bar{\mu}_2$ themselves. This implies that as μ moves from $\bar{\mu}_1$ to $\bar{\mu}_2$, γ^a and γ^b are always proportionate to μ. It follows that γ^a and γ^b lie at the intersection of a line through μ and $(0, 0, 1)$, the dashed grey line in the figure, and a line parallel to the segment connecting $(1, 0, 0)$ and $(0, 1, 0)$, the solid red and blue lines in the figures. $\bar{\gamma}_1^a$ and $\bar{\gamma}_1^b$ in the figure denote the optimal posteriors for $(\bar{\mu}_1, A)$, and $\bar{\gamma}_2^a$ and $\bar{\gamma}_2^b$ the optimal posteriors for $(\bar{\mu}_2, A)$.

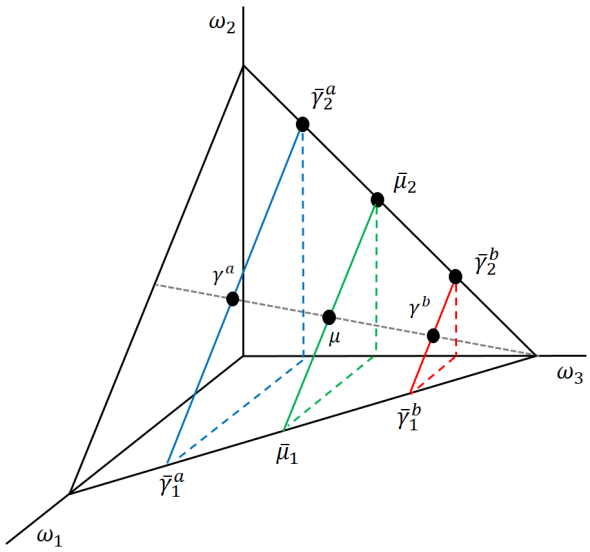

Figure 7: Implications of Compression

These two observations when combined relate the derivatives of $\tilde{T}$ at γ^a and γ^b in the Figure. The Lagrangian Lemma implies that there is a hyperplane tangent to both $N(\gamma^a)$ and $N(\gamma^b)$. Suppose that both $\tilde{T}_{(ij)}(\gamma^a)$ and $\tilde{T}_{(ij)}(\gamma^b)$ exist. Since prize-based expected utility is linear, the difference between $\tilde{T}_{(ji)}(\gamma^a)$ and $\tilde{T}_{(ji)}(\gamma^b)$ must equal $u(a,\omega_i) - u(a,\omega_j) - u(b,\omega_i) + u(b,\omega_j)$. Since shifts in μ from $\bar{\mu}_1$ to $\bar{\mu}_2$, do not affect prize based utility, $\tilde{T}_{(ji)}(\gamma^a) - \tilde{T}_{(ji)}(\gamma^b)$ must be independent of μ whenever both derivatives exist (Lemma 5.13).

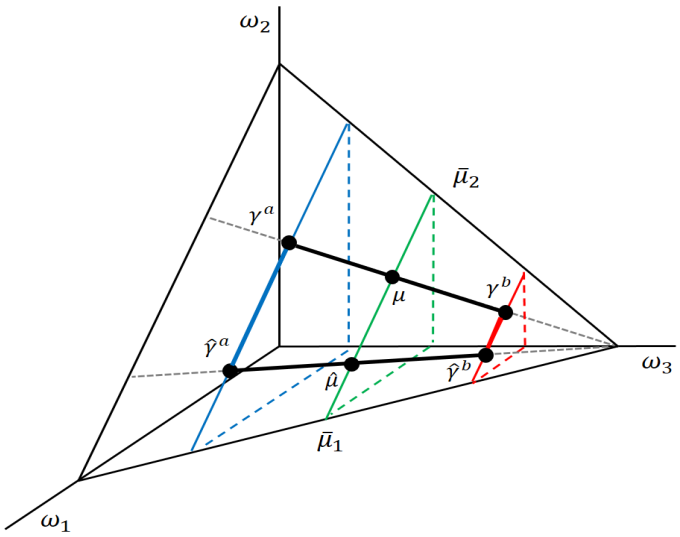

Figure 8: The Trapezoid

Consider now two priors μ and $\hat{\mu}$, each lying between $\bar{\mu}_1$ and $\bar{\mu}_2$. Figure 8 shows that the four

posteriors γ^a,γ^b,$\hat{\gamma}^a$, and $\hat{\gamma}^b$ form a trapezoid. If $\tilde{T}$ is differentiable at all four points we would know that:[12]

$$\tilde{T}_{(ji)}(\gamma^a) - \tilde{T}_{(ji)}(\gamma^b) = \tilde{T}_{(ji)}(\hat{\gamma}^a) - \tilde{T}_{(ji)}(\hat{\gamma}^b). \tag{11}$$

Equation (11) is close to the rectangle condition for additive separability. To apply the rectangle condition, we deform the simplex so that the trapezoid becomes a rectangle, and then return to the simplex. This results in the following characterization of the directional derivative which we state in terms of the dimension J since it requires $J \geq 4$ (Lemma 5.21):

$$\tilde{T}_{(ji)}(\gamma) = \boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right) + \boldsymbol{B}\left(\gamma(2), ..., \gamma(J-1)\right), \tag{12}$$

for some functions $\boldsymbol{A} : \mathbb{R}_+ \longrightarrow \mathbb{R}$ and $\boldsymbol{B} : \mathbb{R}^{J-2} \longrightarrow \mathbb{R}$, and for all $2 \leq i \neq j \leq J-1$. As (11) must hold for a range of $\gamma(1)$ and $\gamma(J)$ we can show that $\boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right)$ must be constant (Lemma 5.22). Symmetry then implies that, if $\tilde{T}_{(ji)}(\gamma)$ does not depend on $\gamma(1)$ and $\gamma(J)$, $\boldsymbol{B}$ cannot depend on any $\gamma(k)$ other than $\gamma(i)$ and $\gamma(j)$ (Lemma 5.24). Finally, we use the fact that $\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{(ki)}(\gamma) - \tilde{T}_{(kj)}(\gamma)$ whenever the latter are well defined to establish that there exists a function f on $(0,1)$ such that for all $\gamma \in \tilde{\Gamma}$ (Lemma 5.29):

$$\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j)).$$

5.4.4 The Second Order PDE and Shannon Entropy

Consider again the problem in Figure 7. The Lagrangian Lemma implies that as we shift μ between $\bar{\mu}_1$ and $\bar{\mu}_2$, the resulting revealed posteriors satisfy $N^a_{(ji)}(\gamma^a(\mu)) = N^b_{(ji)}(\gamma^b(\mu))$. Setting $\mu(t) = t\bar{\mu}_2 + (1-t)\bar{\mu}_1$, we can define $\gamma^a(t) = \gamma^a(\mu(t))$ and $\gamma^b(t) = \gamma^b(\mu(t))$ to be the revealed posteriors associated with $\mu(t)$. Given the twice differentiability of $\tilde{T}$, we have $\frac{d}{dt}N^a_{(ji)}(\gamma^a(t)) = \frac{d}{dt}N^b_{(ji)}(\gamma^b(t))$, and, since ω_1 and ω_2 are redundant prized-based utility does not depend on t, so that

$$\frac{d}{dt}\tilde{T}_{(ji)}(\gamma^a(t)) = \frac{d}{dt}\tilde{T}_{(ji)}(\gamma^b(t)).$$

Finally, note that since $\gamma^a(t)$, $\mu(t)$, and $\gamma^b(t)$ all lie along a line through $(0,0,1)$, a change in t alters γ^a proportionately more than γ^b. The chain rule implies:

$$\gamma^a(1)\tilde{T}_{(ji)(12)}(\gamma^a) = \gamma^b(1)\tilde{T}_{(ji)(12)}(\gamma^b).$$

Since this equation holds for all $\gamma(1)$, both sides must equal some constant κ^J,

$$\gamma(1)\tilde{T}_{(ji)(12)}(\gamma) = \kappa^J,$$

$$\gamma(i)\tilde{T}_{(ji)(li)}(\gamma) = \kappa^J,$$

and, since $\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j))$, taking $j = 1$ implies

$$\gamma(1)f'(\gamma(1)) = \kappa^J.$$

[12] Lemma 5.17 establishes (11) everywhere for the directional derivatives $\tilde{T}_{\overrightarrow{ji}}$ by finding pairs of differentiable points that simultaneously converge to the four posteriors γ^a,γ^b,$\hat{\gamma}^a$, and $\hat{\gamma}^b$.

A particular solution to this equation is $\kappa^J \ln x$. Integrating once more yields the Shannon form:

$$\tilde{T} = \kappa^J \sum_j \gamma(j) \ln(\gamma(j)).$$

Other solutions to these differential equations can be rejected as either irrelevant (they sum to a constant because the $\gamma(j)$ sum to a constant), inconsistent with the dependence of $\tilde{T}_{(ji)}$ on solely on $\gamma(i)$ and $\gamma(j)$, or inconsistent with symmetry.

5.4.5 IUC and Universal Domain

The proof at this stage has three gaps. First, it applies only to interior posteriors. Second, there is no tie between dimensions $J \geq 4$. Third it does not cover lower dimensional cases. We show next that IUC solves all of these.

The first key observation is that, given $J \geq 4$, all optimal strategies are precisely as if κ^J applied to all posteriors $\gamma \in \Gamma$ with $|\Omega(\gamma)| = L \leq J$. Note that, as a convex function, the costs are at least as high as the limit of the costs on the boundary. This limit function is in fact the classical Shannon entropy function,

$$T(\gamma) \geq \kappa^J \sum_{l=1}^{L} \gamma(l) \ln \gamma(l).$$

Even if costs take this minimum value, the known necessary and sufficient conditions for optimality imply that no prior possible states are ever ruled out in an optimal strategies. Hence the behavioral data is precisely as it would be if this function applied to all posteriors, even those that set some prior possible states as impossible.

The final part of the proof uses IUC to iterate down in dimension. To be precise, define K^J to be the Shannon cost function with parameter κ^J for $J \geq 4$ as defined on all posteriors with that state space or below,

$$K^J(\gamma) \equiv \kappa^J \sum_{j \in \Omega(\gamma)} \gamma(j) \ln \gamma(j), \text{ for all } \gamma \in \Gamma \text{ with } |\Omega(\gamma)| \leq J.$$

The precise result we establish is that, given any decision problem $(\mu, A) \in \mathcal{D}$ with a prior of cardinality one lower, $|\Omega(\mu)| = J - 1$,

$$P \in C(\mu, A) \text{ iff } \exists \lambda \in \hat{\Lambda}(\mu, A | K^J) \text{ such that } \mathbf{P}_\lambda = P.$$

Note that establishing this completes the proof of the theorem, since it directly implies that $\kappa^J = \kappa^{J-1}$ for $J \geq 4$, where the Shannon form was already established, and that the Shannon form and the corresponding parameter apply also to $J = 3$, then iteratively to $J = 2$, completing the logic.

6 Existence and Recoverability

As indicated in the introduction, our remaining results establish necessary and sufficient conditions for a UPS representation. In this section we cover the first stage of this three stage process, by

introducing conditions that establish recoverability of the cost function.

6.1 NIAS, NIAC, and Completeness

Our general recoverability result rests on three axioms, all of which are necessary for a PS representation of any kind, and indeed apply even more generally. Our first two axioms are required for existence of any CIR. "No Improving Action Switches" (NIAS), due to Caplin and Martin [2015], is based on utility being maximized at each posterior. It insists that all actions chosen maximize expected utility at the corresponding posterior. "No Improving Attention Cycles" (NIAC), adapted from Caplin and Dean [2015], rules out switching attention strategies across problems in a manner that increases overall utility. It insists that attention strategies cannot be shuffled between decision problems in such a manner as to raise total utility across these decision problems.

Axiom A2 No Improving Action Switches (NIAS): *Given* $(\mu, A) \in D$ *and* $P \in C(\mu, A)$,

$$a \in \mathcal{A}(P) \Longrightarrow \bar{u}(\gamma_P^a, a) = \max_{a \in A} \bar{u}(\gamma, a).$$

Axiom A3 No Improving Attention Cycles (NIAC): *Given* $\mu \in \Gamma$ *and a finite set*

$$\{(A(m), P(m))\}_{1 \leq m \leq M}$$

with $(\mu, A(m)) \in D$, $P(m) \in C(\mu, A(m))$, *and* $(A(1), P(1)) = (A(M), P(M))$,

$$\sum_{m=1}^{M-1} \hat{U}(\mu, A(m), P(m)) \geq \sum_{m=1}^{M-1} \hat{U}(\mu, A(m), P(m+1)),$$

where,

$$\hat{U}(\mu, A, P) \equiv \sum_{\gamma \in \Gamma(P)} \mathbf{Q}_P(\gamma) \left[\max_{a \in A} \bar{u}(\gamma, a) \right]$$

Our third axiom insists that almost all posterior distributions satisfying Bayes' rule can be found in the data for some decision problem. The caveat relates to posteriors that entirely rule out some ex ante possible states of the world. As indicated above, this never happens in the Shannon model.

To state this formally, we let $\Gamma(C, \mu)$ denote all revealed posteriors ever observed in any decision problem with the given prior, and correspondingly $\mathcal{Q}(C, \mu)$ as distributions over posteriors that are observed in the data.

Axiom A4 Completeness: *Given* $\mu \in \Gamma$*:*

1. $\Gamma(C, \mu)$ contains all interior posteriors, $\tilde{\Gamma}(\mu) \subset \Gamma(C, \mu)$.
2. $\Gamma(C, \mu)$ is a convex set.
3. $\Delta(\Gamma(C, \mu)) \cap \mathcal{Q}(\mu) \subset \mathcal{Q}(C, \mu)$.

6.2 Recoverability

The recoverability result rests on A2-A4 alone.

Theorem 2: Given $C \in \mathcal{C}$ satisfying A2-A4, there exists a function $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ for all $(\mu, A) \in \mathcal{D}$. This function is unique on $(\mu, Q) \in \mathcal{F}$ with $Q \in Q(C, \mu)$.

The proof has two key steps. The first establishes existence of a cost function $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ for all $(\mu, A) \in \mathcal{D}$ based on NIAS and NIAC. This proof is essentially the same as that of Caplin and Martin [2015] and Caplin and Dean [2015].[13] In the second step we find a condition that any rationalizing K must satisfy, and show that with A4 this is stringent enough to pin down K uniquely. The second stage is worth sketching out, not only because of its technical importance, but also because it underlies our characterization of PS representations.

The procedure for constructing the cost function involves application of the fundamental theorem of calculus. Given $\mu \in \Gamma$ and $\bar{Q} \in Q(C, \mu)$, we first enumerate the possible posteriors $\bar{\gamma}^n \in \Gamma(\bar{Q})$ for $1 \leq n \leq N = |\Gamma(\bar{Q})|$ and define corresponding fixed probability weights $\bar{Q}^n \equiv \bar{Q}(\bar{\gamma}^n)$. We then construct a path from the prior to the set of posteriors by defining for each n a line

$$\bar{\gamma}_t^n = t\bar{\gamma}^n + (1-t)\mu,$$

so that at $t = 0$ we have $\bar{\gamma}_0^n = \mu$ and at $t = 1$ we have $\bar{\gamma}_1^n = \bar{\gamma}^n$. For each t we consider the distribution $\bar{Q}_t$ in which each $\bar{\gamma}_t^n$ is selected with the same probability as $\bar{\gamma}^n$,

$$\bar{Q}_t(\bar{\gamma}_t^n) = \bar{Q}^n.$$

Note that this construction ensures that the weighted average of the posteriors always averages back to the prior,

$$\sum_n \bar{Q}_t(\bar{\gamma}_t^n)\bar{\gamma}_t^n = \sum_n \bar{Q}^n \left[t\bar{\gamma}^n + (1-t)\mu\right] = \mu,$$

so that $\bar{Q}_t \in \mathcal{Q}(\mu)$.

Since $\bar{Q}_t \in \mathcal{Q}(\mu)$, A4 implies that $\bar{Q}_t \in Q(C, \mu)$. Hence for every $t \in [0, 1]$, there exists a decision problem $(\mu, \bar{A}_t) \in \mathcal{D}$ and observed data $\bar{P}_t \in C(\mu, \bar{A}_t)$ that give rise to the corresponding distribution of revealed posteriors $\mathbf{Q}_{\bar{P}_t} = \bar{Q}_t$.

Given any cost function $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ for all $(\mu, A) \in \mathcal{D}$, we then show that,

$$K(\mu, \bar{Q}_t) \equiv \bar{K}(t).$$

is convex and continuous in $t \in [0, 1]$, and hence almost everywhere differentiable in t with,

$$\bar{K}(t) = \int_0^t \bar{K}'(s)ds, \tag{13}$$

where the integration is over points of differentiability.

Next we characterize $\bar{K}'(t)$. At any point t at which $\bar{K}(t)$ is differentiable, we consider the decision problem $(\mu, \bar{A}_t)$ for which $\bar{Q}_t$ is globally, hence locally, optimal. Thinking of shifting

[13] The richer data also leads us to change proof method, relying in this case on the work of Rochet [1987].

locally to a different posterior distribution Q_s for $s \in (t-\epsilon, t+\epsilon)$ leads to a first-order condition,

$$\bar{K}'(t) = \sum_n \bar{Q}^n \left([\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n) \right). \tag{14}$$

where $\bar{a}_t^n$ is any chosen action associated with $\gamma_t^n \in \Gamma(Q_t)$ and where the dot product $[\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n)$ is defined by,

$$[\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n) \equiv \sum_{\omega \in \Omega(\mu)} [\bar{\gamma}^n(\omega) - \mu(\omega)] \, u(\bar{a}_t^n, \omega).$$

Substituting (14) into (13) yields,

$$K(\mu, \bar{Q}) = \sum_n \bar{Q}^n \, [\bar{\gamma}^n - \mu] \cdot \int_0^1 u(\bar{a}_t^n) dt.$$

Note that, given $(\mu, \bar{Q}) \in \mathcal{F}$ with $\bar{Q} \in Q(C, \mu)$, enumerating the support $\Gamma(\bar{Q}) = \{\bar{\gamma}^n | 1 \leq n \leq N\}$ and using the notation above, this cost function is of the form,

$$K(\mu, \bar{Q}) \equiv \sum_n \bar{Q}(\bar{\gamma}^n) T_\mu^C(\bar{\gamma}^n, \bar{Q}) - T_\mu^C(\mu, \bar{Q}),$$

where $T_\mu^C(\mu, \bar{Q}) = 0$ and,

$$T_\mu^C(\bar{\gamma}^n, \bar{Q}) \equiv [\bar{\gamma}^n - \mu] \cdot \int_0^1 u(\bar{a}_t^n) dt. \tag{15}$$

There are three noteworthy aspects of the result. First, the variational logic reflects the economic intuition that marginal utility of improved information should align with its marginal cost. If a large change in payoffs is required to induce a small change in the optimal posterior, learning is costly on the margin. The second point is that many action sets produce the same distribution of posteriors. For example one could shift up all payoffs by a constant amount. What we know is that (15) must be invariant to the particular action set that generates this posterior distribution. In the particular case of adding a constant to all payoffs, invariance follows because state by state differences between prior and posterior average to zero. What the general result tells us is that the corresponding invariance is fully general once A2 through A4 are assumed.

The third point of interest is that the cost function recovered in this general case has much in common with PS cost functions. The key distinction is that $T_\mu^C(\bar{\gamma}^n, \bar{Q})$ depends not only on the particular posterior $\bar{\gamma}^n$ but also the full distribution of posteriors $\bar{Q}$. Hence the computation for a fixed posterior can be entirely different should the distribution of posteriors change. This differentiates it from the PS form, to which we now turn.

7 PS and UPS Representations

In this section we introduce axioms for PS and UPS representations. The UPS characterization rests on a regularity condition introduced in Definition 13 below.

7.1 Separability

As indicated above, the first key step in the PS proof is to rule out dependence of $T_\mu^C(\gamma^n, \bar{Q})$ in (15) on the distribution of posteriors. Given $\gamma \in \Gamma(\bar{Q}) \cap \Gamma(\bar{Q}')$, we want to ensure that,

$$T_\mu^C(\gamma, \bar{Q}) = T_\mu^C(\gamma, \bar{Q}').$$

This requires an invariance axiom concerning data with shared revealed posteriors. We must be able to find decision problems that produce both distributions using common actions at shared posteriors. The logic of this axiom is demonstrated in Figure 9. Consider again the decision problem $(\mu, \{a, b\})$ of Figure 4. The optimal strategy for this decision problem involves the use of posteriors $\hat{\gamma}^a$ and $\hat{\gamma}^b$ and so these posteriors would be revealed in the data. Our separability axiom demands that for any arbitrary posterior $\hat{\gamma}^c$, such that $\{\hat{\gamma}^b, \hat{\gamma}^c\}$ can be the support for an attention strategy feasible from μ, there must exist a corresponding action c such that this pair of posteriors are revealed in the SDSC data from $(\mu, \{b, c\})$, with $\hat{\gamma}^b$ still the revealed posterior for action b (see Figure 9a).

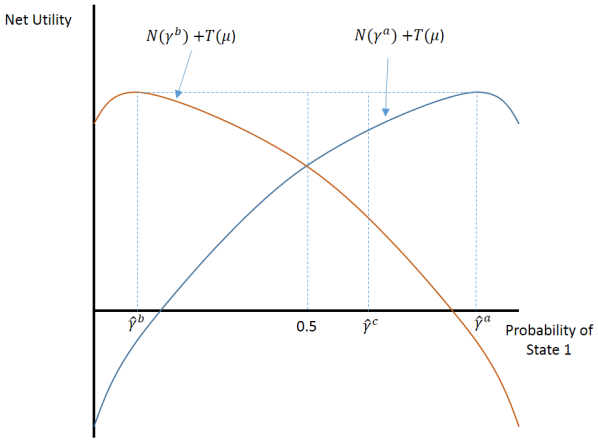

Figure 9a

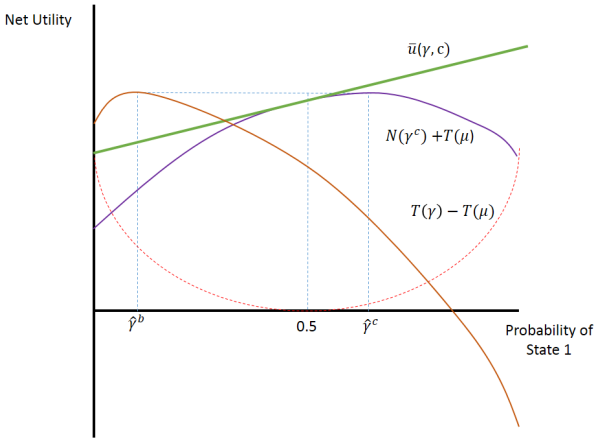

Figure 9b

The necessity of this axiom for our model is illustrated in Figure 9b. We begin with the hyperplane which defines the optimal strategy in problem $(\mu, \{a, b\})$ which is tangent to the net utility function for action a at $\hat{\gamma}^a$ and action b at $\hat{\gamma}^b$. Given the ability to shift and tilt the gross utility line defined by the payoffs, it is always possible to find an action c such that the resulting gross utility function, when combined with the cost curve, gives a net utility function which is tangent to the hyperplane precisely at $\hat{\gamma}^c$. The Lagrangian Lemma then tell us that $\{\hat{\gamma}^b, \hat{\gamma}^c\}$ define the support of an optimal strategy in the resulting decision problem, and so must be observed in the data for $(\mu, \{b, c\})$ as required.

This logic holds more generally, as stated in the following axiom.

Axiom A5 Separability: *Given* $(\mu, A(1)) \in D$, $P(1) \in C(\mu, A(1))$, *and* $Q_2 \in Q(C, \mu)$ *with* $\Gamma(Q_{P(1)}) \cap \Gamma(Q_2) \neq \emptyset$, *there exists* $A(2) \subset \mathcal{A}$ *and* $P(2) \in C(\mu, A(2))$ *satisfying* $Q_{P(2)} = Q_2$

such that $q_{P(1)}(a|\gamma) = q_{P(2)}(a|\gamma)$ *for all* $a \in A(1) \cup A(2)$ *and for all* $\gamma \in \Gamma(Q_{P(1)}) \cap \Gamma(Q_2)$.

The proof that Separability implies existence of function T_μ such that $T_\mu(\gamma) = T_\mu^C(\gamma, Q)$ in equation (15) for all $Q \in Q(C, \mu)$ is straight forward. It involves standard linear algebra arguments as well as our knowledge of the specific structure of the cost function for each fixed posterior distribution as defined by (15).

While the Separability axiom uses the existential qualifier, in the case of Shannon representations one can specify the precise change in actions needed to generate specified changes in the posteriors. This follows from the invariant likelihood ratio property specified in equation (10). This ratio is enough to pin down the action required in $A(2)$ to generate any $\gamma \in \Gamma(Q_2)/\Gamma(\mathbf{Q}_{P(1)})$ using the posteriors in $\Gamma(\mathbf{Q}_{P(1)}) \cap \Gamma(Q_2)$ and their associated actions.

7.2 Convexity Properties

With Separability, we have a rationalizing cost function of the PS form, but without the required strict convexity. In the next stage of the proof we show that there is no loss of generality in assuming the function to be **weakly** convex. In terms of rationality, there is no advantage to deliberately throwing away information, so that, even if they were present, concave portions of the cost function would never be acted on. This aspect of the proof is very much analogous to the result of Afriat [1967] that concavity can be assumed of any utility function recovered from optimizing choice in a linear budget set.

While weak convexity is guaranteed, one cannot guarantee strict convexity without additional assumptions. To this end we introduce a non-linearity axiom which insists that if one revealed posterior is a mixture of two others, then the expected utilities cannot be correspondingly mixed. This directly permits the further step from weak to strict convexity.

Axiom A6 Non-linearity: Given $(\mu, A) \in \mathcal{D}$, $P \in C(\mu, A)$, and distinct $a_1, a_2, a_3 \in \mathcal{A}(P)$ with $\gamma_P^{a_1} \neq \gamma_P^{a_3}$,

$$\gamma_P^{a_2} = \alpha\gamma_P^{a_1} + (1-\alpha)\gamma_P^{a_3} \Longrightarrow \bar{u}(\gamma_P^{a_2}, a_2) \neq \alpha\bar{u}(\gamma_P^{a_1}, a_1) + (1-\alpha)\bar{u}(\gamma_P^{a_3}, a_3).$$

7.3 From Some to All Optima

With axioms A2 through A6, we are able to identify a PS cost function $K \in \mathcal{K}^{PS}$ that rationalizes all observed data, so that $C(\mu, A) \subset \hat{P}(\mu, A|K)$. Two additional axioms are required to establish that all optimal strategies are seen, $C(\mu, A) = \hat{P}(\mu, A|K)$. We first impose a convexity property on the data.

Axiom A7 Convexity: *Given* $(\mu, A) \in D$, $P_l \in C(\mu, A)$ *for* $1 \leq l \leq L$, *and probability weights* $\alpha(l) > 0$, $P_\alpha \in C(\mu, A)$, *where,*

$$P_\alpha(a|\omega) \equiv \sum_{l=1}^{L} \alpha(l) P_l(a|\omega).$$

With this, we first show that an arbitrary optimal strategy can be decomposed (using an appropriate mixture operation) into a set of such strategies $\lambda(l)$ with linearly independent posteriors,

$$\lambda = \sum_{l=1}^{L} \alpha(l)\lambda(l).$$

Caratheodory's theorem plays the key role in this part of the proof. We then show that this mixture operation correspondingly mixes the data, so that if each of the data sets $\mathbf{P}_{\lambda(l)}$ is observed, Convexity implies that $\mathbf{P}_\lambda = \sum_{l=1}^{L} \alpha(l)\mathbf{P}_{\lambda(l)}$ must also be observed.

Our final axiom provides conditions ensuring that each data set $\mathbf{P}_{\lambda(l)}$ with linearly independent posteriors is indeed observed. A key observation in this stage concerns uniqueness of optimal strategies. A uniqueness lemma ensures that any optimal strategy that uses linearly independent posteriors is uniquely optimal provided all available actions are chosen. To apply this to strategy $\lambda(l)$, we diminish the payoffs to all actions that are unchosen in this strategy by an arbitrarily small amount. This marginal change ensures that the uniqueness result applies to all correspondingly perturbed decision problems, for each of which $\lambda(l)$ is therefore uniquely optimal. Uniqueness of optimal strategies in a PS representation implies that the corresponding SDSC data is observed.

Our process of taking perturbations allows us to construct for each $\lambda(l)$ a corresponding sequence of action sets that converges to A in the limit in such a way that $\lambda(l)$ is uniquely optimal, hence observed in the data all the way to the limit. To use convergence of this sequence of decision problems to make a conclusion on the limit problem itself requires a continuity axiom. Given $\mu \in \Gamma$ we define a payoff-based metric[14] on the space of actions,

$$d(a, a') = \left(\sum_{\omega \in \Omega(\mu)} \big(u(a,\omega) - u(a',\omega)\big)^2 \right)^{\frac{1}{2}}.$$

Axiom A8 Continuity: *Consider $I \geq 1$ sequences of actions $a^i(m)$ with $\lim_{m\to\infty} a^i(m) = \bar{a}^i$ for $1 \leq i \leq I$, and define $A(m) = \cup_{i=1}^{I} a^i(m)$ and $\bar{A} = \cup_{i=1}^{I} \bar{a}^i$. Then given $\mu \in \Gamma$ and $P \in \cap_{m=1}^{\infty} C(\mu, A(m))$,*

$$\mathcal{A}(P) \subset \bar{A} \Longrightarrow P \in C(\mu, \bar{A}).$$

This is a very weak condition concerning sequences of choice sets which converge pointwise and which have a subset of actions which remain fixed. If, at every step in the sequence, the same choice behavior is observed (which must therefore only involve choice amongst actions available in all choice sets), then that behavior must also be observed in the limit. In light of our perturbation method, this suffices to establish that all data sets $\mathbf{P}_{\lambda(l)}$ are observed in the original choice set. To complete the proof, we apply the convexity result to show that the data $\mathbf{P}_\lambda$ generated by the original optimal strategy is also observed.

[14] Technically this is a pseudo metric, as actions that differ in payoffs only in states outside $\Omega(\mu)$ will have a distance of 0 from each other. However, Axioms A2-4 guarantee that any two such actions will be treated identically by the DM, and so we will treat them as the same object for the basis of this definition.

7.4 Existence and Simple Recovery

We summarize this discussion in the following theorem.

Theorem 3: Data set $C \in \mathcal{C}$ has a PS representation if and only if it satisfies Axioms A2 through A8.

Given a PS cost function, we show that there is a relatively simple way to recover it. Given $\mu \in \Gamma$ and non-degenerate $\bar{Q} \in Q(C, \mu)$, Corollary 2 establishes existence of a choice set $\bar{A}$ such that an inattentive strategy $\eta \in \Lambda^I(\mu, \bar{A})$ and a strategy $\lambda = (Q_\lambda, q_\lambda) \in \Lambda(\mu, \bar{A})$ with $Q_\lambda(\gamma) = \bar{Q}(\gamma)$ are both optimal, hence have equal expected utility net of attention costs,

$$U(\lambda) - U(\eta) = K(\mu, \bar{Q}) - K(\mu, \eta)$$

By construction, the inattentive strategy is free, $K(\mu, \eta) = 0$, so that indifference implies that $K(\mu, \bar{Q})$ is directly computable as the difference in expected utility,

$$K(\mu, \bar{Q}) = U(\lambda) - U(\eta).$$

7.5 LIP and UPS Theorem

A single invariance axiom takes us from a PS to a UPS representation. Locally Invariant Posteriors (LIP) conveys the idea that, given $P \in C(\mu, A)$, the resulting action-posterior pairs are invariant to various changes in μ and A.

First, if the prior μ changes to μ' such the posteriors revealed in (μ, A) are still feasible, then they must still be observed in $C(\mu', A)$. The necessity of this condition is illustrated in Figure 9, which again builds on the decision problem $(\mu, \{a, b\})$ with $\mu = 0.5$ and optimal posteriors $\hat{\gamma}^a$ and $\hat{\gamma}^b$. Recall that these posteriors are identified as supporting the highest chord above the prior μ. Consider the prior μ' with $\mu'(\omega_1) = 0.3$, and note that precisely the same posteriors support the highest chord above this new prior as well, implying that they remain optimal and so must be

observed for decision problem $(\mu', \{a, b\})$.

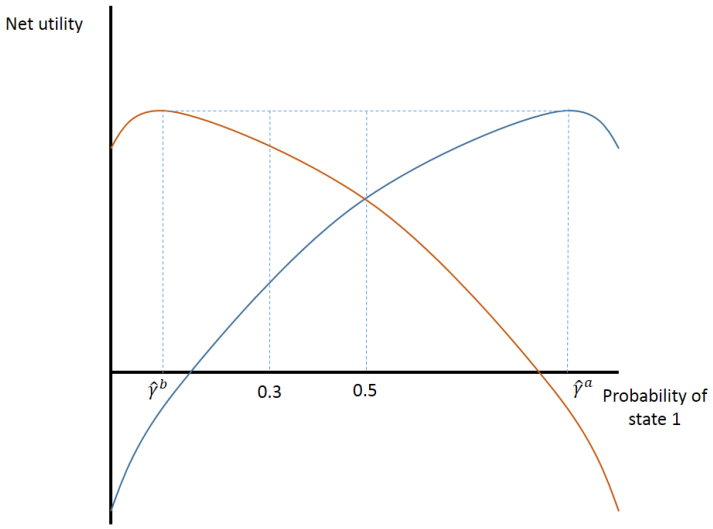

Figure 9: Locally Invariant Posteriors

Note that the Drift Diffusion Model (DDM) which has proven popular in psychology (see Ratcliff *et al.* [2016] for a recent review) satisfies this property. According to the DDM, an agent gathers information about a pair of states and acts only when posterior beliefs reach some threshold values. Since the thresholds do not change as the agent learns, the same posteriors are optimal for any prior that lies in between the posteriors.

LIP also requires that, given $P \in C(\mu, A)$, if a new decision problem is defined by deleting some available actions, then the remaining action-posterior pairs must be observed provided Bayesian consistency is retained.

The following formal definition captures both of these invariance properties.

Axiom A9 Locally Invariant Posteriors (LIP): *Consider* $(\mu, A) \in \mathcal{D}$, $P \in C(\mu, A)$, *and probabilities* $\rho(a) > 0$ *on* $A' \subset \mathcal{A}(P)$ *with* $\sum_{a \in A'} \rho(a) = 1$. *Define* $P' \in \mathcal{P}$ *by* $\mathcal{A}(P') = A'$, $\mathbf{Q}_{P'}(\gamma) = \sum_{\{a \in A' | \gamma_P^a = \gamma\}} \rho(a)$ *and:*

$$\mathbf{q}_{P'}(a|\gamma) = \begin{cases} \frac{\rho(a)}{\mathbf{Q}_{P'}(\gamma)} \text{ if } \gamma_P^a = \gamma; \\ 0 \text{ else.} \end{cases}$$

Then $P' \in C\left(\sum_{a \in A'} \rho(a)\gamma_P^a, A'\right)$.

Our fourth theorem shows essentially that a data set with a PS representation has a UPS representation if and only if it satisfies LIP. There is a caveat. For necessity of LIP (Axiom A9), we

insist on a link between the posteriors $\Gamma(C,\mu_1)$ and $\Gamma(C,\mu_2)$ for distinct priors. If prior μ_2 lies in the convex hull of posteriors that are revealed posteriors from prior μ_1, then these posteriors must be observed from prior μ_2 also. We present a simple example in Appendix 4 in which this does not hold. We define regular data sets as those that have this property globally.

Definition 14 *Data set C is **regular**, $C \in \mathcal{C}^R \subset \mathcal{C}$, if, given $\mu_1 \in \Gamma$ and $Q \in \Delta(\Gamma(\mu_1))$ with $\Gamma(Q) \subset \Gamma(C,\mu_1)$,*

$$\sum_{\gamma \in \Gamma(\mu_2)} \gamma Q(\gamma) = \mu_2 \Longrightarrow \Gamma(Q) \subset \Gamma(C,\mu_2).$$

Note that the Shannon model generates a data set that is regular, as do other standard entropies.

Theorem 4: If $C \in \mathcal{C}$ has a PS representation and satisfies LIP (Axiom A9), it has a UPS representation. If $C \in \mathcal{C}^R$ has a UPS representation then it satisfies LIP.

The proof of theorem 4 is lengthy yet conceptually straight forward. It relies on the Lagrangian Lemma and elementary linear algebra. It also relies on invariance of the cost function under affine transforms of the strictly convex function T_μ.

Note that between them, theorems 1, 3, and 4 show that data set $C \in \mathcal{C}$ has a Shannon representation if and only if it satisfies Axioms A1 through A9. For the sake of completeness, we establish this as Corollary 3 in Appendix 5.

8 Further Results

In this section we provide further results that expand on various features of our representation. We first show how to obtain a representation when the IO observes only a single piece of SDSC for each decision problem - i.e. a choice function rather than a choice correspondence. Second, we describe the relationship between our model and the more traditional model of costly information acquisition in which the DM chooses between information structures consisting of signals, rather than probability distributions over posteriors. Finally we introduce Tsallis entropy (Tsallis [1988]), an alternative formulation to that of Shannon which is of value in describing physical and social systems (see Section 9.3). Costs based on Tsallis entropy fall in the UPS class but do not satisfy IUC, as we demonstrate.

8.1 Choice Functions

To explore the application of our approach to choice functions, we let $\mathcal{C}^F$ be the set of data sets in which there is only one observation of SDSC data for each decision problem,

$$\mathcal{C}^F \equiv \left\{ C^F : \mathcal{D} \to \mathcal{P} | C^F(\mu, A) \in \mathcal{P}(\mu, A) \right\}.$$

Given there will be multiple optima in some decision problems, there are several distinct forms of representation that may be of interest. The most obvious approach captures observation of a selection from the optimal choice correspondence.

Definition 15 *Data set $C^F \in \mathcal{C}^F$ has a **functional costly information representation** (FCIR) $K \in \mathcal{K}$ if, for all $(\mu, A) \in \mathcal{D}$,*

$$C^F(\mu, A) \in \hat{P}(\mu, A|K).$$

It has a FPS/FUPS/F-Shannon representation if it is has an FCIR with $K \in \mathcal{K}^{PS}/\mathcal{K}^{UPS}/K = K^S_\kappa$ for $\kappa > 0$.

We can also consider the case in which the DM mixes among strategies when there are multiple optima, meaning that the observed data falls in the convex hull of the data generated by optimal strategies.

Definition 16 *Data set $C^F \in \mathcal{C}^F$ has a **mixed functional costly information representation** (MCIR) $K \in \mathcal{K}$ if, for all $(\mu, A) \in \mathcal{D}$,*

$$C^F(\mu, A) \in Conv\left\{\hat{P}(\mu, A|K)\right\},$$

It has a MPS/MUPS/M-Shannon representation if it is has an MCIR with $K \in \mathcal{K}^{PS}/\mathcal{K}^{UPS}/K = K^S_\kappa$ for $\kappa > 0$.

Our first observation is that a data set will have a FPS representation if and only if it has an MPS representation. This follows from the fact that, for the PS model, $\hat{P}(\mu, A|K) = Conv\left\{\hat{P}(\mu, A|K)\right\}$. Thus we can concentrate on identifying conditions which allow for the former type of representation.

The key to functional extensions of our approach is a recoverability result in the spirit of that outlined in Section 6, whereby Axioms A2 through A4 alone are enough to uniquely pin down a rationalizing cost function. In the case of a functional representation, we cannot guarantee that all distributions over posteriors will be observed in the data. However, it is the case that all distributions with linearly independent support will be observed, as all such strategies are uniquely optimal in some decision problem if costs are posterior separable. It is therefore possible to uniquely identify costs for all such attention strategies. Moreover, there is a unique way to extend this cost function to all attention strategies in a manner consistent with posterior separability. If we define mixtures of posterior distributions as in Appendix 2,

$$Q = \sum_{l=1}^{L} \alpha_l Q_l \Leftrightarrow Q(\gamma) = \sum_{l=1}^{L} \alpha(l) Q_l(\gamma) \text{ for all } \gamma \in \Gamma(Q),$$

posterior separability of costs implies that,

$$\begin{aligned} Q &= \sum_{l=1}^{L} \alpha_l Q_l \\ &\Rightarrow \quad K(\mu, Q) = \sum_{l=1}^{L} \alpha_l K(\mu, Q_l). \end{aligned}$$

Thus if a data set has a FPS representation then it is possible to uniquely identify those costs K from the data.[15] Having done so, one can then identify all SDSC which are consistent with optimal behavior with respect to this cost function $\hat{P}(\mu, A|K)$. Treating this as a data set, we can then apply the relevant axioms: for an FPS $\hat{P}$ must satisfy Axioms A5-A8, for a FUPS it must also satisfy Axiom A9 and for F-Shannon it must also satisfy Axiom A1. In this way we can construct necessary and sufficient conditions for functional representations.

8.2 Costly Signal Acquisition

The standard approach to modeling optimal acquisition of costly information specifies an information structure, consisting of a joint distribution of signals and states. The DM chooses amongst these structures, which are subject to some cost function (see for example Caplin and Dean [2015]). A signal-based strategy comprises an information structure and a mixed action strategy mapping signals to distributions over chosen actions. As is standard, and as we assume in our posterior-based approach, costs depends only on the information structure, not the action strategy. The DM faced with decision problem $(\mu, A) \in \mathcal{D}$ is modeled as choosing a signal-based strategy to maximize expected utility net of information costs.

The signal-based and posterior-based approaches are equivalent in the sense that a data set can be rationalized by optimal choice of signal-based strategy if and only if it can be rationalized by optimal choice of posterior-based strategies. To go from a CIR in our sense to a corresponding cost function on information structures involves little more than identifying the signals with the posteriors. The argument in the reverse direction involves identifying posteriors associated with the various actions and correspondingly transforming the mixed strategy.

While the data that is characterized is the same using our posterior-based formulation and the standard signal-based formulation, there is a key distinction with regard to testability. Subjective signals are observable only indirectly, through their impact on updating and thereby behavior. From the viewpoint of choice-based analysis, the posterior-based approach has the advantage that it by-passes unobservable signals.

8.3 Tsallis Entropy and Failures of IUC

The IUC property seems sufficiently reasonable as to be more widely true. To understand how IUC fails for cost functions other than Shannon, we show how the condition fails for the class of cost functions associated with entropy functions introduced by Tsallis [1988].

For $\sigma \in \mathbb{R}$, $\sigma \neq 1$, the Tsallis entropy of posterior $\gamma \in \Gamma$ is defined by,

$$TS_\sigma(\gamma) = \frac{1}{\sigma - 1}\left(1 - \sum_{\omega \in \Omega(\gamma)} \gamma(\omega)^\sigma\right) \in \mathbb{R}.$$

As $\sigma \to 1$, Tsallis entropy heads in the limit to Shannon entropy, $H(\gamma)$.

A key property of Tsallis entropy is that it is non-additive. Given two independent probability

[15] With regard to attention strategies with linearly dependent support, one can insist that these are only used when optimal according to the recovered cost function.

distributions γ^1 and γ^2, the entropy of the product distribution can be related to the entropy of the marginal distributions,

$$TS_\sigma(\gamma^1 \times \gamma^2) = TS_\sigma(\gamma^1) + TS_\sigma(\gamma^2) + (1-\sigma)TS_\sigma(\gamma^1)TS_\sigma(\gamma^2).$$

Shannon entropy ($\sigma = 1$) is the special case of additivity.

Given $\mu \in \Gamma$ it is simple to define the Tsallis cost function for information structures with $\Gamma(Q) \subset \tilde{\Gamma}(\mu)$ in a manner completely analogous to the Shannon model. Costs are related to the expected Tsallis entropy of the posteriors less that of the prior, again with multiplicative factor $\kappa > 0$,

$$K_\kappa^{TS_\sigma}(\mu, Q) = -\kappa \left[\sum Q(\gamma) TS_\sigma(\gamma) - TS_\sigma(\mu) \right].$$

Recall that what is costly is reducing entropy so $K_\kappa^{TS_\sigma}$ is decreasing in the entropy of the posteriors. $K_\kappa^{TS_\sigma}(\mu, Q)$ is real-valued for all distributions $Q \in \Delta(\Gamma(\mu))$.[16]

This cost function is a member of the UPS class, and so the resulting behavior satisfies Axioms A2-A9. However it violates IUC. Consider a problem (μ, A) and suppose that states $\omega_1, \omega_2 \in \Omega(\mu)$ are identical in payoff terms, so that, $u(a, \omega_1) = u(a, \omega_2)$ for all $a \in A$. Consider $P \in C(\mu, A)$ and suppose without loss of generality that each action is chosen from one and only one posterior so that $\mathbf{Q}_P(\gamma_P^a) = P(a)$. Now consider $K_\kappa^{TS_\sigma}(\mu, \mathbf{Q}_P)$:

$$\kappa \sum_{\gamma_P^a \in \Gamma(\mathbf{Q}_P)} \mathbf{Q}_P(\gamma_P^a) \sum_{\omega \in \Omega(\mu)} \gamma_P^a(\omega) \left(\frac{\gamma_P^a(\omega)^{\sigma-1} - 1}{\sigma - 1} \right) - \kappa \sum_{\omega \in \Omega(\mu)} \mu(\omega) \left(\frac{\mu(\omega)^{\sigma-1} - 1}{\sigma - 1} \right);$$

where we have pulled out multiplicative factor $\gamma_P^a(\omega)$ to make explicit the relationship to a constant elasticity function. Substituting using Bayes' rule, $\gamma_P^a(\omega) = \frac{P(a|\omega)\mu(\omega)}{P(a)}$ and invoking $\sum_\omega P(a|\omega) = 1$, leads to the following expression for Tsallis costs in terms of SDSC data:

$$\begin{aligned} K_\kappa^{TS_\sigma}(\mu, \mathbf{Q}_P) &= \sum_{a \in \mathcal{A}(P)} \sum_{\omega \in \Omega(\mu)} P(a|\omega)\mu(\omega)^\sigma \left[\frac{(P(a|\omega)/P(a))^{\sigma-1} - 1}{\sigma - 1} \right] \\ &\quad - \sum_{\omega \in \Omega(\mu)} \mu(\omega) \left(\frac{\mu(\omega)^{\sigma-1} - 1}{\sigma - 1} \right) \end{aligned}$$

Now suppose that IUC holds so that $P \in C(\mu, A)$ implies $P(a|\omega_1) = P(a|\omega_2)$ for all $a \in A$. We now focus on the part of this expression associated with a single action $a \in A$ and the two states

[16] A subtle point is that there are cases in which an ex ante possible state may be ruled out, as when $\Omega(\mu) = \{\omega_1, \omega_2, \omega_3\}$ yet $\gamma \in \Gamma(Q)$ has support $\Omega(\gamma) = \{\omega_1, \omega_2\}$. The above formula correctly deals with this case when when $\sigma > 0$ because the contribution of these terms to the sum is zero so that their exclusion is immaterial.

Matters are slightly more complex when $\sigma < 0$. In this case there are infinite costs to ruling out ex ante possible states. This calls for care in specifying the Tsallis attention cost function. Given $\mu \in \Gamma$, the corresponding cost function is:

$$K_\kappa^{TS_\sigma} = \begin{cases} \kappa \left[\sum Q(\gamma) TS_\sigma(\gamma) - TS_\sigma(\mu) \right] & \text{if } \Omega(\gamma) = \Omega(\mu) \text{ all } \gamma \in \Gamma(Q); \\ \infty & \text{if } \Omega(\gamma) \neq \Omega(\mu) \text{ some } \gamma \in \Gamma(Q). \end{cases}$$

The need to depart from the standard specification of Tsallis entropy in the above cases is due to what is essentially a missing argument. The standard Tsallis entropy function makes no explicit reference to the prior. Yet the cost of making an ex ante possible state impossible becomes unboundedly high at the margin when $\sigma < 0$, so that making it free to entirely rule such a state out would be inappropriate.

ω_1 and ω_2:

$$\begin{aligned} & P(a|\omega_1)\mu(\omega_1)^\sigma \frac{(P(a|\omega_1)/P(a))^{\sigma-1}-1}{\sigma-1} + P(a|\omega_2)\mu(\omega_2)^\sigma \frac{(P(a|\omega_2)/P(a))^{\sigma-1}-1}{\sigma-1} \\ = & \left(P(a|\omega_1) \frac{(P(a|\omega_1)/P(a))^{\sigma-1}-1}{\sigma-1} \right) \left[\mu(\omega_1)^\sigma + \mu(\omega_2)^\sigma\right]. \end{aligned}$$

We now compare this to the cost that would be incurred if ω_1 and ω_2 were instead collapsed into the single state ω_1 with prior probability $\mu(\omega_1)+\mu(\omega_2)$. If, as specified by IUC, the choice probabilities remain $P(a|\omega_1)$

$$\left(P(a|\omega_1) \frac{(P(a|\omega_1)/P(a))^{\sigma-1}-1}{\sigma-1} \right) \left[\mu(\omega_1) + \mu(\omega_2)\right]^\sigma .$$

If $\sigma < 1$ then the decision maker finds it more costly to learn about ω_1 and ω_2 separately than together,

$$\mu(\omega_1)^\sigma + \mu(\omega_2)^\sigma > \left(\mu(\omega_1) + \mu(\omega_2)\right)^\sigma .$$

If $\sigma > 1$, the opposite is the case. It is clear that these changes in the marginal cost of information mean that the same $P(a|\omega_1)$ cannot generally be optimal in the original problem and its basic form, leading to a violation of IUC.

Only if $\sigma = 1$ does the DM treat the two scenarios as equivalent. Recall that as $\sigma \to 1$, Tsallis entropy approaches Shannon entropy. Shannon entropy is therefore the special case in which the agent is indifferent between aggregating and separating states. This is the essence of the IUC axiom. With Shannon, the cost of implementing $P(a|\omega)$ rises proportionately with $\mu(\omega)$, whereas with Tsallis entropy costs rise more than proportionately with $\mu(\omega)$ when $\sigma > 1$, and less than proportionately when $\sigma < 1$. The implication is that when $\sigma < 1$, information is proportionately cheaper in more likely states, so that an agent would appear to pay greater attention in such states.

9 Relation to the Literature

9.1 Existing Characterizations of the Shannon Model

Several recent papers have provided insights into the behavior implied by the Shannon model. Matejka and McKay [2015] use first order conditions to provide a generalized logit formula for optimal SDSC probabilities $P(a|\omega)$ in the Shannon model. On its own, this condition is necessary but not sufficient to characterize Shannon-consistent behavior.[17] Subsequent papers (Caplin and Dean [2013], Stevens [2014], and Caplin *et al.* [2018]) show that the addition of appropriate complementary slackness conditions provides both necessity and sufficiency.

[17] Proposition 2 of the same paper shows that if two axioms (IIA Actions and IIA Alternatives) are satisfied then their exists unconditional action probabilities and utilities over payoffs such that choice probabilities are of generalized logit form.

This result is significantly weaker than the characterization presented here for three reasons. First, this proposition refers only to data from a single decision problem - it does not provide conditions under which data from many different decision problems are jointly consistent with Shannon. Second, it does not guarantee that the unconditional action probabilities which rationalize the data are the ones that would emerge from the Shannon model, given utilities and priors. Finally, the generalized logit form is necessary, but not sufficient for data in a given decision problem to be consistent with the Shannon model.

From the starting point of these papers, our work extends understanding of rational inattention in a number of ways. Most obviously, we study and characterize the broader class of PS and UPS models. Furthermore, while insightful, first order conditions for optimality are not directly revealing of the behavioral patterns that the model produces. In contrast, our analysis is of value in this regard, providing a number of benefits. First, we establish the behavioral counterparts to different features of the Shannon cost function: NIAS and NIAC for the data to be consistent with any arbitrary cost function; Separability for costs to be posterior separable; LIP for the same cost function to hold across all priors; and IUC for the Shannon form. This means that behavioral violations of the Shannon model can be attributed to specific features of the cost function, aiding model development. Second, our IUC axiom is of independent interest as it captures the behavioral sense in which Shannon is an idealized model of learning. The fact that it is this feature alone which identifies Shannon within the UPS class cannot be established directly from the first order conditions. Finally, many of the tools we describe - such as the posterior based approach to optimal strategies in the Shannon model and the geometry of net utility functions - have already proved useful in economic research since their introduction in Caplin and Dean [2013] (see for example Caplin *et al.* [2015] and Martin [2017]). For example, in the UPS case, LIP makes it relatively easy to derive comparative static results as priors change.

A more closely related analysis is that of de Oliveira [2014], who uses three axioms to characterize decision making given a Shannon cost function. The key difference is that de Oliveira [2014] places axioms on preference orderings over menus, whereas we place axioms on choices as revealed in SDSC data.[18] Nevertheless some link between the two axiomatizations can be drawn. The Symmetry axiom of de Oliveira [2014] says that states which have the same probability can have their roles exchanged without affecting preferences. This in turn means that costs and optimal information structures are also symmetric, which is implied by the IUC condition. IUC also appears related to de Oliveira [2014]'s independence of orthogonal decision problem (IODP) axiom. IODP involves indifference between solving two decision problems with independent payoffs together or separately. We early on conjectured that we would need both IUC and IODP to generate the Shannon form. We only later realized that IUC alone was sufficient. It is therefore possible that IUC implies IODP. de Oliveira [2014] also does not consider generalizations of the Shannon model.

Pioneering work by Shannon [1948] and Khinchin [1957] provides direct axiomatizations of Shannon entropy. Axioms such as continuity, being maximal at uniformity, being invariant to zero probability events, and satisfaction of additivity conditions are shown to imply the Shannon entropy function for probability distributions. This work is focussed on properties of measures of disorder, rather than understanding the behavioral implications of associated attention cost functions.

To see the difference between the two approaches, it is instructive to compare IUC with Shannon's third axiom, as the properties bear a superficial resemblance. Shannon's axiom states that if the cost of information is invariant to whether it is revealed all at once or in stages. Stated in

[18] There is, however a potential link between the two data sets. A preference over decision problems is closely related to the value function for that decision problem, while state-dependent stochastic choice is closely related to the optimal distribution over posteriors. Hence the two datasets are connected because the optimal distribution over posteriors is identified by the subgradient of the value function. We thank an anonymous referee for pointing this out. Exploring how the model could be jointly axiomatized on the two data sets is a promising avenue for future work.

terms of combining two states the axiom states that:

$$\begin{aligned} & H(\gamma(\omega_1), \gamma(\omega_2), ...) \\ = \; & H(\gamma(\omega_1) + \gamma(\omega_2), \gamma(\omega_3)...) + (\gamma(\omega_1) + \gamma(\omega_2)) H\left(\frac{\gamma(\omega_1)}{\gamma(\omega_1) + \gamma(\omega_2)}, \frac{\gamma(\omega_2)}{\gamma(\omega_1) + \gamma(\omega_2)}\right), \quad (16) \end{aligned}$$

for all distributions γ. (16) states that the entropy of $\{\gamma(\omega_1), \gamma(\omega_2), ...\}$ is equal to the entropy of $\{\gamma(\omega_1) + \gamma(\omega_2), \gamma(\omega_3)...\}$ plus the entropy of breaking the first state into $\{\gamma(\omega_1), \gamma(\omega_2)\}$.

Note several differences between this axiom and IUC. First, (16) is a property of the cost function, whereas IUC is a property of behavior. It is not immediately obvious what restrictions IUC places on the cost function. These implications are all indirect results of the behavioral restrictions. Second, (16) holds for all γ and IUC holds only for decision problems such that the payoff to state to states ω_1 and ω_2 are the same. The hard work in the sufficiency proof is precisely showing why this property has such powerful global implications for the shape of the cost function. Third, (16) concerns two ways of revealing the same γ, whereas IUC compares γ of two different dimensions. In effect, IUC equates behavior under $\{\gamma(\omega_1), \gamma(\omega_2), ...\}$ and $\{\gamma(\omega_1) + \gamma(\omega_2), \gamma(\omega_3)...\}$ which are not the same according to (16) as they differ by the second term.

9.2 Limits to the Shannon Model

Just as the restrictions that a Cobb-Douglas utility function do not apply to all choice settings, the restrictions that the Shannon model places on SDSC data do not hold universally. IUC, in particular, implies that states are defined only by their payoffs. In some cases, however, behavior inconsistent with the Shannon model can be tied directly to the importance of payoff irrelevant information.[19]

One case concerns perceptual distance. Perceptual distance is critical in many every day decisions, as when good decisions require the DM to differentiate between alternative pricing schemes: it seems likely that prices which are closer together will be harder to distinguish than those which are far apart. By way of conceptual confirmation, Dean and Neligh [2017] design an experiment with 100 balls on a screen, of which a random number (between 40 and 60) are red, with the remainder blue. Subjects are tasked with correctly identifying which color ball is in the majority. According to the Shannon model, there are two states: more red balls and more blue balls. The exact number of balls is not payoff relevant. The Shannon model therefore implies that subjects must be just as good at the task when there are 51 red balls on the screen as when there are 60, which is strongly rejected by the data.

Woodford [2012] cites another case. He discusses the experimental results of Shaw and Shaw [1977], in which a subject briefly sees a symbol which may appear at one of a number of locations on a screen. Their task is to accurately report the symbol. According to the Shannon model the state is defined only by the symbol. The location on the screen is payoff irrelevant and therefore should also be irrelevant to task performance. Yet in practice, performance is superior at locations

[19] Caplin and Dean [2013] provide another example of behavior inconsistent with the Shannon model. The functional form of the Shannon model makes precise predictions about the rate at which subjects improve their accuracy in response to improved incentives. Caplin and Dean [2013] show in a simple two state, two action set-up that agents do not pay enough attention at high rewards given the attention paid at low rewards. This behavior could be rationalized by a UPS cost function that is more convex than Shannon or by a cost function in which it its easier to learn posteriors in some neighborhood of the prior. The former case violates IUC. The latter violates LIP.

that occur more frequently.

9.3 PS Models

UPS models were introduced in Caplin and Dean [2013], while this paper is the first to introduce the broader category of PS cost functions. We believe that this class of models provide an attractive combination of tractability and flexibility.

In terms of tractability, the PS class is the broadest that allows solution using the technique of 'concavification', by which optimal behavior can be determined by identifying the tangent to the concavified net utility function. This approach has been widely used since its introduction to the economics literature (Aumann *et al.* [1995]). Most notably, the 'Bayesian Persuasion' literature (Kamenica and Gentzkow [2011]) has used concavification to successfully approach a number of problems in information economics (see Alonso and Câmara [2016] and Ely and Szydlowski [2017] for recent examples). Indeed, since their introduction, several papers have made use of UPS costs functions for rational inattention - see for example Steiner *et al.* [2015], Clark [2016] and Morris and Strack [2017]. Of particular note are papers that have used PS cost functions to examine situations in which both costly information acquisition and persuasion are important (Gentzkow and Kamenica [2014], Matyskova [2018]).

PS cost functions are rich enough to allow for many of the behavioral findings that call the Shannon model into question. With regard to incentives, Caplin and Dean [2013] develop a simple two parameter UPS model that generalizes the Shannon model. They find that the additional degree of freedom leads to a significantly better fit of the data according to the Akaike Information Criterion. With regard to perceptual distance, while rejecting the Shannon model, Dean and Neligh [2017] find (weak) support for LIP, and hence the UPS model. Finally, note that while the results of Shaw and Shaw [1977] are inconsistent with the Shannon model, they are consistent with the UPS model. In the Tsallis model with $\sigma < 1$, for example, learning about unlikely states is proportionately more expensive than about likely states. This produces a commensurately greater error rate, as in the experiment.

One particular extension which is allowed by the PS class is to replace Shannon with other forms of entropic cost. This has proved valuable in other disciplines. There are many settings in which the additional flexibility they allow for leads to a better ability to describe physical and social systems. Examples include internet usage (Tellenbach *et al.* [2009]), machine learning (Maszczyk and Duch [2008]), statistical mechanics (Lenzi *et al.* [2000]), and many other applications in physics (Beck [2009]). See Gell-Mann and Tsallis [2004] for a review. In these cases, the additivity property of Shannon entropy is found to be unhelpful in describing the phenomena of interest.

Interestingly, the literature in information theory and on the design of experiments has also focussed on PS cost functions. For example, the Blackwell-Sherman-Stein Theorem shows that PS functions can be used to characterize the property of statistical sufficiency, and so provide an alternative characterization of Blackwell's theorem. The theorem states that an information structure π is statistically sufficient for π' (i.e. π Blackwell dominates π') if and only if,

$$\sum_{\gamma \in \Gamma(Q_\pi)} Q_\pi(\gamma) T(\gamma) \geq \sum_{\gamma \in \Gamma(Q_{\pi'})} Q_{\pi'}(\gamma) T(\gamma),$$

for every continuous, (weakly) convex T, where Q_π is the distribution over posteriors generated

by π (see for example Le Cam [1996]).[20] Torgersen [1991] further shows that the class of PS cost functions can be characterized by properties of the costs themselves. Specifically, the (weakly convex) PS class of cost function of information structures characterizes monotonicity in Blackwell informativeness and linearity in a natural mixture operation.

Recent work of Hébert and Woodford [2017] provides an entirely different perspective on why UPS cost functions may be of interest. Their paper is similarly motivated to ours: they look to generalize the Shannon cost function to more closely match observation. To arrive at these general forms, they consider models of optimal sequential learning. They model costly information processing with essentially unrestricted flow costs of incremental updating from any given posterior. They allow for differential costs of discriminating among states and analyze the corresponding optimal stopping problem. Despite having an entirely different starting point, their theorem 1 ends up pinpointing static UPS cost functions as being of particular interest. It shows equivalence between the information that is acquired through their process of continuous updating and optimal stopping, and the information acquired in a static model with a cost function in the UPS class. They also show how to derive the particular cost function from the local structure of learning.

The link that Hébert and Woodford [2017] establish between static UPS models and continuous time models of optimal stopping enhances interest both in the broad class and in those functions that capture particular respects in which the Shannon model may be unrealistic in application. Their result also suggests the value of enriching observations to include stopping times. It will be of interest in future to characterize the joint distribution of action choices and stopping times that UPS models produce.

9.4 Alternative Models of Limited Attention

Our work belongs to a recent literature which characterizes the behavior associated with models of incomplete attention - see for example Masatlioglu *et al.* [2012], Manzini and Mariotti [2014] and Steiner and Stewart [2016]. It is also related to significant bodies of work on costly information acquisition with very different forms of cost function. The most ubiquitous such model is search theoretic, involving a fixed cost of uncovering each available option (e.g. Caplin *et al.* [2011]). Other approaches include costly purchase of normal signals (Verrecchia [1982], Llosa and Venkateswaran [2012] and Colombo *et al.* [2014]) and "all or nothing" information costs (Reis [2006]). Even in the rational inattention literature alternative cost functions have been provided. For example, Paciello and Wiederholt [2014] consider costs that are convex in mutual information, while Sims [2003] considers a model in which there is a hard constraint on the amount of mutual information a DM can use. Inspired by the findings of Shaw and Shaw [1977], Woodford [2012] considers a cost function which is linear in Shannon capacity, rather than Shannon mutual information. Another ongoing body of work to which our modeling relates is the sparsity-based model of Gabaix [2014]. This model is based on a distinct form of attention cost function involving fixed costs of comprehending individual characteristics of options. The question of how these other cost functions restrict behavior, and so how they differ from the PS class, remains open.

[20] We thank Daniel Csaba for pointing this out to us.

10 Concluding Remarks

Together our results provide necessary and sufficient conditions for cost functions of increasing specificity. Theorem 3 states that Axioms A2-A8 are necessary and sufficient for the existence of a Posterior Separable attention cost function. In addition, given a Posterior Separable cost function, Theorem 4 states that Locally Invariant Posteriors (Axiom A9) is necessary and sufficient for the existence of a Uniformly Posterior Separable cost function. Finally, given a Uniformly Posterior Separable cost function, Theorem 1 states that Invariance under Compression (Axiom A1) is necessary and sufficient for the cost function to take the Shannon form. In addition, Theorem 2 states that Axioms A2-A4 are sufficient for there to exist a unique attention cost function that represents the data.

11 Notational Glossary

symbol	description	def #	page
$\mathcal{A}$	set of available actions	–	5
$a, b, c...$	generic action in $\mathcal{A}$	–	5
A	generic non-empty finite subset of $\mathcal{A}$	–	5
$\mathcal{A}(\lambda)$	actions chosen with positive probability given strategy λ	3	6
$\mathcal{A}(P)$	actions chosen with positive probability given data P	11	18
$\mathcal{B}$	set of basic decision problems	12	19
$\mathcal{B}(\mu, A)$	set of basic decision problems corresponding to (μ, A)	13	20
C	generic observed data correspondence: mapping from decision problems to subsets of observed data $\mathcal{P}$ such that $C(\mu, A) \subset \mathcal{P}(\mu, A)$	–	17
$C(\mu, A)$	set of observed SCSD data given decision problem (μ, A)	–	17
$\mathcal{C}$	the set of all possible observed data	–	17
$\mathcal{D}$	set of all decision problems	2	5
$\Delta(\Omega)$	set of probability densities over Ω with finite support	1	5
$\Delta(\Gamma(\mu))$	set of probability distributions over $\Gamma(\mu)$ with finite support	3	5
Γ	shorthand for $\Delta(\Omega)$	1	5
γ	generic element of Γ	–	5
$\Gamma(\mu)$	set of probability densities over $\Omega(\mu)$	1	5
$\tilde{\Gamma}(\mu)$	set of probability densities such that $\gamma(\omega) > 0$ for all ω in $\Omega(\mu)$	1	5
$\hat{\Gamma}(\mu\|K)$	set of posteriors optimal for some problem given prior μ and cost function K	5	9
$\Gamma(Q)$	the support of $Q \in \Delta(\Gamma)$ in Γ	3	5
$\Gamma(P)$	set of revealed posteriors given P	11	18
$\Gamma(C, \mu)$	set of posteriors given observed data C and prior μ	–	28
γ_P^a	revealed posterior associated with action a given P	11	18
Γ^{J-1}	corner of the unit $J-1$ dimensional cube	–	14
$\mathcal{F}$	set of all priors and Bayes' consistent posterior distributions: (μ, Q) such that $\mu \in \Gamma$ and $Q \in \mathcal{Q}(\mu)$	4	7
$\hat{\mathcal{F}}(\mu\|K)$	set of all priors and Bayes' consistent posterior distributions: (μ, Q) such that $\mu \in \Gamma$ and $Q \in \hat{\mathcal{Q}}(\mu\|K)$	5	9
$H(\gamma)$	Shannon entropy of γ	–	8
$\mathcal{K}$	set of all attention cost functions $K : \mathcal{F} \to \bar{\mathbb{R}}$ such that inattention is feasible	4	7
$\mathcal{K}^{PS}$	set of all posterior separable cost functions in $\mathcal{K}$	6	9
$\mathcal{K}^{UPS}$	set of all uniformly posterior separable cost functions in $\mathcal{K}$	7	10
K	generic cost function in $\mathcal{K}$	4	7
$K_\kappa^S(\mu, Q)$	Shannon cost of posteriors Q given prior μ and cost parameter κ	–	8
$\Lambda(\mu, A)$	set of feasible posterior-based attention strategies given (μ, A)	3	5
$\hat{\Lambda}(\mu, A\|K)$	set of optimal posterior-based strategies given $(\mu, A) \in \mathcal{D}$ and $K \in \mathcal{K}$	–	8
$\Lambda^I(\mu)$	set of inattentive strategies	4	7

symbol	description	def #	page
λ	generic element of $\Lambda(\mu, A)$	3	6
λ^*	specific choice of $\lambda \in \Lambda(\mu, A)$	–	6
$\boldsymbol{\lambda}_P$	revealed posterior-based attention strategy	11	18
μ	generic prior	–	5
$\mu(\omega)$	prior probability of state ω	–	5
(μ, A)	a decision problem with prior $\mu \in \Gamma$ and choice set $A \subset \mathcal{A}$	2	5
$N^a(\gamma)$	net utility: expected utility to action a given posterior γ less cost of posterior γ in the UPS model		11
$N^a_\mu(\gamma)$	net utility: expected utility to action a given posterior γ less cost of posterior γ in the PS model given μ		12
Ω	set of all conceivable states of the world	–	5
$\Omega(\gamma)$	states in Ω to which γ assigns positive probability	1	5
$\omega, \omega_1, \omega_2$	generic states in Ω	–	5
$\tilde{\Omega}$	a fixed set of states	–	23
P	a mapping from states $\Omega(\mu)$ to action probabilities $\Delta(A)$	8	15
$P(a\|\omega)$	probability of action a in state ω according to $P \in \mathcal{P}(\mu, A)$	8	15
$P(a)$	unconditional action probabilities given $P \in \mathcal{P}(\mu, A)$	11	18
$\mathcal{P}(\mu, A)$	set of all possible SDSC data for decision problem (μ, A)	8	15
$\mathcal{P}$	set of all possible SDSC data	8	15
$\mathbf{P}_\lambda$	SCSD data generated by strategy $\lambda \in \Lambda(\mu, A)$	9	17
$\mathbf{P}_\lambda(a\|\omega)$	action probabilities given state ω generated by $\lambda \in \Lambda(\mu, A)$	9	17
$\mathbf{P}_\lambda(a)$	action probabilities generated by $\lambda \in \Lambda(\mu, A)$	9	17
$\hat{P}(\mu, A\|K)$	SCSD data generated by all $\lambda \in \hat{\Lambda}(\mu, A\|K)$	10	17
$\mathcal{Q}(\mu)$	finite support distributions over $\Delta(\Gamma(\mu))$ satisfying Bayes' rule given μ	3	6
Q_λ	generic element of $\mathcal{Q}(\mu)$ associated wtih $\lambda \in \Lambda(\mu, A)$	3	6
$\mathcal{Q}(C, \mu)$	subset of $\mathcal{Q}(\mu)$ observed in C from prior μ	–	28
$\hat{\mathcal{Q}}(\mu\|K)$	elements of $\mathcal{Q}(\mu)$ that are also distributions over elements of $\hat{\Gamma}(\mu\|K)$	5	8
Q_P	revealed distribution of posteriors	11	18
q_λ	given $\lambda \in \Lambda(\mu, A)$ mapping from $\Gamma(Q_\lambda)$ to $\Delta(A)$	3	6
$q_\lambda(a\|\gamma^a)$	probability of a given γ^a according to q_λ		7
$q_P(a\|\gamma^a_P)$	revealed choice probability given revealed posterior	11	18
T_μ	generic strictly convex function from $\Gamma(\mu)$ to $\bar{\mathbb{R}}$	6	9
T	generic strictly convex function from Γ to $\mathbb{R}$	7	10
$\check{T}$	restriction of T to $\tilde{\Gamma}$	–	23
$T_{\substack{\rightarrow \\ ji}}$	directional derivative of T associated with increasing the ith coordinate and reducing the jth coordinate	–	23
$T_{(ji)}$	two-sided directional derivative of T	–	23
$\theta(j)$	Lagrange multiplier on state j	–	14
$u(a, \omega)$	utility to action $a \in \mathcal{A}$ in state $\omega \in \Omega$	–	5
$\bar{u}(\gamma, a)$	expectation of $u(a, \omega)$ given γ	–	7
$V(\mu, \lambda\|K)$	value of strategy λ given prior μ and cost function K	–	8
$\hat{V}(\mu, A\|K)$	value of optimal strategy given prior μ and cost function K	–	8

References

Sydney N. Afriat. The Construction of Utility Functions from Expenditure Data. *International Economic Review*, 8(1):67–77, 1967.

Marina Agranov and Pietro Ortoleva. Stochastic choice and preferences for randomization. *Journal of Political Economy*, 125(1):40–68, 2017.

Ricardo Alonso and Odilon Câmara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706, 2016.

Jose Apesteguia, Miguel A Ballester, and Jay Lu. Single-crossing random utility models. *Econometrica*, 85(2):661–674, 2017.

Robert J Aumann, Michael Maschler, and Richard E Stearns. *Repeated games with incomplete information*. MIT press, 1995.

Vojtěch Bartoš, Michal Bauer, Julie Chytilová, and Filip Matějka. Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, 106(6):1437–1475, 2016.

Christian Beck. Generalised information and entropy measures in physics. *Contemporary Physics*, 50(4):495–510, 2009.

David Blackwell. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, pages 93–102, 1951.

Henry David Block and Jacob Marschak. *Contributions to Probability and Statistics*, volume 2, chapter Random orderings and stochastic theories of responses, pages 97–132. Stanford University Press, 1960.

Andrew Caplin and Mark Dean. Behavioral implications of rational inattention with shannon entropy. NBER Working Papers 19318, National Bureau of Economic Research, Inc, August 2013.

Andrew Caplin and Mark Dean. Revealed preference, rational inattention, and costly information acquisition. *The American Economic Review*, 105(7):2183–2203, 2015.

Andrew Caplin and Daniel Martin. A testable theory of imperfect perception. *The Economic Journal*, 125(582):184–202, 2015.

Andrew Caplin, Mark Dean, and Daniel Martin. Search and satisficing. *The American Economic Review*, 101(7):2899–2922, 2011.

Andrew Caplin, John Leahy, and Filip Matějka. Social learning and selective attention. Technical report, National Bureau of Economic Research, 2015.

Andrew Caplin, Mark Dean, and John Leahy. Rational Inattention, Optimal Consideration Sets, and Stochastic Choice. *The Review of Economic Studies*, 07 2018.

Raj Chetty, Adam Looney, and Kory Kroft. Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–1177, 2009.

Aubrey Clark. Contracts for information acquisition. 2016.

Luca Colombo, Gianluca Femminis, and Alessandro Pavan. Information acquisition and welfare. *The Review of Economic Studies*, 81:1438–1483, 2014.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Henrique de Oliveira. Axiomatic foundations for entropic costs of attention. Mimeo, Northwestern University, 2014.

Mark Dean and Nathaniel Neligh. Experimental tests of rational inattention. 2017.

Ambuj Dewan and Nathaniel Neligh. Estimating information cost functions in models of rational inattention. 2017.

Jeffrey C Ely and Martin Szydlowski. Moving the goalposts. Technical report, Working paper, 2017.

Xavier Gabaix. A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710, 2014.

Murray Gell-Mann and Constantino Tsallis. *Nonextensive entropy: interdisciplinary applications*. Oxford University Press, 2004.

Matthew Gentzkow and Emir Kamenica. Costly persuasion. *The American Economic Review*, 104(5):457–462, 2014.

Friedrich August Hayek. Economics and knowledge. *Economica*, 4(13):33–54, 1937.

Friedrich August Hayek. The use of knowledge in society. *The American economic review*, pages 519–530, 1945.

Benjamin Hébert and Michael Woodford. Rational inattention with sequential information sampling. 2017.

Christian Hellwig, Sebastian Kohls, and Laura Veldkamp. Information choice technologies. *The American Economic Review*, 102(3):35–40, 2012.

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

Akovlevich Khinchin. *Mathematical Foundations of Information Theory*, volume 434. Courier Corporation, 1957.

Ian Krajbich and Antonio Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857, 2011.

L Le Cam. Comparison of experiments: A short review. *Lecture Notes-Monograph Series*, pages 127–138, 1996.

EK Lenzi, RS Mendes, and LR Da Silva. Statistical mechanics based on renyi entropy. *Physica A: Statistical Mechanics and its Applications*, 280(3):337–345, 2000.

Luis Gonzalo Llosa and Venky Venkateswaran. Efficiency with endogenous information choice. *Unpublished working paper. University of California at Los Angeles, New York University*, 2012.

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

Paola Manzini and Marco Mariotti. Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176, 2014.

Paola Manzini and Marco Mariotti. Dual random utility maximisation. 2016.

Daniel Martin. Strategic pricing with rational inattention to quality. *Games and Economic Behavior*, 104:131–145, 2017.

Yusufcan Masatlioglu, Daisuke Nakajima, and Erkut Y Ozbay. Revealed attention. *American Economic Review*, 102(5):2183–2205, 2012.

Tomasz Maszczyk and Włodzisław Duch. Comparison of shannon, renyi and tsallis entropy used in decision trees. In *International Conference on Artificial Intelligence and Soft Computing*, pages 643–651. Springer, 2008.

Filip Matejka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–98, 2015.

Filip Matějka. Rationally inattentive seller: Sales and discrete pricing. *The Review of Economic Studies*, 83(3):1156–1188, 2015.

Ludmila Matyskova. Bayesian persuasion with costly information acquisition. Technical report, Working paper, 2018.

Daniel McFadden. Revealed stochastic preference: A synthesis. *Economic Theory*, 26(2):245–264, 2005.

Jordi Mondria. Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5):1837–1864, 2010.

Stephen Morris and Philipp Strack. The wald problem and the equivalence of sequential sampling and static information costs. 2017.

Henrique Oliveira, Tommaso Denti, Maximilian Mihm, and Kemal Ozbek. Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, 12(2):621–654, 2017.

Luigi Paciello and Mirko Wiederholt. Exogenous information, endogenous information and optimal monetary policy. *The Review of Economic Studies*, 83:356–388, 2014.

Roger Ratcliff, Philip Smith, Scott Brown, and Gail McKoon. Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, 20(4):260–281, April 2016.

Doron Ravid. Bargaining with rational inattention. 2017.

Ricardo Reis. Inattentive producers. *Review of Economic Studies*, 73(3):793–821, 2006.

Marcel K Richter. Revealed preference theory. *Econometrica: Journal of the Econometric Society*, pages 635–645, 1966.

Jean-Charles Rochet. A necessary and sufficient condition for rationalizability in a quasi-linear context. *Journal of Mathematical Economics*, 16(2):191–200, April 1987.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

M. L. Shaw and P. Shaw. Optimal allocation of cognitive resources to spatial locations. *J Exp Psychol Hum Percept Perform*, 3(2):201–211, May 1977.

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

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

Jakub Steiner and Colin Stewart. Perceiving prospects properly. *American Economic Review*, 2016.

Jakub Steiner, Colin Stewart, and Filip Matejka. *Rational inattention dynamics: Inertia and delay in decision-making*. Centre for Economic Policy Research, 2015.

Luminita Stevens. Coarse pricing policies. *Available at SSRN 2544681*, 2014.

Bernhard Tellenbach, Martin Burkhart, Didier Sornette, and Thomas Maillart. Beyond shannon: Characterizing internet traffic with generalized entropy metrics. In *International Conference on Passive and Active Network Measurement*, pages 239–248. Springer, 2009.

Erik Torgersen. *Comparison of statistical experiments*. Number 36. Cambridge University Press, 1991.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1-2):479–487, 1988.

Robert Verrecchia. Information acquisition in a noisy rational expectations economy. *Econometrica*, 50(6):1415–1430, 1982.

Michael Woodford. Information constrained state dependent pricing. *Journal of Monetary Economics*, 56(S):S100–S124, 2009.

Michael Woodford. Inattentive valuation and reference-dependent choice. Mimeo, Columbia University, 2012.

Ming Yang. Coordination with flexible information acquisition. *Journal of Economic Theory*, 158:721–738, 2015.