Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy

Andrew Caplin

New York University and NBER

Mark Dean

Columbia University

John Leahy

University of Michigan and NBER

We introduce three new classes of attention cost functions: *posterior separable, uniformly posterior separable* and *invariant posterior separable.* As with the Shannon cost function, all can be solved using Lagrangean methods. *Uniformly posterior separable* cost functions capture many forms of sequential learning, hence play a key role in many applications. *Invariant posterior separable* cost functions make learning strategies depend exclusively on payoff uncertainty. We introduce two behavioral axioms, *Locally Invariant Posteriors* and *Only Payoffs Matter,* which identify posterior separable functions respectively as uniformly and invariant posterior separable. In combination they pinpoint the Shannon cost function.

We thank three anonymous for their extraordinaily helpful guidance. We also thank George-Marios Angeletos, Sandro Ambuehl, Dirk Bergemann, Daniel Csaba, Tommaso Denti, Henrique de Oliveira, Xavier Gabaix, Sen Geng, Andrei Gomberg, Benjamin Hebert, Jennifer La'O, Michael Magill, Daniel Martin, Filip Matejka, Alisdair McKay, Stephen Morris, Efe Ok, Larry Samuelson, and Michael Woodford for their constructive contributions. Andrew Caplin thanks the NOMIS and Slan Foundations for their support. 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. This paper was edited by Harald Uhlig.

1 Introduction

Understanding limits on private information has been central to economic analysis since the pioneering work of Hayek [1937, 1945]. Recent years have seen renewed interest in the endogeneity of information, and the importance of information acquisition costs in determining outcomes. Sims [1998, 2003] kicked off this modern literature by introducing a model of rational attention based on Shannon mutual information in which both the extensive margin (how much is learned) and the intensive margins (precisely what is learned) are intimately shaped by incentives. In part due to its analytic tractability (Matějka and McKay [2015], Caplin, Dean, and Leahy [2019]), the ensuing period has seen applications to such diverse subjects as stochastic choice (Matějka and McKay [2015]), investment decisions (Mondria [2010]), global games (Yang [2015]), pricing decisions (Woodford [2009], Mackowiak and Wiederhold [2009], and Matějka [2015]), dynamic learning (Steiner, Stewart and Matějka [2015]) and social learning (Caplin, Leahy, and Matějka [2015]). These analyses show a tight link between decision making incentives and the structure of learning and illustrate sophisticated and realistic behaviors such as incomplete consideration of options (Caplin, Dean, and Leahy [2019]), attentional discrimination (Bartoš, *et al.* [2016]), and mental accounting (Kőszegi and Matějka [2020]).

As the Shannon model[1] has become increasingly well understood, it has become clear that its behavioral implications are counter-factual in many applications. For example, the Shannon model implies that all states are equally easy to identify and to discriminate among, a result that runs counter to the experimental evidence of Dewan and Neligh [2020] (see also Hébert and Woodford [2018] and Dean and Neligh [2017]). It also implies that attention is not focussed on states that are a priori more likely, a result that runs counter to the experimental evidence cited in Woodford [2012]. The Shannon model also makes specific predictions regarding the elasticity of choice in response to incentives which fail in the experiments of Caplin and Dean [2013]. In some sense, these failures are not surprising. The Shannon model is analytically simple because it places severe restrictions on behavior.

In this paper, we introduce three new classes of attention cost functions that generalize the Shannon model, share many of its attractive features, yet are flexible enough to cover essentially all behaviors uncovered to date that contradict it. We also provide solution methods that can be applied to all three models, aiding their use in applied work. All of our models can be solved using the same Lagrangian approach as the Shannon model, in which optimal behavior can be determined by identifying the tangent to the "concavified" net utility function (Aumann, Maschler, and Stearns [1995], Kamenica and Gentzkow [2011], Gentzkow and Kamenica [2014], Alonso and Câmara [2016], Ely and Szydlowski [2017], and Matyskova [2018]). Finally, we provide characteri-

[1] We use the term "Shannon model" to refer to a the model of rational inattention in which costs are linear in the mutual information between signal and state. Other variants of the model impose an upper limit on the amount of mutual information (e.g. Sims [2003]) or allow for a non-linear relationship between mutual information and cost (e.g. Mackowiak and Wiederholt [2009])

zations of the behavior implied by each cost function, aiding understanding of where each is likely to be behaviorally appropriate and suggesting model tests.

The most general class that we introduce, which we call *posterior separable*, is additively separable across the chosen posteriors. Posterior-separable models not only cover rich behaviors in any given decision problem, but also allow these costs to depend on prior beliefs and hence vary from context to context. The other two classes of cost functions are more specialized than the posterior-separable model, yet generalize in different respects the Shannon model. *Uniformly posterior-separable* cost functions impose additive separability in the prior in addition to additive separability in the posteriors. We will be interested in a particular generalization of uniform posterior-separable cost functions which we call *weakly uniformly posterior-separable.* These cost functions are additively separable in the prior so long as the support of the prior remains unchanged. Allowing the cost function to depend on the support of the prior allows us to consider cost functions in which the cost of ruling out a state depends on whether or not the prior places positive probability on that state. Uniformly posterior-separable cost functions, weakly or not, have all of the flexibility of the posterior-separable form for a given prior belief, but impose strong cross prior restrictions. They have found many applications since they were introduced in Caplin and Dean [2013], which is subsumed into this paper. For example the neighborhood-based cost function of Hébert and Woodford [2020], which captures the idea that some pairs of states are more easily distinguished than others, is in this class. Morris and Strack [2017], Hébert and Woodford [2019], and Bloedel and Zhong [2021] show that cost functions in this class are consistent with models of optimal sequential learning. Miao and Xing [2020] place uniformly posterior-separable cost functions into a dynamic discrete choice setting.

Our second specialization of the posterior-separable model is the *invariant posterior-separable* class. These cost functions relate the cost of learning about an event to the prior beliefs over the states that make up the event. A cost function is invariant if the least costly way to learn about an event is to learn nothing additional about the relative probabilities of states that make up the event and is independent of the prior over these states. Invariance rules out situations in which a subset of states are particularly easy to discern. Again, these are in active use. Angeletos and Sastry [2019] and Hébert and La'O [2020] study conditions under which equilibria in markets or games with incomplete information are efficient. Invariance is an important prerequisite. When invariance holds it is impossible to create payoff irrelevant signals that simultaneously reduce information costs and coordinate behavior in suboptimal ways.

In addition to introducing these classes of cost functions, we characterize their behavioral properties in state dependent stochastic choice data.[2] This data set treats both the payoff determining states of the world and the distribution of choices in these states as observable. It is routinely

[2] In this paper, we begin with a posterior separable representation and show that two behavioral axioms yield a Shannon representation. In the working paper, Caplin, Dean and Leahy [2017] we also present axioms that characterize a posterior separable representation.

gathered in many settings, such as simple production tasks in which production quality is monitored (e.g. Kaur, *et al.* [2019]). It rests on the idea that attentional constraints do not apply to an ideal observer. For example, while buyers may have difficulty assessing whether or not sales tax is included in the price paid at the register, the econometrician knows (Chetty, Looney and Kroft [2009]). Recent research shows that costs can be recovered from sufficiently rich data of this form, and that data of this form permits welfare analysis of mistaken decisions (Caplin and Martin [2020], Caplin, Csaba, Leahy, and Nov [2020]).

Our first behavioral characterization introduces one axiom, *Locally Invariant Posteriors,* which identifies weakly uniformly posterior-separable cost functions among the class of posterior separable cost functions. The axiom states that if a given set of posteriors is used in one decision problem, then those same posteriors are chosen in any decision problem which shares the same payoffs and has a prior that lies in the convex hull of the initially chosen posteriors. In essence, if the payoffs are unchanged and the posteriors are feasible, then the posteriors remain optimal. This behavioral axiom underlies the Drift Diffusion Model which has proven popular in psychology (see Ratcliff, *et al.* [2016] for a recent review) and also in economics (Fehr and Rangel [2011]). According to the Drift Diffusion Model, an agent gathers information about a pair of states and acts only when posterior beliefs reach some predetermined threshold values. Since the thresholds do not change as the agent learns, the chosen thresholds do not vary with the prior. This explains the link with optimal sequential sampling and hence why the models of Morris and Strack [2017], Hébert and Woodford [2019], and Bloedel and Zhong [2021] are in the uniformly posterior-separable class.

Our second behavioral characterization identifies invariant posterior-separable cost functions from the class of posterior separable cost functions. We provide two equivalent behavioral characterizations. The first, which we call *Only Payoffs Matter,* captures the notion that economically meaningful states are defined by payoffs, a common assumption in many economic models. We say that two decision problems are payoff equivalent if each vector of action payoffs appears with the same probability in the two problems. The axiom states that given two payoff equivalent decision problems the frequency with which actions are observed in the two problems depends only on the payoffs to actions and not the exact mapping between payoffs and states. The second version, which we call *Invariance under Compression,* states that all states with a common payoff profile can be compressed into a single state without altering observed behavior. There is an attentional rationale for these axioms. There is no functional value for the decision maker in distinguishing between states that assign the same payoffs to all actions. Hence an ideally designed machine for learning about states would not waste any of its scarce resources on this task. Our axioms make this statement precise in the language of behavior.

Our final result takes us all the way back to the Shannon model. To our surprise, the two behavioral axioms that we introduce in combination characterize the Shannon model among all posterior-separable models. One part of this result is relatively straight forward: that the Shannon cost function gives rise to behavior that satisfies both Locally Invariant Posteriors and Only Payoffs

Matter. Far less obvious is the converse. If a data set can be rationalized by a posterior-separable cost function and satisfies both Locally Independent Posteriors and Only Payoffs Matter, then this cost function must be linear in the expected reduction in Shannon entropy. The fact that it is solvable by Lagrangian methods and uniquely has these conceptually attractive behavioral properties may explain why it has become, and will likely remain, a central model not only in studies of individual decision making but in market settings.[3]

Our results contribute to a growing literature studying the implications of information acquisition for stochastic choice. Caplin and Martin [2015] and Caplin and Dean [2015] present necessary and sufficient conditions for a data set to be represented by optimal choice subject to an information cost. Denti [2020] provides axioms that characterize posterior-separable and uniformly posterior-separable cost functions, while de Olivera [2014] considers the behavioral implications of the Shannon model, but for a data set which consists of observed choices over different menus of alternatives.

Section 2 sets the stage for our analysis. The first subsection presents our idealized state dependent stochastic choice (SDSC) data set. Section 2.2 presents the theoretical model that we use to rationalize the data set. This model consists of an expected utility maximizer subject to an additive cost of acquiring information. We distinguish between attention strategies (the choice of posteriors) and choice strategies (the choice of actions). Information costs depend only on the former. Section 2.3 bridges the gap between observational data and the theoretical representation and shows how to map between the theory and the data.

With these preliminaries out of the way, the heart of the paper lies in Sections 3 through 6. Section 3 introduces posterior-separable cost functions and establishes general applicability of Lagrangian methods of identifying optimal strategies. Section 4 introduces weakly uniformly posterior-separable cost functions and the axiom of Locally Invariant Posteriors. This section culminates in Theorem 1 which states the necessity and sufficiency of the axiom. Section 5 introduces invariant posterior-separable cost functions and the axiom of Only Payoffs Matter. This section concludes with Theorem 2 stating the necessity and sufficiency of this axiom. Section 6 presents Theorem 3 which states that the Shannon cost function is unique among posterior-separable costs functions in that it is both weskly uniformly posterior separable and invariant. Section 7 presents the axiom of Invariance under Compression and shows that it is equivalent to Only Payoffs Matter. This section also discusses some assumptions and extensions. Section 8 outlines the relationship with the prior literature. Section 9 concludes.

[3] Recent examples include Caplin, Leahy, and Matejka [2015], Martin [2017] and Ravid [2020].

2 Preliminaries

2.1 The Data Set

The fundamental object in our analysis is a data set that catalogues the choices made by a decision maker across a large number of choice situations. The fundamental question that we ask concerns the conditions that this data set must satisfy for this decision maker's choices to be observationally equivalent to those of an expected utility maximizer facing a certain cost of information acquisition. In this subsection we present the data set. The next models a decision maker with information costs. We then define observational equivalence before presenting axioms on the data which imply observational equivalence.

Our goal is to understand how much a decision maker attends to their environment. For this it is not enough to know how often each choice is made. We need to know if the decision maker makes the right choice in the right situation. We will therefore need to see pairs of choices and outcomes in our data set. In a pure revealed preference analysis this is all that we would see. From these choices we would recover the decision maker's payoffs to various actions, their beliefs and the cost of gathering information. Our focus in this paper is on the recovery of information costs and we have little to add to the literature on eliciting beliefs or payoffs. We will therefore assume that beliefs and payoffs are also observable and focus on the implications of information costs for observed choice.

We consider a setting in which an agent chooses among actions that are distinguished by their state-contingent payoffs in units of utility. Formally, let Ω denote the set of conceivable states of the world. We assume that Ω is countably infinite. Each action a is a function $a : \Omega \rightarrow \mathbb{R}$ in which $a(\omega)$ is the utility to action a in state ω. Let $\mathcal{A}$ denote the set of potential actions, $\mathcal{A} \equiv \{a : \Omega \rightarrow \mathbb{R}\}$.

A decision problem (μ, A) is the choice over a finite set of actions $A \subset \mathcal{A}$ given the prior probability distribution μ over states in Ω. Throughout we assume that the prior places positive probability on only a finite subset of states. With slight abuse of notation we will let $\Delta\Omega$ denote the set of probability distributions on Ω with finite support (omitting probability distributions with infinite support). This assumption simplifies the measure theoretic aspects of the analysis. The set of potential events associated with any decision problem is therefore the discrete σ-algebra $2^{\text{supp } \mu}$ and the restriction of any action to the support of μ is trivially measurable with respect to this σ-algebra. Again with slight abuse of notation, let $2^{\mathcal{A}}$ denote the set of all non-empty finite subsets of $\mathcal{A}$. Given these notational conventions, the set of all decision problems is $\Delta\Omega \times 2^{\mathcal{A}}$.

Following Matějka and McKay [2015] and Caplin and Martin [2015], we assume that our data comes in the form of a joint distribution between states and actions, which we encode using the marginal distributions of actions conditional on states. Given a decision problem (μ, A) we define **state dependent stochastic choice (SDSC) data** as a mapping from possible states to action probabilities,

$$\mathbf{P}_{(\mu,A)} : \text{supp } \mu \to \Delta A$$

For each decision problem (μ, A), $\mathbf{P}_{(\mu,A)}$ specifies the frequency that the agent chooses each action $a \in A$ in each state in the support of μ. We use the notation $\mathbf{P}_{(\mu,A)}(a|\omega)$ for the conditional probability of observing action a in state ω in decision problem (μ, A) and $\mathbf{P}_{(\mu,A)}(a)$ for the unconditional probability of observing action a.

Our data set is the collection of all SDSC data that a decision maker would willingly choose across all decision problems. Since the agent may be indifferent between multiple patterns of choice, this data set is a correspondence. Let $\mathbf{D}$ denote this correspondence. $\mathbf{D}_{(\mu,A)}$ is the set of observed SDSC functions $\mathbf{P}_{(\mu,A)}$ for the decision problem (μ, A). Throughout we will use bold letters to refer to data objects in order to distinguish them from elements of the theory that represents the data, as well as elements of the environment.

SDSC is by now a common data set with which to test models of information acquisition (see for example Caplin and Martin [2015], Caplin and Dean [2015], and Denti [2020]). Moreover, while our axioms require data to be observed from a rich set of choice problems in order to guarantee sufficiency, each can be falsified using a finite number of observations, and in this sense is analogous to the independence axiom in characterizations of expected utility theory. Our data set is somewhat unusual in that $\mathbf{D}$ is a correspondence: we observe all of the SDSC data $\mathbf{P}_{(\mu,A)}$ that the decision maker is willing to choose in a given decision problem. While standard in characterizations of deterministic choice models (for an early example see Richter [1966]), this assumption is less common in stochastic choice models. While it makes the statement of our theorems more elegant, our results can readily be extended to the case in which only a selection from the maximal SDSC data is observed, essentially because such data is rich enough to uniquely pin down what the rationalizing cost function must be (see Caplin, Csaba, Leahy and Nov [2020]).

We now turn to the class of theoretical models which we will use to rationalize the data set $\mathbf{D}$.

2.2 A Theoretical Model of State Dependent Stochastic Choice

In this section we model an expected utility maximizer subject to an additive cost of acquiring information. We first discuss the strategies available to the decision maker. These include both attention strategies and choice strategies. We then place costs on the attention strategies and present the decision maker's problem.

2.2.1 Feasible Posterior-Based Strategies

Consider a theoretical decision maker facing the decision problem (μ, A). We break this agent's decision into a sequence of two steps. The first step is the attention strategy, the choice of how much to learn. We capture learning by a probability distribution over posteriors $Q \in \Delta\Delta\Omega$. Intuitively, the agent has learned something when the posteriors differ from the prior. Let $Q(\gamma)$ denote the probability of a particular posterior $\gamma \in \Delta\Omega$. We will assume that the support of Q is finite.[4]

We require that all feasible attention strategies Q satisfy Bayes rule so that the $\gamma \in$supp Q average to the prior μ:

$$\mu = \sum_{\gamma \in \text{supp } Q} \gamma Q(\gamma)$$

The set of feasible attention strategies therefore depends on the prior. Given prior μ, let $\mathcal{Q}(\mu)$ denote the set of Q that satisfy Bayes rule. Choosing not to learn is always an option. In this case, the chosen posterior is equal to the prior with probability one, $Q(\mu) = 1$. Note also that Bayes rule imposes the restriction that all $\gamma \in$supp Q are absolutely continuous with respect to μ, so that supp $\gamma \subseteq$supp μ. The only states that matter in decision problem (μ, A) are those that receive positive probability according to the prior μ.

Given a realized posterior γ, the second step is to choose an action $a \in A$. Here we allow for random strategies. We represent action choice by a mapping q from posteriors in the support of Q to probability distributions over the available actions, $q :$ supp $Q \rightarrow \Delta A$.

Given a decision problem (μ, A), the set of **feasible posterior-based strategies** $\Lambda(\mu, A)$ comprises all Bayes-consistent probability distributions over posteriors, $Q \in \mathcal{Q}(\mu)$, and corresponding mixed action strategies, $q :$ supp $Q \rightarrow \Delta A$:

$$\Lambda(\mu, A) \equiv \{(Q, q) | Q \in \mathcal{Q}(\mu),\ q : \text{supp } Q \rightarrow \Delta A\}.$$

While we allow for random strategies, it is clear that the agent will only randomize if they are indifferent between options. At times, it will be useful to eliminate this randomization. To this end, we define a deterministic action choice function $\bar{q} :$ supp $Q \rightarrow A$ as a mapping from posteriors in the support of Q to actions in A (rather than distributions over actions).

2.2.2 Utility, Costs and Optimal Strategies

The goal of the decision maker is to maximize expected utility net of attention costs. Given a decision problem (μ, A) and a feasible strategy $(Q, q) \in \Lambda(\mu, A)$, expected utility is computed in the standard manner,

$$U(Q, q|\mu, A) = \sum_{\gamma \in \text{supp } Q} \sum_{a \in A} Q(\gamma) q(a|\gamma) \left[\sum_{\omega \in \text{supp } \gamma} \gamma(\omega) a(\omega) \right]$$

[4] While it is theoretically possible that choice from a finite set of options is only rationalizable by a model with an uncountable number of optimal posteriors, we show that a finite set of posteriors is sufficient.

where $\gamma(\omega)$ is the probability of state ω given the posterior γ. The term in brackets is the expected value of action a given the posterior γ. This term is multiplied by $q(a|\gamma)$, the probability of choosing action a given γ, and by $Q(\gamma)$, the probability of learning γ. The summation is over all actions in A and posteriors in the support of Q.

We assume that attention costs for strategy $(Q, q) \in \Lambda(\mu, A)$ depend only on the prior μ and the distribution of posteriors Q, and not on the choice set A or the action probabilities q. The domain of the cost function is the set of all priors $\mu \in \Delta\Omega$ and all posterior distributions $Q \in \mathcal{Q}(\mu)$ consistent with that prior:

$$\mathcal{F} = \{(\mu, Q)|\mu \in \Delta\Omega, Q \in \mathcal{Q}(\mu)\}. \tag{1}$$

An **attention cost function** K maps elements of $\mathcal{F}$ into the extended positive real line, $K : \mathcal{F} \to [0, \infty]$. We normalize the cost of learning nothing to zero, so that $K(\mu, Q) = 0$ whenever supp $Q = \{\mu\}$. Allowing $K(\mu, Q)$ to equal infinity for some Q allows for the possibility that some attention strategies in $\mathcal{Q}(\mu)$ are not chosen for any A. For example, there are interesting cases in which it is prohibitively costly to rule out ex ante possible states, so that the decision maker will never choose posteriors on the boundary of $\Delta(\text{supp } \mu)$.

Given an attention cost function K, the value of a feasible strategy $(Q, q) \in \Lambda(\mu, A)$ is computed by subtracting the attention cost from expected utility:

$$V(Q, q|\mu, A, K) \equiv U(Q, q|\mu, A) - K(\mu, Q). \tag{2}$$

The value of an optimal strategy and the set of optimal strategies are defined in the natural manner. We use hats to denote the maximized value function and the set of optimal strategies.[5]

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

Of particular interest is the case in which costs are linear with respect to the expected reduction in Shannon entropy between prior and posterior beliefs. The **Shannon cost function**, $K^S_\kappa(\mu, Q)$, is

$$K^S_\kappa(\mu, Q) \equiv \kappa \left[\sum_{\gamma\in\text{supp } Q} Q(\gamma) \sum_{\omega\in\text{supp } \gamma} \gamma(\omega) \ln \gamma(\omega) - \sum_{\omega\in\text{supp } \mu} \mu(\omega) \ln \mu(\omega)\right]. \tag{3}$$

We will show that the behavior associated with the Shannon cost function can be characterized as the intersection of two axioms, each of which define important classes of cost functions that

[5] While we write $\hat{V}$ as the supremum over all feasible strategies, the data set $\mathbf{D}$ assigns a choice to every decision problem. If K represents $\mathbf{D}$, meaning that the set of policies that maximize choice given K reproduce the data set $\mathbf{D}$ (see Section 2.3.2 below), then $\hat{\Lambda}$ will never be empty, and there will always exists a policy that achieves the supremum.

inherit some of its features.

Our goal is to match properties of data sets $\mathbf{D}$ with the choices implied by classes of cost functions. In the next section we show how to move back and forth between state dependent stochastic choice data and posterior-based strategies.

2.3 Mapping the Theory to the Data

Our data is in terms of state dependent stochastic choice, whereas our theory delivers posterior-based strategies. In this section we discuss the mapping between the two. We begin with the mapping from posterior-based strategies to state dependent stochastic choice. This mapping gives the SDSC data generated by our model. Our representation theorems all state that this generated data is equivalent to the observed data set $\mathbf{D}$. We then look at the mapping between state dependent stochastic choice and posterior-based attention strategies. This mapping gives the revealed posterior-based attention strategies. Our axioms all place restrictions on these revealed strategies.

2.3.1 From Theory to Generated Data

We begin with the generation of SDSC data from our theoretical model. We translate each strategy into its observable counterpart in SDSC data. Given a feasible strategy $(Q, q) \in \Lambda(\mu, A)$ we define the **generated** SDSC data $P_{(\mu,A|Q,q)}$:supp $\mu \rightarrow \Delta A$ and the corresponding action choice probabilities $P_{(\mu,A|Q,q)}(a)$ on $a \in A$ by:

$$P_{(\mu,A|Q,q)}(a) = \sum_{\gamma \in \text{supp } Q} Q(\gamma) q(a|\gamma); \tag{4}$$

$$P_{(\mu,A|Q,q)}(a|\omega) = \frac{\sum_{\gamma \in \text{supp } Q} Q(\gamma) q(a|\gamma) \gamma(\omega)}{\mu(\omega)}.$$

The first equation sums the probability of the occurrence of action a across posteriors. The second equation follows from Bayes rule. The numerator is the joint probability of action a and state ω, and the denominator is the probability of the state ω.

As the pattern of choice $P_{(\mu,A|Q,q)}$ is completely determined by the strategy (Q, q), we will write $P_{(Q,q)}$ when no confusion would result from not specifying the decision problem (μ, A).

2.3.2 Representations

Given a decision problem (μ, A), $P_{(Q,q)}$ is the theoretical counterpart to the data object $\mathbf{P}_{(\mu,A)}$. Both $P_{(Q,q)}$ and $\mathbf{P}_{(\mu,A)}$ are mappings from the support of μ to ΔA. Our representations are based on the equivalence of these two objects. We say that K represents $\mathbf{D}$ if the observed SDSC data set $\mathbf{D}$ comprises all SDSC data generated by the optimal policies of an expected utility maximizer facing the information cost function K.

Definition 1 *The cost function K* ***represents*** *the data set $\mathbf{D}$ if for all $(\mu, A) \in \Delta\Omega \times 2^{\mathcal{A}}$,*

$$\mathbf{D}_{(\mu,A)} = \{P_{(\mu,A|Q,q)} | (Q, q) \in \hat{\Lambda}(\mu, A|K)\}.$$

Note that our data set is rich enough that the recovery theorem in Caplin, Csaba, Leahy, and Nov [2020] implies that any such representation will be unique.

2.3.3 From Data to Revealed Strategy

While the state dependent stochastic choice data associated with any strategy (Q, q) is unique, there are many strategies (Q, q) that could have generated any given $\mathbf{P}_{(\mu,A)}$. Caplin and Dean [2015], however, show that there is always a unique least Blackwell informative strategy consistent with the data, and that this strategy treats every action as chosen at one and only one posterior. This observation allows us to associate $\mathbf{P}_{(\mu,A)}$ with a unique "revealed" strategy $(\mathbf{Q}, \mathbf{q})$.

Consider a decision problem (μ, A) and observed state dependent stochastic choice data $\mathbf{P}_{(\mu,A)}$. If a is chosen with positive probability, we can use the fact that a is chosen from only one posterior to derive the "revealed" posterior γ^a associated with action a using Bayes rule

$$\gamma^a(\omega) = \frac{\mathbf{P}_{(\mu,A)}(a|\omega)\mu(\omega)}{\mathbf{P}_{(\mu,A)}(a)}$$

where $\mathbf{P}_{(\mu,A)}(a|\omega)$ is the observed probability of a in state ω and $\mathbf{P}_{(\mu,A)}(a)$ is the observed frequency with which a is chosen.

We can then define the **revealed posterior-based strategy** $(\mathbf{Q}, \mathbf{q})$. The revealed probability of choosing posterior γ is the sum of the revealed probabilities of choosing actions a with revealed posteriors γ^a equal to γ,

$$\mathbf{Q}(\gamma) = \sum_{\{a \in A | \gamma = \gamma^a\}} \mathbf{P}_{(\mu,A)}(a),$$

and the revealed probability of choosing an action a conditional on γ^a is equal to the revealed probability of choosing a divided by the revealed probability of choosing γ^a,[6]

$$\mathbf{q}(a|\gamma^a) = \frac{\mathbf{P}_{(\mu,A)}(a)}{\mathbf{Q}(\gamma^a)}$$

Note that $\mathbf{P}_{(\mu,A)}$, $\mathbf{Q}$, $\mathbf{q}$, and γ^a are all bold as these are data objects.

[6] Note that since we are looking for a posterior based strategy in which a is only chosen at most one posterior, $\mathbf{q}(a|\gamma) = 0$ if $\gamma \neq \gamma^a$ or if $\mathbf{P}_{(\mu,A)}(a) = 0$.

3 Posterior-Separable Cost Functions

The starting point in this paper is a data set that may be represented by a cost function that is additively separable across the chosen posteriors. We call such cost functions posterior separable. This class includes most of the cost functions considered in the literature, including mutual information (Sims [1998]), expected Tsallis entropy (Caplin, Dean and Leahy [2017]), the neighborhood-based cost function of Hébert and Woodford [2018], and the Log-Likelihood Ratio cost function of Pomatto, Strack and Tamuz [2018]. Posterior separability is also generally assumed in the literature on Bayesian persuasion (Kamenica and Gentzkow [2011]). Caplin, Dean and Leahy [2017] provide behavioral axioms that characterize these cost functions. In this section, we formally define these cost functions and present several properties that will be useful in proving the representation theorems that follow.

3.1 Definition

Definition 2 *An attention cost function K is* ***posterior separable*** *if, given $\mu \in \Delta\Omega$ and any Bayes consistent posteriors $Q \in \mathcal{Q}(\mu)$*

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

for some convex function $T_\mu : \Delta(supp\ \mu) \to \bar{\mathbb{R}}$ such that $T_\mu(\gamma) \geq 0$ with $T_\mu(\mu) = 0$.

Equation (5) captures the essence of posterior separability. It states the cost function is equal to the expectation of some function of the individual posteriors, T_μ. Note that this function is allowed to depend in an arbitrary way on the prior. Different priors can lead to very different functions T_μ. In addition, the definition states that learning nothing is costless ($T_\mu(\mu) = 0$) and learning something is weakly costly ($T_\mu(\gamma) \geq 0$). These properties define T_μ as a divergence extended to the boundary of the simplex.[7]

The posterior separable cost functions that we will study satisfy a number of additional properties.

Assumption 1: Let K be a posterior separable cost function with $K(\mu, Q) = \sum_{\gamma \in \text{supp } Q} Q(\gamma) T_\mu(\gamma)$. We assume $T_\mu(\gamma)$ is strictly convex and continuous in γ and that $T_\mu(\gamma) < \infty$ on $\text{int}\Delta(\text{supp } \mu)$.

[7] A divergence is a weak notion of the distance between probability distributions. Given an arbitrary state space S, a divergence $D(p||q)$ is a function from $\text{int}\Delta S \times \text{int}\Delta S$ to $\bar{\mathbb{R}}$ satisfying only that $D(p||q) \geq 0$ and $D(q||q) = 0$. A prominent example is the Kullback-Leibler divergence.

Assumption 1 adds strict convexity, continuity and finiteness to the definition of a posterior separable cost function. The assumption that $T_\mu(\gamma)$ is finite on the interior of the support of μ together with the convexity of T_μ ensures that all posteriors in the interior of the support of μ are optimal for some decision problem.[8] Strict convexity simplifies the analysis by eliminating ties. Since T_μ is convex and finite-valued, it is continuous on the interior of $\Delta(\text{supp } \mu)$. Assumption 1 adds that T_μ is continuous on the boundary. We will discuss relaxing these assumptions in Section 8 below. In Section 8, we also prove that the assumption that T_μ is continuous on the boundary is without loss of generality.

From this point on, whenever we refer to a posterior separable cost function, we will assume that the additional restrictions of Assumption 1 hold unless otherwise stated. Rather than repeat Assumption 1 each time we state a theorem, we define a posterior-separable representation for the purposes of this paper to include the assumption.

Definition 3 *A data set* $\mathbf{D}$ *has a **posterior-separable representation** if it is represented by a posterior-separable cost function that satisfies Assumption 1.*

3.2 Properties of Posterior-Separable Models

Several properties of posterior-separable cost functions will prove useful in what follows. The first is that decision problems with posterior-separable attention costs can be solved with Lagrangian methods.

Because both the cost function and expected utility are additively separable in the posteriors, we can rewrite the value of any given strategy (Q, q), collecting terms specific to each chosen posterior and each action associated with that posterior:

$$V(Q, q|\mu, A, K) = \sum_{\gamma \in \text{supp } Q} \sum_{a \in A} Q(\gamma) q(a|\gamma) \left[\sum_{\omega \in \text{supp } \gamma} \gamma(\omega) a(\omega) - T_\mu(\gamma) \right]$$

Whenever $q(a|\gamma) > 0$, the term in brackets is the **net utility** of action a and posterior γ:

$$N_\mu^a(\gamma) \equiv \sum_{\omega \in \text{supp } \gamma} \gamma(\omega) a(\omega) - T_\mu(\gamma) \tag{6}$$

Note that since $T_\mu(\gamma)$ depends on μ so does $N_\mu^a(\gamma)$.

Writing the maximization problem in terms of net utilities allows a simple geometric interpretation of the solution. Maximizing the value function is equivalent to maximizing a Bayes consistent

[8] Since the support of μ is finite, $\Delta(\text{supp } \mu)$ is homeomorphic to the simplex of dimension $|\text{supp } \mu|$-1. The interior of $\Delta(\text{supp } \mu)$ is therefore the interior of the simplex, i.e. the set of probability distributions that place positive probability on all states in the support of μ. Similarly, the boundary of $\Delta(\text{supp } \mu)$ is the set of probability measures that place zero probability on some state.

weighted average of net utilities. This method is illustrated in Figure 1 for a simple decision problem with two states $\{\omega_1, \omega_2\}$ and two actions $\{a, b\}$. The horizontal axis in the figure records the probability of the state ω_1. The solid and dashed curves depict the net utilities to actions a and b respectively. These net utility curves are concave since net utility is the difference between expected utility which is linear in the probabilities and the cost of information which is convex. These curves differ in the utility that they assign to each state and action. In the figure action a pays off relatively more in state ω_2 and action b pays off relatively more in state ω_1. The prior μ is depicted on the horizontal axis.

Any pair of posteriors that span the prior are feasible in this problem. For example, γ^a and γ^b are feasible. The value of choosing γ^a and γ^b is given by the height of the chord connecting $N^a_\mu(\gamma^a)$ and $N^b_\mu(\gamma^b)$ as it passes over the prior μ. This is point A in the figure. This policy is not optimal. The optimal policy is $\hat{\gamma}^a$ and $\hat{\gamma}^b$ which returns the height of point B in the figure.

[Figure 1 approximately here]

The shaded area in Figure 1 is the lower epigraph of the concavified net utility functions, defined as the minimal concave function that majorizes all net utilities (Rockafellar [1970]). It is clear that the highest chord lies on the line tangent to the lower epigraph at the prior and that the optimal posteriors are the points at which this tangent line meets the net utilities with the value of the net utilities at all other points (weakly) below this tangent line.

This geometric intuition is general. We can always find the optimal posteriors by considering the hyperplane tangent to the lower epigraph at the prior and finding where this supporting hyperplane meets the net utilities. Suppose that the support of the prior μ has J distinct states and label these states ω_j for $j = \{1, \dots J\}$. Given an optimal posterior $\hat{\gamma}$, we can write the supporting hyperplane as the set of potential net utility levels N such that

$$N = N^a_\mu(\hat{\gamma}) + \sum_{j=1}^{J-1} \theta_j(\gamma(\omega_j) - \hat{\gamma}(\omega_j))$$

where the Lagrange multipliers θ_j for $j = 1 \dots J-1$ capture the change in net utility as the posterior $\gamma(\omega_j)$ is raised at the expense of reducing $\gamma(\omega_J)$. For example, in Figure 1 $J = 2$ and θ_1 is the slope of the chord connecting $N^a_\mu(\gamma^a)$ and $N^b_\mu(\gamma^b)$. Note that it does not matter which optimal posterior is used in this construction as all combinations of optimal posteriors $\hat{\gamma}$ and their corresponding net utility levels $N^a_\mu(\hat{\gamma})$ lie on this plane. Net utilities for all other posteriors lie weakly below the plane, that is $N^a_\mu(\gamma) \leq N^a_\mu(\hat{\gamma}) + \sum_{j=1}^{J-1} \theta_j(\gamma(\omega_j) - \hat{\gamma}(\omega_j))$. This result is summarized in the following lemma. All results are proved in the on-line appendix.

Lemma 1 (The Lagrangian Lemma): Given a posterior-separable cost function K and decision problem (μ, A) with dimension $J = |\text{supp } \mu|$, $(Q, q) \in \hat{\Lambda}(\mu, A|K)$ if and only if there exists $\theta \in \mathbb{R}^{J-1}$ such that, given $\gamma \in$supp Q and $a \in A$ with $q(a|\gamma) > 0$,

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

for all $\gamma' \in$supp μ and $a' \in A$.

This lemma characterizes optimal strategies, and opens up standard methods of model solution. It should be noted that the Lagrangian Lemma does not require Assumption 1. It requires only that $T_\mu(\gamma)$ is a proper convex function.

A second useful property of posterior-separable cost functions is that they imply that almost all sets of feasible posteriors are solutions to some decision problem. To see this consider again the net utilities in Figure 1. Fix the tangent line running through point B. Since net utility is the difference between a linear function and a convex function, we can adjust the linear portion of net utility to "rotate" net utility in any way that we see fit. In this way we can choose payoffs so that this line is tangent to net utility at any pair of feasible posteriors. There is one exception to this rule. While the construction is always possible for beliefs in which all states in the support of the prior are possible, there are important cases, such as the Shannon cost function, in which the slope of the net utility function approaches infinity at the boundaries of Δ(supp μ). In such cases there may be no choice of payoffs to actions a or b that will make it optimal to choose posteriors that set the probability of either state equal to zero. To formalize the precise limit on the domain, we let $\hat{\Gamma}(\mu|K)$ denote the set posteriors at which T_μ is subdifferentiable.[9] With this we can formally state our result identifying all sets of optimal posteriors in posterior-separable models.

Lemma 2 (Feasibility Implies Optimality): Fix $\mu \in \Delta\Omega$ and a posterior-separable cost function K that satisfies Assumption 1, then

1. Given $Q \in \mathcal{Q}(\mu)$ with supp $Q \subset \hat{\Gamma}(\mu|K)$, there exist a choice set A and a deterministic choice function $\bar{q}$ such that $(Q, \bar{q}) \in \hat{\Lambda}(\mu, A|K)$.
2. Given $\gamma \in \hat{\Gamma}(\mu|K)$, there exists $a \in \mathcal{A}$ such that $N_\mu^a(\gamma) = 0$ and $N_\mu^a(\gamma') \leq 0$ for all $\gamma' \in \Delta$supp μ.
3. int(Δ(supp μ)) $\subset \hat{\Gamma}(\mu|K)$.

Point 1 says that for any feasible set of posteriors we can find a decision problem such that this set of posteriors is optimal and that choice is deterministic conditional on the resulting posterior.

[9] A convex function $f(x)$ is subdifferentiable at $\bar{x}$ if there exists a vector θ such that $f(x) \geq f(\bar{x}) + \theta \cdot (x - \bar{x})$. A proper convex function will always be subdifferentiable on its relative interior. It may or may not be subdifferentiable elsewhere. In our case, T_μ will be subdifferentiable on the interior of Δ(supp μ). T_μ may or may not be subdifferentiable on the boundary. It will not be subdifferentiable on the boundary if T_μ is infinite, discontinuous, or if the derivative of T_μ approaches infinity. See Section 23 of Rockafellar [1970].

An immediate implication is that the set of posteriors that are observed as part of some optimal attention strategy given μ is the same as the set of posteriors at which T_μ is subdifferentiable. Part 2 says that we can take our tangent plane to be $\theta \cdot \gamma = 0$ and manipulate expected utility so that net utility is tangent to this plane at any posterior which is chosen. This result makes it easy to construct decision problems in which a given set of posteriors are chosen. Part 3 says that every posterior interior to $\Delta(\text{supp } \mu)$ is optimal in some decision problem. It follows from the assumption that T_μ is finite on the interior of $\Delta(\text{supp } \mu)$ and that convex functions are subdifferentiable on their relative interior (Rockafellar [1970, Theorem 23.4]). The only posteriors that might not be chosen as part of an optimal policy are those that lie on the boundary of $\Delta(\text{supp } \mu)$.

4 Weakly Uniformly Posterior-Separable Cost Functions

The next three sections present the main results of the paper. Each section considers a different subclass of posterior-separable cost functions. We begin each section by defining the subclass. We then present axioms that are necessary and sufficient for a posterior-separable representation to lie in this subclass. We begin, in this section, with weakly uniformly posterior-separable cost functions.

4.1 Definition

In the posterior-separable model, if the prior μ changes, K can change in arbitrary ways. If the data set has a weakly uniformly posterior-separable representation, however, K is in an important sense independent of μ: it depends on μ only through the support of μ. A cost function K is weakly uniformly posterior separable, if it is posterior separable and the strictly convex function T_μ depends only on the set of possible states.[10]

Definition 4 *A posterior-separable cost function K is* ***weakly uniformly posterior separable,*** *if for each finite subset $\bar{\Omega} \subset \Omega$ there exists a strictly convex function $T_{\bar{\Omega}} : \Delta(\bar{\Omega}) \to \bar{\mathbb{R}}$ such that, for all $\mu \in \Delta(\Omega)$ and $Q \in \mathcal{Q}(\mu)$,*

$$K(\mu, Q) = \sum_{\gamma \in supp\ Q} Q(\gamma) T_{supp\ \mu}(\gamma) - T_{supp\ \mu}(\mu). \tag{7}$$

The defining characteristic of a weakly uniformly posterior-separable cost function is that it is additively separable in both the priors and the posterior. $T_{\text{supp } \mu}(\gamma)$ depends on μ only through the support of μ, not the value of μ itself. Allowing $T_{\text{supp } \mu}$ to depend on the support of μ allows

[10] We label this definition "weak" uniform posterior separability because it is weaker than the definition introduced in Caplin, Dean and Leahy [2017]. In that paper, we insist on a single function $T : \Delta\Omega \to \bar{\mathbb{R}}$ rather than a set of functions $T_{\bar{\Omega}}$ that depend on the support of the prior. The new definition greatly simplifies the analysis. We discuss this issue further in Section 7.

us to handle cases in which the cost of setting $\gamma(\omega) = 0$ depends on whether or not $\mu(\omega)$ is zero. For example, it is costless to rule out ex ante impossible states, but might be extremely costly to rule out ex ante possible ones. Subtracting off $T_{\text{supp } \mu}(\mu)$ is a normalization that ensures that it is costless not to learn anything. This normalization has no effect on choice.[11]

4.2 Locally Invariant Posteriors

To understand the behavioral implications of weak uniform posterior separability, note that so long as the support of μ remains unchanged, μ enters additively in (7) and changes in μ do not affect the relative cost of any posterior. The only role that the prior plays is to determine what posteriors are consistent with Bayes rule. This has strong implications for the structure of optimal policies. Consider again Figure 1. If K is weakly uniformly posterior separable, then changes in μ shift the two net utility curves up or down by the same amount. Hence $\hat{\gamma}^a$ and $\hat{\gamma}^b$ remain optimal so long as they are still feasible, and they are feasible so long as the prior lies in the open interval $(\hat{\gamma}^a, \hat{\gamma}^b)$. So if a data set has a uniformly posterior-separable representation and $\hat{\gamma}^a$ and $\hat{\gamma}^b$ are revealed posteriors for the problem (μ, A), they must also be observed in any problem (μ', A) in which $\mu' \in (\hat{\gamma}^a, \hat{\gamma}^b)$. We call this property *Locally Invariant Posteriors.* Importantly, this property is not only necessary but also sufficient for a data set with a posterior-separable representation to have a weakly uniformly posterior-separable representation.

Axiom 1 Locally Invariant Posteriors: Consider any decision problem (μ, A) and state dependent stochastic choice data $\mathbf{P}_{(\mu,A)} \in \mathbf{D}_{(\mu,A)}$ with revealed strategy $(\mathbf{Q}, \bar{\mathbf{q}})$ such that $\bar{\mathbf{q}}$ is a deterministic action choice function. Consider (Q', q') with $\sum_{\gamma \in \text{ supp } Q'} \gamma Q'(\gamma) = \mu'$. If supp $Q' \subset$supp $\mathbf{Q}$, supp μ' =supp μ,and $q'(\gamma) = \bar{\mathbf{q}}(\gamma)$ for all $\gamma \in$supp Q', then $P_{(Q',q')} \in \mathbf{D}_{(\mu',A)}$.

The axiom states that if an attention strategy $\mathbf{Q}$ is observed for the decision problem (μ, A), then the posteriors in $\mathbf{Q}$ remain optimal for (μ', A) where μ' lies in the convex hull of supp $\mathbf{Q}$. Note that we require that μ' does not place zero probability on any state in the support of μ. The reason is that if supp $\mu' \neq$supp μ, then there is no reason to expect that $T_{\text{supp } \mu'} = T_{\text{supp } \mu}$ and no reason to expect that the optimal policy in the problem (μ', A) will bear any relation to the optimal policy in (μ, A).

[11] Comparing equations (5) and (7), one might be tempted to equate T_μ and $T_{\text{supp } \mu}(\gamma) - T_{\text{supp } \mu}(\mu)$. This is not the case, however. The precise relationship is

$$T_\mu(\gamma) = T_{\text{supp } \mu}(\gamma) - T_{\text{supp } \mu}(\mu) - \alpha \cdot (\gamma - \mu)$$

where α is an element of the subdifferential of $T_{\text{supp } \mu}$ at μ. Since $\sum Q(\gamma)\gamma = \mu$, the linear term drops out of the cost function K. It's only role is to ensure that $T_\mu(\gamma) \geq 0$, as required by the definition of posterior-separability.

4.3 From Posterior Separable to Uniformly Posterior Separable

Our first theorem states that a data set with a posterior-separable representation has a weakly uniformly posterior-separable representation if and only if it satisfies Locally Invariant Posteriors (Axiom 1).

Theorem 1: A data set **D** with a posterior-separable representation has a weakly uniformly posterior-separable representation if and only if it satisfies Locally Invariant Posteriors (Axiom 1).

The theorem captures the tight connection between weakly uniformly posterior-separable cost functions and Locally Invariant Posteriors. Locally Invariant Posteriors is the behavioral manifestation of weak uniform posterior separability. The sufficiency part of the proof uses the Lagrangian Lemma to show that, if the same posteriors are optimal in decision problems that differ only in their prior, then the associated posterior separable cost functions must be affine transforms of each other. It is then without loss of generality to assume that they are the identical. Necessity again follows from the Lagrangian Lemma: if two priors have the same cost function, then the same Lagrange multipliers can be used to determine optimal strategies in decision problems that differ only in these priors. This in turn implies that the posteriors that are optimal in one will also be optimal in the other, assuming they are still Bayes feasible.

5 Invariant Posterior-Separable Cost Functions

We begin this section by defining a second subclass of posterior-separable attention cost functions: invariant posterior-separable cost functions. We then present a related behavioral invariance axiom: Only Payoffs Matter. We close the section by stating Theorem 2, which establishes a tight link between these concepts.

5.1 Definition

Invariant cost functions impose two conditions on the cost of learning about states and about events.[12] Both involve a fixed state space $\bar{\Omega} \subset \Omega$ of finite size, and a partition $\{\bar{\Omega}_z\}_{z=1,...Z}$ of $\bar{\Omega}$. Consider any prior $\mu \in \text{int}\Delta\bar{\Omega}$ and corresponding attention strategy $Q \in \mathcal{Q}(\mu)$. For each $\gamma \in$supp Q, construct γ' such that γ' assigns the same probability as γ to each subset $\bar{\Omega}_z$,

$$\gamma'(\bar{\Omega}_z) \equiv \sum_{\omega \in \bar{\Omega}_z} \gamma'(\omega) = \gamma(\bar{\Omega}_z) \equiv \sum_{\omega \in \bar{\Omega}_z} \gamma(\omega), \tag{8}$$

[12] Our characterization of invariance is related to the definition in Hébert and La'O [2019] who build on the work of Chentsov [1982] and Amari and Nagaoka [2000].

and, within each subset $\bar{\Omega}_z$, the conditional probability of each state is equal to that of the prior,

$$\gamma'(\omega|\bar{\Omega}_z) = \mu(\omega|\bar{\Omega}_z) \tag{9}$$

Finally let

$$Q'(\gamma') = Q(\gamma) \tag{10}$$

where γ is the posterior in Q used in the construction of γ'.

The first defining feature of invariance involves the sense in which Q' represents less learning than Q. When it comes to the question of whether or not a state is in one of the sets $\bar{\Omega}_z$, Q' assigns the same probabilities as Q, but when it comes to understanding the states within the subset $\bar{\Omega}_z$, Q' is no different than the prior. Q' captures the learning in Q about the partition but not the learning within the partition. It is therefore not unreasonable to assume that the attention strategy Q' is less costly,

$$K(\mu, Q) \geq K(\mu, Q') \tag{11}$$

Versions of equation (11) are often referred to as information monotonicity (See Amari, 2016, 51-54). While, intuitively appealing, this inequality is not without content. Consider, for example, a decision problem with three states $\bar{\Omega} = \{\omega_1, \omega_2, \omega_3\}$. Consider a partition into two sets $\bar{\Omega}_1 = \{\omega_1, \omega_2\}$ and $\bar{\Omega}_2 = \{\omega_3\}$. Suppose that the prior is uniform, and consider two attention strategies: Q is comprised of the posteriors $\gamma_1 = (2/3, 1/3, 0)$ and $\gamma_2 = (0, 1/3, 2/3)$ with equal probability (where the ith element of the vector denotes the probability of state ω_i), while Q' is comprised of $\gamma'_1 = (1/2, 1/2, 0)$ and $\gamma'_2 = (1/6, 1/6, 2/3)$. Note that the two attention strategies agree on the probability of the partition: $\gamma_1(\bar{\Omega}_1) = \gamma'_1(\bar{\Omega}_1)$ and $\gamma_2(\bar{\Omega}_1) = \gamma'_2(\bar{\Omega}_1)$. Q and Q' therefore satisfy (8). Also γ'_1 and γ'_2 are uniform conditional on $\bar{\Omega}_1$. Hence Q' satisfies (9). Monotonicity would therefore imply that Q' is less costly than Q. Suppose, however, that ω_1 is very easy to distinguish from ω_3, and that ω_2 is very difficult to distinguish from ω_3. In this case, it is entirely possible that Q is the less costly attention strategy. Monotonicity rules out this sort of asymmetry.

The second defining feature of invariance relates to any alternative prior $\bar{\mu}$ which assigns the same probabilities as the original prior μ to each subset $\bar{\Omega}_z$ in the partition,

$$\bar{\mu}(\bar{\Omega}_z) = \mu(\bar{\Omega}_z).$$

For each $\gamma \in$supp Q, construct $\bar{\gamma}$ as we did γ'. For each $\bar{\Omega}_z$, set

$$\bar{\gamma}(\bar{\Omega}_z) = \gamma(\bar{\Omega}_z)$$

and for each $\omega \in \bar{\Omega}_z$ set

$$\bar{\gamma}(\omega|\bar{\Omega}_z) = \bar{\mu}(\omega|\bar{\Omega}_z)$$

Finally, let $\bar{Q}(\bar{\gamma}) = Q(\gamma)$. There is a sense in which Q' and $\bar{Q}$ represent the same amount of attention. The posterior distributions over the partition $\{\bar{\Omega}_z\}$ are the same for each attention strategy, and in each case the conditional distribution over states within the partition is equal to the prior. In each case, the same information is learned about the partition and nothing else. An invariant cost function imposes equal costs on the two strategies:

$$K(\mu, Q') = K(\bar{\mu}, \bar{Q}) \tag{12}$$

We now define an invariant cost function formally.[13] Before doing so, it is useful to define the operation that took us from Q to Q' and $\bar{Q}$. Given Q let Q_μ denote the attention strategy defined by the operations (8), (9), and (10) so that $Q' = Q_\mu$ and $\bar{Q} = Q_{\bar{\mu}}$.

Definition 5 *A cost function K is* ***invariant*** *if for all finite sets of states $\bar{\Omega} \subset \Omega$, all partitions of $\bar{\Omega}$, all pairs of priors μ and $\bar{\mu}$ that place equal probability on each partition subset, and all feasible strategies $Q \in \mathcal{Q}(\mu)$:*

$$K(\mu, Q) \geq K(\mu, Q_\mu)$$

and

$$K(\mu, Q_\mu) = K(\bar{\mu}, Q_{\bar{\mu}}).$$

A cost function is ***invariant posterior separable*** *if it is both invariant and posterior separable.*

We say that a data set $\mathbf{D}$ has an **invariant posterior-separable representation** if it can be represented by an invariant posterior-separable cost function.

To understand the behavioral implications of invariance, suppose that the cost function is invariant and consider a decision problem (μ, A) in which two states ω_1 and ω_2 are redundant in the sense that for all $a \in A$ we have $a(\omega_1) = a(\omega_2)$. In this case, the expected utility of any policy (Q, q) depends only on the sum $\gamma(\omega_1) + \gamma(\omega_2)$ for each $\gamma \in$supp Q, as the payoffs in these states are identical. The cost of attention, however, depends individually on $\gamma(\omega_1)$ and $\gamma(\omega_2)$. With an invariant cost function, however, it will always be cheaper to choose $\gamma(\omega_1)$ and $\gamma(\omega_2)$ so that the conditional probability of ω_1 given the event $\{\omega_1, \omega_2\}$ is equal to the prior probability of ω_1 conditional on $\{\omega_1, \omega_2\}$. There is no gain to learning the relative frequency of redundant states. It follows immediately that $\gamma(\omega_1)$ and $\gamma(\omega_2)$ are proportionate to $\mu(\omega_1)$ and $\mu(\omega_2)$ so that

$$\frac{\gamma(\omega_1)}{\mu(\omega_1)} = \frac{\gamma(\omega_2)}{\mu(\omega_2)}$$

[13] Note that we define invariance of $K(\mu, Q)$. Hébert and La'O [2020] define invariance of $T_\mu(\gamma)$. It is certainly the case that invariance of T_μ implies invariance of K. We conjecture that the converse is true, but we do not have a proof. If so, then the two definitions are equivalent.

Given $q(a|\gamma) > 0$, Bayes rule implies that:

$$P(a|\omega_1) = \frac{\gamma(\omega_1)P(a)}{\mu(\omega_1)} = \frac{\gamma(\omega_2)P(a)}{\mu(\omega_2)} = P(a|\omega_2)$$

Invariance therefore implies that the frequency with which each action is chosen is equalized across states which provide the same payoff to all actions. The two states ω_1 and ω_2 may without loss of generality be considered a single state. The "states" that matter for choice are not the states ω but the partition of the state space defined by the payoffs.

5.2 Only Payoffs Matter

Our axiom of Only Payoffs Matter captures this idea that the true economically relevant states are determined by the payoffs to actions. Given a set of actions A, let $\overrightarrow{A} = \{a^{(1)}, \ldots a^{(n)}\}$ define an ordered labeling of the actions. Note that this labeling is arbitrary. Expected utility only depends on the payoffs to actions and information costs depend only on posterior beliefs. Neither depends on the labeling of actions. It follows permuting $\overrightarrow{A}$ does not affect behavior.

Given a labeling $\overrightarrow{A} = \{a^{(1)}, \ldots a^{(n)}\}$, define the payoff profile $\overrightarrow{A}(\omega)$ in state ω as the vector $(a^{(1)}(\omega), \ldots a^{(n)}(\omega)) \in \mathbb{R}^n$. We say that two decision problems (μ_1, A_1) and (μ_2, A_2) are **payoff equivalent** if there exist labelings $\overrightarrow{A}_1$ and $\overrightarrow{A}_2$, such that $\mu_1\{\omega|\overrightarrow{A}_1(\omega) = f\} = \mu_2\{\omega|\overrightarrow{A}_2(\omega) = f\}$ for all observed payoff vectors $f \in \mathbb{R}^n$. Payoff equivalent decision problems assign the same probability to each payoff vector. They differ only in the mapping between these payoffs and the states that generate them. The labeling orders the actions so that the ith action in $\overrightarrow{A}_1$ has the same distribution of payoffs as the ith action in $\overrightarrow{A}_2$. Note that given the fixed length of the vector f, two payoff equivalent decision problems must have the same number of actions.

Given this notion of payoff equivalence, we can define what it means for behavior to depend only on payoffs. Consider two payoff equivalent decision problems (μ_1, A_1) and (μ_2, A_2) and associated labelings $\overrightarrow{A}_1$ and $\overrightarrow{A}_2$ such that $\mu_1\{\omega|\overrightarrow{A}_1(\omega) = f\} = \mu_2\{\omega|\overrightarrow{A}_2(\omega) = f\}$. Suppose that we have state dependent stochastic choice data $\mathbf{P}_{(\mu_1,A_1)}$ for (μ_1, A_1) and $\mathbf{P}_{(\mu_2,A_2)}$ for (μ_2, A_2). For behavior to depend only on payoffs, it must be the case that the probability of choosing the ith action from $\overrightarrow{A}_1$ must be the same as the probability of choosing the ith action from $\overrightarrow{A}_2$ across states with the identical payoff profiles. Formally, given $a_1^{(i)} \in \overrightarrow{A}_1$ and $a_2^{(i)} \in \overrightarrow{A}_2$ such that $a_1^{(i)}$ and $a_2^{(i)}$ are each the ith element in their respective labelings and given $\omega_1 \in$supp μ_1 and $\omega_2 \in$supp μ_2 such that $\overrightarrow{A}_1(\omega_1) = \overrightarrow{A}_2(\omega_2)$, we must have $\mathbf{P}_{(\mu_2,A_2)}(a_2^{(i)}|\omega_2) = \mathbf{P}_{(\mu_1,A_1)}(a_1^{(i)}|\omega_1)$. Note that this condition implies that the probability of choosing an action a from $\overrightarrow{A}_1$ must be equal across states with the same profile of payoffs, as we can match any two states in which $\overrightarrow{A}_1(\omega) = f$ to a single state in which $\overrightarrow{A}_2(\omega) = f$.

We now state the axiom.

Axiom 2 Only Payoffs Matter: Given any two payoff equivalent decision problems (μ_1, A_1) and (μ_2, A_2) and associated labelings $\overrightarrow{A}_1$ and $\overrightarrow{A}_2$ such that $\mu_1\{\omega|\overrightarrow{A}_1(\omega) = f\} = \mu_2\{\omega|\overrightarrow{A}_2(\omega) = f\}$ then $\mathbf{P}_{(\mu_1,A_1)} \in \mathbf{D}_{(\mu_1,A_1)}$ if and only if there exists $\mathbf{P}_{(\mu_2,A_2)} \in \mathbf{D}_{(\mu_2,A_2)}$ such that given $a_1^{(i)} \in \overrightarrow{A}_1$ and $a_2^{(i)} \in \overrightarrow{A}_2$ such that $a_1^{(i)}$ and $a_2^{(i)}$ are each the ith element in their respective labelings and $\omega_1 \in$supp μ_1 and $\omega_2 \in$supp μ_2 such that $\overrightarrow{A}_1(\omega_1) = \overrightarrow{A}_2(\omega_2)$, $\mathbf{P}_{(\mu_2,A_2)}(a_2^{(i)}|\omega_2) = \mathbf{P}_{(\mu_1,A_1)}(a_1^{(i)}|\omega_1)$.

5.3 From Posterior Separable to Invariant Posterior Separable

Only Payoffs Matter is a very powerful behavioral axiom. Theorem 2 establishes that it takes us from a data set with a posterior-separable representation to one with an invariant posterior-separable representation.

Theorem 2: A data set $\mathbf{D}$ with a posterior-separable representation has an invariant posterior-separable representation if and only if it satisfies Only Payoffs Matter (Axiom 2).

The proof of Theorem 2 involves two steps. The first step is to show that Only Payoffs Matter is identical to another axiom which we call Invariance under Compression. Invariance under Compression is similar to Only Payoffs Matter except that it applies to a fixed set of states and a fixed set of actions. Only Payoffs Matter, on the other hand applies to all payoff equivalent decision problems. We discuss Invariance under Compression in Section 7 below and establish its equivalence to Only Payoffs Matter. It is then relatively straightforward to show that invariance implies Invariance under Compression. Showing that Invariance under Compression implies invariance, however, is a bit more involved. One complication is that for each attention strategy Q we need to see the strategy Q_μ in the data in order to apply Invariance under Compression. This may be problematic if μ lies on the boundary of the simplex. Feasibility Implies Optimality (Lemma 2) and the assumption that T_μ in a posterior separable cost function is continuous prove useful in this regard.

6 Shannon Cost Functions and Representations

Recall from Section 2.2 that the Shannon cost function takes the form

$$K^S_\kappa(\mu, Q) \equiv \kappa \left[\sum_{\gamma \in \text{supp } Q} Q(\gamma) \sum_{\omega \in \text{supp } \gamma} \gamma(\omega) \ln \gamma(\omega) - \sum_{\omega \in \text{supp } \mu} \mu(\omega) \ln \mu(\omega) \right].$$

This function is weakly uniformly posterior separable with $T_{\text{supp }\mu}(\gamma)$ proportionate to the negative of the Shannon entropy of γ :

$$T_{\text{supp }\mu}(\gamma) = \kappa \sum_{\omega \in \text{supp }\gamma} \gamma(\omega) \ln \gamma(\omega).$$

$\kappa > 0$ is the only free parameter, it translates the cost of information in nats into units of utility. To see that K^S_κ is invariant, note that K^S_κ can be rewritten as the expectation (across posteriors) of the Kullback-Leibler divergence, $\sum_{\omega \in \text{supp }Q} \gamma(\omega) \ln \frac{\gamma(\omega)}{\mu(\omega)}$. K^S_κ then satisfies (11) and (12) as

$$\sum_{\omega \in \bar{\Omega}} \gamma(\omega) \ln \frac{\gamma(\omega)}{\mu(\omega)} \geq \gamma(\bar{\Omega}) \ln \frac{\gamma(\bar{\Omega})}{\mu(\bar{\Omega})}$$

for any set of states $\bar{\Omega} \subseteq$supp Q with equality when γ is proportionate to μ over $\omega \in \bar{\Omega}$.

Our last theorem states that the Shannon cost function is the unique invariant and weakly uniformly posterior-separable cost function.

Theorem 3: *The Shannon cost function is unique in that it is invariant and weakly uniformly posterior separable.*

We say that data set **D** has a Shannon representation if it can be represented by K^S_κ for $\kappa > 0$. It follows immediately from Theorems 1-3, that a data set with a posterior-separable representation has a Shannon representation if and only if it satisfies both Axioms 1 and 2.

Corollary: *A data set* **D** *with a posterior-separable representation has a Shannon representation if and only if it satisfies both Locally Invariant Posteriors (Axiom 1) and Only Payoffs Matter (Axiom 2).*

That the Shannon cost function is invariant and uniformly posterior separable is easy to establish. The complication is in showing that these properties define the Shannon cost function. The hard part of the proof is showing that $T_{\text{supp }\mu}(\gamma)$ is differentiable in γ, which allows us to use results from information geometry (See Section 8.4). The key insights are, first, that Only Payoffs Matter allows for the construction of decision problems in which the chosen posteriors are proportionate to the prior, and, second, that the Lagrangian Lemma (Lemma 1) relates the subdifferential of $T_{\text{supp }\mu}(\gamma)$ at these chosen posteriors.

7 Discussion and Extensions

We begin this section with an alternative statement of Axiom 2 and show that the two axioms are equivalent. This alternative statement simplifies the proofs of Theorems 2 and 3 because it only requires comparing payoff equivalent decision problems that share the same set of states and actions.

We also consider Assumption 1. Our starting point is a data set with a posterior-separable representation. We assume that this representation is continuous, strictly convex and finite on the interior of the support of the prior. We chose this starting point because we know that there are behavioral axioms that yield this starting point (Caplin, Dean and Leahy [2017]). We discuss briefly the implications of relaxing each of these assumptions.

7.1 Invariance under Compression

Given a labeled set of actions $\overrightarrow{A} = \{a^{(1)}, \dots a^{(n)}\}$, recall that a payoff profile $\overrightarrow{A}(\omega)$ is the vector $(a^{(1)}(\omega), \dots a^{(n)}) \in \mathbb{R}^n$. We say that a decision problem $(\bar{\mu}, A)$ is a **reduction of** (μ, A) if (1) the support of $\bar{\mu}$ is contained in the support of μ, supp $\bar{\mu}$ ⊂supp μ, and (2) the probability of each payoff profile is the same under both priors, $\bar{\mu}\{\omega|\overrightarrow{A}(\omega) = f\} = \mu\{\omega|\overrightarrow{A}(\omega) = f\}$ for all observed payoff vectors $f \in \mathbb{R}^n$. The idea of a reduction is that we have reduced the number of states with strictly positive probability (point 1) without altering the frequency with which any vector of payoffs is observed (point 2). We say that a decision problem (μ, A) is **basic** if there exists no decision problem $(\bar{\mu}, A)$ that is a reduction of (μ, A). Basic decision problems are those in which no two states have the same profile of payoffs. Let $\mathcal{B}(\mu, A)$ denote the set of basic decision problems that are reductions of (μ, A). We will call elements of $\mathcal{B}(\mu, A)$ **basic forms** of (μ, A).

Our Invariance under Compression axiom insists that patterns of choice are equivalent in all decision problems with a common basic form. Given a decision problem (μ, A) and a basic form $(\bar{\mu}, A) \in \mathcal{B}(\mu, A)$, we can define a **reduction mapping** ξ :supp μ →supp $\bar{\mu}$ which assigns each state in the support of μ to the unique state in of the support of $\bar{\mu}$ with the same payoff profile: $\overrightarrow{A}(\omega) = f$ implies $\overrightarrow{A}(\xi(\omega)) = f$. Invariance under Compression states that the observed frequency of action a in state ω in the decision problem (μ, A) is the same as the observed frequency of action a in state $\xi(\omega)$ in the basic form $(\bar{\mu}, A)$.

Axiom 2a Invariance under Compression: *Given any* (μ, A) and $(\bar{\mu}, A)$ such that $(\bar{\mu}, A) \in \mathcal{B}(\mu, A)$, *then* $\mathbf{P}_{(\mu,A)} \in \mathbf{D}_{(\mu,A)}$ if and only if there exists $\bar{\mathbf{P}}_{(\bar{\mu},A)} \in \mathbf{D}_{(\bar{\mu},A)}$ s.t. $\mathbf{P}_{(\mu,A)}(a|\omega) = \bar{\mathbf{P}}_{(\bar{\mu},A)}(a|\xi(\omega))$, for all ω ∈supp μ, *where* ξ :supp μ →supp $\bar{\mu}$ is the corresponding reduction mapping.

Invariance under compression compares two decision problems with identical actions A and a fixed set of states contained in the support of μ. This greatly simplifies the analysis. Only Payoffs Matter, on the other hand, compares any two decision problems with the same frequency of payoff. In spite of these differences, the two axioms are equivalent.

Proposition 1: A data set $\mathbf{D}$ satisfies Invariance under Compression if and only if it satisfies Only Payoffs Matter.

Only Payoffs Matter relates any two decision problems in which payoff profiles occur with the same frequency. Invariance under Compression only compares decision problems which share the same basic form. Only Payoffs Matter therefore directly implies Invariance under Compression. While we state the Invariance under Compression in terms of a decision problem and its basic form, it indirectly links problems that share a basic form. Given (μ_1, A) and (μ_2, A) with a common basic form $(\bar{\mu}, A)$ and the associated reduction mappings ξ_1 and ξ_2, the axiom implies that $\mathbf{P}_{(\mu_1,A)}(a|\omega) = \mathbf{P}_{(\mu_2,A)}(a|\omega')$ whenever $\xi_1(\omega) = \xi_2(\omega')$. This insight can be used to show that Invariance under Compression implies Only Payoffs Matter.

7.2 Continuity

Assuming posterior separable cost functions are continuous helps to alleviate a potential indeterminacy in the definition of T_μ. Observed behavior can only characterize $T_\mu(\gamma)$ over the set of revealed posteriors. If a particular posterior $\bar{\gamma}$ on the boundary of $\Delta(\text{supp } \mu)$ is never chosen in any decision problem and if $T_\mu(\bar{\gamma})$ is finite, then raising $T_\mu(\bar{\gamma})$ further will not have any effect on optimal choice. Continuity pins down $T_\mu(\gamma)$ at such points.

It turns out that given strict convexity, the assumption of continuity is without loss of generality.

Lemma 3 (Continuity): Consider a data set $\mathbf{D}$ that is represented by a cost function

$$\hat{K}(\mu, Q) = \sum_{\gamma \in \text{supp } Q} Q(\gamma)\hat{T}_\mu(\gamma).$$

where $\hat{T}_\mu$ is discontinuous at the boundary of $\Delta(\text{supp } \mu)$. Let

$$K(\mu, Q) = \sum_{\gamma \in \text{supp } Q} Q(\gamma)T_\mu(\gamma).$$

where $T_\mu(\bar{\gamma}) = \hat{T}_\mu(\bar{\gamma})$ for all $\bar{\gamma} \in \text{int}\Delta(\text{supp } \mu))$ and for all $\bar{\gamma}$ on the boundary,

$$T_\mu(\bar{\gamma}) = \lim_{\gamma \to \bar{\gamma}} \hat{T}_\mu(\gamma) \tag{13}$$

where the limit is taken with respect to $\gamma \in \text{int}(\Delta(\text{supp } \mu))$. Then K also represents $\mathbf{D}$.

Note that (13) allows $T_\mu(\bar{\gamma}) = \infty$.

The intuition behind the Lemma is simple. First note that since $\hat{T}_\mu$ is convex and real valued on the interior of $\Delta(\text{supp } \mu)$, the theorem of Gale, Klee and Rockafellar (1968) states that the extension to the boundary T_μ exists and is unique. Now consider any posterior-separable cost function $\hat{K}$ and suppose that the associated $\hat{T}_\mu$ is discontinuous at some posterior $\bar{\gamma}$ which lies on the boundary of $\Delta(\text{supp } \mu)$. Note $\bar{\gamma} \notin \hat{\Gamma}(\mu|K)$ since there are other posteriors in the neighborhood of $\bar{\gamma}$ which yield almost the same utility at strictly lower cost. Now replacing $\hat{T}_\mu(\bar{\gamma})$ with its continuous extension either has no effect at all on choice or there exists some decision problem in which the optimal policy changes. In the latter case, the new policy must involve $\bar{\gamma}$, since this is the only change to the cost function, and the new policy must have a value at least as high as the original policy or else $\bar{\gamma}$ would not be chosen. In this case, the strict convexity of the cost function implies that we can find a third policy, which is a mixture of the new policy and the original policy, that improves on the original policy in the original problem. this contradiction establishes the result.

The proof of this Lemma makes use of a special feature of revealed preference analysis, namely that we observe choices in all decision problems. Therefore all decision problems have solutions. Hence even when $T_\mu(\gamma)$ jumps at $\bar{\gamma}$, some other posterior must be chosen. The usual problem that occurs with discontinuous payoffs – the non-existence of policy that achieves the supremum – does not arise.

7.3 Strict Convexity

Assumption 1 restricts our posterior separable cost functions to be strictly convex. This is where we had left off in Caplin, Dean and Leahy [2017]. Many of the results, however, would go through in amended form if the cost function were weakly convex. The one place that we use strict convexity is in proving that continuity at the boundary is without loss of generality. This is no longer the case if the cost function is weakly convex.

Note that in our definition of a representation it is not without loss of generality to assume that the cost function is even weakly convex. This is because we assume that all strategies that achieve the optimum in the theoretical model are observed in the data set. If a non-convex cost function is replaced by a convex one, then the observed data will be contained among the theoretical optima, but there may be theoretically optimal strategies that are not observed.

7.4 Real-valued

Assumption 1 states that T_μ is finite on the interior of $\Delta\text{supp } \mu$. Allowing T_μ to equal infinity on the interior would make it possible to model situations in which certain posteriors are unlearnable even if they do not rule out any a priori possible state. This change would have no effect on Lemmas 1 and 2, except for point 3 of Lemma 2, which would no longer hold. These results only need that $T_\mu(\gamma)$ is a proper convex function.

Allowing T_μ to equal infinity on the interior of $\Delta\text{supp } \mu$ would affect Theorems 1 and 3. The key to proving these theorems is that we can find decision problems that link the cost function at different priors. In the case of Theorem 1, we need to find a decision problem for one prior such that the convex hull of the revealed posteriors contains another prior. This allows us to invoke Locally Invariant Posteriors. In the case of Theorem 3, we need to find decision problems in which the chosen posteriors are proportionate to the prior on some pair of states. This allows us to invoke Invariance under Compression. If T_μ is allowed to equal infinity on the interior, it is easy to construct examples in which no such problems exist. The simplest example is one in which learning is impossible: T_μ is infinite unless $\gamma = \mu$. More complex examples allow for Locally Invariant Posteriors to hold within partitions of the simplex, with costs infinite outside of a given partition. Additional assumptions would need to be made that knit the simplex together.

7.5 Strong and Weak Uniform Posterior Separability

In this paper we introduce the concept of weak uniform posterior separability, in which the cost of posterior beliefs is independent of the prior, as long as the support of the prior does not change. It is also possible to define other, stronger versions of uniform posterior separability. For example, the cost of posteriors in the Shannon model is completely independent of the prior regardless of the support of the distribution. This, however, is not necessarily true of all cost functions that one might want to include in the uniform posterior separable class, specifically those that make it prohibitively costly to rule out ex-ante possible states. Consider, for example, costs based on Tsallis entropy with $\sigma < 0$:

$$TS_\sigma(\gamma) = \frac{1}{\sigma - 1}\left(1 - \sum_{\omega \in \text{supp}(\mu)} \gamma(\omega)^\sigma\right)$$

The cost of posteriors that set $\gamma(\omega) = 0$ for some $\omega \in$supp(μ) is infinite. Yet the same posterior belief can have finite cost if ω is not in the support of μ. This cost function therefore depends on the support of the prior.[14] To include such cost functions, Caplin, Dean and Leahy [2017] defined an intermediate version of uniform posterior separability in which $T(\gamma)$ was independent of the prior for information structures that are optimal in some decision problem, while allowing for this set of information structures that are ever used to depend on the prior.

The current weak version of the uniform posterior separability simplifies the analysis relative to Caplin, Dean and Leahy [2017]. The version in that paper did not lead to a simple and straightforward "if and only if" characterization of the Shannon cost function. To prove that that version of uniform posterior separability implied locally invariant posteriors, we needed an additional "Regularity" assumption that linked the sets of posteriors observed in decision problems that with overlapping supports. In this paper, we use invariance to link the cost function across priors with

[14] It is also true that the Shannon cost function implies that the set of posteriors that will be chosen in some decision problem depends on the support of the prior, but since Shannon entropy is equal to zero on the boundary, this does not necessitate that the cost function depend on the support of the prior.

different supports.

Note many papers in the literature effectively use the weak version of uniform posterior separability. These include Denti [2020] and Bloedel and Zhong [2020], both of whom use a fixed state space and prior beliefs on the interior of the simplex.

7.6 Monotonicity vs Blackwell Ordering

Information monotonicity is a very different concept than a Blackwell ordering. The example given in Section 5.1 provides two information structures Q and Q' that are monotonic but not Blackwell ordered. To see that Q' cannot be more Blackwell informative than Q, introduce a decision problem A with two actions: action b has utility 1 in all states and action a has utility 0 in all states except ω_1 in which it has a payoff u_1 such that a is optimal when the probability of ω_1 is $2/3$ or higher but not $1/2$ or lower. For example $u_1 = 7/4$. In this case it is optimal to choose b in for both γ_1' and γ_2', whereas a is chosen for γ_1 and b for γ_2. It follows that $U(Q, q|\mu, A) > U(Q', q'|\mu, A)$ were q and q' are the action choices just described. Q' therefore cannot be more Blackwell informative than Q. To see that the converse is true, consider the decision problem $A' = \{a', b\}$ with action a' that has utility 0 in all states except ω_2 in which it has a payoff u_2' such that it is optimal to choose a' when the probability of ω_2 is $1/2$ but not $1/3$ or lower. For example $u_2' = 5/2$. In this case it is optimal to choose a' for γ_1' and b for γ_2', whereas b is chosen for both γ_1 and γ_2. It follows that $U(Q', q'|\mu, A) > U(Q, q|\mu, A)$ so that Q cannot be Blackwell more informative than Q.

It is also possible that two distributions are Blackwell ordered but not monotonic. With two equiprobable states, perfect knowledge is more Blackwell informed than any other distribution but there is no distribution that satisfies the restrictions of monotonicity: this is general for any number of states.

8 Literature

We outline connections between our research and related research in four main areas. Before doing this it is important to understand what is new in this paper relative to earlier versions it incorporates. Uniformly posterior-separable models were introduced in Caplin and Dean [2013], while Caplin, Dean, and Leahy [2017] introduced the broader category of posterior-separable cost functions. For that reason uniformly posterior-separable cost functions are the best studied. After reviewing recent research on cost functions in this class, we pull back in section 9.2 to the broader literature on costs of information. A key feature of our approach is our focus on the behavioral imprint of cost functions in rich choice data. In section 9.3 we therefore place our work in relation to other research of this form. In section 9.4 we relate our work to the notion of invariance in information geometry. Finally in section 9.5 we relate it to other axiomatic treatments of Shannon entropy and the Shannon cost function.

8.1 Uniformly Posterior-Separable Cost Functions

Uniformly posterior-separable models have been studied in part because there are settings in which Locally Invariant Posteriors is intuitively reasonable. For example this behavioral axiom underlies the Drift Diffusion model which has proven popular in psychology (see Ratcliff, *et al.* [2016] for a recent review) and economics (Fehr and Rangel [2011]). According to the basic version of this model, an agent gathers information to resolve prior uncertainty over two ex ante possible 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 used for any prior that lies between the posteriors.

Recent work of Morris and Strack [2017], Hébert and Woodford [2019], and Bloedel and Zhong [2021] provides a related perspective on why uniformly posterior-separable cost functions may be of interest. The work of Hébert and Woodford [2019] in particular 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. Their theorem 1 pinpoints static uniformly posterior-separable 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 uniformly posterior-separable class. They also show how to derive the particular cost function from the local structure of learning. The link between static uniformly posterior-separable 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.

Given the reasonable patterns of behavior they produce and their tractability, a number of papers have made use of uniformly posterior-separable costs. Of particular note are papers that have used posterior-separable cost functions to examine situations in which both costly information acquisition and persuasion are important (Gentzkow and Kamenica [2014], Matyskova [2018]). Moreover recent work by Miao and Xing [2019] has shown how a posterior-based approach lends itself naturally to dynamic programming as the chosen posteriors today become the priors tomorrow. Zhong [2017] also considers dynamic information acquisition under posterior separability.

There is also a stream of research explicitly aimed at introducing uniformly posterior-separable cost functions that capture features of behavior ruled out by Only Payoffs Matter and Invariance under Compression. The strong symmetry properties that these behavioral axiom convey often fails in practice. To date, three such violations have been addressed by replacing the Shannon cost function with other uniformly posterior-separable cost functions.

As noted in section 5, the symmetry imposed by invariance rules out perceptual distance as a factor in discriminating between states. Yet it is critical in many every day decisions: prices which are closer together are harder to distinguish than those which are far apart. By way of confirmation, Dean and Neligh [2017] design an experiment that highlights precisely this failing of the Shannon model. A recent paper by Hébert and Woodford [2020] provides a family of uniformly posterior-separable cost functions that can accommodate the notion of perceptual distance. They propose a class of "neighborhood-based" cost functions. In order to construct these costs, the state space is divided into I "neighborhoods" $\Omega_1 ... \Omega_I$. A posterior is assigned a cost for each neighborhood based on some convex function of the distribution conditional on being in that neighborhood. The total cost of the posterior is then the sum of costs across all neighborhoods. Hence neighborhood based cost functions allow for it to be more expensive to differentiate between some states than others: the cost of differentiating between two states depends on which neighborhoods they share. Hence states that share more neighborhoods can be more costly to distinguish. Following Hébert and Woodford [2020], Dean and Neligh [2017] consider a neighborhood-based cost function that does a good job of fitting the data from an experimental design in which perceptual distance plays a critical role in learning.

A second feature of the Shannon model, again implied by Invariance under Compression, is that behavior should be invariant to changes in prior beliefs that move probabilities between payoff equivalent states. Woodford [2012] cites experimental evidence that challenges this implication. 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. Since the location on the screen is payoff irrelevant, Invariance under Compression implies that it should also be irrelevant to task performance. Yet in practice, performance is superior at locations that occur more frequently. Caplin, Dean and Leahy [2017] show that a cost function based on Tsallis entropy (Tsallis [1988]) is sufficiently flexible to allow for this.

Third, the Shannon model makes precise predictions about the rate at which subjects improve their accuracy in response to improved incentives: essentially, the observation of behavior in any given decision problem pins down the model's one free parameter, and so behavior in any other decision problem. Caplin and Dean [2013] show in a simple two state, two action set-up that agents are not responsive enough to changes in incentives: they do not pay enough attention at high rewards given the attention paid at low rewards. Relaxing either Only Payoffs Matter or Locally Invariant Posteriors can address this problem. Caplin and Dean [2013] and Dean and Neligh [2017] relax Only Payoffs Matter and consider uniformly posterior-separable cost functions based on generalized entropy (Shorrocks [1980]). Dean and Neligh [2017] show that statistical tests on their experimental data favor a model with generalized entropy over the Shannon model. Dean and Neligh [2017] also consider relaxing Locally Invariant Posteriors, by considering non-linear transforms of Shannon mutual information.

These last two examples show that the uniformly posterior-separable class allows us to replace

Shannon with other forms of entropic cost, which has proven valuable in other disciplines. Examples include internet usage (Tellenbach, *et al.* [2009]), machine learning (Maszczyk and Duch [2008]), statistical mechanics (Lenzi, Mendes and Da Silva [2000]), and many other applications in physics (Beck [2009]). See Gell-Mann and Tsallis [2004] for a review.

8.2 Posterior-Separable and Invariant Cost Functions

There are well motivated cost functions in the literature that are posterior separable but not uniformly posterior separable, such as the Log-Likelihood Ratio cost function of Pomatto, Strack and Tamuz [2018]. Yet the economic research in this area remains in its infancy. To date, most of the literature has focused on uniformly posterior-separable costs and failings of Only Payoffs Matter. While there is as yet little direct research demonstrating failings of Locally Invariant Posteriors, we believe that this will change as cost functions become more widely studied. As these violations are noted, posterior-separable cost functions will provide an attractive combination of tractability and flexibility. In analytic terms, all of our results are made possible by the Lagrangian Lemma, which show that the model can be solved by identifying the tangent to the concavified net utility function. This approach has been widely used since its introduction to the economics literature (Aumann, Maschler, and Stearns [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).

While new in the economic literature, the literature in information theory and on the design of experiments has also focused on posterior-separable cost functions. For example, the Blackwell-Sherman-Stein Theorem shows that posterior-separable cost functions can be used to characterize the property of statistical sufficiency, and so provide an alternative characterization of Blackwell's theorem. Given prior μ the theorem states that an information structure Q_1 is statistically sufficient for Q_2 (i.e. Q_1 Blackwell dominates Q_2) if and only if,

$$\sum_{\gamma \in \text{supp } Q_1} Q_1(\gamma) T_\mu(\gamma) \geq \sum_{\gamma \in \text{supp } Q_2} Q_2(\gamma) T_\mu(\gamma),$$

for every continuous, (weakly) convex T_μ (see for example Le Cam [1996]).[15] Torgersen [1991] further shows that the class of posterior-separable cost functions can be characterized by properties of the costs themselves. Specifically, the (weakly convex) posterior-separable class of cost function of information structures characterizes monotonicity in Blackwell informativeness and linearity in a natural mixture operation.

Invariant posterior-separable cost functions arose independently in the work of Angeletos and Sastry [2019] and Hébert and La'O [2020]. The former study the efficiency of Walrasian equi-

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

libria with learning and the latter ask how the endogeneity of information affects the efficiency of games with strategic interactions. In both cases, invariance is a sufficient condition for equilibrium attention strategies to be efficient in settings in which the full information equilibrium is efficient. Without invariance, one can often construct signals that are uncorrelated with payoffs, reduce the cost of information, and coordinate actions in suboptimal ways. The link between invariant posterior-separable cost functions and the Invariance under Compression axiom (introduced in Caplin, Dean and Leahy [2017]) is new to this paper, as is the formulation of this axiom in terms of Only Payoffs Matter. It is noteworthy that this is the axiom that tends to fail in the experimental literature, so the efficiency of equilibria is far from assured.

8.3 Revealed Preference and Imperfect Information

The profession has long been grappling with a basic identification problem: that of separating beliefs and preferences. Our work belongs to a recent body of literature addressed to this challenge. One approach involves modeling processes of learning that have clear implications for standard choice data. For example Masatlioglu, Nakajima, and Ozbay [2012] introduce a model of limited attention that can be identified in standard choice data. Their method of identification requires strong assumptions on the nature of inattention: options are either considered perfectly or not at all, and the removal of an alternative that is not considered from a choice set does not affect what is considered.

Another approach to identification involves data enrichment. In their pioneering work on stochastic choice data and random utility, Block and Marschak [1959] noted that their data set was inadequate to separate random perception from random utility. They proposed development of new data sets to accomplish this separation.

As has been noted, the data enrichment on which our research lies involves observing not only choices, but also underlying facts about the environment. The extent to which these underlying realities impact choice is encoded in state dependent stochastic choice data. This data set turns out to be useful for revealed preference analysis. Caplin and Martin [2015] establish “no improving action switches” (NIAS) as the behavioral signature of Bayesian expected utility maximization in this data, while Caplin and Martin [2020] use it to characterize robust welfare rankings of different framings of one and the same decision problem. de Clippel and Rozen [2020] study the interplay between costly perception and strategic inferences, both experimentally and theoretically. They do so by characterizing, in state dependent stochastic choice data, the testable implications of equilibrium play in a class of games with strategic communication. With regard to costs, Caplin and Dean [2015] show that NIAS and an additional behavioral axiom, “no improving action cycles” (NIAC) characterize data sets in which learning is optimized for a fixed information cost function. Caplin, Csaba, Leahy, and Nov [2020], provide a simple method of recovering costs from this data set. Chambers, Liu, and Rehbeck [2020] relax the assumption that information costs are additively separable from the gross utility resulting from choice.

Our contribution in this paper is to show that state dependent stochastic choice data is of great value in identifying qualitative as well as quantitative aspects of attention costs. Yet our analysis leaves many open questions in this regard. Most pertinently, the analysis in this paper starts with a posterior-separable cost function. In the working paper version of this paper (Caplin, Dean and Leahy [2017]) we provided additional behavioral axioms characterizing cost functions with this property: these are neither intuitive nor easy to test. Recently Denti [2020] has provided a simpler treatment given a finite number of decision problems, involving a "no improving posterior cycles" axiom. Other open questions relate to the behavioral signatures of non posterior-separable cost functions involving: costly purchase of normal signals (Verrecchia [1982], Llosa and Venkateswaran [2012] and Colombo, Femminis, and Pavan [2014]); "all or nothing" information costs (Reis [2006]); costs involving a hard constraint (Sims [2003]) or strictly convex in mutual information (Paciello and Wiederholt [2014]); costs linear in Shannon capacity Woodford [2012]; or costs covered by the sparsity-based model of Gabaix [2014]. How these cost functions restrict behavior, and so how they differ from the posterior-separable class, remains open.

8.4 Information Geometry

Information geometry studies the geometry of divergences. A divergence is a weak notion of the distance between probability distributions. Given an arbitrary state space S, a divergence $D(p||q)$ is a function from $\text{int}\Delta S \times \text{int}\Delta S$ to $\bar{\mathbb{R}}$ satisfying only that $D(p||q) \geq 0$ and $D(q||q) = 0$. $T_\mu(\gamma)$ in a posterior-separable representation is a divergence on $\text{int}(\Delta\text{supp }\mu)$ extended to γ on the boundary of $\Delta\text{supp }\mu$. By Lemma 3 we may take this extension to be continuous. Our definition of an invariant cost function is a direct adaptation of the definition of an invariant divergence to priors and attention strategies (see Amari [2016]).

At various parts of the proof of Theorem 3 we make use of connections with the information geometry. The most closely related result shows that the Kullback-Leibler divergence is unique in that it is at once invariant and a Bregman divergence (see Jiao, *et al.* [2014]). A Bregman divergence specifies the distance between γ and μ as the distance between γ and the hyperplane tangent to a differentiable, convex function at μ. It takes the form

$$D_B(\gamma||\mu) = f(\gamma) - f(\mu) - \nabla f(\mu) \cdot (\gamma - \mu)$$

where f is the differentiable, convex function. The class of weakly uniformly posterior-separable cost functions is the class of cost functions K that are expectations of Bregman divergences. To see this note that since $\sum Q(\gamma)\gamma = \mu$, the linear term $\nabla f(\mu) \cdot (\gamma - \mu)$ drops out when we calculate the cost of an attention strategy $Q \in \mathcal{Q}(\mu)$. We are left with a weakly uniformly posterior-separable cost function with with $T_{\bar{\Omega}}$ equal to f.

A major portion of the proof of Theorem 3 involves showing that together invariance and weak uniform posterior separability imply that $T_{\bar{\Omega}}$ is differentiable on the interior of $\Delta\bar{\Omega}$. With differentiability in hand, we are able to show that the invariance of K implies that the Bregman divergence $T_{\bar{\Omega}}(\gamma) - \nabla T_{\bar{\Omega}}(\mu) \cdot (\gamma - \mu)$ is invariant. The result of Jiao, *et al.* [2014] implies that this divergence is the Kullback-Leibler divergence, and hence T_μ is the Kullback-Leibler divergence. We complete the proof by linking together different state spaces. As a first step, we note that Axiom 2 implies that the cost function has the same form for all state spaces with the same cardinality, so that the particular states in $\bar{\Omega}$ do not matter. We then show that invariance equates the coefficient κ across sets of states with different cardinality. Intuitively, Invariance under Compression links behavior in a decision problem with behavior in the basic version which may be of lower dimension.[16]

Optimization plays an essential role in the proof of differentiability. The Lagrangian Lemma, which is entirely based on optimization theory, is essential in this proof. It relates the subdifferential of $T_{\bar{\Omega}}$ at different chosen posteriors and places structure on the set of posteriors at which $T_{\bar{\Omega}}$ is differentiable. Invariance under Compression, by placing structure on the set of posteriors that are solutions to decision problems, also places structure on the set of posteriors at which $T_{\bar{\Omega}}$ is differentiable. There is no parallel to these results in the information geometry literature because this literature is not built around decisions and behavior.

8.5 Foundations of the Shannon Model

This paper contributes to the literature on the axiomatic foundations of the Shannon information cost function. de Olivera [2014] provides an alternative axiomatization of the Shannon cost function, one that places axioms on preference orderings over menus. One of the axioms is a "symmetry axiom" in which states that 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 Only Payoffs Matter. Invariance under Compression also appears related to de Olivera's Independence of Orthogonal Decision Problems axiom. This axiom involves indifference between solving two decision problems with independent payoffs together or separately. We originally had both Invariance under Compression and a version of Independence of Orthogonal Decision Problems. We found, however, that we could dispense with the latter.

In addition to these characterizations, several recent papers have provided insights into the behavior implied by the Shannon model. Matějka 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. Subsequent papers (Stevens [2019], Caplin, Dean and Leahy [2019]) show that the addition of appropriate complementary slackness conditions provides both necessity and sufficiency.

[16] In Caplin, Dean and Leahy [2017] we provide an alternate proof of Theorem 3 that does not rely on the result of Jiao *et al [2014]*.

9 Concluding Remarks

We introduce three new classes of attention cost function: *posterior separable, weakly uniformly posterior separable* and *invariant posterior separable.* As with the Shannon cost function, we show that they can all be solved using Lagrangian methods. *Uniformly posterior-separable* cost functions capture many forms of sequential learning, hence play a key role in many applications. *Invariant posterior-separable* cost functions make learning strategies depend exclusively on payoff uncertainty. We characterize the resulting behavior in state dependent stochastic choice data. We show that two behavioral axioms, *Locally Invariant Posteriors* and *Only Payoffs Matter* (or *Invariance under Compression),* define posterior-separable functions respectively as uniformly and invariant posterior separable. We show that in combination they pinpoint the Shannon cost function.

References

[1] Alonso, Richard and Odilon Câmara. 2016. “Bayesian Persuasion with Heterogeneous Priors.” *Journal of Economic Theory* 165, 672–706.

[2] Amari, Shun-ichi. 2016. *Information Geometry and Its Applications*. Springer Japan.

[3] Amari, Shun-ichi, and Hiroshi Nagaoka. 2000. *Methods of Information Geometry.* Oxford: Oxford University Press.

[4] Angeletos, George-Marios, and Karthik Sastry. 2019. “Inattentive Economies.” NBER Working Paper No. 26413.

[5] Aumann, Robert, Michael Maschler, and Richard E Stearns. 1995. *Repeated Games with Incomplete Information.* Cambridge: MIT press.

[6] Bartoš, Vojtěch, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. “Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition.” *American Economic Review* 106, 1437–1475.

[7] Beck, Christian. 2009. “Generalised Information and Entropy Measures in Physics.” *Contemporary Physics* 50, 495–510.

[8] Block, Henry David, and Jacob Marschak. 1959. “Random Orderings and Stochastic Theories of Response.” Cowles Foundation Discussion Paper No. 66.

[9] Bloedel, Alex and Weijie Zhong. 2020. “The Cost of Optimally Acquired Information.” Working paper.

[10] Caplin, Andrew, Daniel Csaba, John Leahy, and Oded Nov. 2020. “Rational Inattention, Competitive Supply, and Psychometrics.” *Quarterly Journal of Economics* 135, 1681-1724.

[11] Caplin, Andrew, and Mark Dean. 2013. “Behavioral Implications of Rational Inattention with Shannon Entropy.” NBER Working Paper No. 19318.

[12] Caplin, Andrew, and Mark Dean. 2015. “Revealed Preference, Rational Inattention, and Costly Information Acquisition, *American Economic Review* 105, 2183–2203.

[13] Caplin, Andrew, and Daniel Martin. 2015. “A Testable Theory of Imperfect Perception.” *Economic Journal* 125, 184–202.

[14] Caplin, Andrew, and Daniel Martin. 2020. “Framing, Information, and Welfare." NBER Working Paper No. 27265.

[15] Caplin, Andrew, John Leahy, and Filip Matějka. 2015. “Social Learning and Selective Attention.” NBER Working Paper No. 21001.

[16] Caplin, Andrew, Mark Dean, and John Leahy. 2017. "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy." NBER Working Paper No. 23652.

[17] Caplin, Andrew, Mark Dean, and John Leahy. 2019. "Rational Inattention, Optimal Consideration Sets, and Stochastic Choice." *Review of Economic Studies* 86, 1061-1094.

[18] Chambers, Christopher, Ce Liu, and John Rehbeck. 2020. "Costly Information Acquisition." *Journal of Economic Theory* 186(C).

[19] Chentsov, N.N.. 1982. "Statistical Decision rules and Optimal Inference." American Mathematical Society.

[20] Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99, 1145–1177.

[21] Colombo, Luca, Gianluca Femminis, and Alessandro Pavan. 2014. "Information Acquisition and Welfare." *Review of Economic Studies* 81, 1438–1483.

[22] de Clippel, Geoffroy, and Kareen Rozen. 2020. "Communication, Perception and Strategic Obfuscation." Working Paper.

[23] de Oliviera, Henrique. 2014. "Axiomatic Foundations for Entropic Costs of Attention." Working Paper.

[24] Dean, Mark and Nathaniel Neligh. 2017. "Experimental Tests of Rational Inattention." Working Paper.

[25] Denti, Tomasso. 2020. "Posterior-Separable Cost of Information." Working Paper.

[26] Dewan, Ambuj, and Nathaniel Neligh. 2020. "Estimating Information Cost Functions in Models of Rational Inattention." Journal of Economic Theory 187, 1-32.

[27] Ely, Jeffrey, and Martin Szydlowski. 2017. "Moving the Goalposts." Working Paper.

[28] Fehr, Ernst, and Antonio Rangel. 2011. "Neuroeconomic Foundations of Economic Choice – Recent Advances." *Journal of Economic Perspectives* 25, 3-30.

[29] Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129, 1661–1710.

[30] Gale, David, Victor Klee and R. T. Rockafellar. 1968. "Convex Functions on Convex Polytopes." *Proceedings of the American Mathematical Society* 19, 867-873.

[31] Gell-Mann, Murray, and Constantino Tsallis. 2004. *Nonextensive Entropy: Interdisciplinary Applications*, Oxford: Oxford University Press.

[32] Gentzkow, Matthew, and Emir Kamenica. 2014. "Costly Persuasion." *American Economic Review* 104, 457–462.

[33] Hayek, Friedrich. 1937. "Economics and Knowledge." *Economica* 4, 33–54.

[34] Hayek, Friedrich. 1945. "The Use of Knowledge in Society." *American Economic Review* 35, 519–530.

[35] Hébert, Benjamin, and Jennifer La'O. 2020. "Information acquisition, Efficiency, and Non-Fundamental Volatility." NBER Working Paper No. 26771.

[36] Hébert, Benjamin, and Michael Woodford. 2018. "Information Costs and Sequential Information Sampling." NBER Working Paper No. 25316.

[37] Hébert, Benjamin, and Michael Woodford. 2019. "Rational Inattention when Decisions take Time." NBER Working Paper No. 26415.

[38] Hébert, Benjamin, and Michael Woodford. 2020. "Neighborhood-Based Information Costs." NBER Working Paper No. 26743.

[39] Jiao, Jiantao, Thomas Courtade, Albert No, Kartik Venkat, and Tsachy Weissman. 2014. "Information Measures: the Curious Case of the Binary Alphabet." *IEEE Transactions on Information Theory* 60, 7616–7626.

[40] Kamenica, Emir, and Matthew Gentzkow. 2011. "Bayesian Persuasion." *American Economic Review* 101, 2590–2615.

[41] Kaur, Supreet, Sendhil Mullainathan, Suanna Oh, and Frank Schilbach. 2019. "Does Nancial Strain Lower Productivity?" Working Paper.

[42] Le Cam, L.. 1996. "Comparison of Experiments: A Short Review." in *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, Institute of Mathematical Statistics, Lecture Notes-Monograph Series 30, 127-138.

[43] Lenzi, EK, RS Mendes, and LR Da Silva. 2000. "Statistical Mechanics Based on Renyi Entropy." *Physica A: Statistical Mechanics and its Applications*, 280, 337–345.

[44] Llosa, Luis Gonzalo and Venky Venkateswaran. 2012. "Efficiency with Endogenous Information Choice." Working Paper.

[45] Kőszegi, Boton and Matějka, Filip. 2020. "Choice simplification: A theory of mental budgeting and naive diversification." *Quarterly Journal of Economics* 135, 1153-1207.

[46] Mackowiak, Bartosz, and Mirko Wiederholt. 2009. "Optimal Sticky Prices under Rational Inattention." *American Economic Review* 99, 769–803.

[47] Martin, Daniel. 2017. "Strategic Pricing with Rational Inattention to Quality." Games and Economic Behavior 104, 131–145.

[48] Masatlioglu, Yusufcan, Daisuke Nakajima, Erkut Ozbay. 2012. "Revealed Attention." American Economic Review 102, 2183-2205.

[49] Maszczyk, Tomasz and Włodzisław Duch. 2008. "Comparison of Shannon, Renyi and Tsallis Entropy used in Decision Trees." *International Conference on Artificial Intelligence and Soft Computing*, 643–651.

[50] Matějka, Filip. 2015. "Rationally Inattentive Seller: Sales and Discrete Pricing." *Review of Economic Studies* 83, 1156–1188.

[51] Matějka, Filip, and Alisdair McKay. 2015. "Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model." *American Economic Review* 105, 272–98.

[52] Matyskova, Ludmila. 2018. "Bayesian Persuasion with Costly Information Acquisition." Working Paper.

[53] Miao, Jianjun, and Hao Xing. 2019. "Dynamic Rationally Inattentive Discrete Choice: A Posterior-Based Approach." Working Paper.

[54] Mondria, Jordi. 2010. "Portfolio Choice, Attention Allocation, and Price Comovement." *Journal of Economic Theory* 145, 1837–1864.

[55] Morris, Stephen, and Philipp Strack. 2017. "The Wald Problem and the Equivalence of Sequential Sampling and Static Information Costs." Working Paper.

[56] Paciello, Luigi, and Mirko Wiederholt. 2014. "Exogenous Information, Endogenous Information and Optimal Monetary Policy." *Review of Economic Studies* 83, 356–388.

[57] Painsky, Amichai, and Gregory Wornell. 2020. "Bregman Divergence Bounds and Universality Properties of the Logarithmic Loss." *IEEE Transactions on Information Theory* 66, 1658-1673.

[58] Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2018. "The Cost of Information." Working Paper.

[59] Ratcliff, Roger, Philip Smith, Scott Brown, and Gail McKoon. 2016. "Diffusion Decision Model: Current Issues and History." *Trends in Cognitive Sciences* 20, 260–281.

[60] Ravid, Doron. 2020. "Ultimatum Bargaining with Rational Inattention." *American Economic Review* 110, 2948-63.

[61] Reis, Ricardo,. 2006. "Inattentive Producers." *Review of Economic Studies* 73, 793–821.

[62] Richter, Marcel. 1966. "Revealed Preference Theory." *Econometrica* 34,635–645.

[63] Rockafellar, R. Tyrrell. 1970. *Convex Analysis,* Princeton: Princeton University Press.

[64] Shaw, M. L., and P. Shaw. 1977. "Optimal Allocation of Cognitive Resources to Spatial Locations." Journal of Experimental Psychology: Human Perception and Performance 3, 201–211.

[65] Shorrocks, Anthony. 1980. "The Class of Additively Decomposable Inequality Measures." *Econometrica* 48, 613–625.

[66] Sims, Christopher. 1998. "Stickiness." *Carnegie-Rochester Conference Series on Public Policy* 49, 317–356.

[67] Sims, Christopher. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50, 665–690.

[68] Steiner, Jakub, Colin Stewart, and Filip Matějka. 2015. "Rational Inattention Dynamics: Inertia and Delay in Decision-Making." Working Paper.

[69] Stevens, Luminita. 2019. "Coarse Pricing Policies." *Review of Economic Studies*, forthcoming.

[70] Tellenbach, Bernhard, Martin Burkhart, Didier Sornette, and Thomas Maillart. 2009. "Beyond Shannon: Characterizing Internet Traffic with Generalized Entropy Metrics." *International Conference on Passive and Active Network Measurement*, 239–248.

[71] Torgersen, Erik. 1991. *Comparison of Statistical Experiments,* Cambridge: Cambridge University Press.

[72] Tsallis, Constantino. 1988. "Possible Generalization of Boltzmann-Gibbs Statistics." *Journal of Statistical Physics* 52, 479–487.

[73] Verrecchia, Robert. 1982. "Information Acquisition in a Noisy Rational Expectations Economy." *Econometrica* 50, 1415–1430.

[74] Woodford, Michael. 2009. "Information Constrained State Dependent Pricing." *Journal of Monetary Economics* 56(S. S100–S124.

[75] Woodford, Michael. 2012. "Inattentive Valuation and Reference-Dependent Choice." Working Paper.

[76] Yang, Ming. 2015. "Coordination with Flexible Information Acquisition." *Journal of Economic Theory* 158, 721–738.

[77] Zhong, Weijie. 2017. "Optimal Dynamic Information Acquisition." Working Paper.