Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy[*]

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

February 2019

Abstract

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2 Attention Strategies and Costs

2.1 A Note on Notation

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

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

We let Γ denote the set of all distributions over the set of states of the world Ω. $\Gamma(\cdot)$ is then used throughout to restrict Γ in a way implied by its argument. For example, given μ, a distribution over states of the world, $\Gamma(\mu)$ is the restriction to distributions that are absolutely continuous with respect to μ. Given Q, a distribution over distributions, $\Gamma(Q)$ is the support of Q.

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

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

2.2 Posterior-Based Attention Strategies

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

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

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

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

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

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

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

simple probability distributions over posteriors and corresponding mixed action strategies,

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

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

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

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

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

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

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

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

[6] We use the notation

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

to describe probability distributions.

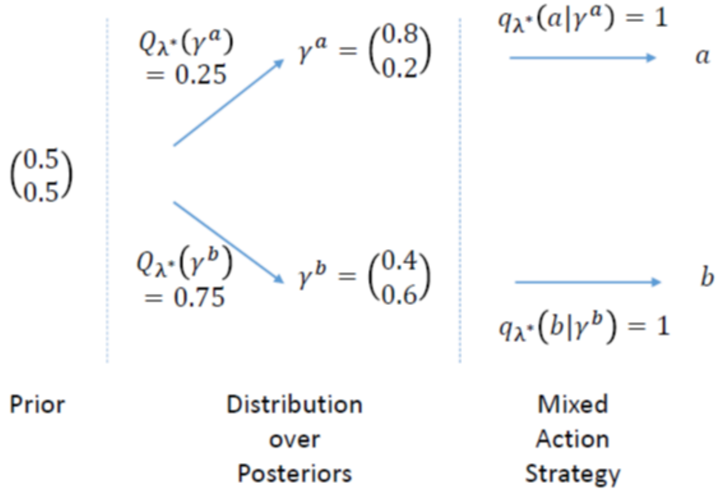
$\binom{0.5}{0.5}$
$Q_{\lambda^*}(\gamma^a) = 0.25$
$\gamma^a = \binom{0.8}{0.2}$
$Q_{\lambda^*}(\gamma^b) = 0.75$
$\gamma^b = \binom{0.4}{0.6}$
$q_{\lambda^*}(a|\gamma^a) = 1$
a
b
$q_{\lambda^*}(b|\gamma^b) = 1$
Prior
Distribution over Posteriors
Mixed Action Strategy

Figure 1: Strategy λ^*

2.3 Utility, Costs and Optimal Strategies

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

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

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

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

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

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

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

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

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

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

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

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

2.4 The Shannon Cost Function

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

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

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

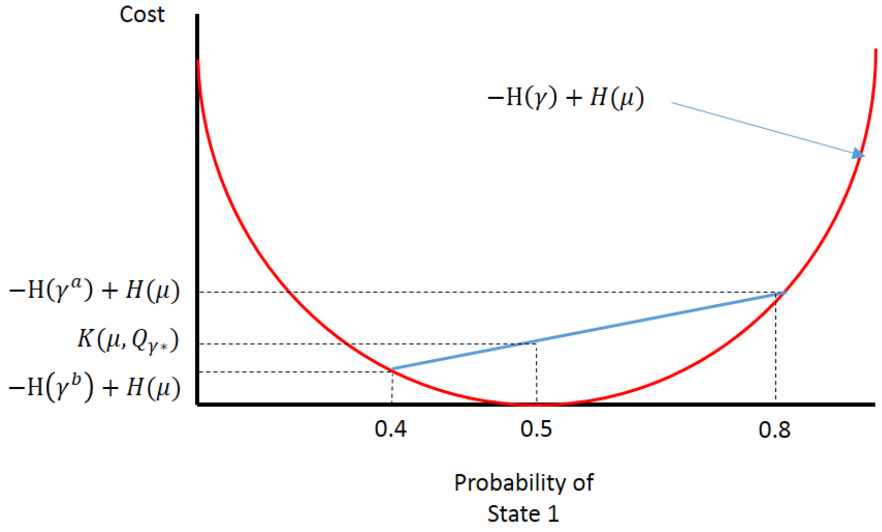

Figure 2: Cost of Strategy λ^*

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

2.5 PS Cost Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.6 UPS Cost Functions

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

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

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

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

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

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

3 PS Models, Optimal Strategies, and Lagrangians

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

3.1 Net Utility

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

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

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

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

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

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

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

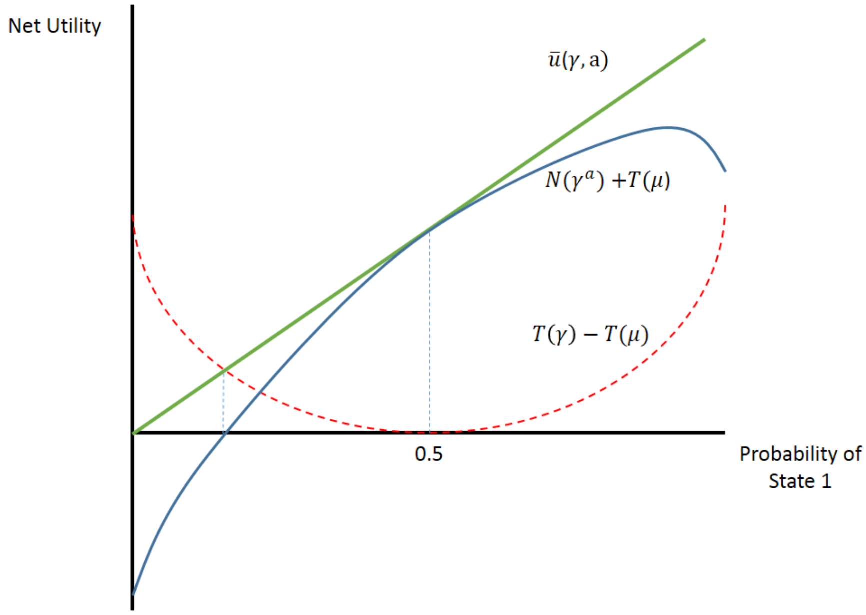

Figure 3: Net Utility of Action a.

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

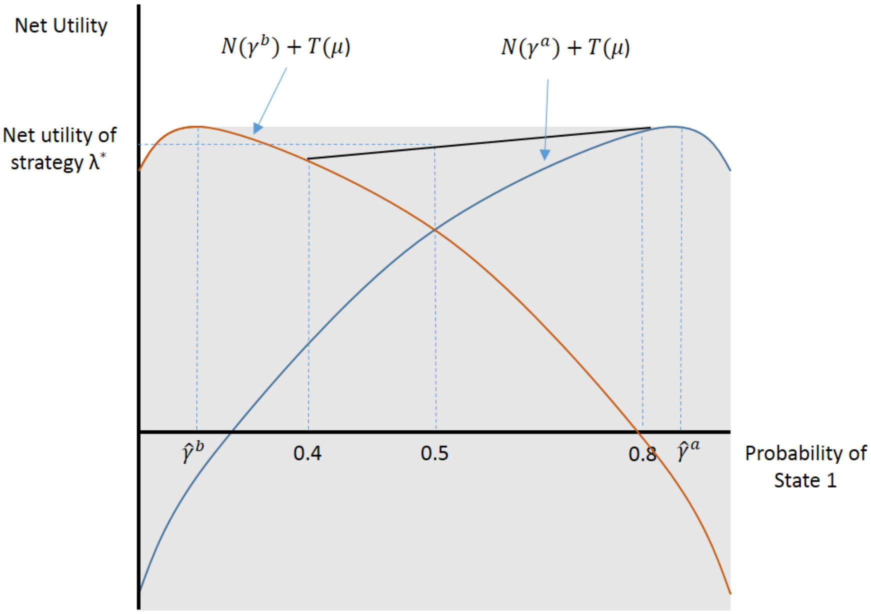

Figure 4: Net Utility of Strategy λ^*

3.2 Lagrange Multipliers and the PS model

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

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

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

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

denote the corresponding space of probability distributions:

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

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

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

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

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

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

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

4 SDSC and Representations

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

4.1 State Dependent Stochastic Choice Data

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

Our goal is to specify observable patterns in choice data that narrow down the theories of inattentive choice. Before we begin, however, we must first specify exactly what sort of data is sufficient for this task. An important first point is that standard stochastic choice data, in which one only observes the unconditional likelihood of each choice, is fundamentally inadequate for capturing attentional constraints. To see this, consider the two action decision problem illustrated in Figure 4. Note that the symmetry of the decision problem implies that the optimal strategy results in each action being chosen equally often. In the particular strategy chosen, this reflects partial information. Yet the same unconditional probabilities are also consistent with perfect information, with each action chosen precisely when it is optimal. These probabilities are also consistent with completely inattentive choice, with a fair coin flipped to decide which action is taken. Unconditional choice probabilities in no way reflect the extent to which behavioral patterns are impacted by reality. One must also know how well the action suited reality.

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

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

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

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

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

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

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

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

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

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

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

4.2 From Strategy to Data

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

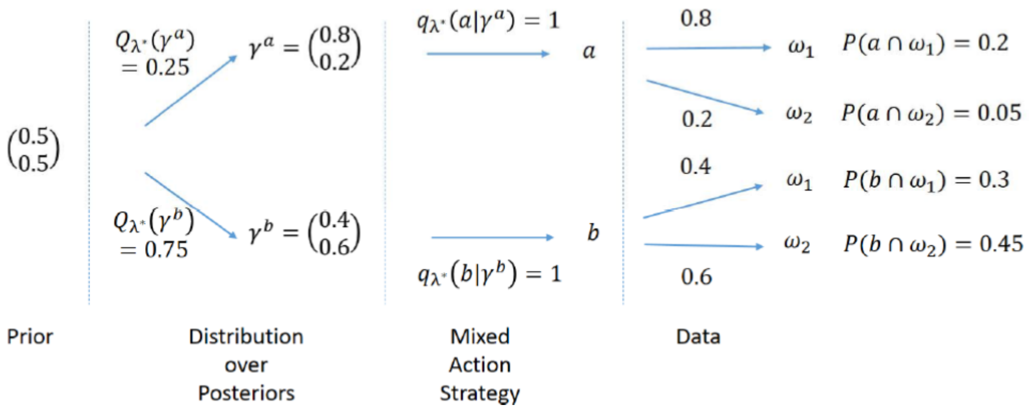
$Q_{\lambda^*}(\gamma^a) = 0.25$
$\gamma^a = \binom{0.8}{0.2}$
$q_{\lambda^*}(a|\gamma^a) = 1$
a
0.8
ω_1
$P(a \cap \omega_1) = 0.2$
0.2
ω_2
$P(a \cap \omega_2) = 0.05$
$\binom{0.5}{0.5}$
0.4
ω_1
$P(b \cap \omega_1) = 0.3$
$Q_{\lambda^*}(\gamma^b) = 0.75$
$\gamma^b = \binom{0.4}{0.6}$
b
ω_2
$P(b \cap \omega_2) = 0.45$
$q_{\lambda^*}(b|\gamma^b) = 1$
0.6
Prior
Distribution over Posteriors
Mixed Action Strategy
Data

Figure 5: Data Generated by Strategy λ^*

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

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

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

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

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

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

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

4.3 Choice Correspondence and Representations

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

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

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

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

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

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

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

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

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

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

4.4 The Revealed Strategy

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

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

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

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

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

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

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

Similarly

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

We can then calculate the revealed strategy as involving

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

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

Hence,

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

Furthermore,

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

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

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

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

5 Compression and the Shannon Model

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

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

5.1 Basic Decision Problems and Basic Forms

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

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

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

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

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

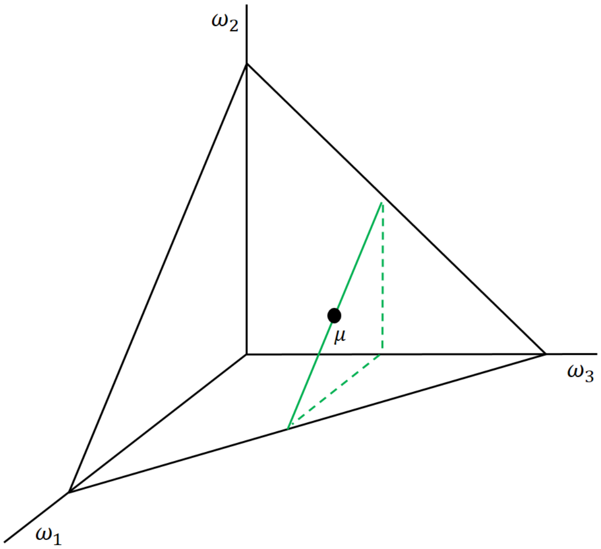

Figure 6: Basic Forms of Decision Problem (μ, A)

We now provide the general technical definitions.

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

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

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

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

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

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

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

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

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

5.2 Invariance Under Compression

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

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

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

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

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

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

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

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

5.3 Necessity

That IUC is necessary for a Shannon representation follows directly from the posterior-based characterization of the solution to the Shannon model. Caplin and Dean [2013] provide an "invariant likelihood ratio" condition for optimality. This states that $P \in C(\mu, A)$ is consistent with optimality for a cost function K_κ^S if and only if:

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

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

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

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

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

5.4 Guide to the Sufficiency Proof

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

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

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

5.4.1 Symmetry

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

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

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

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

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

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

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

5.4.2 Differentiability

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

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

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

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

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

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

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

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

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

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

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

5.4.3 Additive Separability

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

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

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

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

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

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

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

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

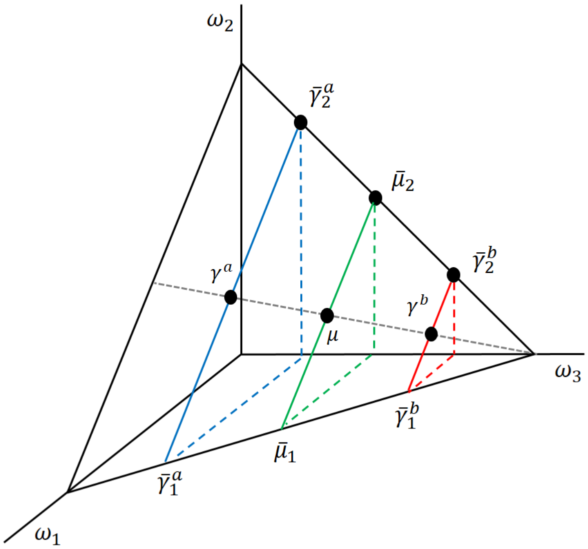

Figure 7: Implications of Compression

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

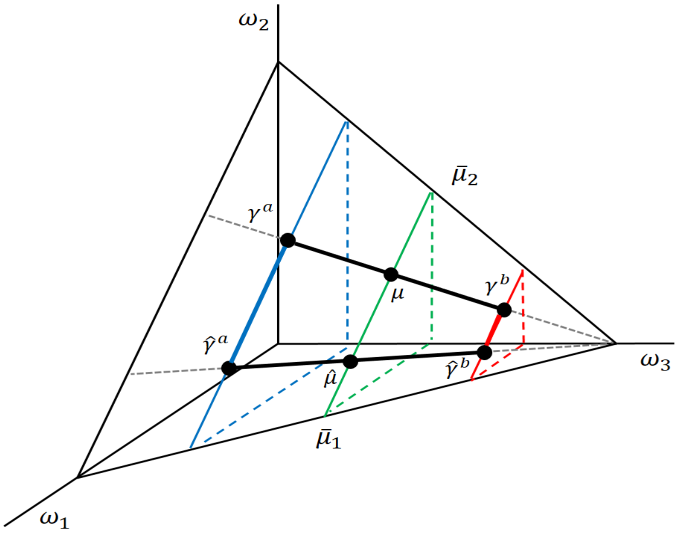

Figure 8: The Trapezoid

Consider now two priors μ and $\hat{\mu}$, each lying between $\bar{\mu}_1$ and $\bar{\mu}_2$. Figure 8 shows that the four posteriors γ^a,γ^b,$\hat{\gamma}^a$, and $\hat{\gamma}^b$ form a trapezoid. If $\tilde{T}$ is differentiable at all four points we would know that:[12]

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

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

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

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

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

5.4.4 The Second Order PDE and Shannon Entropy

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

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

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

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

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

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

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

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

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

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

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

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

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

5.4.5 IUC and Universal Domain

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

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

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

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

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

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

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

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

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

6 Existence and Recoverability

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

6.1 NIAS, NIAC, and Completeness

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

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

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

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

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

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

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

where,

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

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

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

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

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

6.2 Recoverability

The recoverability result rests on A2-A4 alone.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where the integration is over points of differentiability.

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

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

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

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

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

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

Substituting (14) into (13) yields,

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

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

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

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

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

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

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

7 PS and UPS Representations

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

7.1 Separability

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

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

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

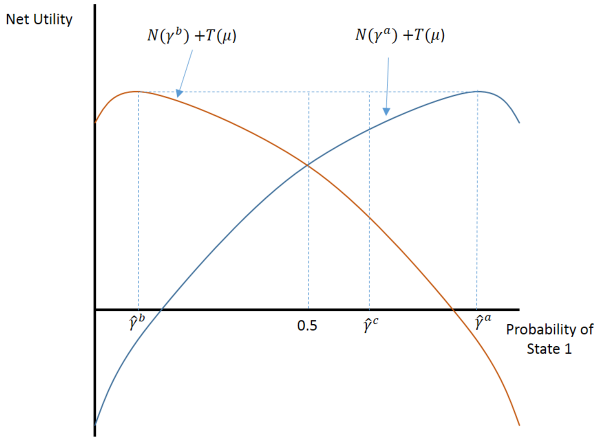

Figure 9a

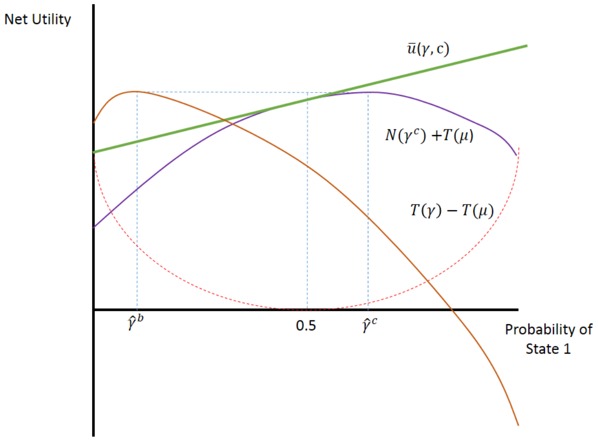

Figure 9b

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

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

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

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

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

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

7.2 Convexity Properties

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

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

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

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

7.3 From Some to All Optima

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

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

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

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

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

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

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

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

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

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

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

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

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

7.4 Existence and Simple Recovery

We summarize this discussion in the following theorem.

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

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

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

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

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

7.5 LIP and UPS Theorem

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

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

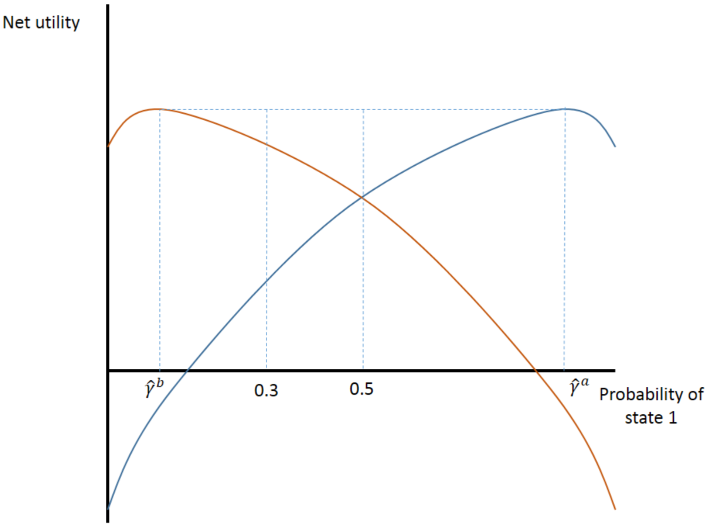

Figure 9: Locally Invariant Posteriors

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

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

The following formal definition captures both of these invariance properties.

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

$$\mathbf{Q}_{P'}(\gamma) = \sum_{\{a \in A' | \gamma_P^a = \gamma\}} \rho(a) \textit{ and:}$$

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

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

Our fourth theorem shows essentially that a data set with a PS representation has a UPS representation if and only if it satisfies LIP. There is a caveat. For necessity of LIP (Axiom A9), we insist on a link between the posteriors $\Gamma(C,\mu_1)$ and $\Gamma(C,\mu_2)$ for distinct priors. If prior μ_2 lies in the convex hull of posteriors that are revealed posteriors from prior μ_1, then these posteriors must be observed from prior μ_2 also. We present a simple example in Appendix 4 in which this does not hold. We define regular data sets as those that have this property globally.

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

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

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

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

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

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

8 Further Results

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

8.1 Choice Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

posterior separability of costs implies that,

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

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

8.2 Costly Signal Acquisition

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

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

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

8.3 Tsallis Entropy and Failures of IUC

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

ω_1 and ω_2:

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

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

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

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

$$
\mu(\omega_1)^{\sigma}+\mu(\omega_2)^{\sigma} > (\mu(\omega_1)+\mu(\omega_2))^{\sigma}\,.
$$

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

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

9 Relation to the Literature

9.1 Existing Characterizations of the Shannon Model

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

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

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

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

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

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

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

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

terms of combining two states the axiom states that:

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

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

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

9.2 Limits to the Shannon Model

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

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

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

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

that occur more frequently.

9.3 PS Models

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

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

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

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

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

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

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

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

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

9.4 Alternative Models of Limited Attention

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

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

10 Concluding Remarks

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

11 Notational Glossary

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Aubrey Clark. Contracts for information acquisition. 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Doron Ravid. Bargaining with rational inattention. 2017.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy

Andrew Caplin[*], Mark Dean[†], and John Leahy[‡]

Online Appendix

There are in total five appendices. The first establishes theorem 2, the recoverability theorem, which is a stand-alone result. The second establishes a series of lemmas on optimal strategies for PS cost functions and additional lemmas that link strategies and data. These are employed in all subsequent proofs. The third appendix proves theorem 3 which characterizes PS cost functions. The fourth establishes theorem 4 which characterizes UPS cost functions. The final appendix, which is significantly the longest, establishes theorem 1. While it is presented first in the paper, theorem 1 is proved last since it builds on all earlier results.

For brevity we have ommitted the proofs of some lemmas that we deemed to be immediate. Further details of these proofs can be obtained from the authors on request.

Appendix 1: Theorem 2

A1.1: NIAS, NIAC, and Existence

The first step in the proof of theorem 2 uses NIAS (A2) and NIAC (A3) to arrive at a cost function $K(\mu, Q)$ that rationalizes all observed data. The proof joins the methods of Caplin and Martin [2015] and Caplin and Dean [2015] for finitely many observations with the corresponding methods for unrestricted data introduced by Rochet [1987] and Rockafellar [1970]. This step involves entirely different logic than the second step and is separated out as a lemma that may itself be of independent interest. In the second step we find a condition that any rationalizing K must satisfy, and show that with A4 this is stringent enough to pin down K uniquely.

Lemma 1.1: (Existence of Rationalizing Cost Function) Given $C \in \mathcal{C}$ satisfying A2 and A3, there exists $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$.

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

[†] Department of Economics, Columbia University. Email: mark.dean@columbia.edu

[‡] Department of Economics, Universtity of Michigan and NBER. Email: jvleahy@umich.edu

Proof. The first step in the proof is to define the maximal utility that can be obtained given an arbitrary set of available actions A and posterior distribution $Q \in \mathcal{Q}(\mu)$,

$$\hat{G}(A,Q) \equiv \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\hat{u}(\gamma, A), \tag{1}$$

where $\hat{u}(\gamma, A) \equiv \max_{a \in A} \bar{u}(\gamma, a)$, as defined in equation (1) in the main document. We use a constructive procedure to find $K \in \mathcal{K}$ such that, given $(\mu, A) \in \mathcal{D}$ and $P \in C(\mu, A)$,

$$\hat{G}(A, \mathbf{Q}_P) - K(\mu, \mathbf{Q}_P) \geq \hat{G}(A,Q) - K(\mu, Q), \tag{2}$$

all $Q \in \mathcal{Q}(\mu)$. The second step shows that this function is a rationalizing cost function in the sense of this Lemma.

It simplifies the construction of the function for which inequality (2) holds to introduce the set of all pairs of decision problems and associated chosen data,

$$\mathcal{B} = \{b = (A_b, P_b) | (\mu, A_b) \in \mathcal{D} \text{ and } P_b \in C(\mu, A_b)\}.$$

Since the proof works prior by prior, we simplify from now on by fixing $\mu \in \Gamma$ in the background and treating it implicitly unless confusion would result. For example we define $\mathcal{Q}(\mu) \equiv \mathcal{Q}$, $\Omega(\mu) = \Omega$, $C(\mu, A) = C(A)$, etc., and show how to identify the corresponding section of the cost function $K(Q) \equiv K(\mu, Q)$ on $Q \in \mathcal{Q}(\mu)$.

An important construction involves associating a value with switching choice data across decision problems. Specifically, for $b, c \in \mathcal{B}$ we define $G(b,c)$ to be the maximum value associated with action set A_b and $\mathbf{Q}_{P_c}$ the revealed posterior distribution in c,

$$G(b,c) \equiv \hat{G}(A_b, \mathbf{Q}_{P_c}). \tag{3}$$

so that $G : \mathcal{B} \times \mathcal{B} \to \mathbb{R}$ due to the finiteness of all action sets. In what follows it is important to note from (1) that this connects directly to a utility calculation introduced in the definition of the NIAC axiom in equation (14) in the main document,

$$G(b,c) = \sum_{\gamma \in \Gamma(\mathbf{Q}_{Pc})} \mathbf{Q}_{P_c}(\gamma)\hat{u}(\gamma, A_b) = \hat{U}(A_b, P_c). \tag{4}$$

Rather than directly establish existence of a qualifying cost function, we use the indirect approach of Rochet [1987] and Rockafellar [1970]. What we identify is a function $t : \mathcal{B} \to \mathbb{R}$ such that $\forall\ b, b' \in \mathcal{B}$,

$$G(b,b) - t(b) \geq G(b,b') - t(b') \tag{5}$$

If such a function can indeed be identified, it allows us to define a candidate cost function $K(Q)$ on $Q \in \mathcal{Q}$ satisfying (2). Concretely,

$$K(Q) = \begin{cases} t(b) \text{ for any } b \in \mathcal{B} \text{ such that } Q = \mathbf{Q}_{P_b}; \\ \infty \text{ if } \nexists\ b \in \mathcal{B} \text{ such that } Q = \mathbf{Q}_{P_b}. \end{cases} \tag{6}$$

Note that existence of $t : \mathcal{B} \to \mathbb{R}$ satisfying (5) ensures that this is well-defined: $t(b) = t(c)$ for any $b, c \in \mathcal{B}$ with $\mathbf{Q}_{P_b}$=$\mathbf{Q}_{P_C}$. Assume to the contrary that there exists $b, c \in \mathcal{B}$ with $\mathbf{Q}_{P_b}$=$\mathbf{Q}_{Pc}$ yet with

$t(b) > t(c)$. This implies that,

$$G(b,b) - t(b) = \hat{G}(A_b, \mathbf{Q}_{P_b}) - t(b) < \hat{G}(A_b, \mathbf{Q}_{P_c}) - t(c) = G(b,c) - t(c)$$

a contradiction of condition (5). Note that K is not necessarily a qualifying cost function, as inattentive strategies are not guaranteed to be zero cost. However, following Caplin and Dean [2015] it is always possible to renormalize any cost function to make this hold, so without loss of generality we proceed assuming that K has this property.

To establish that $K \in \mathcal{K}$ as defined in (6) satisfies inequality (2), the key observation relates to inequality (5). Note first that the infinite cost of posterior distributions that are not the revealed posteriors (6) for some observed data means that the only possible reversals of inequality (2) derive from a distinct observed posterior distribution. Now consider two observations $b = (A_b, P_b), c = (A_c, P_c) \in \mathcal{B}$ and note that, by direct substitution of the definitions of G and K into (3),

$$\hat{G}(A_b, \mathbf{Q}_{P_b}) - K(\mathbf{Q}_{P_b}) = G(b,b) - t(b) \geq G(b,c) - t(c) = \hat{G}(A_b, \mathbf{Q}_{Pc}) - K(\mathbf{Q}_{Pc}),$$

establishing (2).

To construct $t : \mathcal{B} \to \mathbb{R}$ satisfying (5), we first define, for any $x, y \in \mathcal{B}$ the set $L(x,y)$ of all finite sequences starting at $x \in \mathcal{B}$ and ending at y, which we refer to as chains from x to y. Generic element $l \in L(x,y)$, comprises an ordered list of finite length $N(l)+1$ from $\mathcal{B}$, $(b_0^l, b_1^l, ...b_{N(l)}^l) \in \mathcal{B}^{N(l)+1}$ with $b_0^l = x$ and $b_{N(l)}^l = y$. We define the value of such a chain $v : L(x,y) \to \mathbb{R}$ as,

$$v(l) = \sum_{n=0}^{N(l)} \left[G(b_{n+1}^l, b_n^l) - G(b_n^l, b_n^l) \right]; \tag{7}$$

and define also the corresponding supremal value,

$$V(x,y) = \sup_{l \in L(x,y)} v(l).$$

The function $V(x,y)$ has important qualitative properties. First among these is that, due to NIAC (A2), the supremal value of all chains that have the same start and end point is zero,

$$V(x,x) = 0. \tag{8}$$

all $x \in \mathcal{B}$. To prove this, let $(A_n^l, P_n^l) = b_n^l$ be the corresponding action and choice set. Note first that the chain that goes directly from x to x gives a value of zero so that $V(x,x) \geq 0$. To show the opposite inequality, consider an arbitrary chain $(b_0^l, b_1^l, ...b_{N(l)}^l) \in L(\bar{b}_0)$ from x to x and substitute equation (4) into equation (7) to derive,

$$v(l) = \sum_{n=0}^{N(l)} \left[G(b_{n+1}^l, b_n^l)) - G(b_n^l, b_n^l) \right] = \sum_{n=0}^{N(l)} \left[\hat{U}((A_{n+1}^l, P_n^l) - \hat{U}(A_n^l, P_n^l) \right].$$

Note by construction that $A^l_0 = A^l_{N(l)} = A_x$. Given that $P^l_n \in C(A^l_n)$, NIAC (A2) directly implies that,

$$\sum_{n=0}^{N(l)} \hat{U}((A^l_{n+1}, P^l_n) \leq \sum_{n=0}^{N(l)} \hat{U}(A^l_n, P^l_n).$$

Hence indeed

$$\sum_{n=0}^{N(l)} \Big[G(b^l_{n+1}, b^l_n)) - G(b^l_n, b^l_n) \Big] \leq 0.$$

As the inequality holds for every element in the set it must also hold for the supremum, completing the proof of (8).

We now show that $V(x, y)$ for general $x, y \in \mathcal{B}$ is real-valued by providing real upper and lower bounds. The lower bound derives from the direct chain $\bar{l} = (x, y) \in \mathcal{B}^2$ for which,

$$V(x, y) \geq v(\bar{l}) = G(y, x) - G(x, x) \in \mathbb{R}. \tag{9}$$

We adapt this reasoning to provide an upper bound on $V(x, y)$. Specifically, given an arbitrary $l \in L(x, y)$, we define $l' \in L(x, x)$ as the chain $(b^l_0, b^l_1, ...b^l_{N(l)}, x) \in \mathcal{B}^{N(l))+2}$. Note that since $l' \in L(x, x)$ we know that it achieves no higher than the supremal value of zero,

$$v(l') \leq V(x, x) = 0. \tag{10}$$

Note also that the difference between $v(l)$ and $v(l')$ is defined by function G as:

$$v(l) + G(x, y) - G(y, y) = v(l') \leq 0;$$

where the final inequality follows direction from (10). Hence,

$$v(l) \leq G(y, y) - G(x, y).$$

Since this applies to arbitrary $l \in L(x, y)$, it applies also to the supremal value,

$$V(x, y) \leq G(y, y) - G(x, y). \tag{11}$$

An extension of this reasoning shows that, for any $x, y, z \in \mathcal{B}$,

$$V(x, y) - V(x, z) \geq G(y, z) - G(z, z). \tag{12}$$

To see this, note that for, given an arbitrary $\bar{l} \in L(x, z)$, we can define the new chain $\bar{l}' \in L(x, y)$ as the chain $(b^l_0, b^l_1, ...b^l_{N(l)}, y) \in \mathcal{B}^{N(l))+2}$. Note that the difference between $v(\bar{l})$ and $v(\bar{l}')$ is defined by function G as:

$$v(\bar{l}') = v(\bar{l}) + G(z, y) - G(z, z); \tag{13}$$

Taking the supremum on $\bar{l} \in L(x, z)$ on the RHS we arrive at,

$$\sup_{\bar{l} \in L(x,z)} v(\bar{l}) + G(z, y) - G(z, z) = V(x, z) + G(z, y) - G(z, z).$$

Applying equation (13) to a sequence of strategies $\bar{l}(n) \in L(x, z)$ converging to the supremum, we note that the corresponding sequence $\bar{l}'(n) \in L(x, y)$ can achieve no more than the corresponding supremal value, $V(x,y)$. Hence,

$$V(x,y) \geq V(x,z) + G(y,z) - G(z,z),$$

establishing (12).

Finally, we are in position to define the sought for function $t : \mathcal{B} \to \mathbb{R}$ such that (5) holds. Specifically we fix a reference element $z \in \mathcal{B}$ and define,

$$t(b) \equiv G(b,b) - V(z,b), \tag{14}$$

on $b \in \mathcal{B}$. To validate (5), we set $x = b' \in \mathcal{B}$ and $y = b$ and note by direct substitution of (14) that,

$$t(b') - t(b) = G(b',b') - G(b,b) + V(z,b) - V(z,b')$$

We now apply inequality (12) to replace the value functions on the RHS,

$$\begin{aligned} t(b') - t(b) &\geq G(b',b') - G(b,b) + G(b,b') - G(b',b') \\ &= G(b,b') - G(b,b), \end{aligned}$$

establishing (5).

We have now completed construction of a qualifying cost function $K \in \mathcal{K}$ satisfying (2). Expanding the inequality out, note that, given $(\mu, A) \in \mathcal{D}$ and $P \in C(\mu, A)$ and associated $\mathbf{Q}_P$,

$$\sum_{\gamma \in \Gamma(\mathbf{Q}_P)} \mathbf{Q}_P(\gamma)\hat{u}(\gamma, A) - K(\mathbf{Q}_P) \geq \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\hat{u}(\gamma, A) - K(Q),$$

all $Q \in \mathcal{Q}$.

To complete the proof of Lemma 1.1 we need to show that this cost function represents the data in the sense of the Lemma: given $(\mu, A) \in \mathcal{D}$ and $P \in C(\mu, A)$ we can find $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}(\mu, A|K)$ such that $P = \mathbf{P}_\lambda$. The proof is constructive. Our candidate strategy is the revealed attention strategy $\boldsymbol{\lambda}(P) = (\mathbf{Q}_P, \mathbf{q}_P)$. Since (2) holds, it suffices first to show that $(\mathbf{Q}_P, \mathbf{q}_P)$ obtains $\hat{G}(A, \mathbf{Q}_P)$, and then that it generates the data, $P = \mathbf{P}_{\boldsymbol{\lambda}(P)}$.

With regard to the first step, note first that $(\mathbf{Q}_P, \mathbf{q}_P) \in \Lambda(\mu, A)$, since, given $\omega \in \Omega(\mu)$,

$$\begin{aligned} \sum_{\{\gamma \in \Gamma(P)\}} \mathbf{Q}_P(\gamma)\gamma(\omega) &= \sum_{\{\gamma \in \cup_{a \in \mathcal{A}(P)} \bar{\gamma}_P^a\}} \sum_{\{a \in \mathcal{A}(P) | \bar{\gamma}_P^a = \gamma\}} P(a)\gamma(\omega) = \sum_{a \in \mathcal{A}(P)} P(a)\bar{\gamma}_P^a(\omega) \\ &= \sum_{a \in \mathcal{A}(P)\}} \mu(\omega)P(a|\omega) = \mu(\omega) \sum_{a \in \mathcal{A}(P)\}} P(a|\omega) = \mu(\omega). \end{aligned}$$

Note further that for any $a \in A$ with $\mathbf{q}_P(a|\gamma) > 0$ for some $\gamma = \bar{\gamma}_P^a \in \Gamma(\mathbf{Q}_P)$, we know that $a \in \mathcal{A}(P)$. Hence by NIAS (A2) we know that,

$$\sum_{\omega \in \Omega(\mu)} \bar{\gamma}_P^a(\omega)u(a,\omega) = \hat{u}(\gamma, A).$$

Hence,

$$\sum_{\gamma\in\Gamma(\mathbf{Q}_P)} \mathbf{Q}_P(\gamma) \sum_{a\in A} \mathbf{q}_P(a|\gamma) \left(\sum_{\omega\in\Omega} \gamma(\omega) u(a,\omega) \right) = \sum_{\gamma\in\Gamma(\mathbf{Q}_P)} \mathbf{Q}_P(\gamma) \hat{u}(\gamma, A)$$
$$= \hat{G}(A, \mathbf{Q}_P).$$

Hence indeed $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}(\mu, A|K)$.

It remains only to show that $P = \mathbf{P}_{\boldsymbol{\lambda}(P)}$, which is established in Lemma 2.13 in Appendix 2.[1] ■

A1.2: Completeness and Uniqueness

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

Proof. By Lemma 1.1, we know that with A2 and A3 there exists $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$. With the addition of Completeness, A4, we will show that any function with this property is recoverable directly from the data, hence unique.

We show how to establish the costs associated with some arbitrary $\bar{Q} \in Q^C(\mu)$. The recovery method involves constructing a parametrized path of distributions over posteriors $\bar{Q}_t \in \mathcal{Q}^C(\mu)$ for $t \in [0, 1]$, all of which are also observed in the data. As a first step in the construction, we index by n for $1 \leq n \leq N = |\Gamma(\bar{Q})|$ the posteriors $\bar{\gamma}^n \in \Gamma(\bar{Q})$. For each n we define a linear path from prior to posterior,

$$\bar{\gamma}^n_t = t\bar{\gamma}^n + (1-t)\mu, \tag{15}$$

on $t \in [0, 1]$. By construction $\bar{\gamma}^n_0 = \mu$ and $\bar{\gamma}^n_1 = \bar{\gamma}^n$. We define $\bar{\Gamma}_t = \cup_{n=1}^N \bar{\gamma}^n_t$ as the corresponding set of such posteriors. Finally we define the parametrized path of posterior distribution of interest $\bar{Q}_t \in \mathcal{Q}(\mu)$ by setting,

$$\bar{Q}_t(\bar{\gamma}^n_t) = \bar{Q}(\bar{\gamma}^n) \equiv \bar{Q}^n \tag{16}$$

for all $t \in [0, 1]$ and $1 \leq n \leq N$.

We wish to show that $\bar{Q}_t \in \mathcal{Q}^C(\mu)$ all $t \in [0, 1]$. By construction $\bar{Q}_1 = \bar{Q} \in Q^C(\mu)$. Also by construction each posterior distribution satisfies the Bayesian constraint,

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

Finally, note that for all $t \in [0, 1)$, $\Gamma(\bar{Q}_t) \subset \bar{\Gamma}(\mu)$, so that by Completeness (A4), $\bar{Q}_t \in Q^C(\mu)$, as required.

Given $K \in \mathcal{K}$ and our parameterized posterior distributions Q_t, we define corresponding cost functions,

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

[1] The proof of Lemma 2.13 is independent of the proof of Lemma 1.1, so no circularity arises.

By definition the posterior distribution at $t = 0$ is inattentive and that at $t = 1$ generates $\bar{Q}$. Hence, by normalization,

$$\bar{K}(0) = 0 \text{ and } \bar{K}(1) = K(\mu, \bar{Q}).$$

Beyond this, key observations are that $\bar{K}(t)$ is continuous and convex in t. In proving this it is convenient to simplify notation. As in the last proposition, since the proof works prior by prior, it can be suppressed in all ensuing statements unless this would cause confusion.

Since $\bar{Q}_t \in \mathcal{Q}^C$, there is, for each $t \in [0, 1]$, an action set that has the given posterior distribution as its revealed posterior distribution.

Technically, Completeness (A4) implies that given $t \in [0, 1]$ there exists $(\mu, \bar{A}_t) \in \mathcal{D}$ and $\bar{P}_t \in C$ $(\mu, \bar{A}_t)$ such that $\mathbf{Q}_{\bar{P}_t} = \bar{Q}_t$, and we introduce just such a path through action sets. We now compute expected utility that is derived by setting each action as in set $\bar{A}_t$ and using it at the posteriors corresponding to all different values $s \in [0, 1]$. Note first that, while there may be several actions in principle in each set $\bar{A}_s$ that have the same revealed posterior, they all must have the same expected utility. For any $\bar{\gamma}^n_s \in \Gamma(\bar{Q}_s)$ and any $a, a' \in \bar{A}_s$ that are possibly chosen, $\min\{q(a|\bar{\gamma}^n_s), q(a'|\bar{\gamma}^n_s)\} > 0$, NIAS (A2) implies that the corresponding expected utility is equal,

$$\bar{u}(a, \bar{\gamma}^n_s) = \sum_{\omega \in \Omega} u(a, \omega)\bar{\gamma}^n_s(\omega) = \sum_{\omega \in \Omega} u(a', \omega)\bar{\gamma}^n_s(\omega) = \bar{u}(a', \bar{\gamma}^n_s). \tag{17}$$

Hence with regard to the computation of expected utility we can WLOG select one chosen action and designate it as the unique chosen action $\bar{a}^n_t \in \bar{A}_t$ for computing maximized expected utility.

The specific path of expected utility that we compute involves fixing $t \in [0, 1]$ and using the action $\bar{a}^n_t \in \bar{A}_t$ at all posteriors γ^n_s for $s \in [0, 1]$. We compute the corresponding expected utility for all pairings of parameterized action sets $\bar{A}_t$ and posterior distributions $\bar{Q}_s$ on the defined path,

$$H(t, s) = \sum_n \bar{Q}_s(\bar{\gamma}^n_s) \left(\sum_{\omega \in \Omega} \bar{\gamma}^n_s(\omega) u(\bar{a}^n_t, \omega) \right) \equiv \sum_n \bar{Q}^n \bar{u}(\bar{a}^n_t, \bar{\gamma}^n_s). \tag{18}$$

A simple observation is that $H(t, s)$ is linear on $s \in [0, 1]$. Given $\alpha, s_1, s_2 \in [0, t]$,

$$\alpha H(t, s_1) + (1 - \alpha) H(t, s_2) = H(t, \alpha s_1 + (1 - \alpha) s_2). \tag{19}$$

This follows directly from (18) since,

$$\begin{aligned} H(t, \alpha s_1 + (1 - \alpha) s_2) &= \sum_n \bar{Q}^n \bar{u}(\bar{a}^n_t, \bar{\gamma}^n_{\alpha s_1 + (1-\alpha) s_2}) \\ &= \sum_n \bar{Q}^n \left[\alpha \bar{u}(\bar{a}^n_t, \bar{\gamma}^n_{s_1}) + (1 - \alpha) \bar{u}(\bar{a}^n_t, \bar{\gamma}^n_{s_2}) \right] = \alpha H(t, s_1) + (1 - \alpha) H(t, s_2) \end{aligned}$$

As a linear function, $H(t, s)$ is differentiable in $s \in [0, 1]$. The corresponding partial derivative is of interest since we will consider a related optimization problem,

$$H_2(t, s) = \sum_n \bar{Q}^n \left(\sum_{\omega \in \Omega} \frac{\partial \bar{\gamma}^n_s(\omega)}{\partial s} u(\bar{a}^n_t, \omega) \right)$$

Substituting in for $\bar{\gamma}_s^n(\omega)$ from definition (15) that

$$\frac{\partial \bar{\gamma}_s^n(\omega)}{\partial s} = \left[\bar{\gamma}^n(\omega) - \mu(\omega)\right].$$

Hence,

$$H_2(t,s) = \sum_n \bar{Q}^n \left(\sum_{\omega \in \Omega} \left[\bar{\gamma}^n(\omega) - \mu(\omega)\right] u(\bar{a}_t^n, \omega) \right).$$

Given its importance in what follows it is valuable to simplify the notation for the dot product between the state specific vector of changes from posterior to prior and the corresponding state specific vector of utilities.

$$[\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n) \equiv \sum_{\omega \in \Omega} \left[\bar{\gamma}^n(\omega) - \mu(\omega)\right] u(\bar{a}_t^n, \omega). \tag{20}$$

The optimization problem that we study relies on that observation that, for all $t \in [0,1]$,

$$H(t,t) - \bar{K}(t) \geq H(t,s) - \bar{K}(s), \tag{21}$$

all $s \in [0,1]$. This follows directly from inequality (19), since we know that $(\mu, \bar{A}_t) \in \mathcal{D}$ and $\bar{P}_t \in C(\mu, \bar{A}_t)$, so that the corresponding revealed posterior $\mathbf{Q}_{P_t} = \bar{Q}_t$ maximizes expected utility net of costs of attention costs and

$$\begin{aligned} H(t,t) - \bar{K}(t) &= \hat{G}(\bar{A}_t, \bar{Q}_t) - K(\bar{Q}_t) \geq \hat{G}(\bar{A}_t, \bar{Q}_s) - K(\bar{Q}_s) \\ &\geq H(t,s) - \bar{K}(s). \end{aligned}$$

In essence the left-hand side expression is the optimized expected utility at t in the observed data under the CIR, while the RHS represents expected utility for a policy that would be feasible in this set of using posterior distribution Q_s and at each γ_s^n picking $\bar{a}_t^n \in \bar{A}_t$, which may or may not in fact be optimal.

We use the function H to prove continuity and convexity of $\bar{K}(t)$. With regard to convexity, consider $t_1 \neq t_2 \in [0,1]$ and their average

$$\bar{t} = \frac{t_1 + t_2}{2}.$$

Direct application of inequality (21) establishes that,

$$H(\bar{t}, \bar{t}) - K(\bar{t}) \geq 0.5 \left[H(\bar{t}, t_1) - \bar{K}(t_1) + H(\bar{t}, t_2) - \bar{K}(t_2) \right] \tag{22}$$

From (19) it is clear that,

$$0.5 \left[H(\bar{t}, t_1) + H(\bar{t}, t_2) \right] = H(\bar{t}, \bar{t}).$$

Substitution in (22) yields,

$$\bar{K}(\bar{t}) \leq 0.5 \left[\bar{K}(t_1) + \bar{K}(t_2) \right]$$

implying that $\bar{K}(t)$ is convex as claimed.

Given that $\bar{K}(t)$ is convex, it is continuous on its interior, $t \in (0,1)$. Moreover the only possible discontinuities at the boundary point involve an increase in costs,

$$\bar{K}(0) > \lim_{t \downarrow 0} \bar{K}(t) \text{ or } \bar{K}(1) > \lim_{t \uparrow 1} \bar{K}(t). \tag{23}$$

To see that these cannot hold, apply (21) at the corresponding end-point,

$$\begin{array}{rcl} H(0,0) - \bar{K}(0) & \geq & H(0,\epsilon) - \bar{K}(\epsilon); \\ H(1,1) - \bar{K}(1) & \geq & H(1,1-\epsilon) - \bar{K}(1-\epsilon) \end{array}$$

all $\epsilon > 0$. (19) implies continuity of $H(0,\epsilon)$ and $H(0,\epsilon)$ in ϵ, we conclude therefore that $\bar{K}(0) \leq \lim_{\epsilon \downarrow 0} \bar{K}(\epsilon)$ and $\bar{K}(1) \leq \lim_{\epsilon \uparrow 1} \bar{K}(\epsilon)$, both of which directly contradict (23).

Given that $\bar{K}(t)$ is convex and continuous, it can have at most a countable number of non-differentiable points (Rockafellar [1970]) and hence is integrable. By the fundamental theorem of calculus $\bar{K}(t)$ can be reconstructed from its derivative, $\bar{K}'(t)$, which is defined except on a set of measure zero. Hence since $\bar{K}(0) = 0$, and we can recover its final value focusing only on points of differentiability as,

$$K(\bar{Q}) = \bar{K}(1) = \int_0^1 \bar{K}'(t) dt, \tag{24}$$

We now show how to characterize the derivative $\bar{K}'(t)$ from the path of utilities identified above, noting that $H(t,s)$ satisfies (19) and is hence everywhere differentiable in s. Consider a point $t \in (0,1)$ of differentiability of $\bar{K}$ and problem $(\mu, \bar{A}_t)$ for which t therefore maximizes,

$$\max_{s \in [0,1]} H(t,s) - \bar{K}(s) = \sum_n \bar{Q}^n \bar{u}(\bar{a}_t^n, \gamma_s^n) - \bar{K}(s). \tag{25}$$

Hence the corresponding first order condition for maximizing (25) at $s = t$ is,

$$\bar{K}'(t) = H_2(t,t) = \sum_n \bar{Q}^n \frac{\partial \bar{u}(\bar{a}_t^n, \gamma_s^n)}{\partial s}. \tag{26}$$

Substituting the definition in (17),

$$\bar{u}(\bar{a}_t^n, \gamma_s^n) = \sum_{\omega \in \Omega(\mu)} \bar{\gamma}_s^n(\omega) u(\bar{a}_t^n, \omega) = \sum_{\omega \in \Omega(\mu)} [s\bar{\gamma}^n(\omega) + (1-s)\mu(\omega)] u(\bar{a}_t^n, \omega).$$

Hence the chain rule yields,

$$\frac{\partial \bar{u}(\bar{a}_t^n, \gamma_s^n)}{\partial s} = \sum_{\omega \in \Omega} [\bar{\gamma}^n(\omega) - \mu(\omega)] \, u_2(\bar{a}_t^n, \omega) = [\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n),$$

where the last equation uses the simpler notation for the dot product introduced in (20).

The corresponding first order condition for maximizing (25) at $s = t$ is,

$$\bar{K}'(t) = H_2(t,t) = \sum_n \bar{Q}^n \left([\bar{\gamma}^n - \mu] \cdot u(\bar{a}_t^n) \right). \tag{27}$$

Combining (27) and(24),

$$\begin{aligned} K(\mu,\bar{Q}) &= \int_0^1 \sum_n \bar{Q}^n \left\{[\bar{\gamma}^n - \mu]\cdot u(\bar{a}_t^n)\right\} dt \\ &= \sum_n \bar{Q}^n \left[\bar{\gamma}^n - \mu\right] \cdot \int_0^1 u(\bar{a}_t^n) dt, \end{aligned}$$

since the dot product survives under integration. This completes the constructive procedure for computing the cost function. ■

We state the form of this computation as a corollary since it is the jumping off point for the proof of theorem 3.

Corollary 1 *Given $C \in \mathcal{C}$ satisfying A2-A4, the unique function $K \in \mathcal{K}$ such that $C(\mu,A) \subset \hat{P}(\mu,A|K)$ all $(\mu,A) \in \mathcal{D}$ can be computed for each $(\mu,\bar{Q}) \in \mathcal{F}$ with $\bar{Q} \in \mathcal{Q}^C(\mu)$ by enumerating the support $\Gamma(\bar{Q}) = \{\bar{\gamma}^n | 1 \leq n \leq N\}$ and computing,*

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

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

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

Appendix 2: The PS Model and Convex Analysis

For the analysis of the PS model, we fix a specific prior $\bar{\mu} \in \Gamma$. Costs are then defined by a strictly convex function $T_{\bar{\mu}} : \Gamma(\bar{\mu}) \to \bar{\mathbb{R}}$, real valued on $\tilde{\Gamma}(\bar{\mu})$, such that, given $Q \in \mathcal{Q}(\bar{\mu})$,

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

We define a "$\bar{\mu}$-based" net utility function for strategies $\lambda \in \Lambda(\mu,A)$ for $\mu \in \Delta(\Gamma(\bar{\mu}))$ using the costs associated with $\bar{\mu}$,

$$N_{\bar{\mu}}(\mu,\lambda) \equiv U(\lambda) - \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) T_{\bar{\mu}}(\gamma) = V(\bar{\mu},\lambda) - T_{\bar{\mu}}(\bar{\mu}). \tag{28}$$

We introduce this function because it is useful in the course of the proofs below to assign net utilities to strategies that are not directly feasible for the initially given prior and therefore whose costs cannot be evaluated directly. By the time we use the results that this "$\bar{\mu}$-based" fictional cost makes simpler to state, we reweight the other strategies to arrive back at $\bar{\mu}$ itself.

Another key function is the "net utility" of choosing action $a \in A$ given posterior $\gamma \in \Gamma(\bar{\mu})$,

$$N_{\bar{\mu}}^a(\gamma) \equiv \sum_{\omega \in \Gamma(\bar{\mu})} u(a,\omega)\gamma(\omega) - T_{\bar{\mu}}(\gamma) = \bar{u}(\gamma,a) - T_{\bar{\mu}}(\gamma). \tag{29}$$

We introduce also a mixture operation.

Definition 1 *Given any finite set of strategies, $\{\lambda(l) = (Q_l, q_l) \in \Lambda(\mu(l)\}_{1 \leq l \leq L}$, and strictly positive probability weights $\{\alpha(l)\}_{1 \leq l \leq L}$, define the corresponding* ***mixture strategy*** $\lambda(\alpha) = (Q_\alpha, q_\alpha) \in \Lambda$ *by:*

$$\begin{aligned} Q_\alpha(\gamma) &= \sum_l \alpha(l) Q_l(\gamma) \text{ all } \gamma \in \Gamma(Q_\alpha); \\ q_\alpha(a|\gamma) &= \frac{\sum_l \alpha(l) q_l(a|\gamma) Q_l(\gamma)}{Q_\alpha(\gamma)} \text{all } \gamma \in \Gamma(Q_\alpha),\ a \in \mathcal{A}(\lambda(\alpha)); \end{aligned}$$

where $\Gamma(Q_\alpha) = \cup_l \Gamma(Q_l)$ and $\mathcal{A}(\lambda(\alpha)) = \cup_l \mathcal{A}(\lambda(l))$. Define $\mu(\alpha) = \sum_l \alpha(l)\mu(l)$ as the corresponding weighted average of priors.

A2.1: Linearity and Uniqueness

Lemma 2.1 (Linearity under Mixing): Given $K \in \mathcal{K}^{PS}$, $(\bar{\mu}, A) \in \mathcal{D}$, and, for $1 \leq l \leq L$, strategies $\lambda(l) = (Q_l, q_l) \in \Lambda(\mu(l), A)$ and probability weights $\alpha(l)$,

$$N_{\bar{\mu}}(\mu(\alpha), \lambda(\alpha)) = \sum_l \alpha(l) N_{\bar{\mu}}(\mu(l), \lambda(l)).$$

Proof. Immediate. ■

Lemma 2.2: (Mixing and Optimality) Given $K \in \mathcal{K}^{PS}$, $(\bar{\mu}, A) \in \mathcal{D}$, and strategies $\lambda \in \Lambda(\bar{\mu}, A)$ and $\lambda(l) = (Q_l, q_l) \in \Lambda(\bar{\mu}, A)$ for $1 \leq l \leq L$, together with strictly positive probability $\alpha(l) > 0$ such that $\lambda = \sum_{l=1}^{L} \alpha(l)\lambda(l)$,

$$\lambda \in \hat{\Lambda}(\bar{\mu}, A|K) \Longleftrightarrow \lambda(l) \in \hat{\Lambda}(\bar{\mu}, A|K) \text{ all } l.$$

Proof. Immediate. ■

Lemma 2.3 (Unique Posterior Lemma): Given $K \in \mathcal{K}^{PS}$, $(\bar{\mu}, A) \in \mathcal{D}$, $\lambda \in \hat{\Lambda}(\bar{\mu}, A|K)$, and $a \in \mathcal{A}(\lambda)$, there exists a unique posterior $\gamma \in \Gamma(Q_\lambda)$ such that $q_\lambda(a|\gamma) > 0$. We denote this posterior γ_λ^a.

Proof. Immediate. ■

Lemma 2.4 (Unique Optimal Strategy): Given $K \in \mathcal{K}^{PS}$, $(\bar{\mu}, A) \in \mathcal{D}$, and $\lambda \in \hat{\Lambda}(\bar{\mu}, A|K)$ with $\Gamma(Q_\lambda)$ linearly independent,

$$|\Gamma(Q_\lambda)| = |A| \Longrightarrow \lambda = \hat{\Lambda}(\bar{\mu}, A|K).$$

Proof. The first observation is that, with $\Gamma(Q_\lambda)$ linearly independent, no strict subset of the posteriors is even feasible. To see this, consider $\lambda' \in \Lambda(\bar{\mu}, A)$ with $\Gamma(Q_{\lambda'}) \subset \Gamma(Q_\lambda)$. Note that for feasibility,

$$\Sigma_{\gamma \in \Gamma(Q_\lambda)} \gamma Q_\lambda(\gamma) = \Sigma_{\gamma \in \Gamma(Q_{\lambda'})} \gamma Q_{\lambda'}(\gamma) = \mu,$$

Subtraction yields,

$$\Sigma_{\gamma \in \Gamma(Q_\lambda)} \gamma \left[Q_\lambda(\gamma) - Q_{\lambda'}(\gamma)\right] = 0,$$

whereupon linear independence implies that $Q_\lambda(\gamma) = Q_{\lambda'}(\gamma)$ all $\gamma \in \Gamma(Q_\lambda)$.

By Lemma 2.3, note that no action is chosen at more than one posterior in an optimal strategy. Hence with $|\Gamma(\lambda)| = |A|$ this means that each action is chosen at only one posterior, at which it is chosen deterministically. Note that if there were two distinct deterministic strategies using the same set of posteriors, mixing them would be optimal by Lemma 2.2, yet would involve the same action at two distinct posteriors, which is inconsistent with Lemma 2.3. Hence changing action choices in any way at the given posteriors must strictly lower the payoff. Hence there is no alternative optimal strategy that involves retaining posterior set $\Gamma(Q_\lambda)$ yet changing action choices.

The final possibility for generating multiplicity is if there exists some $\lambda' \in \hat{\Lambda}(\bar{\mu}, A|K)$ with some posterior $\gamma' \in \Gamma(Q_{\lambda'})$ that is not in $\Gamma(Q_\lambda)$. Since $\mathcal{A}(\lambda) = A$, we can identify $a' \in A \cap \mathcal{A}(\lambda')$ with $q(a'|\gamma') > 0$. By Lemma 2.2, the strategy $\frac{\lambda}{2} + \frac{\lambda'}{2}$ is also optimal. This identifies a supposedly optimal attention strategy in which a' is chosen at two distinct posteriors, contradicting Lemma 2.3 and completing the proof. ■

A2.2: Lagrangian Analysis

Our next series of results relate to the lower epigraph of $N_{\bar{\mu}}(\mu, \lambda)$. To study this, we reduce the dimension of the state space by defining $J = |\Omega(\bar{\mu})|$, correspondingly labeling states, and letting $\Gamma^{J-1}(\bar{\mu})$ denote the space of distributions of interest,

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

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

Lemma 2.5 (Convexity): The lower epigraph of $N_{\bar{\mu}}(\mu, \lambda)$,

$$\mathcal{E}(\bar{\mu}, A) \equiv \left\{ (y, \mu) \in \mathbb{R} \times \Gamma^{J-1}(\bar{\mu}) \; | y \leq N_{\bar{\mu}}(\mu, \lambda) \text{ some } \lambda \in \Lambda(\mu, A) \right\},$$

is a convex set.

Proof. Immediate. ■

Lemma 2.6 (Lagrangean): Given $K \in \mathcal{K}^{PS}$ and $(\bar{\mu}, A) \in \mathcal{D}$, $\lambda \in \hat{\Lambda}(\bar{\mu}, A|K)$ if and only if

$\exists \theta \in \mathbb{R}^{J-1}$ s.t.,

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

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

Proof. If $\lambda \in \hat{\Lambda}(\bar{\mu}, A)$, we know that $(N_{\bar{\mu}}(\bar{\mu}, \lambda), \bar{\mu})$ is an upper boundary point of the convex set $\mathcal{E}(\bar{\mu}, A)$. Hence there exists a supporting hyperplane $\theta(j)$ for $0 \leq j \leq J-1$ with $\theta(j) \neq 0$ for some j such that,

$$\theta(0)y - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq \theta(0) N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j),$$

all $(y, \gamma) \in \mathcal{E}(\bar{\mu}, A)$.

We show now that $\theta(0) > 0$. Suppose to the contrary that $\theta(0) < 0$. In this case we can renormalize to $\theta(0) = -1$ to conclude that,

$$-y - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq -N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j), \tag{30}$$

which is a clear contradiction, since the left hand side is unbounded as we lower y arbitrarily. Finally we show that $\theta(0) \neq 0$. Note that by definition $\bar{\mu}(j) > 0$ all j,

$$\min\left\{ \min_{1 \leq j \leq J-1} \{\bar{\mu}(j)\}, 1 - \sum_{j=1}^{J-1} \bar{\mu}(j) \right\} > 0.$$

Finally suppose that $\theta(0) = 0$, so that,

$$\sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j) \leq \sum_{j=1}^{J-1} \theta(j)\gamma(j).$$

all $\gamma \in \Gamma(\bar{\mu})$. To minimize the expression on the RHS, one can set $\gamma(\bar{j}) = 1$, where the index $\bar{j}$ is chosen so that,

$$\theta(\bar{j}) = \min_{1 \leq j \leq J-1} \{\theta(j)\}.$$

Hence the inequality can be valid only if $\theta(j) = \theta(\bar{j}) = \bar{\theta}$ for all j. Hence what is required is,

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

all $\gamma \in \Gamma(\bar{\mu})$. Since $0 < \sum_{j=1}^{J-1} \bar{\mu}(j) < 1$ while $\sum_{j=1}^{J-1} \gamma(j)$ has a range that includes 0 and 1, this is impossible unless $\bar{\theta} = 0$, a contradiction to the non-zero separating plane.

Given that $\theta(0) > 0$, we renormalize to $\theta(0) = 1$ and conclude that,

$$y - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j).$$

Given $\gamma' \in \Gamma$ and $a' \in A$, we know that $(N_{\bar{\mu}}^{a'}(\gamma'), \gamma') \in \mathcal{E}(\bar{\mu}, A)$, so that,

$$N_{\bar{\mu}}^{a'}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) \leq N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j). \tag{31}$$

Now consider any decomposition of the optimal strategy, $\lambda = \sum_{l=1}^{L} \alpha(l)\lambda(l)$ for a finite set $\{\lambda(l) = (Q_l, q_l) \in \Lambda(\mu(l), A)\}_{1 \leq l \leq L}$, and probability weights $\{\alpha(l)\}_{1 \leq l \leq L}$. Lemma 2.1 implies,

$$\begin{aligned} N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j) &= \sum_{l} \alpha(l) N_{\bar{\mu}}(\mu(l), \lambda(l)) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j) \\ &= \sum_{l} \alpha(l) \left[N_{\bar{\mu}}(\mu(l), \lambda(l)) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) \right], \end{aligned}$$

since $\bar{\mu}(j) = \sum_l \alpha(l)\mu_l(j)$. By inequality (31) none of the terms in the weighted average on the RHS can be higher than the LHS since $(N_{\bar{\mu}}(\mu(l), \lambda(l)) \in \mathcal{E}(\bar{\mu}, A)$. Hence they are all equal to it,

$$N_{\bar{\mu}}(\mu_l, \lambda(l)) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) = N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j). \tag{32}$$

We now provide a simple decomposition of strategy λ using only inattentive strategies. We index possible posteriors $\gamma \in \Gamma(Q_\lambda)$ as γ^l and define the inattentive strategy $\lambda(l) \in I(\gamma^l)$ by setting $q_{\lambda(l)}(a|\gamma^l) = q_\lambda(a|\gamma^l)$. Setting the weights as $\alpha(l) = Q_\lambda(\gamma^l)$ accomplishes this decomposition. The special feature of such inattentive strategies is that,

$$N_{\bar{\mu}}(\mu_l, \lambda(l)) = \sum_{a \in A} q_\lambda(a|\gamma^l) \left[\bar{u}(a, \gamma^l) - T_{\bar{\mu}}(\gamma^l) \right] = \sum_{a \in A} q_\lambda(a|\gamma^l) N_{\bar{\mu}}^{a}(\gamma^l).$$

Hence,

$$\begin{aligned} N_{\bar{\mu}}(\mu_l, \lambda(l)) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) &= \sum_{a \in A} q_\lambda(a|\gamma^l) N_{\bar{\mu}}^{a}(\gamma^l) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) \\ &= \sum_{a \in A} q_\lambda(a|\gamma^l) \left[N_{\bar{\mu}}^{a}(\gamma^l) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) \right] \end{aligned}$$

where the first line follows directly, the second follows because $\sum\limits_{a \in A} q_\lambda(a|\gamma^l) = 1$. Hence equation (32) implies,

$$\sum_{a \in A} q_\lambda(a|\gamma^l) \left[N^a_{\bar{\mu}}(\gamma^l) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) \right] = N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j);$$

Hence by (31), given $\gamma^l \in \Gamma(Q_\lambda)$ and $q_\lambda(a|\gamma^l) > 0$,

$$N^a_{\bar{\mu}}(\gamma^l) - \sum_{j=1}^{J-1} \theta(j)\mu_l(j) = N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j),$$

and again applying (31) completes the proof of necessity.

With regard to sufficiency, consider $\lambda \in \Lambda(\bar{\mu}, A)$ for which there exists $\theta(j)$ such that, given $\gamma \in \Gamma(Q_\lambda)$ and $q_\lambda(a|\gamma) > 0$,

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

all $\gamma' \in \Gamma$ and $a' \in A$. Carry out the decomposition called for above into inattentive strategies indexing $\gamma \in \Gamma(Q_\lambda)$ as γ^l, defining $\lambda(l) \in I(\gamma^l)$ by setting $q_{\lambda(l)}(a|\gamma^l) = q_\lambda(a|\gamma^l)$, and using the linearity lemma to conclude that,

$$N_{\bar{\mu}}(\bar{\mu}, \lambda) = \sum_l Q_\lambda(\gamma^l) \sum_a q_{\lambda(l)}(a|\gamma^l) N^a_{\bar{\mu}}(\gamma^l).$$

Hence, since at all possible posteriors $\gamma \in \Gamma(Q_\lambda)$ and corresponding actions achieve the same value of $N^a_{\bar{\mu}}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j)$, this precise value also applies to strategy λ, so that,

$$\begin{aligned} N^a_{\bar{\mu}}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) &= \sum_l Q_\lambda(\gamma^l) \sum_a q_{\lambda(l)}(a|\gamma^l) \left[N^a_{\bar{\mu}}(\gamma^l) - \sum_{j=1}^{J-1} \theta(j)\gamma^l(j) \right] \\ &= N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j) \left[\sum_l Q_\lambda(\gamma^l)\gamma^l(j) \right] = N_{\bar{\mu}}(\bar{\mu}, \lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j), \end{aligned}$$

where $\sum_l Q_\lambda(\gamma^l)\gamma^l(j) = \bar{\mu}(j)$ by Bayes' rule. Hence,

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

all $\gamma' \in \Gamma$ and $a' \in A$.

Now consider an arbitrary strategy $\eta \in \Lambda(\bar{\mu}, A)$. Repeat precisely the corresponding decomposition into inattentive strategies indexing $\gamma \in \Gamma(Q_\eta)$ as $\tilde{\gamma}^l$, defining $\eta(l) \in I(\tilde{\gamma}^l)$ by setting

$q_{\eta(l)}(a|\tilde{\gamma}^l) = q_\eta(a|\gamma^l)$ to conclude that,

$$N_{\bar{\mu}}(\bar{\mu},\eta) = \sum_l Q_\eta(\tilde{\gamma}^l) \sum_a q_{\eta(l)}(a|\tilde{\gamma}^l) N^a_{\bar{\mu}}(\tilde{\gamma}^l).$$

Using a similar decomposition to the above we have that

$$N_{\bar{\mu}}(\bar{\mu},\eta) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j) = \sum_l Q_\eta(\tilde{\gamma}^l) \sum_a q_{\eta(l)}(a|\tilde{\gamma}^l) \left[N^a_{\bar{\mu}}(\tilde{\gamma}^l) - \sum_{j=1}^{J-1} \theta(j)\tilde{\gamma}^l(j) \right].$$

Since in addition,

$$N^a_{\bar{\mu}}(\tilde{\gamma}^l) - \sum_{j=1}^{J-1} \theta(j)\tilde{\gamma}^l(j) \leq N_{\bar{\mu}}(\bar{\mu},\lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j),$$

for all a and $\tilde{\gamma}^l$ we conclude that,

$$N_{\bar{\mu}}(\bar{\mu},\eta) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j) \leq N_{\bar{\mu}}(\bar{\mu},\lambda) - \sum_{j=1}^{J-1} \theta(j)\bar{\mu}(j).$$

Hence $N_{\bar{\mu}}(\bar{\mu},\eta) \leq N_{\bar{\mu}}(\bar{\mu},\lambda)$, establishing optimality. ■

Lemma 2.7 (Feasibility Implies Optimality - FIO): Given $\bar{\mu} \in \Gamma$ and $K \in \mathcal{K}^{PS}$, there exists a one-to-one function on the optimal posterior set $\hat{\Gamma}(\mu|K)$, $f_{\bar{\mu}} : \hat{\Gamma}(\mu|K) \to \mathcal{A}$, with range $\mathcal{A}_{\bar{\mu}}$ such that, given a set $A \subset \mathcal{A}_{\bar{\mu}}$ with $(\bar{\mu}, A) \in \mathcal{D}$ and given $\lambda = (Q_\lambda, q_\lambda) \in \Lambda(\bar{\mu}, A)$,

$$q_\lambda(f_{\bar{\mu}}(\gamma)|\gamma) = 1 \text{ for all } \gamma \in \Gamma(Q_\lambda) \Longrightarrow \lambda \in \hat{\Lambda}(\bar{\mu}, A|K).$$

Moreover $\tilde{\Gamma}(\mu) \subset \hat{\Gamma}(\mu|K)$.

Proof. Given $K \in \mathcal{K}^{PS}$ and $\bar{\mu} \in \Gamma$, we define for each $\bar{\gamma} \in \hat{\Gamma}(\mu|K)$ a particular action $f_{\bar{\mu}}(\bar{\gamma})$ with the defining property that, using a strictly convex function $T_{\bar{\mu}} : \Gamma(\bar{\mu}) \to \mathbb{R}$ associated with $K \in \mathcal{K}^{PS}$, the corresponding function $N_{\bar{\mu}}$ has maximal value of zero at $\bar{\gamma}$:

$$N^{f_{\bar{\mu}}(\bar{\gamma})}_{\bar{\mu}}(\phi) = \bar{u}(\phi, f_{\bar{\mu}}(\bar{\gamma})) - T_{\bar{\mu}}(\phi) \leq N^{f_{\bar{\mu}}(\bar{\gamma})}_{\bar{\mu}}(\bar{\gamma}) = 0, \tag{33}$$

all $\phi \in \Gamma(\bar{\mu})$. Note that this is sufficient to establish the result, defining $\mathcal{A}_{\bar{\mu}}$ to be the union of $f_{\bar{\mu}}(\bar{\gamma})$. In this case if we consider any set $A \subset \mathcal{A}_{\bar{\mu}}$ with $(\bar{\mu}, A) \in \mathcal{D}$ and a strategy $\lambda \in \Lambda(\bar{\mu}, A)$ such that $q_\lambda(f_{\bar{\mu}}(\gamma)|\gamma) = 1$ for all $\gamma \in \Gamma(Q_\lambda)$, we can conclude that this strategy satisfies the sufficient conditions for optimality in Lemma 2.6 when we set all multipliers to zero, $\theta(j) = 0$, establishing that $\lambda \in \hat{\Lambda}(\bar{\mu}, A)$.

We define the function satisfying equation (33) in two phases. First we consider interior beliefs $\bar{\gamma}$ with $\bar{\gamma}(\omega) > 0$ all $\omega \in \Omega(\bar{\mu})$. We note first that since $-T_{\bar{\mu}}(\gamma)$ is a strictly concave function, there exist multipliers $\bar{\beta}(j)$ on $0 \leq j \leq J-1$, not all zero, such that,

$$-\bar{\beta}(0)T_{\bar{\mu}}(\phi) - \sum_{j=1}^{J-1} \bar{\beta}(j)\phi(j) \leq -\bar{\beta}(0)T_{\bar{\mu}}(\bar{\gamma}) - \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j), \tag{34}$$

all $\phi \in \Gamma(\bar{\mu})$. Given that $\bar{\gamma}$ satisfies $\bar{\gamma}(\omega) > 0$ all $\omega \in \Omega(\bar{\mu})$ we can mimic the proof in the second paragraph of Lemma 2.6 above to establish that $\bar{\beta}(0) \neq 0$. It is also not possible that $\bar{\beta}(0) < 0$. To see this suppose this were so. In this case we could renormalize to $\bar{\beta}(0) = -1$ in (34),

$$T_{\bar{\mu}}(\phi) - \sum_{j=1}^{J-1} \bar{\beta}(j)\phi(j) \leq T_{\bar{\mu}}(\bar{\gamma}) - \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j), \tag{35}$$

all $\phi \in \Gamma(\bar{\mu})$. Given that $\bar{\gamma}(\omega) > 0$ all $\omega \in \Omega(\bar{\mu})$, we can find two distinct beliefs $\phi_1, \phi_2 \in \Gamma(\bar{\mu})$ such that $\bar{\gamma} = \frac{\phi_1+\phi_2}{2}$. Averaging inequality (35) applied to each of ϕ_1, ϕ_2 separately, we conclude that,

$$\begin{aligned} \frac{T_{\bar{\mu}}(\phi_1) - \sum_{j=1}^{J-1} \bar{\beta}(j)\phi_1(j) + T_{\bar{\mu}}(\phi_2) - \sum_{j=1}^{J-1} \bar{\beta}(j)\phi_2(j)}{2} &= \frac{T_{\bar{\mu}}(\phi_1) + T_{\bar{\mu}}(\phi_2)}{2} - \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j) \\ &\leq T_{\bar{\mu}}(\bar{\gamma}) - \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j). \end{aligned}$$

We conclude therefore that,

$$0.5T_{\bar{\mu}}(\phi_1) + 0.5T_{\bar{\mu}}(\phi_2) \leq T_{\bar{\mu}}(\frac{\phi_1+\phi_2}{2}), \tag{36}$$

which contradicts strict convexity of T.

With $\bar{\beta}(0) > 0$, we can renormalize to $\bar{\beta}(0) = 1$ in (34) and flip signs to conclude that,

$$T_{\bar{\mu}}(\phi) + \sum_{j=1}^{J-1} \bar{\beta}(j)\phi(j) \geq T_{\bar{\mu}}(\bar{\gamma}) + \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j). \tag{37}$$

Now define action $f_{\bar{\mu}}(\bar{\gamma})$ so that, for $1 \leq k \leq J$,

$$u(f_{\bar{\mu}}(\bar{\gamma}), k) = \left\{ \begin{array}{c} \left[T_{\bar{\mu}}(\bar{\gamma}) + \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j) \right] - \bar{\beta}(k) \text{ for } 1 \leq k \leq J-1; \\ \left[T_{\bar{\mu}}(\bar{\gamma}) + \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j) \right] \text{ for } k = J. \end{array} \right. .$$

By construction, given $\phi \in \Gamma(\bar{\mu})$ and so

$$\begin{aligned} N_{\bar{\mu}}^{f_{\bar{\mu}}(\bar{\gamma})}(\phi) &\equiv \sum_{k=1}^{J} u(f_{\bar{\mu}}(\bar{\gamma}), k)\phi(k) - T_{\bar{\mu}}(\phi) = \\ &= \left[T_{\bar{\mu}}(\bar{\gamma}) + \sum_{j=1}^{J-1} \bar{\beta}(j)\bar{\gamma}(j) \right] - \left[T_{\bar{\mu}}(\phi) + \sum_{k=1}^{J-1} \bar{\beta}(k)\phi(k) \right] \leq 0. \end{aligned}$$

with $N_{\bar{\mu}}^{f_{\bar{\mu}}(\bar{\gamma})}(\bar{\gamma}) = 0$, where the last inequality derives directly from (37).

Given a boundary posterior $\bar{\gamma} \in \Gamma(Q_\lambda)$ with $\bar{\gamma}(\omega) = 0$ some $\omega \in \Omega(\bar{\mu})$ we cannot guarantee that the multiplier $\beta(0)$ in (34) is non-zero (Shannon is a counterexample). The remaining cases therefore involve boundary posteriors that are part of an optimal strategy for some decision problem - i.e. $\gamma \in \hat{\Gamma}(\bar{\mu}|K)$. By definition there exists an optimal strategy $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}(\bar{\mu}, A)$ with $\bar{\gamma} \in \Gamma(Q_\lambda)$, and so by Lemma 2.6, there exists $\bar{\theta} \in \mathbb{R}^{J-1}$ such that, for $\bar{a}$ with $q_{\bar{\lambda}}(\bar{a}|\bar{\gamma}) > 0$, then,

$$\begin{aligned}
\sum_{j=1}^{J} u(\bar{a}, j)\phi(j) - T_{\bar{\mu}}(\phi) - \sum_{j=1}^{J-1} \bar{\theta}(j)\phi(j) &= N_{\bar{\mu}}^{\bar{a}}(\phi) - \sum_{j=1}^{J-1} \bar{\theta}(j)\phi(j) = \\
&\leq N_{\bar{\mu}}^{\bar{a}}(\bar{\gamma}) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j) = \sum_{j=1}^{J} u(\bar{a}, j)\bar{\gamma}(j) - T_{\bar{\mu}}(\bar{\gamma}) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j)
\end{aligned}$$

all $\phi \in \Gamma(\bar{\mu})$. Rearrangement yields,

$$\sum_{j=1}^{J} u(\bar{a}, j)\phi(j) - \sum_{j=1}^{J-1} \bar{\theta}(j)\phi(j) - \left[\sum_{j=1}^{J} u(\bar{a}, j)\bar{\gamma}(j) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j)\right] \bar{\gamma}(j) + T_{\bar{\mu}}(\bar{\gamma}) - T_{\bar{\mu}}(\phi) \leq 0.$$

Now define action $f_{\bar{\mu}}(\bar{\gamma}) \in \mathcal{A}$ so that, for $1 \leq k \leq J$,

$$u(f_{\bar{\mu}}(\bar{\gamma}), k) = \begin{cases} u(\bar{a}, k) - \bar{\theta}(k) - \left[\sum_{j=1}^{J} u(\bar{a}, j)\bar{\gamma}(j) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j) - T_{\bar{\mu}}(\bar{\gamma})\right] & \text{for } 1 \leq k \leq J-1; \\ u(\bar{a}, J) - \left[\sum_{j=1}^{J} u(\bar{a}, j)\bar{\gamma}(j) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j) - T_{\bar{\mu}}(\bar{\gamma})\right] & \text{for } k = J. \end{cases}$$

By construction, given $\phi \in \Gamma(\bar{\mu})$ and defining $\phi(J) = 1 - \sum_{j=1}^{J-1} \phi(j) \geq 0$ on $\phi \in \Gamma(\bar{\mu})$,

$$N_{\bar{\mu}}^{f_{\bar{\mu}}(\bar{\gamma})}(\phi) = \sum_{k=1}^{J} u(\bar{a}, k)\phi(k) - \sum_{k=1}^{J-1} \bar{\theta}(k)\ \phi(k) - \left[\sum_{j=1}^{J} u(\bar{a}, j)\bar{\gamma}(j) - \sum_{j=1}^{J-1} \bar{\theta}(j)\bar{\gamma}(j) - T_{\bar{\mu}}(\bar{\gamma})\right] - T_{\bar{\mu}}(\phi) \leq 0,$$

with $N_{\bar{\mu}}^{f_{\bar{\mu}}(\bar{\gamma})}(\bar{\gamma}) = 0$.

To complete the proof, note that the 1-1 nature of $f_{\bar{\mu}}(\bar{\gamma})$ follows since otherwise there exists an optimal strategy that selects the same action at two different posteriors, which would contradict Lemma 2.3. ■

A2.3: Preservation of Optimality

We have already seen that mixing preserves optimality. There are other important operations that ensure preservation of optimality.

Lemma 2.8: (Perturbed Payoffs and Optimality) Consider $(\bar{\mu}, A) \in \mathcal{D}$ and $\lambda \in \hat{\Lambda}(\bar{\mu}, A|K)$. Given any unchosen action $b \in A\backslash\mathcal{A}(\lambda)$, consider $h(b) \in \mathcal{A}$ with $u(h(b), \omega) < u(b, \omega)$ all

$\omega \in \Omega(\bar{\mu})$ and define $A' = \mathcal{A}(\lambda) \cup_{b \in A/\mathcal{A}(\lambda)} h(b)$. Then

$$\lambda' \in \hat{\Lambda}(\bar{\mu}, A'|K) \Longrightarrow \mathcal{A}(\lambda') \subset \mathcal{A}(\lambda)$$

Proof. Immediate. ■

Lemma 2.9: (Intersecting Posteriors and Intersecting Actions) Given $(\bar{\mu}, A_1) \in \mathcal{D}$, $\lambda_1 = (Q_1, q_1) \in \hat{\Lambda}(\bar{\mu}, A_1|K)$, and $Q_2 \in \hat{\mathcal{Q}}(\bar{\mu})$ with $\Gamma(Q_1) \cap \Gamma(Q_2) \neq \emptyset$, there exists $(\bar{\mu}, A_2) \in \mathcal{D}$ and $\bar{\lambda}(2) = (\bar{Q}_2, \bar{q}_2) \in \hat{\Lambda}(\bar{\mu}, A_2)$ with $\bar{Q}_2(\gamma) = Q_2(\gamma)$ and,

$$\bar{q}_2(a|\gamma) = q_1(a|\gamma),$$

all $\gamma \in \Gamma(Q_1) \cap \Gamma(Q_2)$.

Proof. Consider $(\bar{\mu}, A_1) \in \mathcal{D}$, and $\lambda(1) = (Q_1, q_1) \in \hat{\Lambda}(\bar{\mu}, A_1|K)$. By the Lagrangian Lemma there exists $\theta \in \mathbb{R}^{J-1}$ s.t.,

$$N_{\bar{\mu}}^{a}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq \sup_{a' \in A_1, \gamma' \in \Gamma(\bar{\mu})} N_{\bar{\mu}}^{a'}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) \equiv \bar{N}, \tag{38}$$

all $\gamma \in \Gamma(\bar{\mu})$ and $a \in A_1$, with equality if $\gamma \in \Gamma(Q_1)$ and $q_{\lambda(1)}(a|\gamma) > 0$. To simplify notation in this step we define subsets $A_1(\gamma) \subset A_1$ on $\gamma \in \Gamma(Q_1)$ by the condition,

$$a \in A_1(\gamma) \Longleftrightarrow q_{\lambda(1)}(a|\gamma) > 0,$$

By (38), we know that, given $\gamma, \gamma' \in \Gamma(Q_1)$,

$$N_{\bar{\mu}}^{a_1(\gamma)}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) = N_{\bar{\mu}}^{a_1(\gamma')}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) = \bar{N}.$$

for any $a_1(\gamma) \in A_1(\gamma)$ and $a_1(\gamma') \in A_1(\gamma')$.

We now associate with each remaining possible posterior $\gamma \in \Gamma(Q_2)/\Gamma(Q_1)$ an action $a_2(\gamma)$. In defining these payoffs, we make essential use of the function $f_{\bar{\mu}}(\gamma) \in \mathcal{A}$ from Lemma 2.7 which is well defined on $\Gamma(Q_2)$ since $Q_2 \in \hat{\mathcal{Q}}(\bar{\mu})$. We make use also of the Lagrangians $\theta(j)$ and the net utility functions and value $\bar{N}$ in (38). Specifically, we define $a_2(\gamma)$ on $\gamma \in \Gamma(Q_2)/\Gamma(Q_1)$ to have state dependent payoffs,

$$u(a_2(\gamma), j) = \begin{cases} \bar{N} + u(f_{\bar{\mu}}(\gamma), j) + \theta(j) \text{ for } 1 \leq j \leq J-1 \\ \bar{N} + u(f_{\bar{\mu}}(\gamma), J) \end{cases}$$

We define the set of such actions, as well as their union with actions selected in the first step:

$$\begin{aligned} B_2 &= \{a_2(\gamma) | \gamma \in \Gamma(Q_2)/\Gamma(Q_1)\}; \\ A_2 &= B_2 \cup \{A_1(\gamma) | \gamma \in \Gamma(Q_1) \cap \Gamma(Q_2)\}. \end{aligned}$$

We now construct the strategy of interest $\bar{\lambda}(2) = (\bar{Q}_2, \bar{q}_2)$ according to the prescription in the statement of the Lemma. We first specify $\bar{Q}_2(\gamma) = Q_2(\gamma)$, so that $\Gamma(\bar{Q}_2) = \Gamma(Q_2)$. With regard to

$\bar{q}_2(\gamma)$, it is specified differently according to whether or not $\gamma \in \Gamma(Q_1) \cap \Gamma(Q_2)$:

$$\bar{q}_2(a|\gamma) = \begin{cases} q_1(a|\gamma) \text{ for } \gamma \in \Gamma(Q_1) \cap \Gamma(Q_2); \\ 1 \text{ if } \gamma \in \Gamma(Q_2)/\Gamma(Q_1) \text{ and } a = a_2(\gamma); \\ 0 \text{ if } \gamma \in \Gamma(Q_2)/\Gamma(Q_1) \text{ and } a \neq a_2(\gamma). \end{cases}$$

Note by construction that $\mathcal{A}(\bar{\lambda}(2)) = A_2$, and also that, since $Q_2 \in \hat{\mathcal{Q}}(\bar{\mu})$,

$$\sum_{\gamma \in \Gamma(\bar{Q}_2)} \gamma \bar{Q}_2(\gamma) = \sum_{\gamma \in \Gamma(Q_2)} \gamma Q_2(\gamma) = \bar{\mu},$$

so that $\bar{\lambda}(2) \in \Lambda(\bar{\mu}, A_2)$.

It remains to show that $\bar{\lambda}(2) \in \hat{\Lambda}(\bar{\mu}, A_2)$. To establish this we use the sufficiency aspect of the Lagrangian Lemma. Specifically, we use the original Lagrangians $\theta(j)$ in (38) and show that,

$$N_{\bar{\mu}}^{a}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq \sup_{a' \in A_2, \gamma' \in \Gamma(\bar{\mu})} N_{\bar{\mu}}^{a'}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) = \bar{N} \tag{39}$$

all $\gamma \in \Gamma(\bar{\mu})$ and $a \in A_2$, with equality if $\gamma \in \Gamma(\bar{Q}_2)$ and $\bar{q}_2(a|\gamma) > 0$.

The relevant equality for $a \in A_1(\gamma)$ for $\gamma \in \Gamma(Q_1) \cap \Gamma(Q_2)$ is directly implied by (38). We now consider $\gamma \in \Gamma(Q_2)/\Gamma(Q_1)$ and the corresponding chosen action $a_2(\gamma)$. By construction,

$$N_{\bar{\mu}}^{a_2(\gamma)}(\gamma) = \bar{N} + N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma) + \sum_{j=1}^{J-1} \theta(j)\gamma(j).$$

Hence,

$$\begin{aligned} N_{\bar{\mu}}^{a_2(\gamma)}(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) &= \bar{N} + N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma) + \sum_{j=1}^{J-1} \theta(j)\gamma(j) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \\ &= \bar{N} + N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma) = \bar{N}, \end{aligned}$$

since $N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma) = 0$, confirming the requisite equality.

It remains to show that the inequality aspect of (39) holds,

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

all $a \in A_2$ and $\gamma \in \Gamma(\bar{\mu})$. That this holds for $a \in A_1(\gamma)$ for $\gamma \in \Gamma(Q_1) \cap \Gamma(Q_2)$ is directly implied by (38). It remains to confirm this for $a = a_2(\gamma) \in B_2$ for $\gamma \in \Gamma(Q_2)/\Gamma(Q_1)$ and $\gamma' \in \Gamma(\bar{\mu})$. In this case, the result follows from the defining properties of $N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}$. Given $\gamma' \in \Gamma(\bar{\mu})$,

$$N_{\bar{\mu}}^{a_2(\gamma)}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) = \bar{N} + N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma') \leq \bar{N},$$

since $N_{\bar{\mu}}^{f_{\bar{\mu}}(\gamma)}(\gamma') \leq 0$ all $\gamma' \in \Gamma(\bar{\mu})$. This completes the proof. ■

Lemma 2.10 (Decomposition and Uniqueness): Given $\lambda \in \hat{\Lambda}(\bar{\mu}, A)$ there exist strategies $\lambda^*(l) = (Q_l^*, q_l^*) \in \hat{\Lambda}(\bar{\mu}, A)$ for $1 \leq l \leq L$ and corresponding probability weights $\alpha(l) > 0$ such that,

$$\lambda \equiv \sum_{l=1}^{L} \alpha(l)\lambda^*(l),$$

with each strategy $\lambda^*(l)$ uniquely optimal with regard to the chosen actions $\mathcal{A}[\lambda^*(l)]$.

$$\hat{\Lambda}(\bar{\mu}, \mathcal{A}[\lambda^*(l)]) = \lambda^*(l).$$

Proof. We first show that there exists a decomposition $\lambda(l) = (Q_{\lambda(l)}, q_{\lambda(l)}) \in \hat{\Lambda}(\bar{\mu}, A)$ for $1 \leq l \leq L$ and corresponding probability weights $\alpha(l)$ such that,

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

with each set $\Gamma(Q_{\lambda(l)})$ linearly independent. The proof is constructive. If $\Gamma(Q_\lambda)$ is linearly independent, we are done. If not, then we know from Caratheodory's theorem that since $\Gamma(Q_\lambda)$ contains $\bar{\mu}$ in its convex hull, there exists a linearly independent set $\Gamma(1) \subset \Gamma(Q_\lambda)$ with $|\Gamma(1)| < |\Gamma(Q_\lambda)|$ that also has $\bar{\mu}$ in its convex hull. Hence there exist strictly positive probability weights $Q_1^{LI}(\gamma) > 0$ on $\gamma \in \Gamma(1)$ (extended to $\Gamma(Q_\lambda)$ by setting probabilities of excluded posteriors to zero) such that $\bar{\mu} = \sum\limits_{\gamma \in \Gamma(1)} \gamma Q_1^{LI}(\gamma)$. We define strategy $\lambda(1) \in \Lambda(\bar{\mu}, A)$ to satisfy $\Gamma(Q_{\lambda(1)}) = \Gamma(1)$ with precisely this distribution of posteriors,

$$Q_{\lambda(1)}(\gamma) = Q_1^{LI}(\gamma),$$

and with the same mixed strategy action choice as in strategy λ,

$$q_{\lambda(1)}(a|\gamma) = q_\lambda(a|\gamma).$$

We now identify the smallest scalar $\pi(1) \in (0, 1)$ such that,

$$\pi(1)Q_1^{LI}(\gamma) = Q_\lambda(\gamma),$$

some $\gamma \in \Gamma(Q_{\lambda(1)})$. That such a scalar exists follows from the fact that,

$$\sum_{\gamma \in \Gamma(Q_{\lambda(1)}))} Q_1^{LI}(\gamma) = \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) = 1,$$

with all components in both sums strictly positive and with $\left|\Gamma(Q_{\lambda(1)})\right| < |\Gamma(Q_\lambda)|$.

Define $\hat{\Gamma}(1) = \Gamma(Q_\lambda)$ and $Q_1 = Q_\lambda$ to start the iteration. We now define function $Q_2(\gamma)$ on $\gamma \in \Gamma(Q_\lambda)$ by,

$$Q_2(\gamma) = \frac{Q_1(\gamma) - \pi(1)Q_{\lambda(1)}(\gamma)}{1 - \pi(1)} \geq 0,$$

Note that these define a probability distribution on $\Gamma(Q_\lambda)$,

$$\sum_{\gamma\in\Gamma(Q_\lambda)} Q_2(\gamma) = \frac{\sum_{\gamma\in\Gamma(Q_\lambda)} Q_1(\gamma) - \pi(1)\sum_{\gamma\in\Gamma(Q_\lambda)} Q_1^{LI}(\gamma)}{1-\pi(1)} = 1.$$

Correspondingly, we define,

$$\tilde{\Gamma}(2) = \{\gamma \in \Gamma | Q_2(\gamma) > 0\},$$

noting that $\left|\tilde{\Gamma}(2)\right| < \left|\hat{\Gamma}(1)\right|$, since by construction there exists $\gamma \in \Gamma(Q_\lambda)$ with $\pi(1)Q_{\lambda(1)}(\gamma) = Q_\lambda(\gamma)$ so that $Q_2(\gamma) = 0$. Note that the mean is preserved,

$$\begin{aligned}\sum_{\gamma\in\tilde{\Gamma}(2)} \gamma Q_2(\gamma) &= \sum_{\gamma\in\Gamma(Q_\lambda)} \gamma Q_2(\gamma) = \sum_{\gamma\in\Gamma(Q_\lambda)} \gamma \left[\frac{Q_1(\gamma)-\pi(1)Q_1^{LI}(\gamma)}{1-\pi(1)}\right] \\ &= \frac{1}{1-\pi(1)}\left[\sum_{\gamma\in\Gamma(Q_\lambda)} \gamma Q_1(\gamma) - \pi(1)\sum_{\gamma\in\Gamma(Q_\lambda)} \gamma Q_1^{LI}(\gamma)\right] \\ &= \frac{\mu}{1-\pi(1)}\left[1-\pi(1)\right] = \bar{\mu}.\end{aligned}$$

We define strategy $\tilde{\lambda}(2) \in \Lambda(\bar{\mu}, A)$ to involve precisely these posteriors, $Q_{\tilde{\lambda}(2)}(\gamma) = Q_2(\gamma)$ on $\gamma \in \Gamma(2) = \Gamma(Q_{\tilde{\lambda}(2)})$, with the same mixed action strategies as in λ,

$$q_{\tilde{\lambda}(2)}(a|\gamma) = q_\lambda(a|\gamma).$$

If set $\tilde{\Gamma}(2)$ is linearly independent we define $\lambda(2) = \tilde{\lambda}(2) \in \Lambda(\bar{\mu}, A)$ and stop the iteration. If not, we reapply Caratheodory's theorem and identify a linearly independent set $\Gamma(2) \subset \tilde{\Gamma}(2)$ that retains $\bar{\mu}$ in its convex hull, hence for which there exist strictly positive probability weights $Q_2^{LI}(\gamma) > 0$ on $\gamma \in \Gamma(2)$ such that $\bar{\mu} = \sum\limits_{\gamma\in\Gamma(2)} Q_2^{LI}(\gamma)\gamma$. In this case, we define strategy $\lambda(2) \in \Lambda(\bar{\mu}, A)$ to involve precisely these posteriors, $\Gamma(\lambda(2)) = \Gamma(2)$ with the corresponding probability weights, $Q_{\lambda(2)}(\gamma) = Q_2^{LI}(\gamma)$, and again with the same mixed action choice as in strategy λ, $q_{\lambda(2)}(a|\gamma) = q_\lambda(a|\gamma)$. Rounding out the iterative process, we then define $\pi(2) \in (0,1)$ as the smallest number such that,

$$\pi(2)Q_{\lambda(2)}(\gamma) = Q_2(\gamma),$$

some $\gamma \in \Gamma(Q_{\lambda(2)})$. Finally, we define $Q_3(\gamma)$ on $\gamma \in \Gamma(Q_\lambda)$ by,

$$Q_3(\gamma) = \frac{Q_2(\gamma) - \pi(2)Q_{\lambda(2)}(\gamma)}{1-\pi(2)} \geq 0,$$

and $\tilde{\Gamma}(3) = \{\gamma \in \Gamma | Q_3(\gamma) > 0\}$. We continue in iterative fashion defining non-empty sets of posterior $\tilde{\Gamma}(l)$, linearly independent subsets $\Gamma(l) \subset \tilde{\Gamma}(l)$ and corresponding strategies $\lambda(l) \in \Lambda(\bar{\mu}, A)$. This iteration is completed in a finite number of steps, $L \in \mathbb{N}$, since $\left|\tilde{\Gamma}(l+1)\right| < \left|\tilde{\Gamma}(l)\right|$.

The above construction provides us with a set of strategies $\{\lambda(l)\}_{1\leq l\leq L}$ that are feasible, $\lambda(l) \in \Lambda(\bar{\mu}, A)$, and that have linearly independent posteriors. By construction, the distribution of the posteriors for Lth strategy are given by

$$Q_{\lambda(L)}(\gamma)=\frac{Q_\lambda(\gamma)}{\Pi_{l=1}^{L-1}(1-\pi(l))}-\sum_{l=1}^{L-1}\frac{\pi(l)Q_{\lambda(l)}(\gamma)}{\Pi_{k=l}^{L-1}(1-\pi(l))}$$

We now show that we can reverse engineer the construction to identify probability weights $\alpha(l)$ such that

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

Specifically, we define

$$\alpha(l)=\Pi_{k=1}^{l-1}\left[1-\pi(k)\right]\pi(l)$$

for $1\leq l\leq L-1$ (using the convention that $\alpha(1)=\Pi_{k=1}^{0}\left[1-\pi(k)\right]\pi(1)=\pi(1)$), and

$$\alpha(L)=\Pi_{k=1}^{L-1}\left[1-\pi(k)\right]$$

First note that these weights sum to 1, as

$$\begin{aligned}&\alpha(1)+\alpha(2)+...+\alpha(L-1)+\alpha(L)\\=\;&\pi(1)+\pi(2)(1-\pi(1))+...+\pi(L-1)\Pi_{k=1}^{L-2}\left[1-\pi(k)\right]+(1-\pi(L-1))\Pi_{k=1}^{L-2}\left[1-\pi(k)\right]\end{aligned}$$

The final two terms collapse to $\Pi_{k=1}^{L-2}\left[1-\pi(k)\right]$, which can then be combined with the term from $\alpha(L-2)$ in order to give $\Pi_{k=1}^{L-3}\left[1-\pi(k)\right]$. Iterating on this process leaves eventually

$$\pi(1)+(1-\pi(1))=1$$

To confirm that indeed

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

we need to show only that the unconditional posterior probabilities are the same,

$$Q_{\lambda(\alpha)}(\gamma)=\sum_{l}\alpha(l)Q_{\lambda(l)}(\gamma)=Q_\lambda(\gamma);$$

since the construction ensures that all conditional action strategies are identical,

$$q_{\lambda(l)}(a|\gamma)=q_\lambda(a|\gamma),$$

all l. Note that,

$$\begin{aligned}&\sum_{l=1}^{L}\alpha(l)Q_{\lambda(l)}(\gamma)\\=\;&\sum_{l=1}^{L-1}\Pi_{k=1}^{l-1}\left[1-\pi(k)\right]\pi(l)Q_{\lambda(l)}(\gamma)+\Pi_{k=1}^{L-1}(1-\pi(k))\left[\frac{Q_\lambda(\gamma)}{\Pi_{l=1}^{L-1}(1-\pi(l))}-\sum_{l=1}^{L-1}\frac{\pi(l)Q_{\lambda(l)}(\gamma)}{\Pi_{k=l}^{L-1}(1-\pi(l))}\right]\\=\;&Q_\lambda(\gamma).\end{aligned}$$

Given that $\lambda \equiv \sum_{l=1}^{L} \alpha(l)\lambda(l)$ and $\lambda \in \hat{\Lambda}(\bar{\mu}, A)$, we apply Lemma 2.2 to conclude that $\lambda(l) \in \hat{\Lambda}(\bar{\mu}, A)$ all l. We now take each strategy $\lambda(l)$ in turn. We move to a pure strategy versions $\lambda^*(l,m) = (Q^*_{l,m}, q^*_{l,m})$. For each such strategy we set $Q^*_{l,m}(\gamma) = Q_{\lambda(l)}(\gamma)$ and then take each possible posterior $\gamma \in \Gamma(l)$, selecting one action $a \in A$ that is chosen with positive probability at that posterior, $q_{\lambda(l)}(a|\gamma) > 0$, and setting its probability to 1,

$$q^*_{l,m}(a|\gamma) = 1.$$

Note that $N_{\bar{\mu}}(\bar{\mu}, \lambda(l)) = N_{\bar{\mu}}(\bar{\mu}, \lambda^*(l,m))$ since optimality implies that all options chosen at any given posterior produce the same value, so that $\lambda^*(l,m) \in \hat{\Lambda}(\bar{\mu}, A)$ all l and m. We repeat this exercise for all possible combinations of actions chosen according to $\lambda(l)$ at posteriors $\gamma \in \Gamma(Q_l)$, using M to denote the number of such combinations, then appropriately weight these strategies together with weights such that $\sum_{m=1}^{M} \alpha^*(l,m) = \alpha(l)$ and

$$\sum_{m=1}^{M} \frac{\alpha^*(l,m)}{\alpha(l)} \lambda^*(l,m) = \lambda(\lambda).$$

To complete the proof we consider the cardinality of the set of chosen actions, $\mathcal{A}[\lambda^*(l)] \subset A$. Note by the Lemma 2.3 that since $\lambda^*(l) \in \hat{\Lambda}(\bar{\mu}, A)$, each chosen action is associated with a unique posterior, so that,

$$|\Gamma(Q^*_l)| = |\mathcal{A}[\lambda^*(l)]|.$$

Together these put us in position to apply Lemma 2.4 to complete the proof: since $\lambda^*(l) \in \hat{\Lambda}(\bar{\mu}, A)$, $\Gamma(Q_{\lambda^*(l)}) = \Gamma(Q_{\lambda(l)}) \subset \Gamma$ is linearly independent, and $\left|\Gamma\left(Q_{\lambda^*(l)}\right)\right| = \mathcal{A}[\lambda^*(l)]$, the optimal strategy is unique,

$$\lambda^*(l) = \hat{\Lambda}(\bar{\mu}, \mathcal{A}[\lambda^*(l)]).$$

■

A2.4: Generalized PS Models

From this point forward through the remainder of this appendix, there is no need to consider changes in the prior within a given proof, so that we remove the over-bar, using μ in place of $\bar{\mu}$, and correspondingly defining the generic decision problem to be (μ, A) rather than $(\bar{\mu}, A)$.

At a key point in the proof of theorem 2 we need to consider variants of the PS cost function in which the T function is not strictly convex. To simplify the proof, it is convenient to consider T functions that take infinite value on unchosen posteriors.

Definition 2 *We define $K \in \mathcal{K}$ to be **generalized** PS, $K \in \mathcal{K}^{GPS}$, if, $\forall \mu \in \Gamma$, $\hat{\Gamma}(\mu|K)$ is convex set and $\exists$ $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$ such that,*

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

and $Q \in \mathcal{Q}(\mu)$. We define the corresponding convexified cost function $K^{CONV} \in \mathcal{K}$ by defining

T_μ^{CONV} *to be the* ***convex hull*** *of* T_μ. *This is defined as the greatest convex function majorized by* T_μ, *and is shown by Rockafellar [1970] (page 36) to be equal to,*

$$T_\mu^{CONV}(\gamma) = \inf\{\sum_{m=1}^{M} \alpha(m)T(\gamma(m))| \sum_{m=1}^{M} \alpha(m)\gamma(m) = \gamma\},$$

over weights $\alpha(m) > 0$ *satisfying* $\sum \alpha(m) = 1$. *We then define the cost function*

$$K^{CONV}(\mu, Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma)T_\mu^{CONV}(\gamma) - T_\mu^{CONV}(\mu).$$

Lemma 2.11: Convexification and Optimal Strategies Given $C \in \mathcal{C}$ with a CIR representation $K \in \mathcal{K}^{GPS}$,

$$\hat{P}(\mu, A|K) = \hat{P}(\mu, A|K^{CONV})$$

all $(\mu, A) \in \mathcal{D}$.

Proof. To show that $\hat{P}(\mu, A|K) \subset \hat{P}(\mu, A|K^{CONV})$, it suffices to show that any strategy $\lambda \in \hat{\Lambda}(\mu, A|K)$ is also optimal with the convexified cost function, $\lambda \in \hat{\Lambda}(\mu, A|K^{CONV})$, since then the corresponding data $\mathbf{P}_\lambda$ is in both sets. The first step is to show that the value function is no higher for the convexified function,

$$\hat{V}(\mu, A|K^{CONV}) \leq \hat{V}(\mu, A|K). \tag{40}$$

Consider an arbitrary strategy $\eta \in \Lambda(\mu, A)$ and index the finite set of possible posteriors $\bar{\eta}(n) \in \Gamma(\eta)$ for $1 \leq n \leq N$. By construction of the lower semi-continuous hull of T_μ (Rockafellar page 36),

$$T_\mu^{CONV}(\gamma) = \inf\{\sum_{m=1}^{M} \alpha(m)T(\gamma(m))| \sum_{m=1}^{M} \alpha(m)\gamma(m) = \gamma\}.$$

Hence we know that for each posterior $\bar{\eta}(n)$ and $p \in \mathbb{N}$ there exists a finite set of posteriors $\eta(n, m, p)$ for $1 \leq m \leq M(p)$ and corresponding weights $\alpha(n, m, p) > 0$ with $\sum_{m=1}^{M_n} \alpha(n, m, p) = 1$ such that.

$$\sum_{m=1}^{M(n,p)} \sum_{n=1}^{N} \alpha(n, m, p)\eta(n, m, p) = \bar{\eta}(n);$$

and such that the corresponding weighted average value of $T\left[\eta(n, m, p)\right]$ is no more than $\frac{1}{p}$ above $T_\mu^{CONV}\left[\bar{\eta}(n)\right]$, so that, for $1 \leq n \leq N$,

$$T_\mu^{CONV}\left[\bar{\eta}(n)\right] \geq \sum_{m=1}^{M_n} \alpha(n, m, p)T\left[\eta(n, m, p)\right] - \frac{1}{p}.$$

For each $p \in \mathbb{N}$ we introduce a corresponding strategy $F(\eta, p) = (Q_{F(\eta,p)}, q_{F(\eta,p)})$ with possible posteriors,

$$\Gamma(Q_{F(\eta,p)}) = \{\eta(n, m, p)|1 \leq n \leq N \text{ and } 1 \leq m \leq M(n, p)\}.$$

Specifically, the strategy is defined by:

$$\begin{aligned} Q_{F(\eta,p)}\left[\eta(n,m,p)\right] &= \alpha(n,m,p)Q_{\eta}(\bar{\eta}(n)); \\ q_{F(\eta,p)}\left[a|\eta(n,m,p)\right] &= q_{\eta}\left[a|\bar{\eta}(n)\right]. \end{aligned}$$

The first key observation is that this strategy is feasible, $F(\eta,p) \in \Lambda(\mu,A)$. It is immediate that $\mathcal{A}(F(\eta,p)) = \mathcal{A}(\eta) \subset A$. That $F(\eta,p) \in \Lambda(\mu,A)$ requires first that it is a strategy, which means that the probabilities over possible posteriors add to 1. This follows directly from the definition,

$$\begin{aligned} \sum_{\eta \in \Gamma(Q_{F(\eta,p)})} Q_{F(\eta,p)}\left[\eta(n,m,p)\right] &= \sum_{n=1}^{N}\sum_{m=1}^{M_n} \alpha(n,m,p)Q_{\eta}(\bar{\eta}(n)) \\ &= \sum_{n=1}^{N} Q_{\eta}(\bar{\eta}(n)) \sum_{m=1}^{M_n} \alpha(n,m,p) = \sum_{n=1}^{N} Q_{\eta}(\bar{\eta}(n)) = 1 \end{aligned}$$

To complete this part of the proof requires confirmation of Bayes' rule. This again is definitional,

$$\begin{aligned} \sum_{\eta \in \Gamma(Q_{F(\eta,p)})} \eta(n,m,p)Q_{F(\eta,p)}\left[\eta(n,m,p)\right] &= \sum_{n=1}^{N}\sum_{m=1}^{M_n} \alpha(n,m,p)\eta(n,m,p)Q_{\eta}(\bar{\eta}(n)) \\ &= \sum_{n=1}^{N} Q_{\eta}(\bar{\eta}(n)) \sum_{m=1}^{M_n} \alpha(n,m,p)\eta(n,m,p) \\ &= \sum_{n=1}^{N} \bar{\eta}(n)Q_{\eta}(\bar{\eta}(n)) = \mu. \end{aligned}$$

The second key observation is that $F(\eta,p)$ using K achieves utility net of attention costs within $\frac{1}{p}$ of that η achieves using K^{CONV}. To see this, consider first the expected prize utility as defined by the probability distribution over rewards,

$$U(F(\eta,p)) = \sum_{\gamma \in \Gamma(Q_{F(\eta,p)})} \sum_{a \in A} Q_{F(\eta,p)}(\gamma)q_{F(\eta,p)}(a|\gamma)\bar{u}(\gamma,a).$$

Defining the relevant set of indices,

$$\mathcal{I} \equiv \{(n,m,p)|1 \leq n \leq N \text{ and } 1 \leq m \leq M(n,p)$$

Note by direct substitution that,

$$\begin{aligned}
\sum_{\gamma\in\Gamma(Q_{F(\eta,p)})}\sum_{a\in A} Q_{F(\eta,p)}(\gamma)q_{F(\eta,p)}(a|\gamma)\bar{u}(\gamma,a) &= \sum_{\mathcal{I}}\sum_{a\in A}\alpha(n,m,p)Q_\eta(\bar{\eta}(n))q_{F(\eta,p)}\left[a|\eta(n,m,p)\right]\bar{u}(\bar{\eta}(n),a) \\
&= \sum_{\mathcal{I}}\sum_{a\in A}\alpha(n,m,p)Q_\eta(\bar{\eta}(n))q_\eta\left[a|\bar{\eta}(n)\right]\bar{u}(\bar{\eta}(n),a) \\
&= \sum_{1\le n\le N}\sum_{a\in A}\sum_{1\le m\le M(n,p)\}}\alpha(n,m,p)Q_\eta(\bar{\eta}(n))q_\eta\left[a|\bar{\eta}(n)\right]\bar{u}(\bar{\eta}(n),a) \\
&= \sum_{1\le n\le N}\sum_{a\in A}Q_\eta(\bar{\eta}(n))q_\eta\left[a|\bar{\eta}(n)\right]\bar{u}(\bar{\eta}(n),a) = U(\eta).
\end{aligned}$$

With regard to the costs, note by construction that,

$$T^{CONV}\left[\bar{\eta}(n)\right] \ge \sum_{m=1}^{M(n,p)}\alpha(n,m,p)T\left[\eta(n,m,p)\right] - \frac{1}{p}.$$

Hence,

$$\begin{aligned}
K^{CONV}(Q_\eta) + T^{CONV}\left[\mu\right] &= \sum_{n=1}^{N}Q_\eta(\bar{\eta}(n))T^{CONV}\left[\bar{\eta}(n)\right] \\
&\ge \sum_{n=1}^{N}\sum_{m=1}^{M_n}Q_\eta(\bar{\eta}(n))\alpha(n,m,p)T\left[\eta(n,m,p)\right] - \frac{1}{p} \\
&= \sum_{n=1}^{N}\sum_{m=1}^{M_n}Q_{F(\eta)}\left[\eta(n,m,p)\right]T\left[\eta(n,m,p)\right] - \frac{1}{p} = K(Q_{F(\eta,p)}) - \frac{1}{p}.
\end{aligned}$$

Hence,

$$V(\mu,\eta|K^{CONV}) = U(\eta) - K^{CONV}(Q_\eta) \le U(F(\eta,p)) - K(Q_{F(\eta,p)}) + \frac{1}{p} = V(\mu,F(\eta,p)|K) + \frac{1}{p}.$$

By increasing p without bound, we establish that,

$$V(\mu,\eta|K^{CONV}) \le \sup_{\lambda\in\Lambda(\mu,A)}V(\mu,\lambda|K) = \hat{V}(\mu,A|K).$$

Since $\eta\in\Lambda(\mu,A)$ is arbitrary, this ensures that the supremum of all values is correspondingly bounded above,

$$\hat{V}(\mu,A|K^{CONV}) \equiv \sup_{\eta\in\Lambda(\mu,A)}V(\mu,\eta|K^{CONV}) \le \hat{V}(\mu,A|K),$$

completing the proof of (40).

To show that $\hat{P}(\mu,A|K)\subset\hat{P}(\mu,A|K^{CONV})$, note that since C has a CIR, we know that $\hat{\Lambda}(\mu,A|K)\neq\emptyset$ for any $(\mu,A)\in\mathcal{D}$. Now consider any optimal strategy $\lambda\in\hat{\Lambda}(\mu,A|K)$ which therefore achieves the value

$$\hat{V}(\mu,A|K) = V(\mu,\lambda|K) = U(\lambda) - K(\mu,Q_\lambda)$$

By definition of the convexification operation, note that $T_{\mu}^{CONV}(\gamma) \leq T_{\mu}\ (\gamma)$ all $\gamma \in \Gamma^{C}(\mu)$, so that,

$$K^{CONV}(\mu, Q_{\lambda}) \leq K(\mu, Q_{\lambda}).$$

Hence,

$$\begin{aligned} \hat{V}(\mu, A|K) &= U(\lambda) - K(\mu, Q_{\lambda}) \leq U(\lambda) - K^{CONV}(\mu, Q_{\lambda}) \leq V(\mu, A|K^{CONV}) \\ &\leq \hat{V}(\mu, A|K^{CONV}) \leq \hat{V}(\mu, A|K), \end{aligned}$$

making all equalities, so that indeed $\lambda \in \hat{\Lambda}(\mu, A|K^{CONV})$, completing the proof that $\hat{P}(\mu, A|K) \subset \hat{P}(\mu, A|K^{CONV})$.

To complete the proof, we establish the converse set inclusion. The first key observation is that $\hat{\Gamma}(\mu|K^{CONV}) \subset \hat{\Gamma}(\mu|K)$. To see this, first note that by definition, $\gamma' \notin \hat{\Gamma}(\mu|K)$ implies $T(\gamma') = \infty$. Moreover, because $\hat{\Gamma}(\mu|K)$ is convex by the definition of a generalized CIR, any weighted average of posteriors for which $\sum_{m=1}^{M} \alpha(m)\gamma(m) = \gamma'$ must involve at least one posterior $\gamma(m) \in \Gamma(\mu)/\hat{\Gamma}(\mu|K)$ for which $T(\gamma(m)) = \infty$. Hence $\sum_{m=1}^{M} \alpha(m)T(\gamma(m)) = \infty$. Thus,

$$T_{\mu}^{CONV}(\gamma') = \inf\{\sum_{m=1}^{M} \alpha(m)T(\gamma(m))|\sum_{m=1}^{M} \alpha(m)\gamma(m) = \gamma'\} = \infty,$$

where the weights $\alpha(m)$ are positive and sum to 1.

Finally, this implies that for any $Q \in \mathcal{Q}(\mu)$ with $\gamma' \in \Gamma(Q)$, the cost $K^{CONV}(\mu, Q) = \infty$, and so cannot be optimal (as the inattentive strategy will always provide a higher payoff). Hence $\gamma' \notin \hat{\Gamma}(\mu|K^{CONV})$.

Next, we show that, given any $\gamma' \in \hat{\Gamma}(\mu|K)$, it must be the case that $T^{CONV}\ [\gamma'] = T[\gamma']$. Assume not, then by definition of $\hat{\Gamma}(\mu|K)$ and K there exists $(\mu, A) \in \mathcal{D}$ and $\lambda = (q_{\lambda}, Q_{\lambda}) \in \hat{\Lambda}(\mu, A|K)$ such that $\gamma' \in \Gamma(Q)$. Note that, by construction $T^{CONV}\ [\gamma'] \leq T[\gamma']$, so assume by way of contradiction that $T^{CONV}\ [\gamma'] < T[\gamma']$. By definition of $T^{CONV}\ [\gamma']$ as the infimum of $\sum_{\xi \in \Gamma(\bar{Q})} T(\xi)\bar{Q}(\xi)$ on the set of posterior distributions that generate γ', this implies that there exists some alternative distribution of posteriors $\bar{Q}(\xi)$ that satisfies two conditions:

$$\begin{aligned} \sum_{\xi \in \Gamma(\bar{Q})} \xi\bar{Q}(\xi) &= \gamma'; \\ \sum_{\xi \in \Gamma(\bar{Q})} T(\xi)\bar{Q}(\xi) &< T[\gamma']. \end{aligned}$$

If such a distribution existed, one could amend strategy $\lambda = (q_{\lambda}, Q_{\lambda}) \in \hat{\Lambda}(\mu, A|K)$ and construct an alternative strategy $\lambda^{*} = (q^{*}, Q^{*}) \in \Lambda(\mu, A)$ that produced strictly higher net utility, contradicting the assumed optimality. One would simply reduce the probability of $\gamma' \in \Gamma(Q_{\lambda})$ by $Q_{\lambda}(\gamma')$ and increase the probability of each $\gamma \in \Gamma(\bar{Q})$ by $\gamma \in Q_{\lambda}(\gamma')\bar{Q}(\gamma)$, adjusting action choice probabilities appropriately to generate the same revealed posteriors as in λ. Effectively the new strategy uses $\sum_{\xi \in \Gamma(\bar{Q})} \xi\bar{Q}(\xi)$ rather than γ', but is otherwise identical. It is straightforward to show that $\lambda^{*} \in \Lambda(\mu, A)$, $U(\lambda) = U(\lambda^{*})$ but $K(\mu, Q^{*}) < K(\mu, Q_{\lambda})$.

To complete the proof of the lemma, we need to show that $\hat{P}(\mu, A|K^{CONV}) \subset \hat{P}(\mu, A|K)$. By construction, for any $P \in \hat{P}(\mu, A|K^{CONV})$, there exists an optimal strategy $\lambda = (q_\lambda, Q_\lambda) \in \hat{\Lambda}(\mu, A|K^{CONV})$ such that,

$$P = \mathbf{P}_\lambda.$$

First, note that, by the first claim above, for any $\gamma \in \Gamma(Q_\lambda)$, it must be the case that $\gamma \in \hat{\Gamma}(\mu|K)$, and so, by the second claim, $T^{CONV}[\gamma'] = T[\gamma']$. This directly implies that

$$\begin{aligned} K^{CONV}(\mu, Q_\lambda) &= \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) T^{CONV}(\gamma) - T^{CONV}(\mu) \\ &= \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma) T(\gamma) - T(\mu) \\ &= K(\mu, \bar{Q}), \end{aligned}$$

and so $V(\mu, \lambda|K^{CONV}) = V(\mu, \lambda|K)$, and as $\hat{V}(\mu, A|K) \leq \hat{V}(\mu, A|K^{CONV}) = V(\mu, \lambda|K^{CONV})$, we have $\lambda \in \bar{\Lambda}(\mu, A|K)$ and so $\mathbf{P}_\lambda \in \hat{P}(\mu, A|K)$, completing the proof. ■

A2.5: Linking Strategies with Data

Lemma 2.12 (Strategies and Revealed Posteriors): Given $\lambda \in \hat{\Lambda}(\mu, A|K)$ for $K \in \mathcal{K}^{PS}$ and some $(\mu, A) \in \mathcal{D}$:

1. $\mathcal{A}(\mathbf{P}_\lambda) = \mathcal{A}(\lambda)$;
2. Given for $\gamma \in \Gamma(Q_\lambda)$ and a such that $q_\lambda(a|\gamma) > 0$
$$\bar{\gamma}^a_{\mathbf{P}_\lambda} = \gamma.$$

Proof. Immediate. ■

Lemma 2.13 (Inverse Operation on Data): Given $P \in \mathcal{P}(\mu, A)$ for some $(\mu, A) \in \mathcal{D}$, $\mathcal{A}(\mathbf{Q}_P) = \mathcal{A}(P)$, and:

1. $\mathbf{P}_{\boldsymbol{\lambda}(P)} = P$.
2. $\gamma^a_{\boldsymbol{\lambda}(P)} = \bar{\gamma}^a_P$.

Proof. Immediate. ■

Lemma 2.14 (Optimal Strategies and Data): Given $\lambda \in \hat{\Lambda}(\mu, A|K)$ for $K \in \mathcal{K}^{PS}$ and some $(\mu, A) \in \mathcal{D}$:

1. $\Gamma(\mathbf{Q}_{\mathbf{P}_\lambda}) = \Gamma(Q_\lambda)$.
2. $\mathbf{Q}_{\mathbf{P}_\lambda}(\gamma) = Q_\lambda(\gamma)$ all $\gamma \in \Gamma(\mathbf{Q}_{\mathbf{P}_\lambda})$.
3. $\mathbf{q}_{\mathbf{P}_\lambda}(a|\gamma) = q_\lambda(a|\gamma)$ all $a \in \mathcal{A}(\mathbf{P}_\lambda)$, $\gamma \in \Gamma(\mathbf{Q}_{\mathbf{P}_\lambda})$.
4. $\boldsymbol{\lambda}(\mathbf{P}_\lambda) = \lambda$.

Proof. Immediate. ■

Lemma 2.15 (Data and Optimal Strategies): Given $C \in \mathcal{C}$ with a PS representation $K \in \mathcal{K}^{PS}$ and $P \in C(\mu, A)$ some $(\mu, A) \in \mathcal{D}$,

$$\boldsymbol{\lambda}(P) \in \hat{\Lambda}(\mu, A|K).$$

Proof. Immediate. ■

Lemma 2.16 (Identical Posteriors): Given $C \in \mathcal{C}$ with a PS representation $K \in \mathcal{K}^{PS}$,

$$\hat{\Gamma}(\mu|K) = \Gamma^C(\mu)$$

all $\mu \in \Gamma$.

Proof. Immediate. ■

Lemma 2.17 (Identical Posterior Distributions): Given $C \in \mathcal{C}$ with a PS representation $K \in \mathcal{K}^{PS}$,

$$\Delta(\hat{\Gamma}(\mu|K)) \cap \mathcal{Q}(\mu) \subset \mathcal{Q}^C(\mu)$$

all $\mu \in \Gamma$.

Proof. Immediate. ■

Lemma 2.18: (Mixtures and Data) Given $\bar{\mu} \in \Gamma$, if strategies $\lambda, \{\lambda(l)\}_{1 \leq l \leq L} \in \Lambda(\bar{\mu}, A)$ are such that $\lambda = \sum_{l=1}^{L} \alpha(l)\lambda(l)$ for probability weights $\{\alpha(l)\}_{1 \leq l \leq L}$, then,

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

Proof. Immediate. ■

Lemma 2.19: (Irrelevance of Impossible Payoffs) Given $C \in \mathcal{C}$ with CIR $K \in \mathcal{K}$, consider $a \in \mathcal{A}(P)$ and $a' \neq a \in \mathcal{A}$ with identical payoffs in possible states $\Omega(\mu)$ according to some μ and define A' to be A with a' replacing a:

$$\begin{aligned} u(a', \omega) &= u(a, \omega) \text{ all } \omega \in \Omega(\mu); \\ A' &= a' \cup A/a. \end{aligned}$$

Then,

$$C(\mu, A') = \left\{P' \in \mathcal{P}(\mu, A') | \exists P \in C(\mu, A) \text{ s.t } P'(a'|\omega) = P(a|\omega) \text{ and } P'(b|\omega) = P(b|\omega) \text{ all } b \in A/a.\right\}$$

Proof. Immediate. ■

Appendix 3: Theorem 3

In this Appendix we prove theorem 3. We start by proving necessity of the axioms, then sufficiency.

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

Proof. Necessity of A2 through A8 for a PS representation:

Necessity of NIAS (A2) and NIAC (A3) is immediate following Caplin and Martin (2015) and Caplin and Dean (2015) respectively. Necessity of Completeness (A4) follows from combining Lemma 2.7 (FIO) with Lemma 2.16 (Identical Posteriors) and 2.17 (Identical Posterior Distribution). First, Lemma 2.7, FIO implies that $\tilde{\Gamma}(\mu) \subset \hat{\Gamma}(\mu|K)$. Given the Identical Posteriors Lemma, it is therefore immediate that $\tilde{\Gamma}(\mu) \subset \Gamma^C(\mu)$, establishing necessity of the first clause in the Completeness axiom. The Identical Posterior Lemma further establishes that $\hat{\Gamma}(\mu|K) = \Gamma^C(\mu)$, whereupon the assumed convexity of $\hat{\Gamma}(\mu|K)$ in a PS representation establishes the convexity clause of the Completeness axiom. Finally, the Identical Posterior Distribution Lemma establishes the third clause

$$\Delta(\Gamma^C(\mu)) = \Delta(\hat{\Gamma}(\mu|K)) = \mathcal{Q}^C(\mu),$$

completing the proof that A4 is necessary,

We now establish necessity of Separability (A5). Consider $(\mu, A(1)) \in \mathcal{D}$, and $P(1) \in C(\mu, A(1))$. Note by Lemma 2.15 that $\boldsymbol{\lambda}(P(1)) = (\mathbf{Q}_{P(1)}, \mathbf{q}_{P(1)}) \in \hat{\Lambda}(\mu, A_1|K)$ and by Lemma 2.13 that $\mathbf{P}_{\boldsymbol{\lambda}(P(1))} = P(1)$.

Now consider $Q_2 \in \mathcal{Q}^C(\mu)$ with $\Gamma(\mathbf{Q}_{P(1)}) \cap \Gamma(Q_2) \neq \emptyset$. By Lemma 2.17, $Q_2 \in \hat{Q}(\mu)$. Now apply Lemma 2.9 to conclude that there exists $(\mu, A_2) \in \mathcal{D}$ and $\lambda(2) = (Q_2, q_2) \in \hat{\Lambda}(\mu, A_2)$ with,

$$q_2(a|\gamma) = \mathbf{q}_{P(1)}(a|\gamma),$$

for $\gamma \in \Gamma(\mathbf{Q}_{P(1)}) \cap \Gamma(Q_2)$. Now apply the $\mathbf{P}$ operator to $\lambda(2)$ conclude that, noting that since this is a PS representation and $\lambda(2) = (Q_2, q_2) \in \hat{\Lambda}(\mu, A_2)$,

$$\mathbf{P}_{\boldsymbol{\lambda}(2)} \equiv P(2) \in C(\mu, A_2).$$

We can apply Lemma 2.14 to conclude that, since $\lambda(2) = (Q_2, q_2) \in \hat{\Lambda}(\mu, A_2)$ that,

$$\mathbf{q}_{P(2)}(a|\gamma) = q_2(a|\gamma).$$

Stringing these together we conclude that indeed,

$$\mathbf{q}_{P(1)}(a|\gamma) = \mathbf{q}_{P(2)}(a|\gamma),$$

all $\gamma \in \Gamma(\mathbf{Q}_{P(1)}) \cap \Gamma(Q_2)$, establishing necessity of separability, and with it the proof of necessity of A5.

To prove necessity of Non-linearity (A6), we consider $(\mu, A) \in \mathcal{D}$, $P \in C(\mu, A)$, and pick a_1, a_2, $a_3 \in A$ such that $\bar{\gamma}_P^{a_1} \neq \bar{\gamma}_P^{a_3}$ and such that,

$$\bar{\gamma}_P^{a_2} = \alpha\bar{\gamma}_P^{a_1} + (1-\alpha)\bar{\gamma}_P^{a_3},$$

for some $\alpha \in (0,1)$. By Lemma 2.15, we know that $\boldsymbol{\lambda}(P) \in \hat{\Lambda}(\mu, A|K)$. We know also that the corresponding revealed posteriors are identical to the posteriors used in the strategy,

$$\gamma^{a_i}_{\boldsymbol{\lambda}(P)} = \bar{\gamma}^{a_i}_P \text{ for } 1 \leq i \leq 3.$$

Thus, by the Lagrangian Lemma there exists $\theta \in \mathbb{R}^{J-1}$ such that:

$$\begin{aligned} \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_2}_P(j) + \bar{u}(\bar{\gamma}^{a_2}_P, a_2) - T_\mu(\bar{\gamma}^{a_2}_P) &= \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_1}_P(j) + \bar{u}(\bar{\gamma}^{a_1}_P, a_1) - T_\mu(\bar{\gamma}^{a_1}_P) \\ &= \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_3}_P(j) + \bar{u}(\bar{\gamma}^{a_3}_P, a_3) - T_\mu(\bar{\gamma}^{a_3}_P). \end{aligned}$$

Hence,

$$\begin{aligned} & \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_2}_P(j) + \bar{u}(\bar{\gamma}^{a_2}_P, a_2) - T_\mu(\bar{\gamma}^{a_2}_P) \\ = \; & \alpha \left(\sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_1}_P(j) + \bar{u}(\bar{\gamma}^{a_1}_P, a_1) - T_\mu(\bar{\gamma}^{a_1}_P) \right) + (1-\alpha) \left(\sum_{j=1}^{J-1} \theta(j)\bar{\gamma}^{a_3}_P(j) + \bar{u}(\bar{\gamma}^{a_3}_P, a_3) - T_\mu(\bar{\gamma}^{a_3}_P) \right). \end{aligned}$$

Rearrangement yields,

$$\begin{aligned} & \sum_{j=1}^{J-1} \theta(j) \left[\bar{\gamma}^{a_2}_P(j) - \alpha\bar{\gamma}^{a_1}_P(j) + (1-\alpha)\bar{\gamma}^{a_3}_P(j)\right] + \bar{u}(\bar{\gamma}^{a_2}_P, a_2) - \left(\alpha\bar{u}(\bar{\gamma}^{a_1}_P, a_1) + (1-\alpha)\bar{u}(\bar{\gamma}^{a_3}_P, a_3)\right) \\ = \; & T_\mu(\bar{\gamma}^{a_2}_P) - \left(\alpha T_\mu(\bar{\gamma}^{a_1}_P) + (1-\alpha) T_\mu(\bar{\gamma}^{a_3}_P)\right). \end{aligned}$$

Since $\bar{\gamma}^{a_2}_P = \alpha\bar{\gamma}^{a_1}_P + (1-\alpha)\bar{\gamma}^{a3}_P$ the first term is equal to zero, and if $\bar{u}(\bar{\gamma}^{a_2}_P, a_2) = \alpha\bar{u}(\bar{\gamma}^{a_1}_P, a_1) + (1-\alpha)\bar{u}(\bar{\gamma}^{a_3}_P, a_3)$, the second term would also be zero. Hence,

$$T_\mu(\bar{\gamma}^{a_2}_P) = \alpha T_\mu(\bar{\gamma}^{a_1}_P) + (1-\alpha) T_\mu(\bar{\gamma}^{a_3}_P),$$

in contradiction of strict convexity of T_μ. This establishes necessity of A6.

To establish necessity of Convexity (A7), consider $(\mu, A) \in \mathcal{D}$, $P_l \in C(\mu, A)$ for $1 \leq l \leq L$, and probability weights $\alpha(l) > 0$. Define the mixture data $P_\alpha \in P(\mu, A)$ by,

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

Convexity requires that $P_\alpha \in C\left(\mu, A\right)$. Note that since the C is a CIR, there exists $\lambda(l) = (Q_l, q_l) \in \hat{\Lambda}\left(\mu, A|K\right)$ for which,

$$\mathbf{P}_{\lambda(l)} = P_l \text{ all } l\ .$$

Define the strategy $\lambda(\alpha)$ to be the corresponding mixture strategy, as defined in Appendix 2. By the Mixing and Optimality Lemma (2.2), $\lambda(\alpha) \in \hat{\Lambda}\left(\mu, A|K\right)$, and since this is a CIR,

$$\mathbf{P}_{\lambda(\alpha)} \in C\left(\mu, A\right).$$

To complete the proof, we apply the Mixtures and Data Lemma (2.18) to confirm that $\lambda(\alpha)$ correspondingly mixes the SDSC data: given $a \in A$ and $\omega \in \Omega(\mu)$,

$$\mathbf{P}_{\lambda(\alpha)}(a|\omega) = \sum_{l=1}^{L} \alpha(l)\mathbf{P}_{\lambda(l)} = \sum_{l=1}^{L} \alpha(l)P_l a|\omega) = P_\alpha(a|\omega) \in C\left(\mu, A\right)$$

We conclude the necessity proof by establishing necessity of Continuity (A8). To this end, we consider $\mu \in \Gamma$ and $K \in \mathcal{K}^{PS}$ together with $I \geq 1$ sequences of actions $a^i(m)$ with $\lim_{m\to\infty} a^i(m) = \bar{a}^i$ for $1 \leq i \leq I$, with $A(m) = \cup_{i=1}^{I} a^i(m)$ and $\bar{A} = \cup_{i=1}^{I}\ \bar{a}^i$. Suppose that there exists $P \in \cap_{m=1}^{\infty} C(\mu, A(m))$ with $\mathcal{A}(P) \subset \bar{A}$. Then by Lemma 2.8, the revealed attention strategy satisfies $\boldsymbol{\lambda}(P) \in \cap_{m=1}^{\infty}\hat{\Lambda}(\mu, A(m)|K)$ and since $\mathcal{A}(P) \subset \bar{A}$, it is also feasible in the limit problem, $\boldsymbol{\lambda}(P) \in \Lambda(\mu, \bar{A})$. What we need to prove is that it is optimal.

Since $\boldsymbol{\lambda}(P) = (\mathbf{Q}_P, \mathbf{q}_P) \in \hat{\Lambda}(\bar{\mu}, A(m)|K)$, we know that it achieves optimal value, which we denote $\hat{V}$,

$$V(\boldsymbol{\lambda}_P) = U(\boldsymbol{\lambda}_P) - K_{\bar{\mu}}(\bar{\mu}, \mathbf{Q}_P) = \hat{V}(\bar{\mu}, A(m)) \equiv \hat{V}.$$

Since $\boldsymbol{\lambda}_P \in \Lambda(\bar{\mu}, \bar{A})$, we know that $\hat{V}(\bar{\mu}, \bar{A}|K) \geq \hat{V}$. Strictness of this inequality is not possible. To see this, consider a purported strategy $\bar{\lambda}' = (\bar{Q}', \bar{q}') \in \Lambda\left(\bar{\mu}, \bar{A}\right)$ that achieves higher value in the limit problem,

$$V(\bar{\lambda}', \bar{\mu}, \bar{A}|K) > \hat{V}.$$

We now construct the corresponding strategy $\lambda'_m = (Q'_m, q'_m) \in \Lambda\left(\bar{\mu}, A(m)\right)$ that uses precisely the same posterior distribution and the correspondingly indexed action choices,

$$Q'_m(\gamma) = \bar{Q}'(\gamma) \text{ and } q'_m(a^i(m)|\gamma^i) = \bar{q}'(\bar{a}^i|\gamma^i).$$

Since $\lim_{m\to\infty} a^i(m) = \bar{a}^i$, we know that

$$\lim_{m\to\infty} U(\lambda'_m) = U(\bar{\lambda}'),$$

hence, as costs depend only on the unchanging distribution over posteriors, the corresponding holds for the valuation,

$$\lim_{m\to\infty} V(\lambda'_m, \mu, A(m)|K) = V(\bar{\lambda}', \mu, \bar{A}|K) > \hat{V},$$

contradicting $\boldsymbol{\lambda}_P \in \hat{\Lambda}(\bar{\mu}, A(m)|K)$. We conclude that $\boldsymbol{\lambda}_P$ is optimal in the limit problem, $\boldsymbol{\lambda}_P \in \hat{\Lambda}(\bar{\mu}, \bar{A}|K)$. Hence, given that this is a CIR, $\mathbf{P}_{\boldsymbol{\lambda}_P} \in C(\bar{\mu}, \bar{A})$. That $\mathbf{P}_{\boldsymbol{\lambda}_P} = P$ follows from Lemma 2.13, completing the proof that $P \in C(\bar{\mu}, \bar{A})$. ■

Proof. Sufficiency of A2 through A8 for a PS representation:

There are three key steps in the sufficiency proof. In the first, we invoke corollary 1 to theorem 2 as established in Appendix 1 which gives a cost function of the PS form, yet in which the distribution of posteriors impacts the computed cost of each posterior. The first point that we establish is that A5, Separability allows us to remove the dependence of the function $T^C_\mu(\cdot, \bar{Q})$ on the particular distribution of posteriors. With this we will know that, given $\mu \in \Gamma$, A2 through A5 imply that there exists Generalized PS cost function $\bar{K} \in \mathcal{K}^{GPS}$ (as defined prior to the proof of Lemma 2.11 in Appendix 2) such that $C(\mu, A) \subset \hat{P}(\mu, A|\bar{K})$ all $(\mu, A) \in \mathcal{D}$ and such that, given $(\mu, Q) \in \mathcal{F}$ with

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

$$\bar{K}(\mu, Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\, \bar{T}_\mu(\gamma) - \bar{T}_\mu(\mu). \tag{41}$$

for some $\bar{T}_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$ (with $\bar{T}_\mu$ real-valued on $\Gamma^C(\mu)$). The second step in the proof applies Lemma 2.11 to show that $\bar{T}_\mu$ can be assumed weakly convex without any loss of generality, and that with A6 (Non-linearity), it must be strictly rather than weakly convex. The final step shows that addition of Axiom A7 (Convexity) and A8 (Continuity) allow us to generate all data, $C(\mu, A) = \hat{P}(\mu, A|\bar{K})$. In this final step, a key role is played by Lemmas that are established in Appendix 2, which is to be expected given that the cost function at this stage is of precisely the PS form and the only remaining question relates to ensuring that all optima are observed in the data.

As noted, in the first step of the proof, we invoke corollary 1 to theorem 2 which we state as follows for current purposes. Given $C \in \mathcal{C}$ satisfying A2-A4, there exists a unique function $K \in \mathcal{K}$ such that, given $\mu \in \Gamma$, $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$ where $K \in \mathcal{K}$ can be computed for $(\mu, \bar{Q}) \in \mathcal{F}$ with $\bar{Q} \in \mathcal{Q}^C(\mu)$ by enumerating the support $\Gamma(\bar{Q}) = \{\bar{\gamma}^n | 1 \leq n \leq N\}$ and using the definitions of $T^C_\mu(\bar{\gamma}^n, \bar{Q})$ and $T^C_\mu(\mu, \bar{Q})$ in the proof of theorem 2

$$T^C_\mu(\bar{\gamma}^n, \bar{Q}) \equiv [\bar{\gamma}^n - \mu] \cdot \int_0^1 u(\bar{a}^n_t) dt.$$

and computing,

$$K(\mu, \bar{Q}) \equiv \sum_n \bar{Q}(\bar{\gamma}^n) T^C_\mu(\bar{\gamma}^n, \bar{Q}) - T^C_\mu(\mu, \bar{Q}). \tag{42}$$

In this stage of the proof we show that A5 enables us to remove the dependence on $\bar{Q}$ and find $\bar{T}^C_\mu$ in the data such that we can set,

$$T^C_\mu(\gamma, Q) = \bar{T}^C_\mu(\gamma),$$

all $Q \in \mathcal{Q}^C(\mu)$.

We set up our candidate function $\bar{T}^C_\mu$ in two steps. In the first step, we select $\bar{Q} \in \mathcal{Q}^C(\mu)$ with $\Gamma(\bar{Q}) = \cup_{1 \leq n \leq N} \bar{\gamma}^n$ with $N = |\Omega(\mu)|$ with the $\{\bar{\gamma}^n\}_{n=1}^N$ being affine independent vectors in $\mathbb{R}^N$ thereby forming a basis for $\Gamma(\mu)$: that this is possible follows from Completeness, whereby the full-dimensional interior posteriors is observed, $\tilde{\Gamma} \subset \Gamma^C(\mu)$. We then apply corollary 1 to establish existence of $T^C_\mu(\bar{\gamma}^n, \bar{Q})$ such that,

$$K(\mu, \bar{Q}) = \sum_{n=1}^N \bar{Q}(\gamma^n) T^C_\mu(\gamma^n, \bar{Q}),$$

where,

$$T^C_\mu(\gamma^n, \bar{Q}) = [\gamma^n - \mu] \cdot \int_0^1 u(\bar{a}^n_t) dt \tag{43}$$

and the $\{\bar{a}^n_t\}_{n=1}^N$ for $t \in [0, 1]$ are the actions that support this construction.

We cost all other posteriors by embedding them in decision problems that include this fixed basis $\Gamma(\bar{Q})$. Picking any other observed posterior $\eta \in \Gamma^C(\mu) \backslash \Gamma(\bar{Q})$, we identify a corresponding distribution $\bar{Q}^\eta$ with full support

$$\Gamma(\bar{Q}^\eta) = \Gamma(\bar{Q}) \cup \eta,$$

that satisfies the Bayesian constraint,

$$\sum_{\gamma\in\Gamma(\bar{Q}^{\eta})}\gamma\bar{Q}^{\eta}\left(\gamma\right)=\mu.$$

That this is possible follows from the fact that $\Gamma(\bar{Q})$ forms a basis for $\Gamma(\mu)$ so that an arbitrarily small probability added to η can be precisely off-set by corresponding reductions in the weights on the basis vectors $\bar{\gamma}^n\in\Gamma(\bar{Q})$ while retaining strict positivity (see Lemma 4.5 for details).

Let $\bar{A}_t=\{\bar{a}_t^n\}_{n=1}^N$ and let $(\mu,\bar{A}_t)\in\mathcal{D}$ and $\bar{P}_t\in C\left(\mu,\bar{A}_t\right)$ such that $\mathbf{Q}_{\bar{P}_t}=\bar{Q}_t$ as in the proof to theorem 2. Let,

$$\eta_t=t\eta+(1-t)\mu$$

on $t\in[0,1]$ and let $\bar{Q}_t(\bar{\gamma}_t^n)=\bar{Q}(\bar{\gamma}^n)\equiv\bar{Q}^n$. By construction $\bar{Q}_t^{\eta}\in\mathcal{Q}^C(\mu)$ all $t\in[0,1]$.

We can now apply the Separability axiom, A5. Since $(\mu,\bar{A}_t)\in\mathcal{D}$, $\bar{P}_t\in C(\mu,\bar{A}_t)$ with $\mathbf{Q}_{\bar{P}_t}=\bar{Q}_t$, and $\bar{Q}_t^{\eta}\in\mathcal{Q}^C(\mu)$ satisfies $\Gamma(\bar{Q}_t^{\eta})\cap\Gamma(\bar{Q}_t)=\Gamma(\bar{Q}_t)$, the axiom asserts existence for each $t\in[0,1]$ of $A_t(\eta)\subset\mathcal{A}$ and $P_t(\eta)\in C(\mu,A_t(\eta))$ with $\mathbf{Q}_{P_t(\eta)}=\bar{Q}_t^{\eta}$ such that for each $\bar{\gamma}_t^n\in\Gamma(\bar{Q}_t)$ there exists an action $a\in\bar{A}_t\cap A_t(\eta)$ with $\bar{\gamma}^a_{P_t(\eta)}=\bar{\gamma}^a_{\bar{P}_t}=\bar{\gamma}^n_t$. We identify just such a choice set and data combination, with $P_t(\eta)\in C(\mu,A_t(\eta))$, and define $a_t(\eta)\in A_t(\eta)$ as the action associated with the new revealed posterior,

$$\eta_t=\bar{\gamma}^{a_t(\eta)}_{P_t(\eta)}$$

According to the prescription in theorem 2 we can now compute the cost function from this data using the posterior-by-posterior approach as,

$$K(\mu,\bar{Q}^{\eta})=\sum_{\gamma\in\Gamma(\bar{Q})\cup\eta}^{N}\bar{Q}^{\eta}(\gamma)T^C_{\mu}(\gamma,\bar{Q}^{\eta}),$$

where

$$T^C_{\mu}(\gamma,\bar{Q}^{\eta})=[\gamma-\mu]\cdot\int_0^1u(\bar{a}_t^n)dt. \tag{44}$$

for each $\gamma\in\Gamma(\bar{Q})$, and

$$T^C_{\mu}(\gamma,\bar{Q}^{\eta})=[\gamma-\mu]\cdot\int_0^1u(a_t(\gamma))dt. \tag{45}$$

for $\gamma=\eta$. Note that since the $\bar{a}_t^n$ are the same in (44) and (43), $T^C_{\mu}(\gamma,\bar{Q}^{\eta})=T^C_{\mu}(\gamma,\bar{Q})$ for all $\gamma\in\Gamma(\bar{Q})$.

We now define our candidate cost function $\bar{T}^C_{\mu}(\gamma)$. Specifically, we repeat the above for all $\eta\in\Gamma(Q)\backslash\Gamma(\bar{Q})$ and set

$$\bar{T}_{\mu}(\gamma)=\begin{cases}T^C_{\mu}(\gamma,\bar{Q}) & \text{for }\gamma\in\Gamma(\bar{Q});\\ T^C_{\mu}(\gamma,\bar{Q}^{\gamma}) & \text{for }\gamma\in\Gamma^C(\mu)\backslash\Gamma(\bar{Q}).\end{cases} \tag{46}$$

The claim is that, for any $Q\in\mathcal{Q}^C(\mu)$,

$$K(\mu,Q)=\sum_{\gamma\in\Gamma(Q)}Q(\gamma)\bar{T}_{\mu}(\gamma) \tag{47}$$

We establish (47) first for all $Q \in \mathcal{Q}^C(\mu)$ such that $\Gamma(\bar{Q}) \subset \Gamma(Q)$. The proof is inductive on cardinality. We first establish that the result holds for any $Q \in \mathcal{Q}^C(\mu)$ with cardinality $N+1$ and $\Gamma(\bar{Q}) \subset \Gamma(Q)$. In this case, there is a unique $\eta \in \Gamma(Q)\backslash\Gamma(\bar{Q})$ and $\Gamma(Q^\eta) = \Gamma(\bar{Q}^\eta)$ where $\bar{Q}^\eta$ is the distribution used the construction of (46). Since $\Gamma(Q^\eta) = \Gamma(\bar{Q}^\eta)$, the Separability axiom says that we can use the same acts to calculate $T^C_\mu(\gamma, Q^\eta)$ as we used to calculate $T^C_\mu(\gamma, \bar{Q}^\eta)$, and since the construction of $T^C_\mu(\gamma, \bar{Q}^\eta)$ in (44) and (45) depends on $\bar{Q}^\eta$ only through the actions, it follows that

$$T^C_\mu(\gamma, Q^\eta) = \bar{T}_\mu(\gamma)$$

for all $\gamma \in \Gamma(Q^\eta)$ and

$$K(\mu, Q) = \sum_{n=1}^{N} Q(\gamma^n)\bar{T}_\mu(\gamma^n) + Q(\eta)\bar{T}_\mu(\eta) \tag{48}$$

as required.

We now suppose that (47) holds for all $Q \in \mathcal{Q}^C(\mu)$ with $\Gamma(\bar{Q}) \subset \Gamma(Q)$ and cardinality $N+m$ and for $m \geq 1$ and show that this extends to $N+m+1$. Consider $\hat{Q} \in \mathcal{Q}^C$ such that $\Gamma(\bar{Q}) \subset \Gamma(\hat{Q})$ and $|\Gamma(\hat{Q})| = N+m+1$. Since $|\Gamma(\hat{Q})| > |\Gamma(\bar{Q})| + 1$, we can find η_1,$\eta_2 \in \Gamma(\hat{Q})\backslash\Gamma(\bar{Q})$. Note that by assumption (47) holds for all $Q_1 \in \mathcal{Q}^C(\mu)$ such that $\Gamma(Q_1) = \Gamma(\hat{Q})\backslash\eta_1$ and some path of actions $A_{\Gamma(\hat{Q})\backslash\eta_1}(t)$ for $t \in [0,1]$. By Separability (A5) we can find, for any $t \in [0,1]$ a corresponding set of action paths $\hat{A}_1(t) = \{A_{\Gamma(\hat{Q})\backslash\eta_1}(t), \hat{a}_{\eta_1}(t)\}$ such that,

$$\begin{aligned} K(\hat{Q}) &= \sum_{\gamma \in \Gamma(\hat{Q})} \hat{Q}(\gamma)T^C_\mu(\gamma, \hat{Q}) \\ &= \sum_{\gamma \in \Gamma(\hat{Q})\backslash\eta_1} \hat{Q}(\gamma)\bar{T}_\mu(\gamma) + \hat{Q}(\eta_1)T^C_\mu(\eta_1, \hat{Q}) \end{aligned}$$

where,

$$T^C(\eta_1, \hat{Q}) = [\eta_1 - \mu] \cdot \int_0^1 \hat{a}_{\eta_1}(t)dt.$$

Similarly, we can find $Q_2 \in \mathcal{Q}^C(\mu)$ such that $\Gamma(Q_2) = \Gamma(\hat{Q})\backslash\eta_2$ and path $\hat{A}_2(t) = \{A_{\Gamma(\hat{Q})\backslash\eta_2}(t), \hat{a}_{\eta_2}(t)\}$ defined by the action paths associated with $\bar{T}(\gamma)$ for $\gamma \neq \eta_2$ such that,

$$K(\hat{Q}) = \sum_{\gamma \in \Gamma(\hat{Q})\backslash\eta_2} \hat{Q}(\gamma)\bar{T}_\mu(\gamma) + \hat{Q}(\eta_2)T^C_\mu(\eta_2, \hat{Q})$$

with

$$T^C(\eta_2, \hat{Q}) = [\eta_2 - \mu] \cdot \int_0^1 \hat{a}_{\eta_2}(t)dt.$$

Comparing the two different expressions for precisely the same cost, we conclude that,

$$\hat{Q}(\eta_1)\bar{T}_\mu(\eta_1) + \hat{Q}(\eta_2)T^C_\mu(\eta_2, \hat{Q}) = \hat{Q}(\eta_1)T^C_\mu(\eta_1, \hat{Q}) + \hat{Q}(\eta_2)\bar{T}_\mu(\eta_2),$$

or,

$$\hat{Q}(\eta_1)\left[\bar{T}_\mu(\eta_1) - T^C_\mu(\eta_1, \hat{Q})\right] = \hat{Q}(\eta_2)\left[\bar{T}_\mu(\eta_2) - T^C_\mu(\eta_2, \hat{Q})\right]. \tag{49}$$

Since $|\Gamma(\hat{Q})| > |\Omega(\mu)|$, and contains a basis, so $\Gamma(\bar{Q}) \subset \Gamma(\hat{Q})$, we can find a distinct distribution of posteriors $\hat{Q}' \in \mathcal{Q}^C$ with $\Gamma(\hat{Q}') = \Gamma(\hat{Q})$, $\hat{Q}'(\eta_1) \neq \hat{Q}(\eta_1)$ yet $\hat{Q}'(\eta_2) = \hat{Q}(\eta_2)$ and run corresponding logic to conclude that,

$$\hat{Q}'(\eta_1)\left[\bar{T}_\mu(\eta_1) - T^C_\mu(\eta_1, \hat{Q}')\right] = \hat{Q}'(\eta_2)\left[\bar{T}_\mu(\eta_2) - T^C_\mu(\eta_2, \hat{Q}')\right]. \tag{50}$$

Since $\Gamma(\hat{Q}') = \Gamma(\hat{Q})$, direct application of Separability to the case in which $\Gamma(Q_{P(1)}) \cap \Gamma(Q_2)$ means that we may use the same action set $\hat{A}_1(t) = \{A_{\Gamma(\hat{Q})\backslash\eta_1}(t), \hat{a}_{\eta_1}(t)\}$ to calculate both $T^C_\mu(\eta_1, \hat{Q}')$ and $T^C_\mu(\eta_1, \hat{Q})$. Similarly for η_2. Hence $T^C_\mu(\eta_1, \hat{Q}') = T^C_\mu(\eta_1, \hat{Q})$ and $T^C_\mu(\eta_2, \hat{Q}') = T^C_\mu(\eta_2, \hat{Q})$. Since $\hat{Q}'(\eta_2) = \hat{Q}(\eta_2)$ by assumption, subtracting (49) from (50) yields,

$$\left[\hat{Q}'(\eta_1) - \hat{Q}(\eta_1)\right]\left[\bar{T}_\mu(\eta_1) - T^C_\mu(\eta_1, \hat{Q})\right] = 0.$$

Since $\hat{Q}'(\eta_1) \neq \hat{Q}(\eta_1)$, it follows that

$$T^C_\mu(\eta_1, \hat{Q}) = \bar{T}_\mu(\eta_1),$$

Since η_1 is arbitrary, this establishes the induction step.

We now consider arbitrary $Q \in \mathcal{Q}^C(\mu)$. In particular, we consider Q such that $\Gamma(\bar{Q}) \not\subset \Gamma(Q)$. The preceding establishes that (47) holds for $Q \cup \bar{Q} \equiv Q'$. Separability (A5) ensures that can use the same actions for Q to finally produce the representation of the desired form,

$$K(Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\bar{T}_\mu(\gamma)$$

for all $Q \in \mathcal{Q}^C$, completing the proof that there exists $K \in \mathcal{K}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$ and such that,

$$K(\mu, Q) = \sum_{\Gamma(Q)} Q\,(\gamma)\,\bar{T}^C_\mu(\gamma).$$

This completes the first step of the proof.

At this stage we set $\bar{T}^C_\mu(\gamma)$ to infinity outside the convex set $\Gamma^C(\mu)$. Note that the construction of K as forming a CIR of the data assumes that subjects use the revealed attention strategy in each decision problem, meaning both that increasing the cost of posteriors outside $\Gamma^C(\mu)$ maintains the CIR and that $\Gamma^C(\mu) \subset \hat{\Gamma}(\mu|K)$. Setting $K(\mu, \gamma) = \infty$ for $\gamma \notin \Gamma^C(\mu)$ therefore makes $\Gamma^C(\mu) = \hat{\Gamma}(\mu|K)$ and so both are convex, by Completeness (A4). This means that $K \in \mathcal{K}^{GPS}$. Direct application of Lemma 2.11 shows that we can replace T_μ with its convexification, T^{CONV}_μ, then define K^{CONV} correspondingly, with assurance that the data is unchanged,

$$\hat{P}(\mu, A|K) = \hat{P}(\mu, A|K^{CONV}).$$

This then implies that,

$$C(\mu, A) \subset \hat{P}(\mu, A|K) \Longrightarrow C(\mu, A) \subset \hat{P}(\mu, A|K^{CONV}).$$

To complete the second step we show that, since $C \in \mathcal{C}$ satisfies A3 and A6, and there exists a convex function $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$ such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$, where,

$$K(\mu, Q) = \sum_{\Gamma(Q)} Q\,(\gamma)\, T_\mu(\gamma),$$

then T_μ is strictly convex.

Assume by way of contradiction that T_μ is convex but not strictly so. Then there exists γ_1, $\gamma_3 \in dom\ T_\mu$ such that $\gamma_2 = \alpha\gamma_1 + (1-\alpha)\gamma_3$ and $T_\mu(\gamma_2) = \alpha T_\mu(\gamma_1) + (1-\alpha) T_\mu(\gamma_3)$. By construction above, $T_\mu(\gamma)$ is real valued only on $\Gamma^C(\mu)$ and so by Axiom A4 (Completeness) we can conclude both that $\gamma_2 \in \Gamma^C(\mu)$ and that there exists a decision problem (μ, A) such that, for some a_1, a_2, $a_3 \in A$, we have $\bar{\gamma}_P^{a_1} = \gamma_1$, $\bar{\gamma}_P^{a_2} = \gamma_2$ and $\bar{\gamma}_P^{a_3} = \gamma_3$. By the Lagrangian Lemma, this implies that there is a strategy $\lambda \in \bar{\Lambda}(\mu, A)$ and corresponding multipliers $\theta \in \mathbb{R}^{J-1}$ such that, since $\gamma_1, \gamma_2, \gamma_3 \in \Gamma(\lambda)$,

$$\begin{aligned}
\sum_{j=1}^{J-1} \theta(j)\gamma_1(j) + \bar{u}(\gamma_1, a_1) - T_\mu(\gamma_1) &= \sum_{j=1}^{J-1} \theta(j)\gamma_2(j) + \bar{u}(\gamma_2, a_2) - T_\mu(\gamma_2); \\
\sum_{j=1}^{J-1} \theta(j)\gamma_3(j) + \bar{u}(\gamma_3, a_3) - T_\mu(\gamma_3) &= \sum_{j=1}^{J-1} \theta(j)\gamma_2(j) + \bar{u}(\gamma_2, a_2) - T_\mu(\gamma_2).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sum_{j=1}^{J-1} \theta(j)\gamma_2(j) + \bar{u}(\gamma_2, a_2) - T_\mu(\gamma_2) \\
= \;& \alpha \left(\sum_{j=1}^{J-1} \theta(j)\gamma_1(j) + \bar{u}(\gamma_1, a_1) - T_\mu(\gamma_1) \right) + (1-\alpha) \left(\sum_{j=1}^{J-1} \theta(j)\gamma_3(j) + \bar{u}(\gamma_3, a_3) - T_\mu(\gamma_3) \right).
\end{aligned}$$

Equivalently,

$$\begin{aligned}
&\sum_{j=1}^{J-1} \theta(j) \left(\gamma_2(j) - (\alpha\gamma_1(j) + (1-\alpha)\gamma_3(j))\right) + \bar{u}(\gamma_2, a_2) - (\alpha\bar{u}(\gamma_1, a_1) + (1-\alpha)\bar{u}(\gamma_3, a_3)) \\
= \;& T_\mu(\gamma_2) - (\alpha T_\mu(\gamma_1) + (1-\alpha) T_\mu(\gamma_3)).
\end{aligned}$$

Since $\gamma_2 = \alpha\gamma_1 + (1-\alpha)\gamma_3$ the first term is equal to zero, and by assumption $T_\mu(\gamma_2) = \alpha T_\mu(\gamma_1) + (1-\alpha) T_\mu(\gamma_3)$, and so the RHS equals zero,

$$\bar{u}(\bar{\gamma}_P^{a_2}, a_2) = \alpha\bar{u}(\bar{\gamma}_P^{a_1}, a_1) + (1-\alpha)\bar{u}(\bar{\gamma}_P^{a_3}, a_3).$$

which directly contradicts A6 (Nonlinearity). This contradiction establishes strict convexity of T^{CONV}, completing the second part of the proof.

The final step is to show that, since we have found the strictly convex function for which $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$, A7 and A8 imply that all optima are seen,

$$C(\mu, A) = \hat{P}(\mu, A|K).$$

Suppose that we have indeed found for $C \in \mathcal{C}$ a strictly convex function function $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$ (real valued on $\tilde{\Gamma}(\mu)$) and corresponding PS cost function **on** (μ, Q) **with** $Q \in Q^C(\mu)$**,**

$$K(\mu, Q) = \sum_{\Gamma(Q)} Q\left(\gamma\right) T_\mu(\gamma) - T_\mu\left(\mu\right),$$

such that $C(\mu, A) \subset \hat{P}(\mu, A|K)$ all $(\mu, A) \in \mathcal{D}$. What we show now is that, if C satisfies A7 and A8, then, given arbitrary $(\mu, A) \in \mathcal{D}$ and corresponding optimal strategy $\lambda \in \hat{\Lambda}(\mu, A|K)$, $P_\lambda \in C(\mu, A)$. To prove this we first invoke Lemma 2.10 (Decomposition and Uniqueness), which implies that there exist strategies $\lambda^*(l) = (Q_l^*, q_l^*) \in \hat{\Lambda}(\mu, A)$ for $1 \leq l \leq L$ and corresponding probability weights $\alpha(l)$ such that,

$$\lambda \equiv \sum_{l=1}^{L} \alpha(l)\lambda^*(l),$$

with each strategy $\lambda^*(l)$ uniquely optimal with regard to the chosen actions $\mathcal{A}\left[\lambda^*(l)\right] \subset A$,

$$\hat{\Lambda}(\mu, \mathcal{A}\left[\lambda^*(l)\right]) = \{\lambda^*(l)\}.$$

For each l and $a \in A$ we now construct sequences of actions $a(l, m)$ for $1 \leq m \leq \infty$ as follows:

$$u(a(l,m), j) = \begin{cases} u(a,j) \text{ if } a \in \mathcal{A}\left[\lambda^*(l)\right]; \\ u(a,j) - \frac{1}{m} \text{ if } a \in A/\mathcal{A}\left[\lambda^*(l)\right]. \end{cases}$$

We now define action sets $A(l, m)$ as the corresponding unions,

$$A(l,m) = \cup_{a \in A} a(l,m).$$

Note that this addition of new actions with lowered payoffs does not expand the set of optimal strategies beyond those in $\mathcal{A}\left[\lambda^*(l)\right]$ by Lemma 2.8. Hence,

$$\hat{\Lambda}(\mu, A(l,m)) = \lambda^*(l).$$

Existence of a CIR and uniqueness of the optimal strategy in the perturbed problems implies that the corresponding data is observed all the way to the limit.

$$\mathbf{P}_{\lambda^*(l)} \in \cap_{m=1}^{\infty} C(\mu, A(l,m)).$$

We are now in position to apply the Axiom A8 (Continuity). By construction, $\lim_{m \to \infty} a(l, m) = a$, $A(m) = \cup_{a \in A} a(l, m)$, and $\mathcal{A}(\mathbf{P}_{\lambda^*(l)}) \subset A$, so that this axiom implies that the data is also observed in the limit problem,

$$\mathbf{P}_{\lambda^*(l)} \in C(\mu, A).$$

Since this is true for all l, we can apply the Axiom A7 (Convexity) to conclude that the convex combination of data corresponding to the given strategy λ is also observed,

$$\sum_{l=1}^{L} \alpha(l)\mathbf{P}_{\lambda^*(l)} \in C(\mu, A).$$

To complete the proof, we note from Lemma 2.18 that, since $\lambda \equiv \sum_{l=1}^{L} \alpha(l)\lambda^*(l)$,

$$\mathbf{P}_{\lambda} = \sum_{l=1}^{L} \alpha(l)\mathbf{P}_{\lambda^*(l)} \in C(\mu, A),$$

as required. This completes the proof that $C(\mu, A) = \hat{P}(\mu, A|K)$ and with it the overall sufficiency proof. ■

A3.1: Recoverability

A second recoverability results follow from the above logic. It establishes a simple method of recovering this cost function.

Corollary 2: If $C \in \mathcal{C}$ has a PS representation $K \in \mathcal{K}^{PS}$, then, given $\mu \in \Gamma$ and non-degenerate $\bar{Q} \in \mathcal{Q}^C(\mu)$, there exists $\bar{A}$ for which there is both an inattentive optimal strategy, $\eta \in \Lambda^I(\mu, \bar{A}) \cap \hat{\Lambda}(\mu, \bar{A})$, and an attentive optimal strategy $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}(\mu, \bar{A})$ with $Q_\lambda(\gamma) = \bar{Q}(\gamma)$ all $\gamma \in \Gamma(\bar{Q})$, so that,

$$K(\mu, Q_\lambda) = U(\lambda) - U(\eta).$$

Proof. If $C \in \mathcal{C}$ has a PS representation $K \in \mathcal{K}^{PS}$, consider $\mu \in \Gamma$ and non-degenerate $\bar{Q} \in \mathcal{Q}^C(\mu)$, and note that $\sum_{\gamma \in \Gamma(Q)} \gamma \bar{Q}(\gamma) = \mu$. By Lemma 2.17, the PS representation ensures $\mathcal{Q}^C(\mu) = \Delta(\hat{\Gamma}(\mu|K))$. The construction of $\bar{A}$ is based entirely on $f_\mu(\gamma)$, the function identified in the Lemma 2.7, FIO,:

$$\bar{A} = \left\{\cup_{\gamma \in \Gamma(\bar{Q})} f_\mu(\gamma)\right\} \cup \left\{f_\mu(\mu)\right\}.$$

The two strategies are defined precisely by deterministic selection of $f_\mu(\gamma)$ at posterior γ, $q_\lambda(f_\mu(\gamma)|\gamma) = 1$. By construction both strategies above satisfy $\lambda, \eta \in \Lambda(\mu, \bar{A})$. Hence we can apply Lemma 2.7, FIO directly to conclude that $\lambda, \eta \in \hat{\Lambda}(\mu, \bar{A}|K)$. Hence they have equal expected utility net of attention costs. Hence the cost difference must be the same as the difference in expected utility,

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

By construction, the inattentive strategy is free, $K(\mu, \eta) = 0$, completing the proof. ■

Appendix 4: Theorem 4

In this section of the appendix we take A2 through A8 and Theorem 3 as our starting point and show that addition of Locally Invariant Posteriors (A9) validates Theorem 4. We first establish some useful Lemmas. We then provide the necessity proof, which follows directly, and finally the sufficiency proof. Before doing this, we restate key definitions to ease reading of the Appendix. A PS cost function $K \in \mathcal{K}^{PS}$ is UPS $K \in \mathcal{K}^{UPS}$, if there exists a strictly convex function $T : \Gamma \to \mathbb{R}$ such that,

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

all $(\mu, Q) \in \mathcal{F}$ such that $Q \in \hat{\mathcal{Q}}(\mu|K) \equiv \mathcal{Q}(\mu) \cap \Delta(\hat{\Gamma}(\mu|K))$, where $\hat{\Gamma}(\mu|K)$ is the optimal posterior set,

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

Recall also that we have special notation for the subset of $\mathcal{F}(\mu)$ consistent with optimality,

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

Given a UPS function, we define also net utility using the common cost function,

$$N^a(\gamma) \equiv \bar{u}(\gamma, a) - T(\gamma). \tag{51}$$

Finally, recall the notation γ^a_λ defined in Lemma 2.3: this is the unique posterior at which any chosen action $a \in \mathcal{A}(\lambda)$ may be chosen, $q_\lambda(a|\ \gamma^a_\lambda) > 0$, in an optimal strategy for a given PS cost function $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}\,(\mu, A|K)$.

A4.1: Lemmas

Lemma 4.1: (UPS Lagrangean Lemma) Given $\lambda = (Q_\lambda, q_\lambda) \in \Lambda\,(\mu, A)$ for $K \in \mathcal{K}^{UPS}$, $\lambda \in \hat{\Lambda}\,(\mu, A|K)$ if and only if $\exists \theta \in \mathbb{R}^{J-1}$ s.t.,

$$N^a(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq \sup_{a' \in A, \gamma' \in \Gamma(\mu)} N^{a'}(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j), \tag{52}$$

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

Proof. By the standard Lagrangian Lemma 2.6 above, since $K \in \mathcal{K}^{UPS}$ implies that $K \in \mathcal{K}^{PS}$, we know that, given $(\mu, A) \in \mathcal{D}$, $\lambda \in \hat{\Lambda}(\mu, A|K)$ if and only if $\exists \theta \in \mathbb{R}^{J-1}$ s.t.,

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

all $\gamma \in \Gamma(\mu)$ and $a \in A$, with equality if $\gamma \in \Gamma(Q_\lambda)$ and $q_\lambda(a|\gamma) > 0$. But with $K \in \mathcal{K}^{UPS}$, we know that, for all μ, this holds also for the fixed function $N^a(\gamma)$ defined in (51), confirming (52). ■

Lemma 4.2: LIP in Optimal Strategies Given $K \in \mathcal{K}^{UPS}$, consider $(\mu, A) \in \mathcal{D}$ and $\lambda =$

$(Q_\lambda, q_\lambda) \in \hat{\Lambda}\,(\mu, A|K)$. Now consider $\rho(a) > 0$ on $A' \subset \mathcal{A}(\lambda)$ with $\sum_{a \in A'} \rho(a) = 1$ and define $\mu' = \sum_{a \in A'} \rho(a)\gamma_\lambda^a$ and $\lambda' = (Q', q') \in \Lambda(\mu', A')$ with $\Gamma(Q') \subset \hat{\Gamma}(\mu'|K)$ by:

$$\begin{aligned} Q'(\gamma) &= \begin{cases} \displaystyle\sum_{\{a \in A' | \gamma_\lambda^a = \gamma\}} \rho(a) \text{ if } \gamma \in \Gamma(Q'); \\ \qquad\qquad 0 \text{ else.} \end{cases} \\ q'(a|\gamma) &= \begin{cases} \frac{\rho(a)}{Q'(\gamma)} \text{ if } \gamma = \gamma_\lambda^a; \\ \quad 0 \text{ else.} \end{cases} \end{aligned}$$

Then $\lambda' \in \hat{\Lambda}(\mu', A'|K)$.

Proof. Consider $(\mu, A) \in \mathcal{D}$ and $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}\,(\mu, A|K)$ for $K \in \mathcal{K}^{UPS}$. Given $K \in \mathcal{K}^{UPS}$ and that $\lambda \in \hat{\Lambda}\,(\mu, A|K)$, we know that $\Gamma(Q_\lambda) \subset \hat{\Gamma}(\mu|K)$ so that $Q_\lambda \in \Delta(\hat{\Gamma}(\mu|K))$. Hence we can use the common strictly convex function $T : \Gamma \to \mathbb{R}$ in computing the corresponding costs,

$$K(\mu, Q_\lambda) = \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma)T(\gamma) - T(\mu).$$

We define N^a to be the corresponding net utility using the common cost function as in (51). Given that $\lambda \in \hat{\Lambda}\,(\mu, A|K)$ for $K \in \mathcal{K}^{PS}$, we can apply Lemma 4.1, the UPS Lagrangian Lemma, to identify multipliers $\theta(j)$ such that,

$$N^b(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \leq N^a(\gamma_\lambda^a) - \sum_{j=1}^{J-1} \theta(j)\gamma_\lambda^a(j),$$

all $\gamma \in \Gamma(\mu)$, $b \in A$, and $a \in \mathcal{A}(\lambda)$. We rewrite the above as an equation and a set of inequalities. Given $a \in \mathcal{A}(\lambda)$,

$$\begin{aligned} N^c(\bar{\gamma}^c) - \sum_{j=1}^{J-1} \theta(j)\gamma_\lambda^c &= N^a(\gamma_\lambda^a) - \sum_{j=1}^{J-1} \theta(j)\gamma_\lambda^a(j) \text{ for } c \in \mathcal{A}(\lambda); \\ N^b(\gamma) - \sum_{j=1}^{J-1} \theta(j)\gamma(j) &\leq N^a(\gamma_\lambda^a) - \sum_{j=1}^{J-1} \theta(j)\gamma_\lambda^a(j) \text{ for } b \in A \text{ and } \gamma \in \Gamma(\mu). \end{aligned}$$

We now consider $\rho(a) > 0$ on $A' \subset \mathcal{A}(\lambda)$ with $\sum_{a \in B} \rho(a) = 1$ and define $\mu' = \sum_{a \in A'} \rho(a)\gamma_\lambda^a$ and $\lambda' = (Q', q') \in \Lambda(\mu', A')$ as in the statement of this Lemma. Given that $\Gamma(Q') \subset \hat{\Gamma}(\mu'|K)$ and $K \in \mathcal{K}^{UPS}$, we know that we can again use the common T function in expressing all net utilities, so that,

$$K(\mu', Q') = \sum_{\gamma \in \Gamma(Q')} Q'(\gamma)T(\gamma) - T(\mu').$$

We now apply the UPS Lagrangian Lemma using the same multipliers. Note that all equalities and inequalities defining of optimality remain valid. The subtlety is that there may be different state spaces, $\Omega(\mu') \neq \Omega(\mu)$, hence the summation and relevant multipliers $\theta(j)$ are only those in the smaller space, $j \in \Omega(\mu')$. The key observation that makes this irrelevant and validates the corresponding inequalities restricted to that subspace is that, for all $\gamma \in \Gamma(Q')$, the posteriors $\gamma(j)$ on all states $j \notin \Omega(\mu')$ are zero. Hence the corresponding terms add nothing to any of the terms on either the left-hand side or right-hand side, leaving the inequalities valid on the smaller state space to conclude that $\lambda' \in \hat{\Lambda}(\mu', A'|K)$, completing the proof. ■

Lemma 4.3: Invariance Under Addition of Affine Functions Consider $K \in \mathcal{K}^{PS}$, $\mu \in \Gamma$, and $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$, such that, given $Q \in \hat{Q}(\mu|K)$,

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

Then if $\tilde{T}_\mu(\gamma) = T_\mu(\gamma) + \alpha + \beta.\gamma$ some $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^{|\Omega(\mu)|}$, then

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

Proof. Immediate. ■

Lemma 4.4: UPS Regularity If $C \in \mathcal{C}$ has a PS representation $K \in \mathcal{K}^{PS}$, and A9 is satisfied, then C is regular, $C \in \mathcal{C}^R$.

Proof. Since $C \in \mathcal{C}$ has a PS representation $K \in \mathcal{K}^{PS}$ we know by Theorem 3 that it satisfies A2 through A8. To establish that satisfaction in addition of A9 yields regularity, we need to show that, in this case, given $\mu_1 \in \Gamma$ and $Q \in \Delta(\Gamma(\mu_1))$ with $\Gamma(Q) \subset \Gamma^C(\mu_1)$,

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

Given such $\mu_1 \in \Gamma$ and $Q \in \Delta(\Gamma(\mu_1))$ with $\Gamma(Q) \subset \Gamma^C(\mu_1)$, we know from Completeness (A4) that there exists a corresponding $(\mu_1, A_1) \in \mathcal{D}$, $P \in C(\mu_1, A_1)$, and $\boldsymbol{\lambda}_P = (\mathbf{Q}_P, \mathbf{q}_P)$ such that $\Gamma(Q)$ is a subset of the support,

$$\Gamma(Q) \subset \Gamma(\mathbf{Q}_P).$$

By construction the probability that action $a \in \mathcal{A}(P) \subset A_1$ is chosen at revealed posterior $\bar{\gamma}_P^a \in \Gamma(\mathbf{Q}_P)$ in strategy $\boldsymbol{\lambda}_P = (\mathbf{Q}_P, \mathbf{q}_P)$ is greater than 0,

$$\mathbf{q}_P(a|\bar{\gamma}_P^a) > 0.$$

Now suppose that $\sum_{\gamma \in \Gamma(\mu_1)} \gamma Q(\gamma) = \mu_2$ and define $A_2 \subset \mathcal{A}(P)$ to comprise the actions chosen at the posteriors in set $\Gamma(Q)$,

$$A_2 = \{a \in A_1 | \bar{\gamma}_P^a \in \Gamma(Q)\}.$$

With LIP (A9) we know that, since $P \in C(\mu_1, A_1)$, $A_2 \subset \mathcal{A}(P)$, and we have found probabilities $Q(\bar{\gamma}_P^a) > 0$ all $a \in A_2$ with $\sum_{a \in A_2} Q(\bar{\gamma}_P^a) = 1$, that the data set $P_2 \in \mathcal{P}\left(\sum_{a \in A_2} \bar{\gamma}_P^a Q(\bar{\gamma}_P^a), A'\right)$ that satisfies $\mathcal{A}(P_2) = A_2$, $\mathbf{Q}_{P_2}(a) = Q(\bar{\gamma}^a_P)$, and $\bar{\gamma}^a_{P_2} = \bar{\gamma}^a_P$ is observed at the corresponding prior $\mu_2 = \sum_{a \in A_2} \bar{\gamma}^a_P Q(\bar{\gamma}^a_P)$,

$$P_2 \in C\left(\mu_2, A'\right).$$

Hence $\Gamma\left(Q\right) \subset \Gamma^C(\mu_2)$, establishing regularity. ∎

Lemma 4.5: Given $C \in \mathcal{C}^R$ with a PS representation $K \in \mathcal{K}^{PS}$,

$$\Omega(\mu_1) = \Omega(\mu_2) \Longrightarrow \Gamma^C(\mu_1) = \Gamma^C(\mu_2).$$

Proof. Given $C \in \mathcal{C}$ with a PS representation, we know by theorem 3 that it satisfies A2 through A8. For the purposes of this proof we set $\Omega(\mu_1) = \Omega(\mu_2) = \Omega$, and know by Completeness (A4) that both contain the common set of interior vectors, from which we correspondingly remove the subscript:

$$\tilde{\Gamma} = \tilde{\Gamma}(\mu_1) = \tilde{\Gamma}(\mu_2) \subset \Gamma^C(\mu_1) \cap \Gamma^C(\mu_2)$$

We now consider an arbitrary posterior $\eta \in \Gamma^C(\mu_1)$ with $\eta(j) = 0$ for some $j \in \Lambda(\mu_1)$ and show that, if $C \in \mathcal{C}^R$, then $\eta \in \Gamma^C(\mu_2)$.

For $1 \leq k \leq J = |\Omega|$ we create corresponding set of interior "basis" vectors constructed in such a manner that they allow us to construct distributions $\bar{Q}_1$ and $\bar{Q}_1$ over them that generate both μ_1 and μ_2. To do this, weight together the unit posteriors $e_k \in \Gamma_1 = \Gamma_2 = \Gamma_{12}$ with 1 in position k and zeroes elsewhere and their average $\bar{e} = \sum_{k=1}^{J} e_k/J$, to arrive at a set of interior posteriors $\bar{\gamma}_k \in \Gamma_{12}$ that span (in the linear algebra sense) the set Γ_{12} and that contain μ_1 and μ_2 in the interior of their convex hull, which is possible since we know that μ_1 and μ_2 are both interior to Γ_{12}. Technically, we find $\delta \in (0, 1)$ such that, when we define the corresponding posteriors $\bar{\gamma}^\delta_k$ for $1 \leq k \leq J$,

$$\bar{\gamma}^\delta_k(j) = \begin{cases} \frac{\delta}{J} + (1-\delta) \text{ if } k = j; \\ \frac{\delta}{J} \text{ if } k \neq j \text{ .} \end{cases}$$

the unique probability weights $\bar{Q}_i(k)$ for $i = 1, 2$ and $1 \leq k \leq J$ with $\sum_{k=1}^{J} \bar{Q}_i(k) = 1$ that re-weight the posteriors to regenerate each prior,

$$\sum_{k=1}^{J} \bar{\gamma}^\delta_k \bar{Q}_i(\bar{\gamma}^\delta_k) = \mu_i;$$

are all strictly positive probability, $\bar{Q}_i(\bar{\gamma}^\delta_k) > 0$ for $i = 1, 2$ and $1 \leq k \leq J$. Going forward we suppress the δ parameter and set $\bar{\gamma}^\delta_k = \bar{\gamma}_k$.

In the next step we adjust the weights $\bar{Q}_2(\bar{\gamma}_k)$ and define a distribution of posteriors $Q_2 \in \Delta\left(\Gamma^C(\mu_1)\right)$ with,

$$\Gamma(Q_2) = \eta \cup \{\bar{\gamma}_k | 1 \leq k \leq J\} \text{ and } \sum_{\gamma \in \Gamma(Q_2)} \gamma Q_2(\gamma) = \mu_2.$$

To accomplish this, we pick $\epsilon > 0$ small enough so that,

$$\max_{k=1..J} \frac{\epsilon\eta(k)}{1-\delta} < \min_{k=1..J} \{\bar{Q}_2(k)\},$$

define $Q_2(\eta) = \epsilon$ and then subtract the corresponding amount from the probability of $\bar{\gamma}_k$,

$$Q_2(\bar{\gamma}_k) = \bar{Q}_2(\bar{\gamma}_k) - \frac{\epsilon}{1-\delta}\left[\eta(k) - \frac{\delta}{J}\right],$$

so that,

$$\begin{aligned} Q_2(\eta) + \sum_{k=1}^{J} Q_2(\bar{\gamma}_k) &= \epsilon + 1 - \frac{\epsilon}{1-\delta}\left[\sum_{k=1}^{J} \eta(k) - \delta\right] \\ &= \epsilon + 1 - \epsilon = \epsilon, \end{aligned}$$

so that this is a probability distribution over posteriors. Note also that,

$$\sum_{\gamma \in \Gamma(Q_2)} \gamma(j) Q_2(\gamma) = \sum_{k=1}^{J} \bar{\gamma}_k(j) \left(\bar{Q}_2(\bar{\gamma}_k) - \left[\frac{\epsilon}{1-\delta}\left(\eta(k) - \frac{\delta}{J}\right)\right]\right) + \eta(j)\epsilon.$$

Note that $\sum_{k=1}^{J} \bar{\gamma}_k(j)\bar{Q}_2(\bar{\gamma}_k) = \mu_2(j)$ so that $\sum_{\gamma\in\Gamma(Q_2)} \gamma(j)Q_2(\gamma) = \mu_2(j)$ if and only if,

$$\sum_{k=1}^{J} \bar{\gamma}_k(j) \left(\left[\frac{1}{1-\delta}\left(\eta(k) - \frac{\delta}{J}\right)\right]\right) = \eta(j). \tag{53}$$

Directly,

$$\begin{aligned} \sum_{k=1}^{J} \bar{\gamma}_k(j) \left(\left[\frac{1}{1-\delta}\left(\eta(k) - \frac{\delta}{J}\right)\right]\right) &= \sum_{k=1}^{J} \frac{\delta}{J}\left[\frac{1}{1-\delta}\left(\eta(k) - \frac{\delta}{J}\right)\right] + (1-\delta)\left[\frac{1}{1-\delta}\left(\eta(j) - \frac{\delta}{J}\right)\right] \\ &= \frac{\delta}{J(1-\delta)} \sum_{k=1}^{J} \left(\eta(k) - \frac{\delta}{J}\right) + (1-\delta)\left[\frac{1}{1-\delta}\left(\eta(j) - \frac{\delta}{J}\right)\right] \\ &= \frac{\delta}{J(1-\delta)}[1-\delta] + \eta(j) - \frac{\delta}{J} = \eta(j), \end{aligned}$$

confirming (53).

Note by construction that $\Gamma(Q_2) \subset \Gamma^C(\mu_1)$. Given that $C \in \mathcal{C}^R$ the fact that $\sum_{\gamma\in\Gamma(Q_2)} \gamma Q_2(\gamma) = \mu_2$ implies that $\Gamma(Q_2) \subset \Gamma^C(\mu_1)$, hence that $\eta \in \Gamma^C(\mu_2)$, so that $\Gamma^C(\mu_1) \subset \Gamma^C(\mu_2)$. Note that the converse argument is identical since the state spaces are identical, so that $\Gamma^C(\mu_1) = \Gamma^C(\mu_2)$, completing the proof. ■

A4.2: Theorem 4

Theorem 4: Necessity If data set $C \in \mathcal{C}^R$ has a UPS representation it satisfies A2 through A9.

Proof. Given that $C \in \mathcal{C}^R$ has a UPS representation, it has a PS representation $K \in \mathcal{K}^{PS}$, and A2-A8 are satisfied. To establish A9, we pick $(\mu, A) \in \mathcal{D}$, $P \in C(\mu, A)$, and probabilities $\rho(a) > 0$ on $A' \subset \mathcal{A}(P)$ with $\sum\limits_{a \in A'} \rho(a) = 1$. Since this is a PS representation and $P \in C(\mu, A)$, the corresponding revealed strategy is optimal by Lemma 2.15,

$$\boldsymbol{\lambda}(P) = (\mathbf{Q}_P, \mathbf{q}_P) \in \hat{\Lambda}(\mu, A|K).$$

By definition, $\boldsymbol{\lambda}(P) = (\mathbf{Q}_P, \mathbf{q}_P)$ is defined by $\Gamma(\mathbf{Q}_P) = \cup_{a \in \mathcal{A}(P)} \bar{\gamma}_P^a$ and:

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

on $\gamma \in \Gamma(\mathbf{Q}_P)$ and $a \in A$. Since $\boldsymbol{\lambda}(P) \in \hat{\Lambda}(\mu, A|K)$, it is definitional that $\Gamma(\mathbf{Q}_P) \subset \Gamma^C(\mu)$.

We now define $P' \in \mathcal{P}$ as in the LIP definition by$\mathcal{A}(P') = A'$, $\mathbf{Q}_{P'}(\gamma) = \sum\limits_{\{a \in A' | \bar{\gamma}_P^a = \gamma\}} \rho(a)$;

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

To show that $P' \in C(\mu', A')$, where,

$$\mu' = \sum_{a \in A'} \rho(a) \bar{\gamma}_P^a,$$

we consider the revealed attention strategy associated with $P' \in C(\mu', A')$, $\boldsymbol{\lambda}(P') = (\mathbf{Q}_{P'}, \mathbf{q}_{P'}) \in \Lambda(\mu, A)$:

$$\begin{aligned} \mathbf{Q}_{P'}(\gamma) &= \begin{cases} \sum\limits_{\{a \in A' | \bar{\gamma}_P^a = \gamma\}} \rho(a) \text{ if } \gamma \in \Gamma(\mathbf{Q}_P); \\ \qquad 0 \text{ else.} \end{cases} \\ \mathbf{q}_{P'}(a|\gamma) &= \begin{cases} \frac{\rho(a)}{Q_\eta(\gamma)} \text{ if } \gamma = \bar{\gamma}_P^a; \\ \quad 0 \text{ else.} \end{cases} \end{aligned}$$

Note that this strategy is derived from $\boldsymbol{\lambda}(P) \in \hat{\Lambda}(\mu, A|K)$ precisely as prescribed in Lemma 4.2, since $\bar{\gamma}_P^a = \gamma_{\boldsymbol{\lambda}(P)}^a \in \Gamma(\mu)$ is indeed the unique posterior with $\mathbf{q}_{P'}(a|\gamma) > 0$ by Lemma 2.12. Hence, provided $\Gamma(\mathbf{Q}_{P'}) \subset \hat{\Gamma}(\mu', K)$ we can conclude from Lemma 4.2 that $\boldsymbol{\lambda}(P') \in \hat{\Lambda}(\mu', A'|K)$. To establish this, note that $\mathbf{Q}_{P'} \in \Delta(\Gamma(\mu))$ satisfies $\Gamma(\mathbf{Q}_{P'}) \subset \Gamma(\mathbf{Q}_P) \subset \Gamma^C(\mu)$, and,

$$\sum_{\gamma \in \Gamma(\mathbf{Q}_{P'})} \gamma \mathbf{Q}_{P'}(\gamma) = \sum_{a \in A'} \rho(a) \bar{\gamma}_P^a = \mu'.$$

Since $C \in \mathcal{C}^R$, we conclude that indeed $\Gamma\left(\mathbf{Q}_{P'}\right) \subset \Gamma^C(\mu')$, so that Lemma 4.2 does apply to ensure that $\boldsymbol{\lambda}(P') \in \hat{\Lambda}(\mu', A'|K)$. Since this is a PS representation, we know further that,

$$\mathbf{P}_{\boldsymbol{\lambda}(P')} \in C(\mu', A').$$

Note finally that by Lemma 2.13,

$$\mathbf{P}_{\boldsymbol{\lambda}(P')} = P',$$

completing the necessity proof. ■

Theorem 4: Sufficiency If data set $C \in \mathcal{C}$ satisfies A2 through A9, it has a UPS representation.

Proof. A2-A8 guarantee existence of a PS representation. To establish existence of a UPS representation, we know that we can identify corresponding functions $K_\mu \in \mathcal{K}^{PS}$ any $\mu \in \Gamma$ and a corresponding strictly convex functions $T_\mu : \Gamma(\mu) \to \bar{\mathbb{R}}$ that is real valued on $\tilde{\Gamma}(\mu)$. We use these functions to define our candidate real-valued function $T(\gamma) \in \mathbb{R}$ on $\gamma \in \Gamma$. Specifically, we define the corresponding uniform prior $\bar{\mu}(\gamma)$ that assigns probability $\frac{1}{|\Omega(\gamma)|}$ to each state in $\Omega(\gamma)$ and specify this as $T(\gamma)$:

$$T(\gamma) \equiv T_{\bar{\mu}(\gamma)}(\gamma), \tag{54}$$

As noted, this is real-valued by definition of a PS representation since $\gamma \in \tilde{\Gamma}(\bar{\mu}(\gamma))$. Note by Lemma 4.3 that $T(\gamma)$ is not unique: but the affine transforms are irrelevant as we will see. We establish now that this definition ensures that the defining property of the UPS representation holds: given $\mu \in \Gamma$,

$$K(\mu, Q) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma)T(\gamma) - T(\mu), \tag{55}$$

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

We establish this result in two stages according to the support. We prove first that (55) would follow provided it held true for the special class of priors that are uniform over a finite set of states. We then establish that indeed (55) does hold for uniform priors.

With regard to the first step, we let μ_1 be a uniform prior over an arbitrary state space, and pick any distinct non-uniform prior with the same state space, so that $\Omega(\mu_1) = \Omega(\mu_2)$. Given that a PS representation exists, we know that we can identify corresponding functions $K_i \equiv K_{\mu_i} \in \mathcal{K}^{PS}$ for $i = 1, 2$. We fix corresponding strictly convex functions $T_i : \Gamma(\mu_i) \to \mathbb{R}$ which have the PS property: for feasible strategies $\lambda = (Q_\lambda, q_\lambda) \in \Lambda(\mu_i)$, using $T_1 = T_{\bar{\mu}(\gamma)}$, the version used in constructing $T(\gamma)$. Correspondingly for $a \in \mathcal{A}$ we define,

$$N_i^a(\gamma) = \sum_{j=1}^{J} u(a, j)\gamma(j) - T_i(\gamma),$$

on $\gamma \in \Gamma(\mu_i)$.

By lemma 4.4 we know that since $C \in \mathcal{C}$ has a PS representation and A9 is satisfied, that C is regular, $C \in \mathcal{C}^R$. With $C \in \mathcal{C}^R$ and $\Omega(\mu_1) = \Omega(\mu_2)$, Lemma 4.5 implies that,

$$\Gamma^C(\mu_1) = \Gamma^C(\mu_2) \equiv \Gamma^C.$$

In light of Lemma 4.3, our goal in this part of the proof is to show that $T_2 : \Gamma(\mu_2) \to \bar{\mathbb{R}}$ represents the addition of an affine function of γ to $T_1 : \Gamma(\mu_2) \to \bar{\mathbb{R}}$, since then the two functions can be reduced to equality.

We first focus on prior μ_1. By Lemma 2.7, FIO, there is a 1-1 function $f_1 : \Gamma^C \to \mathcal{A}$ such that,

$$N_1^{f_1(\gamma)}(\phi) = \sum_{j=1}^{J} u(f_1(\gamma), j)\phi(j) - T_1(\phi) \leq \sum_{j=1}^{J} u(f_1(\gamma), j)\gamma(j) - T_1(\gamma) = N_1^{f_1(\gamma)}(\gamma) \equiv 0, \tag{56}$$

all $\phi, \gamma \in \Gamma$. We find the particular actions associated with the spanning vectors introduced in Lemma 4.5,

$$\bar{a}_k = f_1(\bar{\gamma}_k) \in \mathcal{A},$$

for $1 \leq k \leq J$, where

$$\bar{\gamma}_k(j) = \begin{cases} \frac{\delta}{J} + (1 - \bar{\delta}) \text{ if } k = j; \\ \frac{\bar{\delta}}{J} \text{ if } k \neq j \text{ ;} \end{cases} \tag{57}$$

with $\bar{\delta} \in (0, 1)$ fixed to ensure strict positivity of all weights $\bar{Q}_i(k) > 0$ for $i = 1, 2$ and $1 \leq k \leq J$ with $\sum_{k=1}^{J} \bar{Q}_i(k) = 1$ that re-weight the posteriors to regenerate each prior,

$$\sum_{k=1}^{J} \bar{\gamma}_k \bar{Q}_i(\bar{\gamma}_k) = \mu_i.$$

We define the corresponding action set $\bar{A} = \cup_{k=1}^{J} \bar{a}_k$.

By Lemma 2.7, FIO, we can identify an optimal strategy $\lambda(1) = (Q_1, q_1) \in \hat{\Lambda}(\mu_1, \bar{A}|K_1)$ having posteriors $\Gamma(Q_1) = \cup_{k=1}^{J} \bar{\gamma}_k$, placing probability weights on them according to $\bar{Q}_1$, and involving deterministic choice at each possible posterior of the corresponding action,

$$\begin{aligned} Q_1(\bar{\gamma}_k) &= \bar{Q}_1(k); \\ q_1(\bar{a}_k|\bar{\gamma}_k) &= 1. \end{aligned}$$

We also define strategy $\lambda(2) = (Q_2, q_2)$ as having the same possible posteriors, the same deterministic choice at each possible posterior, yet placing probability weights on them according to Q_2,

$$Q_2(\bar{\gamma}_k) = \bar{Q}_2(k).$$

The key observation is that with A9, $\lambda(2) \in \hat{\Lambda}(\mu_2, \bar{A}|K_2)$. To see this, note first that since this is a CIR $\lambda(1) = (Q_1, q_1) \in \hat{\Lambda}(\mu_1, \bar{A}|K_1)$, the corresponding SDSC data satisfies $\mathbf{P}_{\lambda_1} \in C(\mu_1, \bar{A})$. We now define data set P_2 by,

$$P_2(\bar{a}_k|j) = \frac{\bar{\gamma}_k(j)\bar{Q}_2(k)}{\mu_2(j)},$$

By construction note that this has unconditional action probabilities,

$$P_2(\bar{a}_k) = \sum_{j=1}^{J} \frac{\mu_2(j)\bar{\gamma}_k(j)\bar{Q}_2(k)}{\mu_2(j)} = \sum_{j=1}^{J} \bar{\gamma}_k(j)\bar{Q}_2(k) = \bar{Q}_2(k),$$

and revealed posteriors,

$$\bar{\gamma}_2^k(j) \equiv \bar{\gamma}_{P_2}^{\bar{a}_k}(j) = \frac{\mu_2(j)P_2(\bar{a}_k|j)}{P_2(\bar{a}_k)} = \frac{\mu_2(j)P_2(\bar{a}_k|\omega j)}{\bar{Q}_2(k)} = \bar{\gamma}_k(j).$$

Hence $P_2 \in P(\mu_2, \bar{A})$.

At this point we can apply LIP (A9) to conclude that $P_2 \in C(\mu_2, \bar{A})$ and, by Lemma 2.15, that the related revealed attention strategy $\boldsymbol{\lambda}_{P_2} = \lambda_2$ is optimal, $\lambda_2 \in \hat{\Lambda}(\mu_2, \bar{A}|K_2)$. The Lagrangian Lemma then ensures there are multipliers $\theta \in \mathbb{R}^{J-1}$ s.t., for $1 \leq k, l \leq J$ such that,

$$N_2^{\bar{a}_k}(\bar{\gamma}_k) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_k(j) = N_2^{\bar{a}_l}(\bar{\gamma}_l) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_l(j);$$

or,

$$\sum_{j=1}^{J} u(\bar{a}_k, j)\bar{\gamma}_k(j) - T_2(\bar{\gamma}_k) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_k(j) = \sum_{j=1}^{J} u(\bar{a}_l, j)\bar{\gamma}_l(j) - T_2(\bar{\gamma}_l) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_l(j). \tag{58}$$

By equation (56), we know also that,

$$\begin{aligned} T_1(\bar{\gamma}_k) &= \sum_{j=1}^{J} u(\bar{a}_k, j)\bar{\gamma}_k(j); \text{ and,} \\ T_1(\bar{\gamma}_l) &= \sum_{j=1}^{J} u(\bar{a}_l, j)\bar{\gamma}_l(j). \end{aligned}$$

Substitution in (58) yields,

$$T_1(\bar{\gamma}_k) - T_2(\bar{\gamma}_k) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_k(j) = T_1(\bar{\gamma}_l) - T_2(\bar{\gamma}_l) - \sum_{j=1}^{J-1} \theta(j)\bar{\gamma}_l(j).$$

Hence, for all $1 \leq k, l \leq J$,

$$\sum_{j=1}^{J-1} \theta(j)\left[\bar{\gamma}_k(j) - \bar{\gamma}_l(j)\right] = T_1(\bar{\gamma}_k) - T_2(\bar{\gamma}_k) - T_1(\bar{\gamma}_l) + T_2(\bar{\gamma}_l). \tag{59}$$

The next key claim is that, with Lemma 2.7, FIO, the equation above applies not only to the spanning posteriors but to all pairs of posteriors, $\gamma, \gamma' \in \Gamma$. To see this, set $\gamma = \bar{\gamma}_{J+1}$ and $\gamma' = \bar{\gamma}_{J+2}$ and repeat the above argument to a larger set of posteriors $\cup_{k=1}^{J+2}\bar{\gamma}_k$ and the corresponding actions defined by $f_1 : \Gamma^C \to \mathcal{A}$ as defined by Lemma 2.7, FIO and defined above:

$$\bar{B} = \bar{A} \cup f_1(\gamma) \cup f_1(\gamma').$$

We also find strictly positive probability weights $Q_i'(k)$ for $i = 1, 2$ on $1 \leq k \leq J+2$ that re-weight the posteriors to regenerate each prior,

$$\sum_{k=1}^{J+2} \bar{\gamma}_k Q_i'(k) = \mu_i.$$

This is possible because the vectors $\bar{\gamma}_k$ span Γ_{12}, so that there are weights $\alpha(k)$ and $\alpha'(k)$ on them that average back to each of γ,γ':

$$\begin{aligned} \sum_{k=1}^{J} \alpha(k)\bar{\gamma}_k &= \gamma; \\ \sum_{k=1}^{J} \alpha'(k)\bar{\gamma}_k &= \gamma' \end{aligned}$$

Note also these weights must sum to 1, as

$$1 = \sum_{\omega\in\Omega} \gamma(\omega) = \sum_{\omega\in\Omega}\sum_{k=1}^{J} \alpha(k)\bar{\gamma}_k(\omega) = \sum_{k=1}^{J} \alpha(k) \sum_{\omega\in\Omega} \bar{\gamma}_k(\omega) = \sum_{k=1}^{J} \alpha(k).$$

Moreover, for all $\epsilon > 0$ and for $i = 1, 2$,

$$\epsilon\left(\gamma + \gamma'\right) + \sum_{k=1}^{J} \gamma_k \left[\bar{Q}_i(k) - \epsilon\left[\alpha(k) + \alpha'(k)\right]\right] = \mu_i.$$

Given that $\bar{Q}_i(k) > 0$ all k, we can select ϵ small enough to keep all terms

$$\bar{Q}_i(k) - \epsilon\left[\alpha(k) + \alpha'(k)\right],$$

strictly positive, as required. Thus, we define new weights by setting $Q_i'(k)$ equal to the above expression for $1 \leq k \leq J$, and equal to ϵ for $J+1$ and $J+2$. Repeating the entire remainder of the argument, we apply the Lagrangian Lemma to ensure the existence of multipliers defined by $\eta \in \mathbb{R}^{J-1}$ that produce the corresponding equality for all revealed posteriors, hence in particular for γ and γ':

$$T_1(\gamma) - T_2(\gamma) - \sum_{j=1}^{J-1} \eta(j)\gamma(j) = T_1(\gamma') - T_2(\gamma') - \sum_{j=1}^{J-1} \eta(j)\gamma'(j).$$

A key observation is that the multipliers on the larger set are identical to those on the smaller set, $\eta = \theta$. To see this, note that the equality conditions defining $\theta(j)$ also characterize $\eta(j)$ and have a unique solutions. Specifically, setting $k = j$ and $l = J$, note that posteriors $\bar{\gamma}_j$ and $\bar{\gamma}_J$ differ by $\delta > 0$ in coordinates j and J and are otherwise the same. Hence,

$$\sum_{k=1}^{J-1} \theta(k)\left[\bar{\gamma}_j(k) - \bar{\gamma}_J(k)\right] = \delta\theta(j).$$

This allows us to precisely pin down $\theta(j)$ in terms of the given functions $T_1(\bar{\gamma})$ and $T_2(\bar{\gamma})$ and $\bar{\delta}$ as defined in (57),

$$\theta(j) = \frac{T_1(\bar{\gamma}_j) - T_2(\bar{\gamma}_j) - T_1(\bar{\gamma}_J) + T_2(\bar{\gamma}_J)}{\bar{\delta}},$$

with the corresponding being true for η. This implies that indeed,

$$\begin{aligned} T_2(\gamma) &= T_1(\gamma) - \left[T_1(\gamma') - T_2(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j) \right] - \sum_{j=1}^{J-1} \theta(j)\gamma(j) \\ &= T_1(\gamma) + H_{12}(\gamma') - \theta.\gamma, \end{aligned}$$

where $H_{12}(\gamma') = -\left[T_1(\gamma') - T_2(\gamma') - \sum_{j=1}^{J-1} \theta(j)\gamma'(j)\right] \in \mathbb{R}$ is independent of γ. This establishes $T_2 : \Gamma(\mu_2) \to \bar{\mathbb{R}}$ can be obtained by adding an affine function of γ to $T_1 : \Gamma(\mu_2) \to \bar{\mathbb{R}}$ so that by Lemma 4.3 we can define,

$$T_2'(\gamma) = T_2(\gamma) + H_{12}(\gamma') - \theta.\gamma = T_1(\gamma);$$

without changing the cost of any attention strategies.

What we have now established is that provided the cost function obeys,

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

for all $\bar{\mu} \in \Gamma$ and $(\bar{\mu}, Q) \in \hat{\mathcal{F}}(\bar{\mu}|K)$ that are uniform over some state space, then it holds for all $\mu \in \Gamma$ and $(\mu, Q) \in \hat{\mathcal{F}}(\bar{\mu}|K)$. The subtlety here is that the definition of $T(\gamma)$ adjusts with the cardinality of the support of γ. Hence the question is whether or not one can use the function associated with the lower dimensional prior in characterizing the corresponding cost of $\gamma \in \hat{\Gamma}(\bar{\mu}|K)$ with $|\Omega(\gamma)| < |\Omega(\bar{\mu})|$. This is what we now establish.

Our method of proof ignores the particulars of the state space and involves priors $\mu_1, \mu_2 \in \Omega$ such that $\Omega_2 = \Omega(\mu_2) \subset \Omega(\mu_1)$ with $\Omega(\mu_2) \neq \Omega(\mu_1)$. We define $\Gamma_1 = \Omega(\mu_1)$ and show that we can replace $T_1 : \Gamma_1 \to \bar{\mathbb{R}}$ (real-valued on Γ_1^C) by function $\bar{T}_1 : \Gamma_1 \to \bar{\mathbb{R}}$ that not only retains the *PS* property,

$$K_1(Q_\lambda) = \sum_{\gamma \in \Gamma(Q_\lambda)} Q_\lambda(\gamma)\bar{T}_1(\gamma) - \bar{T}_1(\mu_1), \tag{60}$$

but is also equal to T_2 on posteriors also on $\gamma \in \Gamma_1^C \cap \tilde{\Gamma}(\mu_2)$. We will use Γ_{12}^C to refer to the set of such posteriors,

$$\Gamma_{12}^C = \Gamma_1^C \cap \tilde{\Gamma}(\mu_2).$$

Note first that $\Gamma_{12}^C \subset \Gamma_2^C$, so that,

$$\Gamma_{12}^C \subset \Gamma_1^C \cap \Gamma_2^C.$$

This comes directly from the fact that all $\gamma \in \Gamma_{12}^C$ are interior to Ω_2 by construction, and so, by Completeness (A4) are used in some problem.

It will be convenient to index the states by first assigning indices $1, ..., J_2 - 1$ to states from Ω_2. We next assign indices $J_2 ... J_1 - 1$ to states from $\Omega(\mu_1)/\Omega(\mu_2)$. Finally, we assign to index J_1 the remaining state from Ω_2. This ensures that the "excluded state" from the Lagrangian statements belongs to Ω_2.

In establishing the existence of $\bar{T}_1$ such that (60) holds, we start as in the equal state space case with the function $f_1 : \Gamma_1^C \rightarrow \mathcal{A}$ such that,

$$\sum_{j=1}^{J} u(f_1(\gamma), j)\gamma(j) \equiv T_1(\gamma),$$

all $\gamma \in \Gamma_1$, with $\sum_{j=1}^{J} u(f_1(\gamma), j)\phi(j) \leq T_1(\phi)$ all $\phi \in \Gamma_1$. We also retain the spanning posteriors $\bar{\gamma}_k$ and strictly positive probabilities $\bar{Q}_1(k) > 0$ for $1 \leq k \leq J_1$ that regenerate the prior,

$$\sum_{k=1}^{J_1} \bar{\gamma}_k \bar{Q}_1(k) = \mu_1;$$

and the particular actions $\bar{a}_k = f_1(\bar{\gamma}_k) \in \mathcal{A}$ associated with the spanning vectors.

The key change in the proof is the selection of additional posteriors that sit in Γ_{12}^C and a corresponding set of new strictly positive probability weights. Specifically, we pick a basis for the set Γ_{12}^C and label the finite set of such posteriors as $\Gamma^B \subset \Gamma_{12}^C$. We also associate with these the corresponding average prior, $\bar{\mu}^B$,

$$\bar{\mu}^B(j) = \frac{\sum_{\gamma \in \Gamma^B} \gamma(j)}{|\Gamma^B|},$$

for $1 \leq j \leq J_{12}$. Since $\Omega_2 = \Omega(\bar{\mu}_B)$ we know from the first part of the proof that we can assume that the cost function $K(\bar{\mu}_B, .)$, associated $T_{\bar{\mu}_B}$ and $\Gamma^C(\bar{\mu}_B)$ are identical to $K(\mu_2, .)$, T_2 and $\Gamma^C(\mu_2)$. Below, we will therefore substitute the latter for the former. We identify also the actions $\bar{a}(\gamma) = f_1(\gamma)$ on $\gamma \in \Gamma^B$ and define the larger set of actions,

$$\bar{A}' = \cup_{k=1}^{J_1} \bar{a}_k \cup \{\bar{a}(\gamma) | \gamma \in \Gamma^B\}.$$

By construction, note that $\Omega(\bar{\mu}^B) = \Omega_2$ since a positive posterior probability of a state cannot be generated if all of the basis priors assign it zero probability. The new strictly positive probability weights that we work with place weight $Q_1'(k) > 0$ on all posteriors $\bar{\gamma}_k$ as well as a constant strictly positive weight $\bar{Q}_1'(\gamma) = \epsilon > 0$ on all $\gamma \in \bar{\mu}^B$, while still averaging out to μ_1:

$$\sum_{k=1}^{J_1} \bar{\gamma}_k \bar{Q}_1'(k) + \sum_{\gamma \in \Gamma^B} \epsilon\gamma = \mu_1.$$

The easiest way to do this is to assign a small probability $\delta > 0$ to the mean prior $\bar{\mu}^B$, compensating through appropriate reductions in $\bar{Q}_1(k) > 0$ for $1 \leq k \leq J_1$, as in the construction in Lemma 4.4, while retaining all strictly positive, and thereupon defining $\epsilon = \frac{\delta}{|\Gamma^B|}$.

We now note as before that existence of a PS representation and application of Lemma 2.7, FIO enables us to characterize an optimal strategy $\lambda_1' = (Q_1', q_1') \in \hat{\Lambda}(\mu_1, \bar{A}' | K_1)$ having posteriors $\Gamma(Q_1') = \cup_{k=1}^{J} \bar{\gamma}_k \cup \Gamma^B$, placing probability weights on them according to $\bar{Q}_1'$, and involving deterministic choice at each possible posterior of the corresponding action,

$$\begin{aligned} Q'_1(\bar{\gamma}_k) &= \bar{Q}'_1(k); \\ q'_1(\bar{a}(\gamma)|\gamma) &= 1. \end{aligned}$$

We also define strategy $\lambda'_2 = (Q'_2, q'_2)$ as restricting the posterior set to Γ^B, with equal probability weights, $Q'_2(\gamma) = \frac{1}{|\Gamma^B|}$, with the same deterministic choice at each possible posterior. By construction, $\lambda'_2 \in \Lambda(\bar{\mu}^B, \bar{A}')$. In fact, with exactly the same logic as before, LIP (A9) implies that it is optimal, $\lambda'_2 \in \hat{\Lambda}(\bar{\mu}^B, \bar{A}'|K_2)$. Hence we can repeat the application of the UPS Lagrangian Lemma to identify $\theta \in \mathbb{R}^{J_2-1}$ s.t., for $\bar{\gamma}, \bar{\gamma}' \in \Gamma^B$,

$$T_1(\bar{\gamma}) - T_2(\bar{\gamma}) - \sum_{j=1}^{J_2-1} \theta(j)\bar{\gamma}(j) = T_1(\bar{\gamma}') - T_2(\bar{\gamma}') - \sum_{j=1}^{J_2-1} \theta(j)\bar{\gamma}'(j)$$

The next key claim is that the equation above applies not only to the spanning posteriors $\bar{\gamma}, \bar{\gamma}' \in \Gamma^B$ but to all pairs of posteriors $\gamma, \gamma' \in \Gamma^C_{12}$. To see this, we repeat the above argument on a larger set of actions,

$$\bar{B}' = \bar{A}' \cup f_1(\gamma) \cup f_1(\gamma'),$$

using precisely the same procedure as before. Repeating the entire remainder of the argument, we apply the Lagrangian Lemma to ensure the existence of multipliers defined by $\eta \in \mathbb{R}^{J-1}$ that produce the corresponding equality for all revealed posteriors, hence in particular for γ and γ':

$$T_1(\gamma) - T_2(\gamma) - \sum_{j=1}^{J_2-1} \eta(j)\gamma(j) = T_1(\gamma') - T_2(\gamma') - \sum_{j=1}^{J_2-1} \eta(j)\gamma'(j).$$

For later purposes it is convenient to rewrite this as,

$$T_1(\gamma) - T_2(\gamma) - \left[T_1(\gamma') - T_2(\gamma')\right] = \sum_{j=1}^{J_2-1} \eta(j)\left[\gamma(j) - \gamma'(j)\right].$$

A key observation is that we can set $\eta = \theta$, so that,

$$T_1(\gamma) - T_2(\gamma) - \left[T_1(\gamma') - T_2(\gamma')\right] = \sum_{j=1}^{J_2-1} \theta(j)\left[\gamma(j) - \gamma'(j)\right]. \tag{61}$$

This follows from the fact that both η and θ work for the set of basis posteriors. For $\bar{\gamma}, \bar{\gamma}' \in \Gamma^B$,

$$\begin{aligned} T_1(\bar{\gamma}) - T_2(\bar{\gamma}) - \sum_{j=1}^{J_2-1} \theta(j)\bar{\gamma}(j) &= T_1(\bar{\gamma}') - T_2(\bar{\gamma}') - \sum_{j=1}^{J_2-1} \theta(j)\bar{\gamma}'(j); \\ T_1(\bar{\gamma}) - T_2(\bar{\gamma}) - \sum_{j=1}^{J_2-1} \eta(j)\bar{\gamma}(j) &= T_1(\bar{\gamma}') - T_2(\bar{\gamma}') - \sum_{j=1}^{J_2-1} \eta(j)\bar{\gamma}'(j). \end{aligned}$$

Subtraction yields,

$$\sum_{j=1}^{J_2-1}\left[\eta(j)-\theta(j)\right]\left[\bar{\gamma}(j)-\bar{\gamma}'(j)\right]=0. \tag{62}$$

Since the set Γ^B spans Γ_{12}^C, we know that, given $\gamma,\gamma'\in\Gamma_{12}^C$ there exists weights $\rho(\bar{\gamma})$ and $\rho'(\bar{\gamma})\in\mathbb{R}$ on $\bar{\gamma}\in\Gamma^B$ with $\sum_{\bar{\gamma}\in\Gamma^B}\rho(\bar{\gamma})=\sum_{\bar{\gamma}\in\Gamma^B}\rho'(\bar{\gamma})=1$ and,

$$\gamma=\sum_{\bar{\gamma}\in\Gamma^B}\rho(\bar{\gamma})\bar{\gamma} \text{ and } \gamma'=\sum_{\bar{\gamma}\in\Gamma^B}\rho'(\bar{\gamma})\bar{\gamma}.$$

Hence,

$$\begin{aligned}\sum_{j=1}^{J_2-1}\left[\eta(j)-\theta(j)\right]\left[\gamma(j)-\gamma'(j)\right] &= \sum_{j=1}^{J_2-1}\left[\eta(j)-\theta(j)\right]\left[\sum_{\bar{\gamma}\in\Gamma^B}\left(\rho(\bar{\gamma})\ -\rho'(\bar{\gamma})\right)\bar{\gamma}\right]\\ &= \sum_{\bar{\gamma}\in\Gamma^B}\left(\rho(\bar{\gamma})\ -\rho'(\bar{\gamma})\right)\sum_{j=1}^{J_2-1}\left[\eta(j)-\theta(j)\right]\bar{\gamma}=0.\end{aligned}$$

The final line above follows because all terms $\sum_{j=1}^{J_2-1}\left[\eta(j)-\theta(j)\right]\bar{\gamma}$ on the RHS are equal across $\bar{\gamma}\in\Gamma^B$ (by equation 62) and

$$\sum_{\bar{\gamma}\in\Gamma^B}\left(\rho(\bar{\gamma})\ -\rho'(\bar{\gamma})\right)=0.$$

Hence,

$$\sum_{j=1}^{J_2-1}\eta(j)\left[\gamma(j)-\gamma'(j)\right]=\sum_{j=1}^{J_2-1}\theta(j)\left[\gamma(j)-\gamma'(j)\right].$$

Substitution yields,

$$T_1(\gamma)-T_2(\gamma)=T_1(\gamma')-T_2(\gamma')+\sum_{j=1}^{J_2-1}\theta(j)\left[\gamma(j)-\gamma'(j)\right],$$

hence,

$$T_1(\gamma)-T_2(\gamma)-\sum_{j=1}^{J_2-1}\theta(j)\gamma(j)=T_1(\gamma')-T_2(\gamma')-\sum_{j=1}^{J_2-1}\theta(j)\gamma'(j),$$

verifying equation (61).

As before, we can define $\bar{T}_1:\Gamma^C(\mu_1)\to\mathbb{R}$ as

$$\bar{T}_1(\gamma)=T_1(\gamma)+H_{12}(\gamma')+\theta.\gamma$$

where we define $H_{12}(\gamma')$ as the number $T_2(\gamma')-T_1(\gamma')+\sum_{j=1}^{J_2-1}\theta(j)\gamma'(j)$, and

$$\theta.\gamma=\sum_{j=1}^{J_1-1}\theta(j)\gamma'(j)$$

with $\theta(j) = 0$ for $j > J_2 - 1$. Note that, for $\gamma \in \Gamma_{12}^C$ we have $T_2(\gamma) = \bar{T}_1(\gamma)$, as required. Finally, note that for $i \neq i'$

$$\begin{aligned} K'(\mu_1, Q) &= \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\bar{T}_1(\gamma) - T_1'(\mu_1) = \sum_{\gamma \in \Gamma(Q)} Q(\gamma)\left[T_1(\gamma) + H_{12}(\gamma') - \theta.\gamma\right] - \left[T_1(\mu_1) + H_{12}(\gamma') - \theta.\mu_1\right] \\ &= \sum_{\gamma \in \Gamma(Q)} Q(\gamma)T_1(\gamma) - T_1(\mu_1) - \theta.\left[\sum_{\gamma \in \Gamma(Q)} \gamma Q(\gamma) - \mu_1\right] = K(\mu_1, Q), \end{aligned}$$

as required. This completes the proof that

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

for all $\bar{\mu} \in \Gamma$ and $(\bar{\mu}, Q) \in \hat{\mathcal{F}}(\bar{\mu}|K)$ that are uniform over some state space, and with it the sufficiency proof. ■

0.1 UPS and Non-Regular Data

We illustrate a non-regular data set that suggests a path forward to generalizing the necessity aspect of the UPS theorem to cover all data sets, even those that are not regular. Consider the cost function $K(\mu, Q)$ defined as follows. It is the Tsallis cost function (defined below) with $\sigma = 2$ in all cases except when the prior specifies only two states are possible, and the strategy involves a posterior that rules out one of these states. In such cases the cost is infinite.

$$K(\mu, Q) = \begin{cases} \infty & \text{if } |\Omega(\mu)| = 2 \text{ and } \exists \gamma \in \Gamma(Q) \text{ with } |\Omega(\gamma)| = 1 \\ K_{2}^{TS}(\mu, Q) & \text{otherwise} \end{cases}$$

Note first that the simpler cost function $K_{2}^{TS}(\mu, Q)$ allows the use of all posteriors from any prior, $\Gamma^C(\mu) = \Gamma(\mu)$, and is UPS. The amended cost function is therefore also UPS, as the only change is that now, for priors such that $|\Omega(\mu)| = 2$, the DM will never choose to become fully informed, thus for such priors, strategies that involve degenerate posteriors do not appear in $\mathcal{F}^C(\mu)$.

However, with the amendment to infinite cost, LIP (Axiom A9) is no longer satisfied: the corresponding behavioral data associated with optimal choices violates the axiom. For priors such that $|\Omega(\mu)| > 2$, one can find an action set such that it is optimal to be fully informed, so that all posteriors are the unit vectors, for a prior that makes (say) three states equally likely. LIP (Axiom A9) requires that one see chosen the corresponding certain posteriors in the associated problem with the prior adjusted so that only two states ex ante possible. However this is not consistent with the infinite cost specified for fully informed posteriors when used from such priors.

Thus, this is an example of a UPS model which does not satisfy LIP (Axiom A9). Note, however, that the data produced by this model also violates regularity. As described above, degenerate posteriors which are used when there are three states in the prior and feasible when there are only two states are not used. Regularity rules out exactly this type of problem, and as a result means that LIP (Axiom A9) is implied by the UPS model.

One might at first think that this cost function produces a non-regular data set that satisfies UPS but not LIP, hence ruling out the possibility that the UPS theorem can be generalized. However this is false. In fact, the data set does not permit of a costly information representation at all. There are decision problems for which optimal choices do not exist. Specifically this is the case when the marginal utility of identifying the true state in a two state problem makes it ever more profitable to arrive at certainty, while the discontinuity in cost makes the limit strategy of discretely lower value, giving rise only to ϵ-optimal strategies.

We conjecture that this is a general phenomenon: that any effort to construct a regular data set based on a UPS model that does not satisfy LIP gives rise instead to a model in which there are action sets such that optimal strategies do not exist. We conjecture also that closedness of the convex function (in the sense of Rockafellar) is necessary and sufficient for this.

Appendix 5: Theorem 1

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

In this appendix we go beyond the UPS case and characterize the Shannon function. We take Theorem 4 as established and show that addition of IUC (A1) is equivalent to the UPS representation being of Shannon form. We first show that if C has a Shannon representation, it satisfies IUC: this is a straight forward implication of existing characterizations of optimal strategies. The sufficiency proof is far more involved. Given their centrality we re-state the defining features of basic forms of a decision problem.

Definition 3 *We associate $(\mu, A) \in \mathcal{D}$ with a set of **basic forms** $(\bar{\mu}, A) \in \mathcal{B}$ by:*

1. *Partitioning $\Omega(\mu)$ into L **basic sets** $\left\{\Omega^l(\mu)\right\}_{1 \le l \le L}$ comprising payoff equivalent states, so that, given $\omega \in \Omega^l(\mu)$ and $\omega' \in \Omega^m(\mu)$,*
$$l = m \text{ iff } u(a,\omega) = u(a,\omega') \text{ all } a \in A.$$
2. *For $1 \le l \le L$, defining $I(l) = |\Omega^l(\mu)|$, and indexing by i the states $\omega_i^l \in \Omega^l(\mu)$, so that:*
$$\Omega^l(\mu) = \{\omega_i^l\{\omega_i^l \in \Omega(\mu)|1 \le i \le I(l)\}.$$
3. *Selecting $\bar{\imath}(l) \in \{1,..,I(l)\}$ all l and defining $\bar{\Omega}(\mu) = \cup_{l=1}^L \omega_{\bar{\imath}(l)}^l$.*
4. *Defining $\bar{\mu} \in \Gamma$ by:*
$$\bar{\mu}(\omega_i^l) = \begin{cases} \sum_{j=1}^{I(l)} \mu(\omega_j^l) \text{ if } i = \bar{\imath}(l); \\ 0 \text{ if } i \ne \bar{\imath}(l). \end{cases}$$

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

A5.1: Necessity

Note that the necessity aspect of theorem 1 can be simplified to the statement that a data set with a Shannon representation must satisfy IUC: one need not condition on a UPS representation, since a Shannon representation is a special form of UPS representation.

Theorem 1: Necessity If data set $C \in \mathcal{C}$ has a Shannon representation, it satisfies IUC (A1).

Proof. Consider data set $C \in \mathcal{C}$ that has a Shannon representation K^S_κ, where $\kappa > 0$ is the Shannon multiplicative parameter. Now consider $(\mu, A) \in \mathcal{D}$ and $(\bar{\mu}, A) \in \mathcal{B}(\mu, A)$ for $\bar{\imath}(l)$, $1 \leq l \leq L$,

$$C(\mu, A) = \{P \in \mathcal{P}(\mu, A) | \exists \bar{P} \in C(\bar{\mu}, A) \text{ s.t. } P(a|\omega^l_i) = \bar{P}\left[a|\omega^l_{\bar{\imath}(l)}\right] \text{ all } 1 \leq i \leq I(l),\ 1 \leq l \leq L\}. \tag{63}$$

To establish that IUC holds, we show that the LHS and RHS sets in (63) are mutual subsets. Note the defining feature, which is that utilities to all actions within a given equivalence class are identical in each equivalence class: for each l and for any $1 \leq i, j \leq I(l)$,

$$u(a, \omega^l_i) = u(a, \omega^l_j) \equiv u(a, l).$$

To establish (63) we apply known necessary and sufficient conditions for optimality. Matejka and McKay [2015] (their Corollary 1) show that transformed utilities play a key role,

$$z(a, l) = z(a, \omega^l_i) = \exp^{\frac{u(a, \omega^l_{i\_})}{\kappa}}.$$

The key observation of Matejka and McKay [2015] is that a feasible policy $\lambda \in \Lambda(\mu, A)$ satisfies $\lambda \in \hat{\Lambda}(\mu, A|K^S_\kappa)$ if and only if $\mathbf{P}_\lambda = P$ is a maximizer on $P \in \mathcal{P}(\mu, A)$ of,

$$\sum_{l=1}^{L} \sum_{i=1}^{I(l)} \mu(\omega^l_i) \left(\sum_{a \in A} P(a|\omega^l_i) u(a, \omega^l_i) \right) - \kappa \left[\sum_{l=1}^{L} \sum_{i=1}^{I(l)} \mu(\omega^l_i) \left(\sum_{a \in A} P(a|\omega^l_i) \ln P(a|\omega^l_i) \right) - \sum_{a \in A} P(a) \ln P(a) \right],$$

and,

$$P(a) = \sum_{l=1}^{L} \sum_{i=1}^{I(l)} \mu(\omega^l_i) P(a|\omega^l_i).$$

The necessary (Matejka and McKay [2015]) and sufficient (Caplin *et al.* [2016]) conditions for this are:

$$\sum_{l=1}^{L} \sum_{i=1}^{I(l)} \frac{z(a, \omega^l_i) \mu(\omega^l_i)}{\sum_{b \in A} P(b) z(b, l)} \leq 1, \text{ all } a \in A;$$

with equality for $a \in A$ such that $P(a) > 0$, and,

$$P(a|\omega^l_i) = \frac{P(a) z(a, l)}{\sum_{b \in A} P(b) z(b, l)}.$$

By definition of a Shannon representation,

$$C(\mu, A) = \{P \in \mathcal{P}(\mu, A) | \exists \lambda \in \hat{\Lambda}(\mu, A|K^S_\kappa) \text{ with } P = \mathbf{P}_\lambda\}.$$

To show the set inclusion,

$$C(\mu, A) \subset \{P \in \mathcal{P}(\mu, A) | \exists \bar{P} \in C(\bar{\mu}, A) \text{ s.t. } P(a|\omega_i^l) = \bar{P}(a|\omega_{\bar{\imath}(l)}^l) \text{ all } 1 \leq i \leq I(l),\ 1 \leq l \leq L\},$$

we consider $P \in C(\mu, A)$. Since the data has a Shannon representation, there exists an optimal policy $\lambda \in \hat{\Lambda}(\mu, A|K_\kappa^S)$ with $\mathbf{P}_\lambda = P$ satisfying the optimality conditions. We now define $\bar{P} \in \mathcal{P}(\bar{\mu}, A)$ as above to satisfy the stated condition

$$\bar{P}(a|\omega_{\bar{\imath}(l)}^l) = P(a|\omega_i^l).$$

all $1 \leq i \leq I(l)$, $1 \leq l \leq L$.

Given $a \in A$, we know that

$$\begin{aligned} \bar{P}(a) &= \sum_{l=1}^{L} \bar{\mu}(\omega_{\bar{\imath}(l)}^l) \bar{P}(a|\omega_{\bar{\imath}(l)}^l) = \sum_{l=1}^{L} \sum_{i=1}^{I(l)} \mu(\omega_i^l) P(a|\omega_{\bar{\imath}(l)}^l) \\ &= \sum_{l=1}^{L} \sum_{i=1}^{I(l)} \mu(\omega_i^l) P(a|\omega_i^l) = P(a). \end{aligned}$$

Moreover,

$$\begin{aligned} \sum_{l=1}^{L} \frac{z(a,l)\bar{\mu}(\omega_{\bar{\imath}(l)}^l)}{\sum_{b \in A} \bar{P}(b) z(b,l)} &= \sum_{l=1}^{L} \frac{z(a,l) \sum_{i=1}^{I(l)} \mu(\omega_i^l)}{\sum_{b \in A} P(b) z(b,l)} \\ &= \sum_{l=1}^{L} \sum_{i=1}^{I(l)} \frac{z(a,l)\mu(\omega_i^l)}{\sum_{b \in A} P(b) z(b,l)}. \end{aligned}$$

This implies that both of the conditions defining this as data of the form $\mathbf{P}_{\bar{\lambda}} = \bar{P}$ for an optimal strategy $\bar{\lambda} \in \hat{\Lambda}(\bar{\mu}, A|K_\kappa^S)$ are met:

$$\sum_{l=1}^{L} \frac{z(a,l)\bar{\mu}(\omega_{\bar{\imath}(l)}^l)}{\sum_{b \in A} P(b) z(b,l)} = \sum_{l=1}^{L} \sum_{i=1}^{I(l)} \frac{z(a,l)\mu(\omega_i^l)}{\sum_{b \in A} P(b) z(b,l)} = 1;$$

with the corresponding weak inequality applying to all actions. In fact we can identify an optimal strategy that produces this data as $\boldsymbol{\lambda}(\bar{P}) \in \hat{\Lambda}(\bar{\mu}, A|K_\kappa^S)$. By Lemma 2.13,

$$\mathbf{P}_{\boldsymbol{\lambda}(\bar{P})} = \bar{P}.$$

Since the data has a Shannon representation, this implies that $\bar{P} \in C(\bar{\mu}, A)$.

Analogous reasoning works in the converse direction. We consider $\bar{P} \in C(\bar{\mu}, A)$ and define $P \in \mathcal{P}(\mu, A)$. Since the data has a Shannon representation, there exists an optimal policy $\bar{\lambda} \in \hat{\Lambda}(\mu, A|K_\kappa^S)$ with $\mathbf{P}_{\bar{\lambda}} = \bar{P}$ satisfying the optimality conditions. We now define $P \in \mathcal{P}(\bar{\mu}, A)$ as above to satisfy the stated condition,

$$P(a|\omega_i^l) = \bar{P}(a|\omega_{\bar{\imath}(l)}^l).$$

all $1 \leq i \leq I(l)$, $1 \leq l \leq L$. Given $a \in A$, we run the above equations in reverse to confirm that unconditional probabilities are not affected,

$$P(a) = \bar{P}(a).$$

By precisely the reverse string of equations we find,

$$\sum_{l=1}^{L} \sum_{i=1}^{I(l)} \frac{z(a,l)\mu(\omega_i^l)}{\sum_{b \in A} P(b) z(b,l)} = \sum_{l=1}^{L} \frac{z(a,l)\bar{\mu}(\omega_{\bar{i}(l)}^l)}{\sum_{b \in A} \bar{P}(b) z(b,l)},$$

all $a \in A$. Again this implies that both the equality and the inequality conditions defining this as data of the form $\mathbf{P}_\lambda = P$ for an optimal strategy $\lambda \in \hat{\Lambda}(\mu, A|K_\kappa^S)$ an optimal strategy are met. Again we can identify the optimal strategy that produces this data as that it induces in the data, $\boldsymbol{\lambda}(P) \in \hat{\Lambda}(\mu, A|K_\kappa^S)$, so that $P \in C(\bar{\mu}, A)$ as required. This completes the proof of (63), and with it establishes that any data set $C \in \mathcal{C}$ that has a Shannon representation satisfies IUC (A1). ■

A5.2: Lemmas for Sufficiency

We now develop the machinery required to prove sufficiency: if a data set $C \in C$ with a UPS representation satisfies IUC, it has a Shannon representation. There are many distinct aspects to this proof. In what follows, we will let $K \in \mathcal{K}^{UPS}$ be the UPS representation that is known to exist, and let $T : \Gamma \to \mathbb{R}$ be a strictly convex function such that

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

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

In the first set of lemmas we establish implied symmetry properties. In the second we analyze differentiability and additive separability. We then establish a PDE that characterizes the representation interior to state spaces of dimension 4 or higher. This sets up the proof itself, which analyzes this PDE and considers links between problems of different dimensions. The results display the many different aspects and great power of IUC.

A5.2.1: Symmetry: Definitions and Results

In this subsection we introduce and demonstrate the powerful symmetry implications of IUC (A1). We begin by defining symmetry.

Definition 4 *Beliefs $\gamma_1, \gamma_2 \in \Gamma$ are* ***symmetric****,*

$$\gamma_1 \sim_\Gamma \gamma_2,$$

if there exists a bijection $\sigma : \Omega(\gamma_1) \to \Omega(\gamma_2)$ such that, for all $\omega \in \Omega(\gamma_1)$,

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

Two decision problems are ***symmetric,***

$$(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2),$$

if $\mu_1 \sim_\Gamma \mu_2$ based on bijection $\sigma : \Omega(\gamma_1) \to \Omega(\gamma_2)$ and there exists a bijection $\phi : A_1 \to A_2$ such that,

$$u(a, \omega) = u(\phi(a), \sigma(\omega)),$$

all $\omega \in \Omega(\gamma_1)$. Two decision problems with the same prior $\mu \in \Gamma$ are ***equivalent****, $(\mu, A_1) \equiv_{\mathcal{D}} (\mu, A_2)$, if there exists a bijection $\phi : A_1 \to A_2$ such that,*

$$u(a, \omega) = u(\phi(a), \omega),$$

all $\omega \in \Omega(\mu)$.

All three binary relations are symmetric and transitive.

Lemma 5.1: $\sim_\Gamma$, $\sim_{\mathcal{D}}$, and $\equiv_{\mathcal{D}}$ are symmetric and transitive binary relations.

Proof. Immediate ■

Note that, since payoffs are the same in all possible states, equivalent decision problems differ only in that the actions may have distinct payoffs in impossible states (recall that in our formulation an action specifies payoffs in all states, not just the states possible according to the prior). It is a direct implication of existence of a CIR that equivalent choice data is observed for equivalent decision problems.

Lemma 5.2: If C has a CIR with $K \in \mathcal{K}$ and $(\mu, A_1) \equiv_{\mathcal{D}} (\mu, A_2)$ based on bijection $\phi : A_1 \to A_2$ and $P_1(a|\omega) = P_2(\phi(a), \omega)$ all $a \in A_1$ and $\omega \in \Omega(\mu)$, then,

$$P_1 \in C(\mu, A_1) \Longleftrightarrow P_2 \in C(\mu, A_2), \tag{64}$$

Proof. Immediate. ■

While the equivalence of the data from equivalent decision problems is entirely general, the same is not true for symmetric decision problems. The distinction is that these generally involve learning about distinct states, and there is nothing in Axioms A2 through A9 that imposes symmetry. The fact that Compression does imply symmetry requires more insight. First, we note that symmetry of beliefs survives under taking of particular convex combinations.

Lemma 5.3: Consider $\gamma_1, \gamma_2 \in \Gamma$ satisfying $\gamma_1 \sim_\Gamma \gamma_2$ and $\bar{\gamma}_1, \bar{\gamma}_2 \in \Gamma$ with $\Omega(\bar{\gamma}_i) = \Omega(\gamma_i)$ and $\bar{\gamma}_1 \sim_\Gamma \bar{\gamma}_2$ both based on $\sigma : \Omega(\gamma_1) \to \Omega(\gamma_2)$. Given $\alpha \in (0, 1)$, define,

$$\mu_i = \alpha\gamma_i + (1 - \alpha)\bar{\gamma}_i,$$

for $i = 1, 2$. Then $\mu_1 \sim_\Gamma \mu_2$.

Proof. Immediate. ■

Another important observation is that, if two decision problems are symmetric, $(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2)$, their basic versions are symmetric. It is helpful first to show that symmetric problems can have their actions and states aligned in a natural and useful manner.

Lemma 5.4: Consider $(\mu_1, A_1) \in \mathcal{D}$ with $|A_1| = M$ and $|\Omega(\mu_1)| = J$. Then $(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2)$ if and only if, for $i = 1, 2$, one can index all states $\omega_i(j) \in \Omega(\mu_i)$ and all actions $a_i(m) \in A_i$ so that,

$$\begin{aligned} \mu_1(\omega_1(j)) &= \mu_2(\omega_2(j)) \equiv \mu(j); \text{ and} && (65) \\ u(a_1(m), \omega_1(j)) &= u(a_2(m), \omega_2(j)) \equiv u(m, j). && (66) \end{aligned}$$

Proof. Immediate. ■

Lemma 5.5: If $(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2)$, $(\bar{\mu}_1, A_1) \in \mathcal{B}(\mu_1, A_1)$. and $(\bar{\mu}_2, A_2) \in \mathcal{B}(\mu_2, A_2)$, then,

$$(\bar{\mu}_1, A_1) \sim_{\mathcal{D}} (\bar{\mu}_2, A_2).$$

Proof. Immediate. ■

A key result is that IUC, Axiom A1, implies that all states are equally difficult to learn about, which translates into equivalence of the observed data. This is established in the next lemma.

Lemma 5.6: (Symmetric Data) If $C \in \mathcal{C}$ has a CIR and satisfies IUC, $(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2)$ for bijections $\sigma : \Omega_1 \rightarrow \Omega_2$ and $\phi : A_1 \rightarrow A_2$, $P_1 \in C(\mu_1, A_1)$, and $P_2(\phi(a)|\sigma(\omega)) = P_1(a, \omega)$ for all $a \in A$ and $\omega \in \Omega(\mu_1)$, then $P_2 \in C(\mu_2, A_2)$.

Proof. Immediate. ■

The final symmetry result we establish is that if $C \in \mathcal{C}$ with a UPS representation satisfies IUC, the underlying strictly convex function $T : \Gamma \longrightarrow \mathbb{R}$ is symmetric in the natural sense.

Definition 5 *Strictly convex function $T : \Gamma \longrightarrow \mathbb{R}$ is **symmetric** if it is equal on symmetric beliefs,*

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

Lemma 5.7: (Symmetric Costs) Given $C \in \mathcal{C}$ with a UPS representation satisfying IUC, any function $T : \Gamma \longrightarrow \mathbb{R}$ in a UPS representation $K(Q) = \sum_{\Gamma(Q)} Q(\gamma) T(\gamma)$ must be symmetric.

Proof. Consider $\gamma_1, \gamma_2 \in \Gamma$ satisfying $\gamma_1 \sim_\Gamma \gamma_2$ based on $\sigma : \Omega(\gamma_1) \rightarrow \Omega(\gamma_2)$. If $\gamma_1 = \gamma_2$, $T(\gamma_1) = T(\gamma_2)$ trivially. Therefore consider $\gamma_1 \neq \gamma_2$. Consider now a distinct pair $\bar{\gamma}_1 \neq \bar{\gamma}_2 \in \Gamma$ with $\Omega(\bar{\gamma}_1) = \Omega(\gamma_1)$, $\Omega(\bar{\gamma}_2) = \Omega(\gamma_2)$, and $\bar{\gamma}_1 \sim_\Gamma \bar{\gamma}_2$ based on σ. Define two distinct weighted averages,

$$\begin{aligned} \mu_1 &= \frac{3\gamma_1 + \bar{\gamma}_1}{4} \text{ and } \mu_2 = \frac{3\gamma_2 + \bar{\gamma}_2}{4}; \\ \bar{\mu}_1 &= \frac{\gamma_1 + 3\bar{\gamma}_1}{4} \text{ and } \bar{\mu}_2 = \frac{\gamma_2 + 3\bar{\gamma}_2}{4}. \end{aligned}$$

By Lemma 5.3, $\mu_1 \sim_\Gamma \mu_2$ and $\bar{\mu}_1 \sim_\Gamma \bar{\mu}_2$.

By UPS Feasibility Implies Optimality, there exists $(\mu_1, A_1) \in \mathcal{D}$ such that there exists $\lambda_{A1}, \lambda_{B1} \in \hat{\Lambda}(\mu_1, A_1|K)$ with:

$$\begin{aligned} Q_{\lambda_{A1}}(\gamma_1) &= \frac{3}{4} \text{ and } Q_{\lambda_{A1}}(\bar{\gamma}_1) = \frac{1}{4}; \\ Q_{\lambda_{B1}}(\mu_1) &= 1. \end{aligned}$$

Since $C \in \mathcal{C}$ has a UPS representation, we know that the corresponding data is seen, with $\mathbf{P}_{\lambda_{A1}}, \mathbf{P}_{\lambda_{B1}} \in C(\mu_1, A_1)$. We know by Lemma 2.14 that $\mathbf{Q}_{\mathbf{P}_{\lambda_{A1}}} = Q_{\lambda_{A1}}$ and $\mathbf{Q}_{\mathbf{P}_{\lambda_{B1}}} = Q_{\lambda_{B1}}$.

Create the set of actions A_2, as follows. For each $a \in A_1$, create a corresponding $\phi(a) \in A_2$ such that

$$u(a, \omega) = u(\phi(a), \sigma(\omega))$$

for all $\omega \in \Omega(\gamma_1)$. Since $\sigma : \Omega(\gamma_1) \to \Omega(\gamma_1)$ is a bijection by assumption, $\phi : A_1 \to A_2$ is a bijection by construction, $\Omega(\gamma_1) = \Omega(\mu_1)$ and $\Omega(\gamma_2) = \Omega(\mu_2)$ by construction, and $\mu_1 \sim_\Gamma \mu_2$ from above, we have $(\mu_1, A_1) \sim_{\mathcal{D}} (\mu_2, A_2)$.

By Lemma 5.2, defining $P_2(a|\omega) = P_1(\phi^{-1}(a)|\sigma^{-1}(\omega))$ for all $a \in A_1$ and $\omega \in \Omega(\gamma_2)$,

$$P_1 \in C(\mu_1, A_1) \Longleftrightarrow P_2 \in C(\mu_2, A_2).$$

Hence in particular we can find $P_{A2}, P_{B2} \ \in C(\mu_2, A_2)$ satisfying:

$$\begin{aligned} \mathbf{Q}_{P_{A2}}(\gamma_2) &= \frac{3}{4} \text{ and } \mathbf{Q}_{P_{A2}}(\bar{\gamma}_2) = \frac{1}{4}; \\ \mathbf{Q}_{P_{B2}}(\mu_2) &= 1. \end{aligned}$$

Hence by Lemma 2.15, $\exists \lambda_{A2}, \lambda_{B2} \in \hat{\Lambda}(\mu_2, A_2|K)$ with the corresponding properties,

$$\begin{aligned} Q_{\lambda_{A2}}(\gamma_2) &= \frac{3}{4} \text{ and } Q_{\lambda_{A2}}(\bar{\gamma}_2) = \frac{1}{4}; \\ Q_{\lambda_{B2}}(\mu_2) &= 1. \end{aligned}$$

We repeat the entire structure of the argument to find $\bar{\lambda}_{Ai}, \bar{\lambda}_{Bi} \in \hat{\Lambda}(\bar{\mu}_i, \bar{A}_i|K)$ for $i = 1, 2$ with the reversed probabilities:

$$\begin{aligned} Q_{\bar{\lambda}_{A1}}(\bar{\gamma}_1) &= Q_{\bar{\lambda}_{B1}}(\bar{\gamma}_2) = \frac{3}{4}; \\ Q_{\bar{\lambda}_{A1}}(\gamma_1) &= Q_{\bar{\lambda}_{B1}}(\gamma_2) = \frac{1}{4}; \\ \text{and } Q_{\lambda_{A2}}(\bar{\mu}_1) &= Q_{\lambda_{B2}}(\bar{\mu}_2) = 1. \end{aligned}$$

Now consider any UPS representation $K \in \mathcal{K}^{UPS}$ with $K(Q) = \sum_{\Gamma(Q)} Q(\gamma) T(\gamma)$. To establish symmetry of T, we use simultaneous optimality of an inattentive strategy and an attentive strategy allows us to pin down the difference in costs as based on the difference in expected utility. Taking first $\lambda_{A1}, \lambda_{B1} \in \hat{\Lambda}(\mu_1, A_1|K)$, we note that this implies equality of the corresponding net utilities,

$$\frac{3}{4}\left[\hat{u}(\gamma_1, A_1) - T(\gamma_1)\right] + \frac{1}{4}\left[\hat{u}(\bar{\gamma}_1, A_1) - T(\bar{\gamma}_1)\right] = \hat{u}(\mu_1, A_1). \tag{67}$$

where recall that $\hat{u}(\gamma, A) \equiv \max_{a \in A} \bar{u}(\gamma, a)$.

Taking now $\lambda_{A2}, \lambda_{B2} \in \hat{\Lambda}(\mu_2, A_2|K)$ equality of the corresponding net utilities reduces to,

$$\frac{3}{4}\left[\hat{u}(\gamma_2, A_2) - T(\gamma_2)\right] + \frac{1}{4}\left[\hat{u}(\bar{\gamma}_2, A_2) - T(\bar{\gamma}_2)\right] = \hat{u}(\mu_2, A_2). \tag{68}$$

Given all the symmetries, note that expected utilities are equivalent:

$$\begin{aligned}
\hat{u}(\gamma_1, A_1) &= \hat{u}(\gamma_2, A_2); \\
\hat{u}(\bar{\gamma}_1, A_1) &= \hat{u}(\bar{\gamma}_2, A_2); \\
\hat{u}(\mu_1, A_1) &= \hat{u}(\mu_2, A_2).
\end{aligned}$$

Substitution in (67) and (68) and using the equality between them, we derive,

$$\frac{3}{4}T(\gamma_1) + \frac{1}{4}T(\bar{\gamma}_1) = \frac{3}{4}T(\gamma_2) + \frac{1}{4}T(\bar{\gamma}_2). \tag{69}$$

We follow precisely the same logic with respect to the strategies $\bar{\lambda}_{Ai}, \bar{\lambda}_{Bi} \in \hat{\Lambda}(\bar{\mu}_i, \bar{A}_i|K)$ for $i = 1, 2$ to conclude that the corresponding equation holds with reversed weights

$$\frac{1}{4}T(\gamma_1) + \frac{3}{4}T(\bar{\gamma}_1) = \frac{1}{4}T(\gamma_2) + \frac{3}{4}T(\bar{\gamma}_2). \tag{70}$$

Adding up equations (69) and (70) yields equality of sums, while subtracting them yields equality of differences,

$$\begin{aligned}
T(\gamma_1) + T(\bar{\gamma}_1) &= T(\gamma_2) + T(\bar{\gamma}_2); \\
T(\gamma_1) - T(\bar{\gamma}_1) &= T(\gamma_2) - T(\bar{\gamma}_2).
\end{aligned}$$

Adding these together we finally conclude that $T(\gamma_1) = T(\gamma_2)$, completing the proof. ■

A5.2.2: Directional Derivatives: Basic Results

For the next several sections of the proof, we fix an arbitrary strictly convex function $T : \Gamma \to \mathbb{R}$ for $C \in \mathcal{C}$ with a UPS representation satisfying IUC. In light of Lemma 5.7, it is symmetric, with costs invariant under $\sim_\Gamma$, the equivalence relation on beliefs. We further restrict attention to interior posteriors that place strictly positive probability on a fixed set of underlying states of cardinality 4 or higher. We are particularly interested in those interior posteriors at which this domain-restricted cost function is differentiable.

Definition 6 *We fix a strictly convex and symmetric function $T : \Gamma \to \mathbb{R}$ in a UPS representation. We fix also a set of states $\tilde{\Omega} \subset \Omega$ of cardinality $J \geq 4$, with the states indexed by $1 \leq j \leq J$. We define $\tilde{\Gamma}$ to the the set of posteriors with $\Omega(\gamma) = \tilde{\Omega}$. We define $\tilde{T}$ to be the restriction of the underlying symmetric and strictly convex function on Γ to $\tilde{\Gamma}$,*

$$\tilde{T} : \tilde{\Gamma} \to \mathbb{R},$$

We let $\tilde{\Gamma}' \subset \tilde{\Gamma}$ be posteriors at which $\tilde{T} : \tilde{\Gamma} \to \mathbb{R}$ is differentiable. We let $\tilde{K}(\mu, \lambda) = \sum Q_\lambda(\gamma)\tilde{T}(\gamma)$

be the attention cost function on this limited domain.

The fundamental objects of interest in what follows are certain derivatives of the function $\tilde{T}$ on $\tilde{\Gamma}$. Note that $\tilde{\Gamma}$ does not allow for independent variation in any single state-specific posterior $\gamma(j)$ due to the adding up constraint on probabilities. Hence we use the directional derivatives in what follows. Given convex function $\tilde{T} : \tilde{\Gamma} \to \mathbb{R}$ and $\gamma \in \tilde{\Gamma}$, the directional derivative at $\gamma \in \tilde{\Gamma}$ in direction $y \in \mathbb{R}^J$ is defined as,

$$\tilde{T}'(\gamma|y) = \lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon y) - \tilde{T}(\gamma)}{\epsilon}, \tag{71}$$

if it exists. We use special notation for the directional derivatives of interest.

Definition 7 *Given $\gamma \in \tilde{\Gamma}$ and any pair of states $1 \leq i \neq j \leq J$, we define the* ***one-sided derivative in direction*** *ji, $\tilde{T}_{\overrightarrow{ji}}(\gamma)$, as the directional derivative associated with increasing the ith coordinate and equally reducing the jth:*

$$\tilde{T}_{\overrightarrow{ji}}(\gamma) = \tilde{T}'(\gamma|e_i - e_j) = \lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon}; \tag{72}$$

where $e_k \in \mathbb{R}^J$ is the vector with its only non-zero element being 1 in the k^{th} coordinate. Where it exists, we define the ***two-sided derivative in direction*** *ji, $\tilde{T}_{(ji)}$, by:*

$$\tilde{T}_{(ji)}(\gamma) = \lim_{\epsilon \to 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon}. \tag{73}$$

In what follows we will use standard results of convex analysis, almost all gathered in Rockafellar's comprehensive treatise (Rockafellar, 1970). The first such standard result that we translate to our setting establishes existence of one-sided directional derivatives, as well as an inequality concerning one-sided directional derivatives in opposite directions. For completeness, we note also the standard results that a real-valued convex function is continuous on its relative interior, in this case $\tilde{\Gamma}$.

Lemma 5.8: $\tilde{T}$ is continuous on $\gamma \in \tilde{\Gamma}$, and, given $1 \leq i \neq j \leq J$, $\tilde{T}_{\overrightarrow{ji}}(\gamma)$ exists. Moreover,

$$-\tilde{T}_{\overrightarrow{ij}}(\gamma) \leq \tilde{T}_{\overrightarrow{ji}}(\gamma). \tag{74}$$

Proof. Continuity of $\tilde{T}$ on its relative interior is theorem 10.1 in Rockafellar. By (71), given $\gamma \in \tilde{\Gamma}$,

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

where $y = e_i - e_j$. Rockafellar theorem 23.1 establishes that, since $\tilde{T} : \mathbb{R}^J \to \bar{\mathbb{R}}$ is convex and $\tilde{T}(\gamma)$ is finite at $\gamma \in \tilde{\Gamma}$, for any $y \in \mathbb{R}^J$, the RHS of (71) is a non-decreasing function of $\epsilon > 0$. Hence $\tilde{T}_{\overrightarrow{ji}}(\gamma)$ exists. With regard to (74), note directly from the definition that,

$$\tilde{T}_{\overrightarrow{ij}}(\gamma) = \lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_j - e_i)) - \tilde{T}(\gamma)}{\epsilon} = \tilde{T}'(\gamma| - (e_i - e_j)).$$

Theorem 23.1 in Rockafellar establishes with full generality that,

$$-\tilde{T}'(\gamma|-y) \leq \tilde{T}'(\gamma|y).$$

Applying this to $y = e_i - e_j$ completes the proof of the Lemma,

$$-\tilde{T}_{\overrightarrow{ij}}(\gamma) = -\tilde{T}'(\gamma|e_j - e_i) \leq \tilde{T}'(\gamma|e_i - e_j) = \tilde{T}_{\overrightarrow{ji}}(\gamma).$$

■

We are particularly interested in posteriors $\gamma \in \tilde{\Gamma}$ at which the inequality (74) is replaced with an equality. The next result shows this to be equivalent to existence of the two-sided derivative. We add also the standard result that differentiability of T implies existence of all 2-sided directional derivatives.

Lemma 5.9: $\tilde{T}_{(ji)}(\gamma)$ exists if and only if (74) holds with equality,

$$-\tilde{T}_{\overrightarrow{ij}}(\gamma) = \tilde{T}_{\overrightarrow{ji}}(\gamma), \tag{75}$$

in which case

$$\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{\overrightarrow{ji}}(\gamma) = -\tilde{T}_{\overrightarrow{ij}}(\gamma) = -\tilde{T}_{(ij)}(\gamma). \tag{76}$$

Moreover, given $\gamma \in \tilde{\Gamma}'$, $\tilde{T}_{(ji)}(\gamma)$ exists for all $1 \leq i \neq j \leq J$.

Proof. Note first that if (75) holds so that $\tilde{T}_{\overrightarrow{ji}}(\gamma) = -\tilde{T}_{\overrightarrow{ij}}(\gamma)$, this corresponds to equality of the limits from the left and right

$$\begin{aligned}\lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon} &= \tilde{T}_{\overrightarrow{ji}}(\gamma) = -\tilde{T}_{\overrightarrow{ij}}(\gamma) = -\lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_j - e_i)) - \tilde{T}(\gamma)}{\epsilon} \\ &= -\lim_{\delta = -\epsilon \uparrow 0} \frac{\tilde{T}(\gamma - \delta(e_j - e_i)) - \tilde{T}(\gamma)}{-\delta} = \lim_{\delta \uparrow 0} \frac{\tilde{T}(\gamma + \delta(e_i - e_j)) - \tilde{T}(\gamma)}{\delta}.\end{aligned}$$

It is standard that this implies that the equal left and right limits define the limit itself,

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

establishing equivalence of (75) and existence of $\tilde{T}_{(ji)}(\gamma)$.

Conversely, note that if $\tilde{T}_{(ji)}(\gamma)$ exists,

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

and,

$$\begin{aligned}\tilde{T}_{(ji)}(\gamma) &= \lim_{\epsilon \longrightarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon} = \lim_{\epsilon \downarrow 0} \frac{\tilde{T}(\gamma + \epsilon(e_i - e_j)) - \tilde{T}(\gamma)}{\epsilon} \\ &= -\lim_{\delta = -\epsilon \uparrow 0} \frac{\tilde{T}(\gamma + \delta(e_j - e_i)) - \tilde{T}(\gamma)}{-\delta} = \lim_{\delta \uparrow 0} \frac{\tilde{T}(\gamma + \delta(e_j - e_i)) - \tilde{T}(\gamma)}{\delta} = -\tilde{T}_{\overrightarrow{ij}}(\gamma).\end{aligned}$$

These equations together verify that (75) holds and also that,

$$\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{\overrightarrow{ji}}(\gamma) = -\tilde{T}_{\overrightarrow{ij}}(\gamma). \tag{77}$$

To complete the proof that (76) holds, note that since $\tilde{T}_{\overrightarrow{ji}}(\gamma) = -\tilde{T}_{\overrightarrow{ij}}(\gamma)$, we know from (75) that $\tilde{T}_{(ij)}(\gamma)$ exists, and therefore that it satisfies the corresponding equality,

$$\tilde{T}_{(ij)}(\gamma) = \tilde{T}_{\overrightarrow{ij}}(\gamma) = -\tilde{T}_{\overrightarrow{ji}}(\gamma). \tag{78}$$

In combination, (78) and (77) imply (76).

With regard to the final clause of the Lemma, that $\tilde{T}_{(ji)}(\gamma)$ exists for all $1 \leq i \neq j \leq J$ if $\gamma \in \tilde{\Gamma}'$, Rockafellar theorem 25.2 shows that $\gamma \in \tilde{\Gamma}'$ implies all directional derivatives $\tilde{T}'(\gamma|y)$ exist and are linear in $y = (y(1), ..., y(J))$. Hence they can be written in terms of partial derivatives $\tilde{T}_j(\gamma)$ as,

$$\tilde{T}'(\gamma|y) = \sum_{j=1}^{J} y(j)\tilde{T}_j(\gamma).$$

Hence, given $1 \leq i \neq j \leq J$,

$$-\tilde{T}_{\overrightarrow{ij}}(\gamma) = -\left[\tilde{T}_j(\gamma) - \tilde{T}_i(\gamma)\right] = \tilde{T}_i(\gamma) - \tilde{T}_j(\gamma) = \tilde{T}'(\gamma|e_i - e_j)) = \tilde{T}_{\overrightarrow{ji}}(\gamma),$$

verifying (75) and thereby establishing existence of all 2-sided directional derivatives. ■

Our next preliminary result shows that symmetry of the cost function has implications for directional derivatives. The Lemma specifies the inherited symmetry property precisely.

Lemma 5.10: If $\tilde{T}_{(ji)}(\gamma)$ exists, then for any bijection $\sigma : \{1, .., J\} \to \{1, .., J\}$,

$$\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{(\sigma(j)\sigma(i))}(\gamma^{\sigma}),$$

where,

$$\gamma^{\sigma}(j) = \gamma(\sigma^{-1}(j)).$$

Proof. Suppose that $\tilde{T}_{(ji)}(\gamma)$ exists and note by symmetry of $\tilde{T}$ (Lemma 5.7) and the bijective nature of σ,

$$\tilde{T}(\gamma) = \tilde{T}(\gamma^{\sigma}).$$

Now consider the posterior $\gamma^{\sigma} + \epsilon(e_{\sigma(i)} - e_{\sigma(j)})$ and note that,

$$\tilde{\gamma}(k) = \begin{cases} \gamma^{\sigma}(k) + \epsilon = \gamma(\sigma^{-1}(\sigma(i)) + \epsilon = \gamma(i) + \epsilon \text{ if } k = \sigma(i); \\ \gamma^{\sigma}(k) - \epsilon = \gamma(\sigma^{-1}(\sigma(j)) - \epsilon = \gamma(j) - \epsilon \text{ if } k = \sigma(j); \\ \gamma^{\sigma}(k) \text{ else.} \end{cases}$$

Hence,

$$\gamma^{\sigma} + \epsilon(e_{\sigma(i)} - e_{\sigma(j)}) \sim_{\Gamma} \gamma^{\sigma} + \epsilon(e_i - e_j),$$

so that by the symmetry of $\tilde{T}$,

$$\tilde{T}\left[\gamma^{\sigma}+\epsilon(e_{\sigma(i)}-e_{\sigma(j)})\right]=\tilde{T}(\gamma+\epsilon(e_i-e_j)).$$

Hence,

$$\tilde{T}_{(ji)}(\gamma^{\sigma})=\lim_{\epsilon\to 0}\frac{\tilde{T}\left[\gamma^{\sigma}+\epsilon(e_{\sigma(i)}-e_{\sigma(j)})\right]-\tilde{T}(\gamma^{\sigma})}{\epsilon}=\lim_{\epsilon\to 0}\frac{\tilde{T}\left[(\gamma+\epsilon(e_i-e_j))^{\sigma}\right]-\tilde{T}(\gamma^{\sigma})}{\epsilon}=\tilde{T}_{(ji)}(\gamma),$$

establishing the Lemma. ■

A5.2.3: Lagrangians and Directional Derivatives

The following result explains to some extent the relevance of directional derivatives to our approach. It follows from the Lagrangian Lemma.

Lemma 5.11: (Optimality and Directional Derivatives) Suppose $C\in\mathcal{C}$ has a UPS representation with $K\in\mathcal{K}$, and consider $(\mu,A)\in\mathcal{D}$ and $P\in C(\mu,A)$ with $a,b\in\mathcal{A}(P)$ with $\{\bar{\gamma}^a_P,\bar{\gamma}^b_P\}\subset\tilde{\Gamma}$. Then for all pairs of states $1\le i\neq j\le J$:

1. For $c\in\{a,b\}$, if $\tilde{T}_{(ji)}(\bar{\gamma}^c_P)$ does not exist, $\bar{\gamma}^c_P\in\tilde{\Gamma}\backslash\tilde{\Gamma}'$,

$$-\tilde{T}_{\overrightarrow{ij}}(\bar{\gamma}^c_P)\le u(c,i)-u(c,j)-[\theta(i)-\theta(j)]\le\tilde{T}_{\overrightarrow{ji}}(\bar{\gamma}^c_P);\tag{79}$$

   with at least one inequality strict.

2. For $c\in\{a,b\}$, if $\tilde{T}_{(ji)}(\bar{\gamma}^c_P)$ exists, $\bar{\gamma}^c_P\in\tilde{\Gamma}'$,

$$\tilde{T}_{(ji)}(\bar{\gamma}^c_P)=u(c,i)-u(c,j)-[\theta(i)-\theta(j)];\tag{80}$$

3. If $a,b\in\mathcal{A}(P)$ are such that $\{\bar{\gamma}^a_P,\bar{\gamma}^b_P\}\subset\tilde{\Gamma}'$,

$$\tilde{N}^a_{(ji)}(\bar{\gamma}^a_P)=\tilde{N}^b_{(ji)}(\bar{\gamma}^b_P).\tag{81}$$

Proof. Since $C\in\mathcal{C}$ has a UPS representation and $\bar{\gamma}^a_P,\bar{\gamma}^b_P\in\Gamma(P)$, we know from Lemma 2.15 that there exists an optimal policy $\lambda=(Q_\lambda,q_\lambda)\in\hat{\Lambda}(\mu,A|K)$ with $\bar{\gamma}^a_P,\bar{\gamma}^b_P\in\Gamma(Q_\lambda)$. Given that $J\ge 3$, we apply Lemma 5.10 to ensure that directional derivatives are invariant to re-indexing states if needed to make that state J is neither i nor j. We now apply the UPS Lagrangian Lemma to the decision problem $(\mu,A)\in\mathcal{D}$ to identify corresponding multipliers $\theta(j)$. Introduce the function $F^c(\gamma)$ on $c\in A$ and $\gamma\in\Gamma(\mu)$ and its supremal value:

$$F^c(\gamma)\equiv\tilde{N}^c(\gamma)-\sum_{k=1}^{J-1}\theta(k)\gamma(k);\tag{82}$$

where,

$$\tilde{N}^c(\gamma)=\sum_{k=1}^{J}u(c,k)\gamma(k)-\tilde{T}(\gamma),$$

Note that for $\gamma \in \tilde{\Gamma}'$, net utility $\tilde{N}^c_{(ji)}(\gamma)$ is well-defined since $\tilde{T}_{(ji)}(\gamma)$ is well-defined, and, since the limit operation changes only posteriors i and j,

$$\lim_{\epsilon \to 0} \frac{u(c,i)[\gamma(i) + \epsilon - \gamma(i)] + u(c,j)[\gamma(j) - \epsilon - \gamma(j)]}{\epsilon} = u(c,i) - u(c,j).$$

With $\tilde{T}_{(ji)}(\gamma)$ well-defined, the same holds for $F^c_{(ij)}(\gamma)$, which we can analogously compute from (82) as,

$$F^c_{(ij)}(\gamma) = \tilde{N}^c_{(ij)}(\gamma) - \theta(i) + \theta(j). \tag{83}$$

The Lagrangian Lemma implies that $\hat{F}$, the supremal value of $F^c(\gamma)$,

$$\hat{F} = sup_{c \in A, \gamma \in \Gamma(\mu)} \left[F^c(\gamma) \right]$$

is achieved by the posteriors associated with any optimal policy. By Lemma 2.14 this means that it is achieved both by setting $(c, \gamma) = (a, \bar{\gamma}^a_P)$ and $(c, \gamma) = \left(b, \bar{\gamma}^b_P\right)$. By Lemma 5.9, if $\tilde{T}_{(ji)}(\nu)$ does not exist, we know that the derivative from the left must be non-negative and from the right non-positive and that they cannot be equal. This corresponds precisely to

$$-\tilde{T}_{\overrightarrow{ij}}(\bar{\gamma}^c_P) + u(c,i) - u(c,j) - [\theta(i) - \theta(j)] \leq 0 \leq \tilde{T}_{\overrightarrow{ji}}((\bar{\gamma}^c_P) + u(c,i) - u(c,j) - [\theta(i) - \theta(j)],$$

confirming (79). Conversely, if $\tilde{T}_{(ji)}(\bar{\gamma}^c_P)$ exists for $c \in \{a, b\}$, this maximization implies that the corresponding derivative $\tilde{N}^c(\bar{\gamma}^c_P)$ must equal zero,

$$\tilde{N}^c_{(ji)}(\bar{\gamma}^c_P) = -\tilde{T}_{(ji)}(\bar{\gamma}^c_P) + u(c,i) - u(c,j) - [\theta(i) - \theta(j)] = 0,$$

confirming (80). To complete the proof, note that when $\{\bar{\gamma}^a_P, \bar{\gamma}^b_P\} \subset \tilde{\Gamma}'$, (80) applies at both posteriors. Hence by (83),

$$\tilde{N}^a_{(ji)}(\bar{\gamma}^a_P) = \theta(i) - \theta(j) = \tilde{N}^b_{(ji)}(\bar{\gamma}^b_P),$$

confirming (81) and establishing the Lemma. ■

A5.2.4: Ratio Sets and Linearity of Posteriors

The next key observation is that we can design decision problems in which the derivatives of net utility are profoundly informative about the cost function. To do this we use decision problems for which IUC places strong restrictions on how posteriors change as the prior changes. These are decision problems with two equivalent states and corresponding posteriors satisfying a ratio condition. To show what IUC implies for these problems, we provide a key extension to the Feasibility implies Optimality Lemma. We show that IUC places strong restrictions on the optimal strategies. In particular, it enables us to move the prior between equivalent states without altering the observed state dependent stochastic choice data. Since that the data does not change, Bayes' rule alone determines how changes in the prior impact the observed posteriors. We apply this in the context of a set of parametrized decision problems (μ_t, A) in which the parameter $t \in [0, 1]$ adjusts the weight between the payoff equivalent states $k \neq l \in \Omega(\gamma)$.

Definition 8 *Given $\alpha \in (0, \infty)$ and two states $1 \leq k \neq l \leq J$, we define the corresponding* ***ratio***

set $\Gamma_{kl}(\alpha) \subset \tilde{\Gamma}$ *as the set of posteriors in which α is the ratio between $\gamma(k)$ and $\gamma(l)$:*

$$\Gamma_{kl}(\alpha) = \left\{ \gamma \in \tilde{\Gamma} \,\middle|\, \frac{\gamma(k)}{\gamma(l)} = \alpha \right\}.$$

$\Gamma_{kl}(\alpha)$ is the intersection of $\tilde{\Gamma}$ and a $J-2$ dimensional linear subspace of $\mathbb{R}^J$. It is therefore convex and has dimension $J-2$. Figure 10 depicts $\tilde{\Gamma}$ for $J=4$. The blue triangle represents $\Gamma_{12}(\alpha)$ for $\alpha = 2$ so that $\gamma \in \tilde{\Gamma}$ if and only if $\gamma(1) = 2\gamma(2)$. In the Figure $\Gamma_{12}(2)$ connects all points with $\gamma(1) = \gamma(2) = 0$ (i.e. the line segment connecting $\gamma(3) = 1$ to $\gamma(4) = 1$) to the point with $\gamma(1) = \frac{2}{3}$ and $\gamma(2) = \frac{1}{3}$.

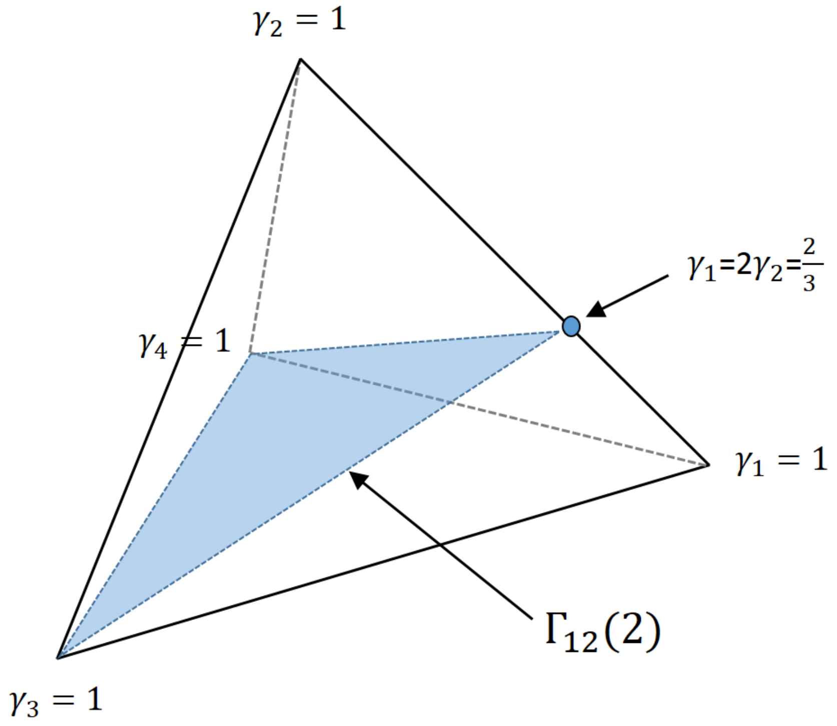

Figure 10

Lemma 5.12: Suppose $C \in \mathcal{C}$ has a UPS representation K and satisfies IUC (A1). Consider $\eta \neq \nu \in \Gamma$ with,

$$\Omega(\eta) = \Omega(\nu) = \{j | 1 \leq j \leq J\},$$

for $J \geq 3$, and $1 \leq k \neq l \leq J$ such that $\eta, \nu \in \Gamma_{kl}(\alpha)$ some $\alpha \in (0, \infty)$,

$$\frac{\eta(k)}{\eta(l)} = \frac{\nu(k)}{\nu(l)} = \alpha. \tag{84}$$

Define the mean belief,

$$\bar{\mu} = \frac{\eta + \nu}{2},$$

and for $t \in [0,1]$ define μ_t, η_t, and ν_t by:

$$\begin{aligned}
\mu_t(j) &= \begin{cases} t[\bar{\mu}(k)+\bar{\mu}(l)] & \text{for } j = k; \\ (1-t)[\bar{\mu}(k)+\bar{\mu}(l)] & \text{for } j = l; \\ \bar{\mu}(j) & \text{otherwise;} \end{cases} && (85)\\
\zeta_t(j) &= \left[\frac{\zeta(j)}{\bar{\mu}(j)}\right]\mu_t(j) \text{ for } 1 \le j \le J; && (86)
\end{aligned}$$

for $\zeta = \eta, \nu$. Then there exists $a, b \in \mathcal{A}$ with $u(a,k) = u(a,l)$ and $u(b,k) = u(b,l)$ such that,

$$C(\mu_t, \{a,b\}) = \{P_t\};\ \bar{\gamma}^a_{P_t} = \eta_t, \text{ and } \bar{\gamma}^b_P = \nu_t. \qquad (87)$$

Specifically, for $\bar{t} = \frac{\bar{\mu}(k)}{\bar{\mu}(k)+\bar{\mu}(l)} \in (0,1)$, $\mu_{\bar{t}} = \bar{\mu}$ and,

$$\zeta_{\bar{t}}(j) = \left[\frac{\zeta(j)}{\bar{\mu}(j)}\right]\bar{\mu}(j) = \zeta(j)$$

for $\zeta = \eta, \nu$.

Proof. Fix $\eta \neq \nu \in \Gamma$ with $\Omega(\eta) = \Omega(\nu) = \{j | 1 \le j \le J\}$ for $J \ge 3$ and $k \neq l \in \Omega(\gamma)$ for which (84) holds. Note that this is only possible if $|\Omega(\eta)| \ge 3$ since otherwise (84) implies $\gamma = \eta$. By construction, note from adding the last two equations that,

$$\frac{\eta_t(j)+\nu_t(j)}{2} = \frac{1}{2}\left[\frac{\eta(j)+\nu(j)}{\bar{\mu}(j)}\right]\bar{\mu}_t(j) = \bar{\mu}_t(j).$$

Now consider the case $t = 1$ so that $\mu_1(l) = \eta_1(1) = \eta_1(l) = 0$, $\mu_1(j) = \bar{\mu}(j)$ for $j \neq k, l$, and:

$$\begin{aligned}
\mu_1(k) &= \bar{\mu}(k)+\bar{\mu}(l);\\
\zeta_1(k) &= \zeta(k)\left[\frac{\mu_1(k)}{\bar{\mu}(k)}\right] = \zeta(k)\left[\frac{\bar{\mu}(k)+\bar{\mu}(l)}{\bar{\mu}(k)}\right] = \zeta(k)\left[1+\frac{\bar{\mu}(l)}{\bar{\mu}(k)}\right]\\
&= \zeta(k)+\zeta(l);
\end{aligned}$$

for $\zeta = \eta, \nu$, where the last line follows (84) and the definition of the mean belief,

$$\frac{\bar{\mu}(l)}{\bar{\mu}(k)} = \frac{\eta(l)+\nu(l)}{\eta(k)+\nu(k)} = \frac{\eta(l)}{\eta(k)} = \frac{\nu(l)}{\nu(k)}.$$

Note that since $\gamma \neq \eta$ and (84) holds, we know that there exists $j \in \Omega(\gamma)\backslash\{k,l\}$ with $\eta(j) \neq \nu(j)$, so that $\eta_1 \neq \nu_1$. Hence we can apply Feasibility Implies Optimality to find action set $\{a,b\}$ such that there is an optimal strategy for the corresponding mean belief,

$$\lambda(1) = (Q_1, q_1) \in \hat{\Lambda}(\mu_1, \{a,b\}|K)$$

in which the only chosen posteriors are η_1 and ν_1, so that $Q_1(\eta_1) = Q_1(\nu_1) = 0.5$. Feasibility Implies Optimality implies also that the deterministic strategy involving each action being chosen deterministically at its corresponding posterior is optimal. In fact. given that the two posteriors are distinct, we know that they are linearly independent, so that the optimal strategy is unique by

Lemma 2.4. WLOG,

$$q_1(a|\eta_1) = q_1(b|\nu_1) = 1.$$

We can readily characterize the corresponding revealed posteriors. Since $C \in \mathcal{C}$ has a UPS representation K and the data corresponding to the optimal strategy is observed, we know by Lemmas 2.13 that $C(\mu_1, \{a,b\}) = \{P_1\}$ has the given revealed posteriors,

$$\bar{\gamma}^a_{P_1} = \eta_1 \text{ and } \bar{\gamma}^b_{P_1} = \nu_1.$$

By Lemma 2.19 note that since $\mu_1(l) = 0$, we can set $u(a,l) = u(a,k)$ and $u(b,l) = u(b,k)$ without affecting the observed pattern of SDSC data, P_1.

We now consider decision problem $(\mu_t, \{a,b\}) \in \mathcal{D}$ noting that $(\mu_1, \{a,b\}) \in \mathcal{B}(\mu_t, \{a,b\})$ all $t \in (0,1)$. Since $C(\mu_1, \{a,b\}) = \{P_1\}$ and C satisfies IUC (A1) , we conclude that $P_t \in C(\mu_t, \{a,b\})$ if and only if it satisfies,

$$P_t(c|j) = \begin{cases} P_1(c|j) & \text{if } j \in \Omega(\eta)\backslash\{k,l\}; \\ P_1(c|k) & \text{if } j \in \{k,l\}; \end{cases} \tag{88}$$

for $c \in \{a,b\}$. We now show that $\bar{\gamma}^a_{P_t} = \eta_t$ and that $\bar{\gamma}^b_{P_t} = \nu_t$ all $t \in (0,1)$.

Note first that Bayes' rule combined with (88) implies that, for all $j \in \Omega(\gamma)\backslash\{k,l\}$, and for $c \in \{a,b\}$,

$$\bar{\gamma}^c_{P_t}(j) = \left[\frac{P_t(c|j)}{P_t(c)}\right]\mu_t(j) = \left[\frac{P_1(c)}{P_t(c)}\right]\left[\frac{P_1(c|j)}{P_1(c)}\right]\mu_1(j) = \left[\frac{P_1(c)}{P_t(c)}\right]\bar{\gamma}^c_{P_1}(j),$$

since $\mu_t(j) = \mu_1(j)$. To compute the unconditional choice probabilities $P_1(c)$, note that $\mu_t(j) = \mu_1(j)$ for $j \neq k,l$ and that, given $t \in (0,1)$,

$$\mu_t(k) + \mu_t(l) = \mu_1(k).$$

Hence,

$$\begin{aligned} P_t(c) &= \sum_{j \neq k,l} P(c|j)\mu_t(j) + P(c|k)\mu_t(k) + P(c|l)\mu_t(l) \\ &= \sum_{j \neq k,l} P_1(c|j)\mu_1(j) + P_1(c|k)\left(\mu_t(k) + \mu_t(l)\right) \\ &= \sum_{j \neq k,l} P_1(c|j)\mu_1(j) + P_1(c|k)\mu_1(k) = P_1(c). \end{aligned}$$

Applying this to actions $c = a,b$ separately we derive that for all $j \in \Omega(\gamma)\backslash\{k,l\}$,

$$\begin{aligned} \bar{\gamma}^a_{P_t}(j) &= \bar{\gamma}^a_{P_1}(j) = \eta_1(j) = \eta(j); \\ \bar{\gamma}^b_{P_t}(j) &= \bar{\gamma}^b_{P_1}(j) = \nu_1(j) = \nu(j) \end{aligned}$$

For $j = k$ we make the corresponding substitutions,

$$\begin{aligned}
\bar{\gamma}^c_{P_t}(k) &= \left[\frac{P_t(c|k)}{P_t(c)}\right]\mu_t(k) = \left[\frac{P_1(c)}{P_t(c)}\right]\left[\frac{P_1(c|k)}{P_1(c)}\right] t\left[\bar{\mu}(k) + \bar{\mu}(l)\right] \\
&= \left[\frac{P_1(c|k)}{P_1(c)}\right] t\mu_1(k) = t\bar{\gamma}^c_{P_1}(k).
\end{aligned}$$

Applying this to actions $c = a, b$ separately we derive,

$$\begin{aligned}
\bar{\gamma}^a_{P_t}(k) &= t\bar{\gamma}^a_{P_1}(k) = t\eta_1(k) = \eta_t(k); \\
\bar{\gamma}^b_{P_t}(k) &= t\bar{\gamma}^b_{P_1}(k) = t\nu_1(k) = \nu_t(k);
\end{aligned}$$

where the final equalities derive from,

$$\frac{\eta_t(k)}{\eta_1(k)} = \frac{\nu_t(k)}{\nu_1(k)} = \frac{\mu_t(k)}{\mu_1(k)} = t.$$

Finally, using the corresponding substitutions for $j = l$ we derive,

$$\begin{aligned}
\bar{\gamma}^c_{P_t}(l) &= \left[\frac{P_t(c|l)}{P_t(c)}\right]\mu_t(l) = \left[\frac{P_1(c)}{P_t(c)}\right]\left[\frac{P_1(c|k)}{P_1(c)}\right](1-t)\left[\bar{\mu}(k) + \bar{\mu}(l)\right] \\
&= \left[\frac{P_1(c|k)}{P_1(c)}\right](1-t)\,\mu_1(k) = (1-t)\bar{\gamma}^c_{P_1}(k).
\end{aligned}$$

Applying this to actions $c = a, b$ separately we derive,

$$\begin{aligned}
\bar{\gamma}^a_{P_t}(l) &= (1-t)\bar{\gamma}^a_{P_1}(k) = (1-t)\eta_1(k) = \eta_t(l); \\
\bar{\gamma}^b_{P_t}(l) &= (1-t)\bar{\gamma}^b_{P_1}(k) = (1-t)\nu_1(k) = \nu_t(l);
\end{aligned}$$

where the final equalities derive from,

$$\frac{\eta_t(l)}{\eta_1(k)} = \frac{\nu_t(l)}{\nu_1(k)} = \frac{\mu_t(l)}{\mu_1(k)} = (1-t).$$

The above concludes the proof that $\bar{\gamma}^a_{P_t} = \eta_t$ and $\bar{\gamma}^b_{P_t} = \nu_t$ all $t \in (0,1)$. To finish the proof of the main clause, note that the result directly holds for $t = 1$. But note that by symmetry that the state labels are irrelevant, so that it also holds for $t = 0$.

The final step in the proof involves direct substitution to show that,

$$\bar{t} = \frac{\bar{\mu}(k)}{\bar{\mu}(k) + \bar{\mu}(l)} \Longrightarrow \mu_{\bar{t}} = \bar{\mu},$$

and correspondingly that,

$$\zeta_{\bar{t}}(j) = \left[\frac{\zeta(j)}{\bar{\mu}(j)}\right]\bar{\mu}(j) = \zeta(j)$$

for $\zeta = \eta, \nu$. ■

A5.2.5: The Trapezoid Condition

We are building towards using the rectangle condition for additive separability. The rectangle condition states that a function $f(a,b)$ on $X_A \times X_B$ is additively separable if

$$f(a_1,b_1) - f(a_2,b_1) = f(a_1,b_2) - f(a_2,b_2).$$

for all $a_1, a_2 \in X_A$ and $b_1, b_2 \in X_B$. Instead of constructing rectangles, we use the ratio sets to construct trapezoids which we then transform into rectangles. To construct these trapezpoids we match sets of posteriors that lie on different ratio sets and differ in two dimensions. We now define the set of posteriors that differ only on states k and l.

Definition 9 *Given $\hat{\gamma} \in \tilde{\Gamma}$ and two states $k, l \in J$, we let $\Phi_{kl}(\hat{\gamma})$ denote the set of posteriors that agree with γ on all states $j \neq k, l$*

$$\Phi_{kl}(\hat{\gamma}) = \{\gamma | \gamma(j) = \hat{\gamma}(j), j \neq k, l\}$$

$\Phi_{kl}(\hat{\gamma})$ represents a line in $\tilde{\Gamma}$ through $\hat{\gamma}$ in the direction $e_k - e_l$ where e_j is the unit vector in $\mathbb{R}^J$ with a one in the jth coordinate.

Figure 11 reproduces Figure 10 and adds the point $\hat{\gamma} \in \Gamma_{12}(2)$. The red line in the Figure represents $\Phi_{12}(\hat{\gamma})$. It is the set of points for on which $\gamma(3) = \hat{\gamma}(3)$ and $\gamma(4) = \hat{\gamma}(4)$. This line segment is parallel to the line connecting $\gamma(1) = \gamma(2) = 1$.

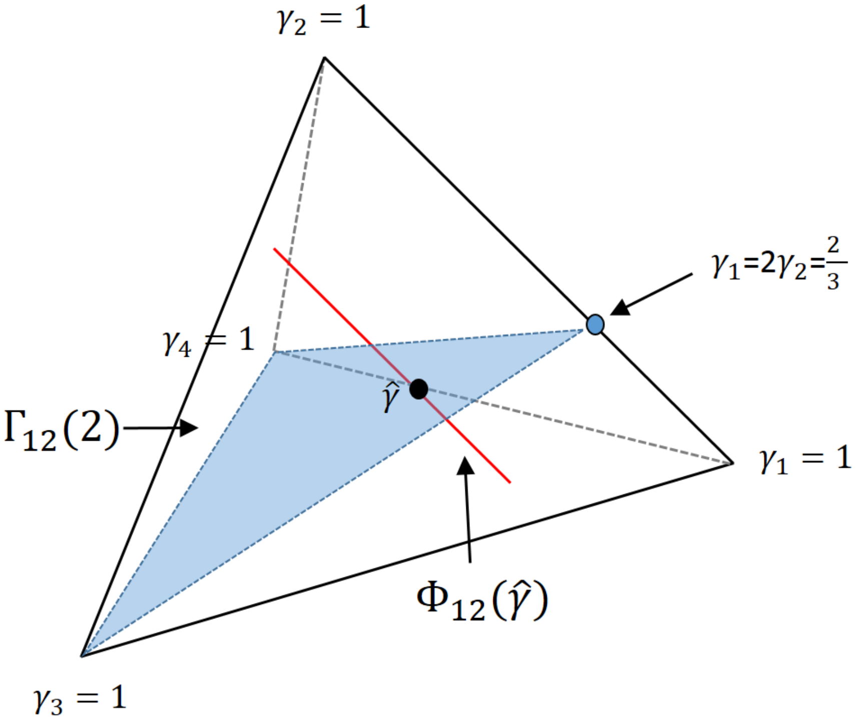

Figure 11

With these definitions we introduce the key sets of four posteriors.

Definition 10 *A set of four distinct posteriors $\eta_1, \eta_2, \nu_1, \nu_2 \in \tilde{\Gamma}$ satisfy the* ***trapezoid condition*** *if there exist distinct $\alpha(1) \neq \alpha(2) > 0$ and $1 \leq k \neq l \leq J$ such that:*

1. *η_1,$\nu_1 \in \Gamma_{kl}(\alpha_1)$,*
$$\frac{\eta_1(k)}{\eta_1(l)} = \frac{\nu_1(k)}{\nu_1(l)} = \alpha_1.$$
2. *η_2,$\nu_2 \in \Gamma_{kl}(\alpha_2)$,*
$$\frac{\eta_2(k)}{\eta_2(l)} = \frac{\nu_2(k)}{\nu_2(l)} = \alpha_2.$$
3. *$\eta_2 \in \Phi_{kl}(\eta_1)$, so that $\eta_2(j) = \eta_1(j)$ for $j \neq k, l$.*
4. *$\nu_2 \in \Phi_{kl}(\nu_1)$, so that $\nu_2(j) = \nu_1(j)$ for $j \neq k, l$.*

Note that the condition $\alpha_1 \neq \alpha_2$ is imposed since with $\alpha_1 = \alpha_2$, the conditions would give rise to $\eta_1 = \eta_2$ and $\nu_1 = \nu_2$, contrary to the defining feature that these are distinct posteriors.

Figure 12 illustrates the Trapezoid Condition. η_1 and ν_1 both lie in $\Gamma_{12}(\alpha_1)$ and η_2 and ν_3 both lie in $\Gamma_{12}(\alpha_2)$. η_1 and η_2 both on the line segment $\Phi_{12}(\eta_1)$ and ν_1 and ν_2 both on the line segment $\Phi_{12}(\nu_1)$. The key observation is that, since the points in $\Phi_{12}(\eta_1)$ and $\Phi_{12}(\nu_1)$ only differ in their first and second coordinate, the two line segments are parallel, so that the points η_1, η_2, ν_1 and ν_2 form a trapezoid. Below we will use this observation to establish the additive separability of $\tilde{T}_{(ji)}(\eta_1)$.

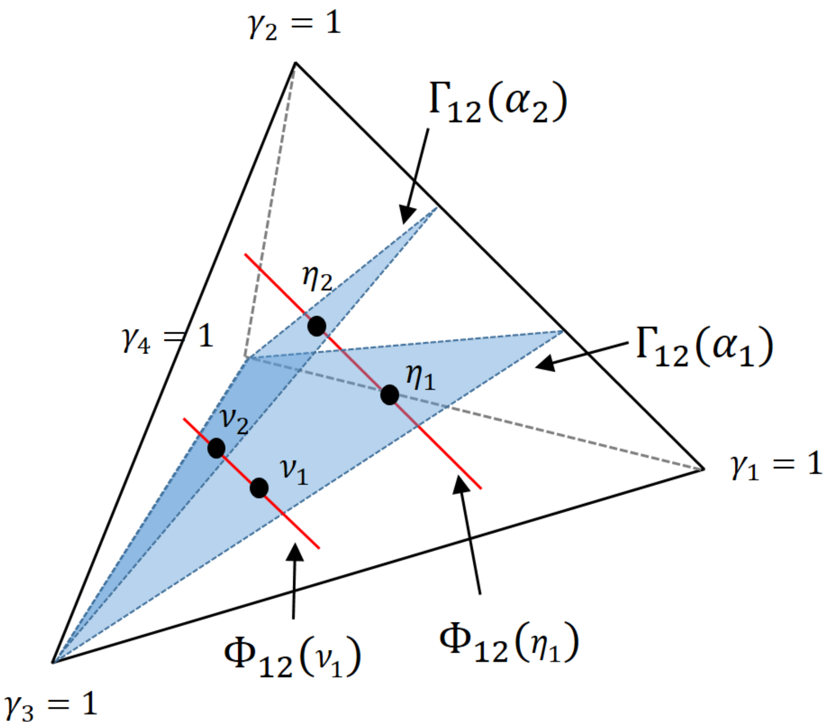

Figure 12

A5.2.6: Equalization of Differences For Two-Sided Directional Derivatives

Our next result relates the Trapezoid Condition, IUC, and the Lagrangian Lemma. If four posteriors η_1, η_2, ν_1, and ν_2 satisfy the Trapezoid Condition and if $\tilde{T}$ is differentiable at each of these points, then IUC and the Lagrangian Lemma relate the change in $\tilde{T}_{(ji)}$ between η_1 and η_2 to that between ν_1 and ν_2.

Lemma 5.13: Suppose $C \in \mathcal{C}$ has a UPS representation $\tilde{T}$ and satisfies IUC (A1). If $\eta_1, \eta_2, \nu_1, \nu_2 \in \tilde{\Gamma}'$ satisfy the Trapezoid Condition for some pair of distinct states $1 \leq k \neq l \leq J$, then

$$\tilde{T}_{(ji)}(\eta_1) - \tilde{T}_{(ji)}(\eta_2) = \tilde{T}_{(ji)}(\nu_1) - \tilde{T}_{(ji)}(\nu_2) \tag{89}$$

for all pairs of distinct states $1 \leq i \neq j \leq J$

Proof. Consider η_1,ν_1,η_2 and ν_2 satisfying the Trapezoid Condition such that $\tilde{T}_{(ji)}$ exists at all four points. By Lemma 5.12, defining,

$$\bar{\mu} = \frac{\eta_1 + \nu_1}{2},$$

there exists $a, b \in \mathcal{A}$ with $u(a,k) = u(a,l)$ and $u(b,k) = u(b,l)$ such that, for $t \in [0,1]$,

$$C(\mu(t), \{a,b\}) = \{P_t\};\ \bar{\gamma}^a_{P_t} = \eta(t),\ \text{and } \bar{\gamma}^b_{P_t} = \nu(t).$$

where,

$$\mu(t,j) = \begin{cases} t[\bar{\mu}(k) + \bar{\mu}(l)] & \text{for } j = k; \\ (1-t)[\bar{\mu}(k) + \bar{\mu}(l)] & \text{for } j = l; \\ \bar{\mu}(j) & \text{otherwise.} \end{cases}$$

and

$$\begin{aligned} \eta(t,j) &= \left[\frac{\eta_1(j)}{\bar{\mu}(j)}\right] \mu(t,j) \text{ for } 1 \leq j \leq J; \\ \nu(t,j) &= \left[\frac{\nu_1(j)}{\bar{\mu}(j)}\right] \mu(t,j) \text{ for } 1 \leq j \leq J; \end{aligned}$$

Where we have placed t as an argument in brackets and avoided the subscript so as to avoid confusion with η_1 and ν_1.

In particular, for $\bar{t}_1 = \frac{\bar{\mu}(k)}{\bar{\mu}(k)+\bar{\mu}(l)} \in (0,1)$,

$$\begin{aligned} \mu(\bar{t}_1) &= \bar{\mu}, \\ \eta(\bar{t}_1) &= \eta_1, \\ \nu(\bar{t}_1) &= \nu_1. \end{aligned}$$

Moreover, since $\tilde{T}$ is differentiable at both η_1 and ν_1, Lemma 5.11, the Optimality and Directional Derivatives Lemma, then implies,

$$\tilde{N}^a_{(ji)}(\eta_1) = \tilde{N}^b_{(ji)}(\nu_1), \tag{90}$$

for all $i, j \in J$.

Now note generally that:

$$\begin{aligned} \eta(t,k) &= \left[\frac{\eta_1(k)}{\bar{\mu}(k)}\right] \bar{\mu}(t,k) \\ &= t\left[\frac{\eta_1(k)}{\bar{\mu}(k)}\right] [\bar{\mu}(k) + \bar{\mu}(l)] \\ &= t\eta_1(k)\left[\frac{\eta_1(k) + \nu_1(k) + \eta_1(l) + \nu_1(l)}{\eta_1(k) + \nu_1(k)}\right] \\ &= t\eta_1(k)\left[\frac{\alpha(1)\eta_1(l) + \alpha(1)\nu_1(l) + \eta_1(l) + \nu_1(l)}{\alpha(1)\eta_1(l) + \alpha(1)\nu_1(l)}\right] \\ &= t\eta_1(k)\left[\frac{\alpha(1)+1}{\alpha(1)}\right]; \end{aligned}$$

where $\alpha_1 = \eta_1(k)/\eta_1(l)$.

Hence if we define $\bar{t}_2$ such that

$$\bar{t}_2 = \frac{\eta_2(k)\alpha_1}{\eta_1(k)(1+\alpha_1)},$$

then,

$$\eta(\bar{t}_2, k) = \eta_2(k).$$

and since $\eta_2(j) \in \Phi_{jk}(\eta_1)$ implies that $\eta(t, j) = \eta_2(j)$ for all $j \neq k, l$,

$$\eta(\bar{t}_2) = \eta_2.$$

Note that since $\eta_2(j) \in \Phi_{jk}(\eta_1)$, $\eta_2(j)$ lies on the line segment connecting $\eta(0)$ and $\eta(1)$, hence $\bar{t}_2 \in [0, 1]$

Similarly, we can show that

$$\nu(t, k) = t\nu_1(k)\left[\frac{\alpha_1 + 1}{\alpha_1}\right].$$

so that

$$\nu(\bar{t}_2, k) = \frac{\eta_2(k)}{\eta_1(k)}\nu_1(k) = \nu_2(k)$$

where the last equality follows from the following line of reasoning:

1. $\eta_2 \in \Phi_{jk}(\eta_1)$ implies $\eta_1(k) + \eta_1(l) = \eta_2(k) + \eta_2(l)$.
2. $\eta_1 \in \Gamma_{12}(\alpha_1)$ and $\eta_2 \in \Gamma_{12}(\alpha_2)$ imply further that $\eta_1(k)[1 + 1/\alpha_1] = \eta_2(k)[1 + 1/\alpha_2]$ for $\alpha_1 = \eta_1(k)/\eta_1(l)$ and $\alpha_2 = \eta_2(k)/\eta_2(l)$.
3. Similarly, $\nu_1(k)[1 + 1/\alpha_1] = \nu_2(k)[1 + 1/\alpha_2]$.
4. So that $\eta_1(k)/\nu_1(k) = \eta_2(k)/\nu_2(k)$ as required.

Therefore given the problem $(\mu(\bar{t}_2), A)$, η_2 is the revealed posterior associated with a and ν_2 is the revealed posterior associated with b. Since $\tilde{T}$ is differentiable at both η_2 and ν_2, Lemma 5.11 then implies that,

$$\tilde{N}^a_{(ji)}(\eta_2) = \tilde{N}^b_{(ji)}(\nu_2) \tag{91}$$

for all $i, j \in J$. Since $\tilde{N}^a(\gamma) = \sum_j u(a, j)\gamma(j) - \tilde{T}(\gamma)$, equations (90) and (91) imply

$$\tilde{T}_{(ji)}(\eta_1) = u(b, i) - u(b, j) - u(a, i) + u(a, j) + \tilde{T}_{(ji)}(\nu_1)$$

and

$$\tilde{T}_{(ji)}(\eta_2) = u(b, i) - u(b, j) - u(a, i) + u(a, j) + \tilde{T}_{(ji)}(\nu_2)$$

Subtracting these two equations yields the desired result. ■

A5.2.7: Monotonicity and Limits

Condition (89) only holds at points at which the two-sided directional derivatives exist. Our next goal is to generalize to one-sided directional derivatives which always exist, thereby extending the result to all sets of posteriors in $\tilde{\Gamma}$ that satisfy the Trapezoid Condition. Our strategy will be to take limits of well-chosen sequences at which the two-sided derivatives exist. Two standard features of the subdifferential map (associating with each posterior $\gamma \in \tilde{\Gamma}$ the full set of corresponding subderivatives of T) allow us to select appropriate sequences. The first is that the sub-differential maps of convex functions are monotone. The second is that they satisfy a form of lower hemi-continuity. The next two lemmas translate these standard results to our setting, starting with the monotonicity lemma.

Lemma 5.14: Given $\gamma \in \tilde{\Gamma}$ and $\epsilon > 0$ such that $\gamma + \epsilon(e_i - e_j) \in \tilde{\Gamma}$,

$$\tilde{T}_{\overrightarrow{ji}}(\gamma + \epsilon(e_i - e_j)) \geq -\tilde{T}_{\overrightarrow{ij}}(\gamma + \epsilon(e_i - e_j)) \geq \tilde{T}_{\overrightarrow{ji}}(\gamma) \tag{92}$$

Proof. The result follows directly from monotonicity properties of the subdifferential maps of convex functions (Rockafellar p. 240). This is particularly simple for one dimensional functions as the general version of the statement that differentiable convex functions have non-decreasing first derivatives. To use in this simple setting, let $\delta > \epsilon > 0$ such that,

$$Y \equiv (\gamma - \delta(e_i - e_j), \gamma + \delta(e_i - e_j)) \subset \tilde{\Gamma},$$

which is possible since $\tilde{\Gamma}$ is relatively open on the line. We then define convex function $G : \mathbb{R} \to \mathbb{R}$ to have value

$$G(\alpha) = \tilde{T}(\gamma + \alpha(e_i - e_j)),$$

on $\alpha \in (-\delta, \delta)$, with its value being infinite elsewhere. It is direct from the definitions that $\tilde{T}_{\overrightarrow{ji}}(\gamma + \alpha(e_i - e_j))$ and the right derivatives of $G(\alpha)$ are equivalent at corresponding points,

$$\begin{aligned} \tilde{T}_{\overrightarrow{ji}}(\gamma + \alpha(e_i - e_j)) &= \lim_{\varsigma \downarrow 0} \frac{\tilde{T}(\gamma + \alpha(e_i - e_j) + \varsigma(e_i - e_j)) - \tilde{T}(\gamma + \alpha(e_i - e_j))}{\varsigma} \\ &= \lim_{\varsigma \downarrow 0} \frac{G(\alpha + \varsigma) - G(\alpha)}{\varsigma} \equiv G'_+(\alpha). \end{aligned}$$

Similarly, $-\tilde{T}_{\overrightarrow{ij}}$ and the left derivatives of $G'_-(\alpha)$ are equivalent where

$$G'_-(\alpha) = -\lim_{\varsigma \uparrow 0} \frac{G(\alpha + \varsigma) - G(\alpha)}{\varsigma}$$

Rockafellar theorem 24.1 establishes monotonicity properties for $G'_+(\alpha)$ and $G'_-(\alpha)$ when G is convex. In particular for $\epsilon > 0$,

$$G'_+(0) \leq G'_-(\epsilon) \leq G'_+(\epsilon).$$

which, given the equivalence between $\tilde{T}$ and G, establishes (92). ■

Lemma 5.15: Given any sequence $\{\epsilon_n\}_{n=1}^{\infty}$ with $\epsilon_n > 0$ and $\lim_{n\to\infty} \epsilon_n = 0$,

$$\lim_{n\to\infty} \tilde{T}_{\overrightarrow{ji}}(\gamma + \epsilon_n(e_i - e_j)) = \tilde{T}_{\overrightarrow{ji}}(\gamma) \tag{93}$$

Proof. By the directional derivative monotonicity lemma 5.14, given $\{\epsilon_n\}_{n=1}^{\infty}$ with $\epsilon_n > 0$, and $\lim_{n\to\infty} \epsilon_n = 0$, such that $\gamma + \epsilon_n(e_i - e_j) \in \tilde{\Gamma}$, (92) implies that the limit exists and that the inequality survives,

$$\lim_{n\to\infty} \tilde{T}_{\overrightarrow{ji}}(\gamma + \epsilon_n(e_i - e_j)) \geq \tilde{T}_{\overrightarrow{ji}}(\gamma). \tag{94}$$

Conversely, Rockafellar 24.5 shows with full generality that given convex function $\tilde{T} : \mathbb{R}^J \to \bar{\mathbb{R}}$, any $\gamma \in \mathbb{R}^J$ at which $\tilde{T}(\gamma)$ is finite, and any sequence $\{\gamma^n\}_{n=1}^{\infty} \to \gamma$ and $y \in \mathbb{R}^J$,

$$\limsup \tilde{T}'(\gamma^n|y) \leq \tilde{T}'(\gamma|y). \tag{95}$$

Defining $\gamma^n = (\gamma + \epsilon_n(e_i - e_j)$ and $y = e_i - e_j$ we get,

$$\lim_{n\to\infty} \tilde{T}_{\overrightarrow{ji}}(\gamma + \epsilon_n(e_i - e_j)) \leq \tilde{T}_{\overrightarrow{ji}}(\gamma). \tag{96}$$

Combining (94) with (96) establishes (93), and completes the proof of the Lemma. ■

A5.2.8: Equal Difference Conditions for One-Sided Directional Differences

To complete the transition from equal difference conditions on two-sided to one-sided directional derivatives, we apply standard results on "almost everywhere" differentiability of convex functions on various important subdomains. These are all convex subdomains of $\tilde{\Gamma}$ whose affine hull is of lower dimension, several of which have appeared already in the proof. When thinking about $\tilde{T}$ even on its full domain $\tilde{\Gamma}$, there is a subtlety in the statement. Since $\tilde{\Gamma}$ respects the adding up constraint on probabilities, it has measure zero as a subset of $\mathbb{R}^J$. For that reason the full measure result applies "relative to" $\tilde{\Gamma}$. This is how we state the corresponding result for more general convex subdomains Y. We note also the preservation of one-sided and two-sided directional derivatives on subdomains. The precise formalism is standard.

Definition 11 *Given a non-empty convex set $Y \subset \mathbb{R}^J$, define $\tilde{\Gamma}(Y)$ to be the corresponding subdomain of $\tilde{\Gamma}$,*

$$\tilde{\Gamma}(Y) \equiv Y \cap \tilde{\Gamma};$$

and define $\tilde{T}^Y : \tilde{\Gamma}(Y) \to \mathbb{R}$ with $\tilde{T}^Y(\gamma) \equiv \tilde{T}(\gamma)$ to be the restriction of $\tilde{T}$ to this domain. We define $\tilde{\Gamma}'(Y) \subset \tilde{\Gamma}(Y)$ to be the set on which $\tilde{T}^Y$ is differentiable. Finally, given $\gamma \in \tilde{\Gamma}(Y)$ and $1 \leq i \neq j \leq J$ such that Y for which there exists $\delta > 0$ such that,

$$[\gamma - \delta(e_i - e_j), \gamma + \delta(e_i - e_j)] \subset \tilde{\Gamma}(Y), \tag{97}$$

we define one and two-sided directional derivative $\tilde{T}^Y_{\overrightarrow{ji}}$ and $\tilde{T}^Y_{(ji)}$ in the standard manner.

We now state the key result about convergence of one-sided directional derivatives for appropriately selected sequences of posteriors.

Lemma 5.16: For any non-empty convex set $Y \subset \mathbb{R}^J$, $\tilde{T}^Y$ is almost everywhere differentiable in the relative interior of $\tilde{\Gamma}(Y)$ and $\tilde{\Gamma}(Y)\backslash\Gamma'(Y)$ is of measure zero, and whenever $\tilde{T}^Y_{\overrightarrow{ji}}(\gamma)$ is well-defined,

$$\tilde{T}^Y_{\overrightarrow{ji}}(\gamma) = \tilde{T}_{\overrightarrow{ji}}(\gamma). \tag{98}$$

Proof. Given that Y is non-empty and convex set, $\tilde{T}^Y$ is a proper convex function. Rockafellar theorem 25.5 translates precisely to the fact that $\tilde{\Gamma}'(Y)$ is dense in $\tilde{\Gamma}(Y)$ and its complement $\tilde{\Gamma}(Y)\backslash\tilde{\Gamma}'(Y)$ is of measure zero in the relative interior of $\tilde{\Gamma}(Y)$. The equality $\tilde{T}^Y_{\overrightarrow{ji}} = \tilde{T}_{\overrightarrow{ji}}$ is definitional given the existence of appropriate convergent sequences in the shared domain and equality of the underlying function. ■

Lemma 5.17: Suppose $C \in \mathcal{C}$ has a UPS representation $\tilde{T}$ and satisfies IUC (A1). Given $\{\eta_1, \eta_2, \nu_1, \nu_2\} \subset \tilde{\Gamma}$ satisfying the Trapezoid Condition for some pair of distinct states $1 \leq k \neq l \leq J$, then

$$\tilde{T}_{\overrightarrow{ji}}(\eta_1) - \tilde{T}_{\overrightarrow{ji}}(\nu_1) = \tilde{T}_{\overrightarrow{ji}}(\eta_2) - \tilde{T}_{\overrightarrow{ji}}(\nu_2), \tag{99}$$

for all pairs of distinct states $i, j \in \{1, \ldots J\} \backslash \{k, l\}$ that are distinct from k and l.

Proof. Consider an arbitrary set of four posteriors $\{\xi_m\} \subset \tilde{\Gamma}$ for $\xi = \eta, \nu$ and $m = 1, 2$ satisfying the Trapezoid Condition. We construct four corresponding sequences of posteriors $\{\xi_m^n\}_{n=1}^{\infty}$ that converge to ξ_m as $n \to \infty$. We ensure that at each n (89) holds and that (89) converges to (99). In light of Lemma 5.14 and 4.15, we ensure that the sequence is picked in a special manner ensuring proper convergence. Specifically, let $\Theta(\xi_m)$ be the set containing posteriors that lie within $\frac{1}{n}$ of ξ_m in the direction $\overrightarrow{ji}$:

$$\Theta(\xi_m) = \{\gamma \in \tilde{\Gamma} | \gamma = \xi_m + \lambda(e_i - e_j) \text{ and } \lambda \in (0, \frac{1}{n})\}.$$

for all $\gamma \in \Theta(\xi_m)$.

Note that $\Theta(\xi_m)$ is a convex subset of $\tilde{\Gamma}$, hence satisfies the conditions of the Lemma 5.16, so that $\tilde{T}_{(ji)}^{\Theta(\xi_m))} = \tilde{T}_{(ji)}$ exists for almost all $\lambda \in (0, \frac{1}{n})$. Let $\Lambda_n(\xi_m)$ denote the set of $\lambda \in (0, \frac{1}{n})$ at which the two-sided directional derivative $\tilde{T}_{(ji)}$ exists,

$$\Lambda_n(\xi_m) = \{\lambda \in (0, \frac{1}{n}) | \tilde{T}_{(ij)}(\gamma) \text{ exists at } \gamma = \xi_m + \lambda(e_i - e_j)\}.$$

It follows that $\Lambda_n(\xi_m)$ has measure $\frac{1}{n}$, as does the corresponding intersection,

$$\Lambda(n) \equiv \cap_{\xi=\eta,\nu} \cap_{m=1,2} \Lambda_n(\xi_m).$$

Select $\bar{\lambda}(n) \in \Lambda(n)$ and correspondingly define,

$$\xi_m^n = \xi_m + \bar{\lambda}(n)(e_i - e_j),$$

for $\xi = \eta, \nu$ and $m = 1, 2$.

By construction, for each n, $\xi_m^n \in \tilde{\Gamma}$ for $\xi = \eta, \nu$ and $m = 1, 2$ satisfy the Trapezoid Condition. To confirm, note that for $\xi = \eta, \nu$ and $m = 1, 2$, $\xi_m^n(k)$ and $\xi_m(k)$ differ only in coordinates i and j. Hence we know that $\xi_m^n(k) = \xi_m(k)$ and $\xi_m^n(l) = \xi_m(l)$. Given that the ξ_m satisfy the Trapezoid Condition, it follows that the ξ_m^n satisfy the Trapezoid Condition as well. Since $\tilde{T}_{(ji)}$ exists at each of the ξ_m^n, and since the ξ_m^n satisfy the Trapezoid Condition, Lemma 5.13 states

$$\tilde{T}_{(ji)}(\eta_1^n) - \tilde{T}_{(ji)}(\nu_1^n) = \tilde{T}_{(ji)}(\eta_2^n) - \tilde{T}_{(ji)}(\nu_2^n) \tag{100}$$

Now consider each element $\tilde{T}_{(ji)}(\xi_m^n)$ and note that, since $\bar{\lambda}(n) > 0$ and

$$\xi_m^n = \xi_m + \bar{\lambda}(n)(e_i - e_j),$$

we can apply Lemma 5.14 directly to conclude that (92) holds,

$$\tilde{T}_{\overrightarrow{ji}}(\xi_m^n) \geq \tilde{T}_{\overrightarrow{ji}}(\xi_m) \tag{101}$$

Since $\bar{\lambda}(n) \to 0$ we know in addition that $\lim_{n \longrightarrow \infty} \xi_m^n = \xi_m$, so that Lemma 5.15 applies to show that,

$$\lim_{n \to \infty} \tilde{T}_{\overrightarrow{ji}}(\xi_m^n) = \lim_{n \to \infty} \tilde{T}_{(ji)}(\xi_m^n) = \tilde{T}_{\overrightarrow{ji}}(\xi_m). \tag{102}$$

Substituting (102) in (100) establishes (99), and completes the proof. ■

A5.2.9: Propagating Existence of Directional Derivatives

A key result shows how to propagate existence of directional derivatives. If the two-sided directional derivative $\tilde{T}_{(ji)}(\eta)$ exists at a point η, then $\tilde{T}_{(ji)}(\nu)$ exists for all $\nu \in \Gamma_{ji}^{\alpha}$ where $\alpha = \frac{\eta(j)}{\eta(i)}$. The intuition for the result is that this set of v can be linked to η by a problem in which the states i and j are redundant. IUC then allows us to alter the prior on i and j and thereby smoothly shift the the resulting posteriors. These posteriors maintain the original ratio between states j and i: $\frac{\nu(j)}{\nu(i)} = \frac{\nu(j)}{\nu(i)}$. If $\tilde{T}(\eta)$ is smooth in the direction (ji), $\tilde{T}(\nu)$ must also be smooth, if the posteriors are to evolve proportionately.

Before proving the result we establish some additional continuity properties that we can import to our apparatus directly from Rockafellar.

Lemma 5.18: Given $\eta \in \tilde{\Gamma}$ and $1 \leq i \neq j \leq J$ such that $\tilde{T}_{(ji)}(\eta)$ exists, then $\tilde{T}_{(ji)}(\nu)$ exists for all $\nu \in \tilde{\Gamma}$ such that:

$$\frac{\nu(j)}{\nu(i)} = \frac{\eta(j)}{\eta(i)}. \tag{103}$$

Proof. The proof is by contradiction. Choose η such that $\tilde{T}_{(ji)}(\eta)$ exists and suppose that there exists ν satisfying (103) such that $\tilde{T}_{(ji)}(\nu)$ does not exist. By Lemma 5.9 above, this means that $-\tilde{T}_{\overrightarrow{jl}}(\nu) \neq \tilde{T}_{\overrightarrow{lj}}(\nu)$.

By Lemma 5.12 there exists $(\mu, A) \in \mathcal{D}$ with $A = \{a, b\}$ such that η is the revealed posterior related to a and ν is the revealed posterior related to b and the states i and j are redundant. The lemma then guarantees that given the parameterized set of problems (μ_t, A) where

$$\mu_t(k) = \begin{cases} t\,[\mu(i) + \mu(j)] & \text{for } k = i; \\ (1-t)\,[\mu(i) + \mu(j)] & \text{for } k = j; \\ \mu(k) & \text{otherwise;} \end{cases}$$

for $t \in (0, 1)$, η_t is the revealed posterior for action a and ν_t is the revealed posterior for action b and

$$\eta_t(j) = \frac{\eta(j)}{\mu(j)}\mu_t(j) \qquad \text{and} \qquad \nu_t(j) = \frac{\nu(j)}{\mu(j)}\mu_t(j)$$

Let $\bar{t}$ be defined by $\mu_{\bar{t}} = \mu$.

By the Lagrangian Lemma, for each t, there exists $\theta_t \in \mathbb{R}^{J-1}$ such that

$$\tilde{N}^a(\eta_t) - \sum_{k=1}^{J-1} \theta_t(k)\eta_t(k) = \tilde{N}^b(\nu_t) - \sum_{k \neq j} \theta_t(k)\nu_t(k) \geq \tilde{N}^c(\gamma) - \sum_{k=1}^{J-1} \theta(k)\gamma(k)$$

for all $\gamma \in \Gamma$ and $c = \{a, b\}$.

Since $\tilde{T}_{(ji)}(\eta)$ exists, the Optimal Directional Derivative Lemma (4.11) tells us that

$$\tilde{T}_{(ji)}(\eta) = u(a,i) - u(a,j) - \theta_{\bar{t}}(i) - \theta_{\bar{t}}(j) \tag{104}$$

and since $\tilde{T}_{(ji)}(\nu)$ does not exist, Lemma 5.11 implies that,

$$-\tilde{T}_{\overrightarrow{ij}}(\nu) - u(b,i) + u(b,J) + \theta_{\bar{t}}(i) + \theta_{\bar{t}}(j) \leq 0 \leq \tilde{T}_{\overrightarrow{ji}}(\nu) - u(b,i) + u(b,J) + \theta_{\bar{t}}(i) + \theta_{\bar{t}}(j),$$

with one of these two inequalities strict. Without loss of generality suppose

$$\tilde{T}_{\overrightarrow{ji}}(\nu) - u(b,i) + u(b,J) + \theta_{\bar{t}}(i) + \theta_{\bar{t}}(j) = \Delta > 0. \tag{105}$$

Define now $Y \subset \mathbb{R}^J$ as all vectors η_t,

$$Y = \{\eta_t \in \tilde{\Gamma} | t \in [0,1]\},$$

noting that, since $\tilde{\Gamma}(Y) = Y$ since $Y \subset \tilde{\Gamma}$. Lemma 5.16 implies that $\tilde{T}^Y$ is differentiable for almost all η_t, so that Lemma 5.9 implies that the two-sided directional derivative,

$$\tilde{T}^Y_{(ji)}(\eta_t) = \tilde{T}_{(ji)}(\eta_t),$$

also exists for almost all $\eta_t \in Y$.

Now consider a sequence $\eta_{t(n)} \to \eta$ such that $t(n) > \bar{t}$ and $\tilde{T}_{(ji)}(\eta_{t(n)})$ exists. Lemma 5.15 implies that

$$\lim_{t(n) \to \bar{t}} \tilde{T}_{(ji)}(\eta_{t(n)}) = \tilde{T}_{(ji)}(\eta).$$

Therefore there exists $t(m) \neq \bar{t}$ such that $\tilde{T}_{(ji)}(\eta_{t(m)}) \in (\tilde{T}_{(ji)}(\eta), \tilde{T}_{(ji)}(\eta) + \Delta)$. Given that $\tilde{T}_{(ji)}(\eta_{t(m)})$ exists,

$$\tilde{T}_{(ji)}(\eta_{t(m)}) = u(a,i) - u(a,j) - \theta_{t(m)}(i) - \theta_{t(m)}(j). \tag{106}$$

Hence, with $\tilde{T}_{(ji)}(\eta_{t(m)}) - \tilde{T}_{(ji)}(\eta) \in (0, \Delta)$, we can subtract the right-hand sides of (104) from (106) to conclude that,

$$\theta_{\bar{t}}(i) + \theta_{\bar{t}}(j) - \theta_{t(m)}(i) - \theta_{t(m)}(j) \in (0, \Delta),$$

so that,

$$-\theta_{t(m)}(i) - \theta_{t(m)}(j) < -\theta_{\bar{t}}(i) - \theta_{\bar{t}}(j) + \Delta. \tag{107}$$

Applying now the Optimal Directional Derivative Lemma (4.11) to $\nu_{t(m)}$ we conclude that,

$$-\tilde{T}_{\overrightarrow{ij}}(\nu_{t(m)}) \leq u(b,i) - u(b,j) - \theta_{t(m)}(i) - \theta_{t(m)}(j) \leq \tilde{T}_{\overrightarrow{ji}}(\nu_{t(m)}).$$

Substitution of (107) thereupon yields,

$$-\tilde{T}_{\overrightarrow{ij}}(\nu_{t(m)}) \leq u(b,i) - u(b,j) - \theta_{t(m)}(i) - \theta_{t(m)}(j) < u(b,i) - u(b,j) - \theta_{\bar{t}}(i) - \theta_{\bar{t}}(j) + \Delta = \tilde{T}_{\overrightarrow{ji}}(\nu)$$

But $t(m) > \bar{t}$, so that the Lemma 5.14 implies directly that $-\tilde{T}_{\overrightarrow{ij}}(\nu_{t(m)}) > \tilde{T}_{\overrightarrow{ji}}(\nu)$. This contradiction establishes the result. ■

A5.2.10: Weak Form Additive Separability

In this section we establish a weak form of additive separability. The basic observation is that (89) is very close to the rectangle condition for this form of additive separability. The difference is that $\eta_1, \eta_2, \nu_1, \nu_2 \in \tilde{\Gamma}$ satisfying the Trapezoid Condition form a trapezoid, not a rectangle. We rectify this problem by deforming $\tilde{\Gamma}$.

Definition 12 *Given $M \in \mathbb{N}$, define Z^M as the strictly positive vectors summing to strictly below 1,*

$$Z^M = \{(z_1, \ldots z_M) \in R^M | z_m > 0 \textit{ all } m \textit{ and } \sum_m z_m < 1\},$$

and define $X = (0,1) \times Z^{J-2}$.

We now deform $\tilde{\Gamma}$ to create rectangles. To simplify the notation, we start with arbitrary states k and l but then re-order the states so that $k = 1$ and $l = J$. Given Lemma 5.7, this is without loss of generality. With this naming convention we will suppress the dependence of the function Ψ on k and l in the following definition.

Definition 13 *Define $\Psi : \tilde{\Gamma} \to X$ with $\Psi(\gamma) = x \in X$ where:*

$$x(j) = \begin{cases} \frac{\gamma(j)}{\gamma(j)+\gamma(J)} & \textit{for } j = 1; \\ \gamma(j) & \textit{for } 2 \leq i \neq j \leq J-1. \end{cases}$$

The next Lemma points out that Ψ is bijective.

Lemma 5.19: $\Psi : \tilde{\Gamma} \to X$ is bijective.

Proof. The mapping Ψ can be constructed as the combination of two mappings each of which we show to be bijective. The first maps the $J-1$ dimensional simplex $\tilde{\Gamma}$ to Z^{J-1} by dropping the coordinate $\gamma(J)$. Given $\gamma \in \tilde{\Gamma}$ define $\Psi_1(\gamma) \in Z^{J-1}$ by:

$$\Psi_1(\gamma(j)) = \gamma(j) \text{ for } 1 \leq j \leq J-1.$$

The second function divides $\gamma(1)$ by,

$$\gamma(1) + \gamma(J) = 1 - \sum_{m=2}^{J-1} \gamma(m).$$

Technically, given $z \in Z^{J-1}$ define $\Psi_2(z) \in X$ by,

$$\Psi_2(z(j)) = \begin{cases} \frac{z(j)}{1-\sum_{m=2}^{J-1} z(m)} & \text{for } j=1; \\ z(j) & \text{for } 2 \leq i \neq j \leq J-1. \end{cases}$$

Clearly $\Psi_1 : \tilde{\Gamma} \to Z^{J-1}$ is bijective. With regard to Ψ_2, note that if that $z_1, z_2 \in Z^{J-1}$ both map to $x \in X$, it is immediate that $z_1 = z_2$. Hence $\Psi_2 : Z^{J-1} \to X$ is injective. To show that it is also surjective, given $x \in X$, define $h(x) \in Z^{J-1}$ by,

$$h(x(j)) = \begin{cases} \left[1 - \sum_{m=2}^{J-1} x(m)\right] x(1) & \text{if } j=1 \\ x(j) & \text{for } 2 \leq j \leq J-1. \end{cases}$$

We now consider $\Psi_2(h(x)) \in X$. By construction, this satisfies:

$$\Psi_2(h(x(j))) = \begin{cases} \frac{h(x(j))}{1-\sum_{m=2}^{J-1} h(x(m))} & \text{for } j=1; \\ x(j) & \text{for } 2 \leq i \neq j \leq J-1. \end{cases}$$

where

$$\frac{h(x(j))}{1-\sum_{m=2}^{J-1} h(x(m))} = \frac{\left[1 - \sum_{m=2}^{J-1} x(m)\right] x(1)}{\left[1 - \sum_{m=2}^{J-1} x(m)\right]} = x(1).$$

Hence $\Psi_2(h(x)) = x$ so that Ψ_2 is surjective. Given that it is also injective, it is bijective.

To complete the proof, we now show that $\Psi = \Psi_2 \circ \Psi_1$ is the composition of these mappings:

$$\begin{aligned} \Psi &= \Psi_2(\Psi_1(\gamma)) = \begin{cases} \frac{\Psi_1(\gamma(j))}{1-\sum_{m=2}^{J-1} \Psi_1(\gamma(m))} & \text{for } j=1; \\ \Psi_1(\gamma(j)) & \text{for } 2 \leq i \neq j \leq J-1. \end{cases} \\ &= \begin{cases} \frac{\gamma(j)}{1-\sum_{m=2}^{J-1} \gamma(m)} & \text{for } j=1; \\ \gamma(j) & \text{for } 2 \leq i \neq j \leq J-1. \end{cases} \end{aligned}$$

This completes the proof that Ψ is bijective. ■

The next lemma shows that, in this space, the Trapezoid Condition transforms into a rectangle condition on X.

Lemma 5.20: Given $\eta_1, \eta_2, \nu_1, \nu_2 \in \tilde{\Gamma}$ that satisfy the Trapezoid Condition for states $k = 1$ and $j = J$, the elements $x_1, x_2, y_1, y_2 \in X$ such that $x_m = \Psi(\eta_m)$ and $y_m = \Psi(\nu_m)$ for $m = 1, 2$, form a rectangle:

$$\begin{aligned} x_1(1) &= x_2(1) \text{ and } y_1(1) = y_2(1); \\ x_1(j) &= y_1(j) \text{ and } x_2(j) = y_2(j); \text{ for } 2 \leq j \leq J-1. \end{aligned}$$

Proof. Consider $x_1, x_2, y_1, y_2 \in X$ such that $x_m = \Psi(\eta_m)$ and $y_m = \Psi(\upsilon_m)$ for $m = 1, 2$. By the Trapezoid Condition and the definition of Ψ, for $2 \leq j \leq J-1$ and $m = 1, 2$,

$$x_m(j) = \Psi(\eta_m(j)) = \eta_m(j) = \nu_m(j) = \Psi(\nu_m(j)) = \nu_m(j).$$

Note also that,

$$\begin{aligned} x_1(1) - x_2(1) &= \frac{\eta_1(1)}{\eta_1(1)+\eta_1(J)} - \frac{\eta_2(1)}{\eta_2(1)+\eta_2(J)} \\ &= \frac{\frac{\eta_1(1)}{\eta_1(J)}}{\frac{\eta_1(1)}{\eta_1(J)}+1} - \frac{\frac{\eta_2(1)}{\eta_2(J)}}{\frac{\eta_2(1)}{\eta_2(J)}+1} = \frac{\alpha}{\alpha+1} - \frac{\alpha}{\alpha+1} = 0 \end{aligned}$$

Similarly,

$$y_1(1) - y_2(1) = \frac{\frac{\nu_1(1)}{\nu_1(J)}}{\frac{\nu_1(1)}{\nu_1(J)}+1} - \frac{\frac{\nu_1(1)}{\nu_1(J)}}{\frac{\nu_1(1)}{\nu_1(J)}+1} = 0,$$

completing the proof. ■

With this we are in position to establish our first version of additive separability.

Lemma 5.21: Suppose $C \in \mathcal{C}$ has a UPS representation and satisfies IUC (A1). Then, given $2 \leq i \neq j \leq J-1$, $\tilde{T}_{\overrightarrow{ji}}(\gamma)$ is additively separable in $\left[\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right]$ and $\{\gamma(j) \mid 2 \leq j \leq J-1\}$ in that there exists $\boldsymbol{A} : \mathbb{R}_+ \longrightarrow \mathbb{R}$ and $\boldsymbol{B} : \mathbb{R}^{J-2} \longrightarrow \mathbb{R}$ such that,

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

Proof. We consider any four posteriors $\eta_1, \eta_2, \nu_1, \nu_2 \in \tilde{\Gamma}$ that satisfy the Trapezoid Condition for states $k = 1$ and $l = J$. We now define $x_1, x_2, y_1, y_2 \in X$ by $x_m = \Psi(\eta_m)$ and $y_m = \Psi(\upsilon_m)$ for $m = 1, 2$. We transfer the directional derivatives to this space by defining the function $\mathcal{T} : X \to \mathbb{R}$ by,

$$\mathcal{T}(x) \equiv \tilde{T}_{\overrightarrow{ji}}(\Psi^{-1}(x)),$$

using the bijective function $\Psi : \tilde{\Gamma} \to X$ introduced above.

Note that the space X is of the cross-product form $X = X_A \times X_B$ with $X_A = (0,1)$ and $X_B = Z^{J-2}$. A standard condition for such an arbitrary function $f : X \longrightarrow \mathbb{R}$ on such a space to be additively,

$$f(a,b) = f_1(a) + f_2(b)$$

is that the rectangle conditions are satisfied: given $a_1, a_2 \in X_A$ and $b_1, b_2 \in X_B$,

$$f(a_1, b_1) - f(a_2, b_1) = f(a_1, b_2) - f(a_2, b_2).$$

To confirm, pick arbitrary $\left(\bar{a}, \bar{b}\right) \in X_A \times X_B$ and note that for any $(a,b) \in X_A \times X_B$,

$$f(a,b) = f(a,\bar{b}) + f(\bar{a},b) - f\left(\bar{a},\bar{b}\right),$$

which is of the additively separable form for $f_1(a) = f(a,\bar{b}) - f\left(\bar{a},\bar{b}\right)$ and $f_2(b) = f(\bar{a},b)$.

Since $\eta_1, \eta_2, \upsilon_1$, and ν_2 satisfy the Trapezoid Condition, Lemma 5.13 states,

$$\tilde{T}_{\overrightarrow{ji}}(\eta_1) - \tilde{T}_{\overrightarrow{ji}}(\nu_1) = \tilde{T}_{\overrightarrow{ji}}(\eta_2) - \tilde{T}_{\overrightarrow{ji}}(\nu_2)$$

By the definition of $\mathcal{T}$ we have,

$$\mathcal{T}(x_1) - \mathcal{T}(y_1) = \mathcal{T}(x_2) - \mathcal{T}(y_2)$$

By Lemma 5.20, $x_m = \Psi(\eta_m)$ and $y_m = \Psi(\upsilon_m)$ for $m = 1, 2$, form a rectangle:

$$\begin{aligned} x_1(1) &= x_2(1) \equiv a_1 \text{ and } y_1(1) = y_2(1) \equiv a_2; \\ x_1(j) &= y_1(j) \text{ and } x_2(j) = y_2(j); \text{ for } 2 \leq j \leq J-1. \end{aligned}$$

Define $b_m \in Z^{J-2}$ for $m = 1, 2$ by,

$$b_m(j) = x_m(j+1) \text{ for } 2 \leq j \leq J-1,$$

substitution yields the rectangle condition,

$$\mathcal{T}(a_1, b_1) - \mathcal{T}(a_2, b_1) = \mathcal{T}(x_1) - \mathcal{T}(y_1) = \mathcal{T}(x_2) - \mathcal{T}(y_2) = \mathcal{T}(a_1, b_2) - \mathcal{T}(a_2, b_2).$$

It follows that $\mathcal{T}$ is additively separable between $a \in X_A = (0,1)$ and $b \in X_B = Z^{J-2}$

$$\mathcal{T}(a, b) = \boldsymbol{A}(a) + \boldsymbol{B}(b)$$

In the final step, we use Ψ to move from $\mathcal{T}$ to $\tilde{T}_{\overrightarrow{ji}}$. Given $x = \Psi(\gamma)$,

$$\begin{aligned} \tilde{T}_{\overrightarrow{ji}}(\gamma) &= \mathcal{T}(\Psi^{-1}(\gamma)) = \mathcal{T}(x) = \boldsymbol{A}(x(1)) + \boldsymbol{B}(x(2), .., x(J-1)) \\ &= \boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1) + \gamma(J)}\right) + \boldsymbol{B}\left(\gamma(2), .., \gamma(J-1)\right), \end{aligned}$$

completing the proof. ■

A5.2.11: Strong Form Additive Separability

In this section we establish a stronger form of additive separability relying on already established symmetry and differentiability properties of the $\tilde{T}$ function.

Lemma 5.22: Suppose $C \in \mathcal{C}$ has a UPS representation and satisfies IUC (A1). If $\gamma \in \tilde{\Gamma}'$, then, given $\gamma \in \tilde{\Gamma}'$ and $1 < i \neq j < J$, there exists $\boldsymbol{B} : \mathbb{R}^{J-2} \longrightarrow \mathbb{R}$ such that

$$\tilde{T}_{(ji)}(\gamma) = \boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right) \tag{108}$$

Proof. We arbitrarily order states, fix states 1 and J, and consider distinct states $2 \leq i \neq j \leq J-1$. By Lemma 5.9, $\gamma \in \tilde{\Gamma}'$ implies $\tilde{T}_{(ji)}(\gamma)$ exists. We set $i = 2$ and $j = 3$. Given Lemma 5.7, this is without loss of generality.

Applying Lemma 5.21

$$\tilde{T}_{(32)}(\gamma) = \boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1) + \gamma(J)}\right) + \boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right).$$

By Lemma 5.9 we also know that,

$$\tilde{T}_{(23)}(\gamma) = -\tilde{T}_{(32)}(\gamma) = -\boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right) - \boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right). \tag{109}$$

Define the mapping $\sigma : \{1, .., J\} \longrightarrow \{1, .., J\}$ that permutes elements 2 and 3:

$$\sigma(k) = \begin{cases} 3 \text{ if } k = 2; \\ 2 \text{ if } k = 3; \\ k \text{ otherwise.} \end{cases}$$

Defining $\gamma^\sigma \in \tilde{\Gamma}$ as the correspondingly permuted posterior, $\gamma^\sigma(j) = \gamma(\sigma^{-1}(j))$. Lemma 5.10 then that, since $\tilde{T}_{(ji)}(\gamma)$ exists,

$$\tilde{T}_{(23)}(\gamma) = \tilde{T}_{(32)}(\gamma^\sigma)$$

Directly by Lemma 5.21,

$$\begin{aligned} \tilde{T}_{(32)}(\gamma^\sigma) &= \boldsymbol{A}\left(\frac{\gamma^\sigma(1)}{\gamma^\sigma(1)+\gamma^\sigma(J)}\right) + \boldsymbol{B}\left(\gamma^\sigma(2), \gamma^\sigma(3), ..., \gamma^\sigma(J-2), \gamma^\sigma(J-1)\right) \\ &= \boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right) + \boldsymbol{B}\left(\gamma(3), \gamma(2), ..., \gamma(J-2), \gamma(J-1)\right). \end{aligned} \tag{110}$$

Since both equal $\tilde{T}_{(23)}(\gamma)$ we know that the right-hand sides of (109) and (110) are equal,

$$2\boldsymbol{A}\left(\frac{\gamma(1)}{\gamma(1)+\gamma(J)}\right) = -\boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right) - \boldsymbol{B}\left(\gamma(3), \gamma(2), ..., \gamma(J-2), \gamma(J-1)\right) \tag{111}$$

By assumption $\tilde{T}_{(32)}(\gamma)$ exists so that by Lemma 5.18 it also exists for all η such that $\eta(2)/\eta(3) = \gamma(2)/\gamma(3)$, including all at which,

$$\rho \equiv \frac{\eta(1)}{\eta(1)+\eta(J)} > 0$$

takes arbitrary values while $\eta(k) = \gamma(k)$ for all $k \neq 1, J$, which by construction differ from i, j. Hence (111) must hold for all $\rho > 0$. Since the right-hand side of the equation is independent of ρ, $\boldsymbol{A}(\rho)$ is independent of ρ,

$$\boldsymbol{A}(\rho) = \bar{\boldsymbol{A}} \in \mathbb{R}.$$

Hence we can add $\bar{\boldsymbol{A}}$ to $\boldsymbol{B}$ and normalize to $\boldsymbol{A}(x) = 0$, completing the proof. ■

In the proceeding, the has been no guarantee that there is a single $\boldsymbol{B}$ that works for all pairs of states. In the next lemma we further restrict the functional dependence of the two-sided directional derivative, and in the process show that there exists a single function $\bar{\boldsymbol{B}}$ that characterizes this derivative.

Lemma 5.23: Suppose $C \in \mathcal{C}$ has a UPS representation and satisfies IUC (A1), then there exists $\bar{\boldsymbol{B}} : (0,1) \times (0,1) \to \mathbb{R}$ such that, given $\gamma \in \tilde{\Gamma}'$, and states $1 \leq i \neq j \leq J$,

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), \gamma(j)). \tag{112}$$

Proof. Given arbitrarily fixed states 1 and J with $J \geq 4$, Lemma 5.22 establishes that if we consider distinct states $i = 2$ and $j = 3$, there exists $\boldsymbol{B} : \mathbb{R}^{J-2} \longrightarrow \mathbb{R}$ such that, given $\bar{\gamma} \in \tilde{\Gamma}'$,

$$\tilde{T}_{(32)}(\gamma) = \boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right)$$

If $J = 4$, then $\boldsymbol{B}$ has only two arguments and is of the desired form,

$$\tilde{T}_{(ji)}(\bar{\gamma}) = \boldsymbol{B}\left(\bar{\gamma}(i), \bar{\gamma}(j)\right) \equiv \boldsymbol{B}\left(\bar{\gamma}(2), \bar{\gamma}(3)\right) \equiv \bar{\boldsymbol{B}}\left(\bar{\gamma}(2), \bar{\gamma}(3)\right).$$

By the symmetry Lemma 5.10, this same function applies regardless of how we label states, completing the proof for $J = 4$.

$$\tilde{T}_{(ji)}(\bar{\gamma}) = \bar{\boldsymbol{B}}(\bar{\gamma}(i), \bar{\gamma}(j))$$

If $J > 4$ we again arbitrarily fixed states 1 and J, and consider state $s \neq i, j$ with $2 \leq s \leq J-1$. Hence by Lemma 5.22 and Lemma 5.10, we can transpose posteriors 1 and s without changing the form of the function, so that,

$$\tilde{T}_{(ji)}(\gamma) = \boldsymbol{B}(\gamma(2), .., \gamma(s-1), \gamma(1), \gamma(s+1), ..., \gamma(J-2), \gamma(J-1)). \tag{113}$$

Raising $\gamma(s)$ and reducing $\gamma(J)$ has no effect on the right hand side of (113), hence no effect on the RHS of (108) so that $\boldsymbol{B}\left(\gamma(2), \gamma(3), ..., \gamma(J-2), \gamma(J-1)\right)$ is independent of $\gamma(s)$. Proceeding in this matter for all $s \neq \{i, j\}$, we have

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), \gamma(j)),$$

where $\bar{\boldsymbol{B}}(\gamma(i), \gamma(j))$ is the common value. To complete the proof, note again that by the symmetry lemma, the same function applies regardless of how we label the states, completing the proof. ■

Note that the function $\bar{\boldsymbol{B}}(\gamma(i), \gamma(j))$ is pinned down only for $\gamma \in \tilde{\Gamma}'$ and not the full domain $(0,1) \times (0,1)$. However we know that it is pinned down on a dense subset of this space, so that it is natural to think of using a limit operation to fill out the function. The next Lemma establishes that this can be done in an unambiguous manner, and characterizes the one-sided directional derivative.

Lemma 5.24: There exists $\bar{\boldsymbol{B}} : (0,1) \times (0,1) \rightarrow \mathbb{R}$ such that, given $\gamma \in \tilde{\Gamma}$,

$$\tilde{T}_{\overrightarrow{ji}}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), \gamma(j)). \tag{114}$$

Proof. Where $\tilde{T}_{(ji)}(\gamma)$, exists, Lemma 5.9 shows that it is equal to $\tilde{T}_{\overrightarrow{ji}}(\gamma)$. Hence the function defined in (112) is of the appropriate form for $\gamma \in \tilde{\Gamma}'$. What is left is to establish that we can define $\bar{\boldsymbol{B}}(\gamma(i), \gamma(j))$ that equals $\tilde{T}_{\overrightarrow{ji}}(\gamma)$ on $\gamma \in \tilde{\Gamma} \backslash \tilde{\Gamma}'$.

Consider $\gamma \in \tilde{\Gamma} \backslash \tilde{\Gamma}'$, and consider any sequence $\{\gamma_n\}_{n=1}^{\infty}$ with $\gamma_n = \gamma + \epsilon_n(e_i - e_j)$ such that $\tilde{T}_{(ji)}(\gamma_n)$ exists for all n and $\epsilon_n \downarrow 0$. To see that such a sequence must exist, let $Y(\gamma, i, j) = \{x \in \mathbb{R}^J | x(k) = \gamma(k) \text{ for all } k \neq i, j\}$. $Y(\gamma, i, j)$ is a convex set, and $\tilde{T}^{Y(\gamma,i,j)} : \tilde{\Gamma}(Y(\gamma, i, j)) \rightarrow \mathbb{R}$ is the restriction of $\tilde{T}$ to $Y(\gamma, i, j) \cap \tilde{\Gamma}$. Lemma 5.16 states that $\tilde{T}^{Y(\gamma,i,j)}$ is almost everywhere differentiable in the relative interior of $\tilde{\Gamma}(Y(\gamma, i, j))$, and that $\tilde{T}^{Y(\gamma,i,j)}_{(ij)} = \tilde{T}_{(ji)}$. We can therefore select the sequence $\{\gamma_n\}_{n=1}^{\infty}$ from $\tilde{\Gamma}(Y(\gamma, i, j))$.

As $\tilde{T}_{(ji)}(\gamma_n)$ exists, Lemma 5.23 implies,

$$\tilde{T}_{(ji)}(\gamma_n) = \bar{\boldsymbol{B}}(\gamma_n(i), \gamma_n(j)).$$

Lemma 5.15 then ensures that,

$$\lim_{n\to\infty} \tilde{T}_{\overrightarrow{ji}}(\gamma + \epsilon_n(e_i - e_j)) = \tilde{T}_{\overrightarrow{ji}}(\gamma).$$

We therefore define $\bar{\boldsymbol{B}}(\gamma)$ on $\gamma \in \tilde{\Gamma}\backslash\tilde{\Gamma}'$ as,

$$\bar{\boldsymbol{B}}(\gamma) \equiv \lim_{n\to\infty} \bar{\boldsymbol{B}}(\gamma_n(i), \gamma_n(j)) = \tilde{T}_{\overrightarrow{ji}}(\gamma), \tag{115}$$

By construction we know that $\tilde{T}_{\overrightarrow{ji}}(\gamma) = \bar{\boldsymbol{B}}(\gamma)$ on the full domain, and that it is of the form $\bar{\boldsymbol{B}}(\gamma(i), \gamma(j))$ on $\gamma \in \tilde{\Gamma}\backslash\tilde{\Gamma}'$. Equation (115) implies that this extends to the limit points, completing the proof of (114) and with it the Lemma. ■

Note that the function $\bar{\boldsymbol{B}} : (0,1) \times (0,1) \to \mathbb{R}$ as introduced above allows for certain jumps at posteriors at the two sided directional derivatives fail to exist. In further characterizing the implications of Compression, such cases will be ruled out.

A5.2.12: Full Additive Separability

We have now established that directional derivatives at any posterior depends only on the probabilities of the two involved states. We now establish that the corresponding function can be defined based on a fixed function of each probability alone. This is what we refer to as full additive separability. The result is connected with a triangular pattern in two-sided directional derivatives. Given $\gamma \in \tilde{\Gamma}$ we know that $\tilde{T}_{(ji)}(\gamma), \tilde{T}_{(ik)}(\gamma), \tilde{T}_{(jk)}(\gamma)$ all exist, and furthermore that they are interdependent,

$$\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{(jk)}(\gamma) + \tilde{T}_{(ki)}(\gamma). \tag{116}$$

In Lemma 5.25 we show that this relationship rests only on existence of any two of these three two-sided directional derivatives. The lemma also uses the negative inverse feature of these directional derivatives to point to the method for identifying the appropriate form of the function that generates the sought after representation.

Lemma 5.25 Given $\gamma \in \tilde{\Gamma}$, suppose that there exist three distinct indices $1 \leq i, j, k \leq J$ such that $\tilde{T}_{(ki)}(\gamma)$ and $\tilde{T}_{(kj)}(\gamma)$ both exist. Then $\tilde{T}_{(ji)}(\gamma)$ exists and,

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), \gamma(k)) - \bar{\boldsymbol{B}}(\gamma(j), \gamma(k)). \tag{117}$$

Proof. Given Lemma 5.7, we may take $i = 1$, $j = 2$ and $k = 3$ without loss of generality.

Given $\gamma \in \tilde{\Gamma}$, define the set X by

$$X \equiv \left\{ x \in \mathbb{R}^2 | x_1, x_2 > 0 \text{ and } x_1 + x_2 < 1 - \sum_{l \neq i,j,k} \gamma(l) \right\}. \tag{118}$$

Define $\eta(x) \in \tilde{\Gamma}$

$$[\eta(x)](l) = \begin{cases} x_1 \text{ if } l = 1; \\ x_2 \text{ if } l = 2; \\ 1 - \sum_{l \neq i,j,k} \gamma(l) - x_1 - x_2 \text{ if } l = 3; \\ \gamma(l) \text{ otherwise} \end{cases} \tag{119}$$

Finally, define $H : X \to \mathbb{R}$ by

$$H(x) = \tilde{T}(\eta(x)) \tag{120}$$

Note that

$$\eta(\gamma(1), \gamma(2)) = \gamma$$

Note also that, given $\tilde{T}_{(32)}(\gamma)$ exists,

$$\begin{aligned} \tilde{T}_{(32)}(\gamma) &= \lim_{\epsilon \to 0} \frac{\tilde{T}(\gamma + \epsilon(e_2 - e_3)) - \tilde{T}(\gamma)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\tilde{T}(\eta(\gamma(1), \gamma(2) + \epsilon)) - \tilde{T}(\eta(\gamma(1), \gamma(2)))}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{H(\gamma(1), \gamma(2) + \epsilon) - H(\gamma(1), \gamma(2))}{\epsilon} \end{aligned}$$

Hence,

$$\tilde{T}_{(32)}(\gamma) = H_2(\gamma(1), \gamma(2)). \tag{121}$$

Analogously,

$$\tilde{T}_{(31)}(\gamma) = H_1(\gamma(1), \gamma(2)).$$

Since both partials exist, note from Rockafellar theorem 25.1 that H is differentiable at γ and from theorem 25.2 that the directional derivative function $H'(\gamma|y)$ is linear in direction $y \in \mathbb{R}^2$. Hence the directional derivative in direction $e_1 - e_2$ is the difference between the partials,

$$H'(\gamma|e_1 - e_2) = H_1(\gamma(1), \gamma(2)) - H_2(\gamma(1), \gamma(2)) = \tilde{T}_{(31)}(\gamma) - \tilde{T}_{(32)}(\gamma). \tag{122}$$

To complete the proof of (117), note directly from the definitions that $H'(\gamma|e_2 - e_1)$ is equal to $\tilde{T}_{(ji)}(\gamma)$,

$$\begin{aligned} H'(\gamma|e_1 - e_2) &= \lim_{\epsilon \to 0} \frac{H(\gamma(1) + \epsilon, \gamma(2) - \epsilon) - H(\gamma(1), \gamma(2))}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\tilde{T}(\gamma + \epsilon(e_j - e_i)) - \tilde{T}(\gamma)}{\epsilon} \equiv \tilde{T}_{(ji)}(\gamma) \end{aligned} \tag{123}$$

Setting the right-hand sides of (122) and (123) to equality establishes that

$$\tilde{T}_{(ji)}(\gamma) = \tilde{T}_{(ki)}(\gamma) - \tilde{T}_{(kj)}(\gamma).$$

In light of Lemma 5.24 this completes the proof of (117). ■

Lemma 5.25 points the way to a possible method for expressing $\tilde{T}_{(ji)}(\gamma)$ in a fully additively separable manner. Given $k \neq i, j$ and $\bar{x} \in (0, 1 - \gamma(i) - \gamma(j))$, Lemma 5.25 implies

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), \bar{x}) - \bar{\boldsymbol{B}}(\gamma(j), \bar{x})$$

so that $\tilde{T}_{(ji)}(\gamma)$ is additively separable in $\gamma(i)$ and $\gamma(j)$. The complication in establishing this form of additive separability is that the requirement that $\bar{x} < 1 - \gamma(i) - \gamma(j)$ means that that no single $\bar{x}$ works for all $\gamma \in \tilde{\Gamma}$. In the following, we establish additive separability on a subset of $\tilde{\Gamma}$ and then drive the value of $\bar{x}$ down to zero to establish additive separability on the whole of $\tilde{\Gamma}$.

Lemma 5.26: Given $\epsilon \in (0, 0.5)$, there exists $x(\epsilon) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$ and a full measure set $I(\epsilon) \subset (0, 1-\epsilon)$ such that, for any distinct states $1 \leq i, k \leq J$, given $\gamma \in \tilde{\Gamma}$ with $\gamma(k) = x(\epsilon)$, $\tilde{T}_{(ik)}(\gamma)$ exists whenever $\gamma(i) \in I(\epsilon)$.

Proof. Pick distinct states $1 \leq i, k \leq J$ and $\epsilon \in (0, 0.5)$. Let $Y(k, \epsilon) = \{z \in \mathbb{R}^J | z(k) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)\}$ denote the set of vectors in $\mathbb{R}^J$ for which $z(k) \in (\frac{\epsilon}{4}, \frac{\epsilon}{2})$ and focus on posteriors with $\gamma(k)$ so restricted,

$$\tilde{\Gamma}(Y(k,\epsilon)) = \left\{\gamma \in \tilde{\Gamma} | \gamma(k) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)\right\}. \tag{124}$$

Per the general prescription, define the restricted function $\tilde{T}^{Y(k,\epsilon)} : \tilde{\Gamma}(Y(k,\epsilon)) \to \mathbb{R}$, and note for arbitrary indices $1 \leq j \neq l \leq J$,

$$\tilde{T}^{Y(k,\epsilon)}_{\overrightarrow{lj}}(\gamma) = \tilde{T}_{\overrightarrow{lj}}(\gamma),$$

given that there is suitable variation of the posterior in all directions.

By Lemma 5.16 we know that $\tilde{T}^{Y(k,\epsilon)}$ is almost every differentiable in the relative interior of $\tilde{\Gamma}(Y(k,\epsilon))$. At any point of differentiability of $\tilde{T}^{Y(k,\epsilon)}$, we know by Lemma 5.9 that all two-sided directional derivatives exists. Hence, we know that $\tilde{T}_{(ik)}(\gamma)$ exists for almost all $\gamma \in \tilde{\Gamma}(Y(k,\epsilon))$. But we already know from Lemma 5.24 that such existence can only depend on the values $\gamma(i)$ and $\gamma(k)$, so that existence is ensured on a full measure subset of the corresponding domain defined by:

$$\gamma(k) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right) \text{ and } \gamma(i) \in (0, 1-\gamma(k)).$$

Now fix $x \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$ and define

$$I(x) = \{y \in (0, 1-x) | \tilde{T}_{(ik)}(\gamma) \text{ exists when } \gamma(k) = x \text{ and } \gamma(i) = y\} \subset (0, 1-x). \tag{125}$$

Note that the union of these sets across $x \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$ is precisely the set of $\gamma \in \tilde{\Gamma}(Y(k,\epsilon))$ on which $\tilde{T}_{(ik)}(\gamma)$ exists, which we know to have the same measure as the relative interior of $\tilde{\Gamma}(Y(k,\epsilon))$. This means that there exists $x(\epsilon) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$ such that the measure of $I(x(\epsilon))$ is $1 - x(\epsilon)$. As $x(\epsilon) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$, take $I(\epsilon) = I(x(\epsilon)) \cap (1, 1-\epsilon)$. This completes the proof. ■

We now show how to define an appropriate fully additively separable function of the form we seek for any given $\epsilon \in (0, 0.5)$.

Lemma 5.27: Given $\epsilon \in (0, 0.5)$ and there exists a dense subset $I(\epsilon) \subset (0, 1-\epsilon)$ and a function $f^\epsilon : I(\epsilon) \to \mathbb{R}$ such that,

$$\tilde{T}_{(ji)}(\gamma) = f^\epsilon(\gamma(i)) - f^\epsilon(\gamma(j)), \tag{126}$$

for all $\gamma \in \tilde{\Gamma}$ such that $\gamma(i), \gamma(j) \in I(\epsilon)$ and $\gamma(i) + \gamma(j) < 1 - \epsilon$.

Proof. Given $\epsilon \in (0, 0.5)$, fix $x(\epsilon) \in \left(\frac{\epsilon}{8}, \frac{\epsilon}{4}\right)$ and the dense subset $I(\epsilon)$ of $(0, 1-\epsilon)$ so that the conditions of the last Lemma are satisfied. Now consider $\tilde{\Gamma}(\epsilon)$, the set of posteriors for which

Lemma 5.26 tells us that both $\tilde{T}_{(ik)}(\gamma)$ and $\tilde{T}_{(jk)}(\gamma)$ are well-defined,

$$\tilde{\Gamma}(\epsilon) = \left\{\gamma \in \tilde{\Gamma} | \gamma(i), \gamma(j) \in I(\epsilon), \gamma(k) = x(\epsilon)\right\},$$

Since both $\tilde{T}_{(ik)}(\gamma)$ and $\tilde{T}_{(jk)}(\gamma)$ exist on $\tilde{\Gamma}(\epsilon)$, by Lemma 5.25,

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), x(k)) - \bar{\boldsymbol{B}}(\gamma(j), x(k)). \tag{127}$$

for $\gamma \in \tilde{\Gamma}(\epsilon)$. We define the candidate function,

$$f^{\epsilon}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), x(\epsilon)).$$

The Lemma requires one more step, which is to remove the condition that $\gamma(k) = x(\epsilon)$, which is absent in the conditions of the Lemma. The key observation here is that according to Lemma 5.23, $\tilde{T}_{(ji)}(\gamma)$ depends only on $\gamma(i)$ and $\gamma(j)$. Hence given γ' such that $\gamma'(i) = \gamma(i)$ and $\gamma'(j) = \gamma\,(i)$,

$$\tilde{T}_{(ji)}(\gamma') = \tilde{T}_{(ji)}(\gamma) = f^{\varepsilon}(\gamma(i)) - f^{\varepsilon}(\gamma(j)).$$

Hence the characterization applies to all $\gamma \in \tilde{\Gamma}$ such that $\gamma(i), \gamma(j) \in I(\epsilon)$ and $\gamma(i) + \gamma(j) < 1 - \epsilon$ as required. ■

Lemma 5.28: There exists $f : \bar{I} \to \mathbb{R}$ with $\bar{I} \subset (0, 1)$ of full measure such that for all $\gamma \in \tilde{\Gamma}$ with $\gamma(i), \gamma(j) \in \bar{I}$, $\tilde{T}_{(ji)}(\gamma)$ exists and,

$$\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j)) \tag{128}$$

Proof. We construct a diminishing sequence $\{\epsilon(n)\}_{n=1}^{\infty} > 0$ with $\epsilon(n+1) < \epsilon(n)$ by setting $\epsilon(1) \in (0.0.5)$ and thereupon successively halving,

$$\epsilon(n+1) = \frac{\epsilon(n)}{2},$$

on $n > 1$. For each n, Lemma 5.26 states that there exists $x(n) \in \left(\frac{\epsilon(n)}{8}, \frac{\epsilon(n)}{4}\right)$ and a set $I(n) \subset (0, 1-\epsilon(n))$ which is dense in $(0, 1-\epsilon(n))$ such that $\tilde{T}_{(ji)}$ exists whenever $\gamma(j) = x(n)$ and $\gamma(i) \in I(n)$.

We now show that $I(n) \subset I(n+1)$. Since $I(n)$ is dense in $(0, 1-\epsilon(n))$ and $I(n+1)$ is dense in $(0, 1-\epsilon(n+1))$ and $(0, 1-\epsilon(n)) \subset (0, 1-\epsilon(n+1))$, $I(n) \cap I(n+1)$ is not empty. Consider $y \in I(n) \cap I(n+1)$ and choose η such that $\gamma(i) = y$, $\gamma(j) = x(n)$. That this is possible follows from the fact that

$$y + x(n) + x(n+1) < y + \frac{\epsilon(n)}{4} + \frac{\epsilon(n+1)}{4} \leq y + \frac{\epsilon(n)}{4} + \frac{\epsilon(n)}{8} < 1$$

Since $\gamma(i) \in I(n)$, $\tilde{T}_{(ji)}$ exists, and since $\gamma(i) \in I(n+1)$, $\tilde{T}_{(ki)}$ exists. It follows from Lemma 5.25, that $\tilde{T}_{(kj)}$ exists. Now consider any $y \in I(n)$, and consider η such that $\gamma(i) = y$, $\gamma(j) = x(n)$, and $\gamma(k) = x(n+1)$. Since $\gamma(i) \in I(n)$, $\tilde{T}_{(ji)}$ exists, and since $\tilde{T}_{(kj)}$ exists, it follows from Lemma 5.25, that $\tilde{T}_{(ki)}$ exists. Hence $y \in I(n+1)$.

By Lemma 5.27,

$$\tilde{T}_{(ji)}(\gamma) = \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(\gamma(j), x(n))$$

for all γ such that $\gamma(i), \gamma(j) \in I(n)$. Fix $z \in I(1)$, since $I(n) \subset I(n+1)$, $z \in I(n)$. For $\gamma(i) \in I(n)$, define

$$\boldsymbol{G}^n(\gamma(i)) = \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(z, x(n))$$

It follows that

$$\begin{aligned}
\tilde{T}_{(ji)}(\gamma) &= \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(\gamma(j), x(n)) \\
&= \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(z, x(n)) - \bar{\boldsymbol{B}}(\gamma(j), x(n)) + \bar{\boldsymbol{B}}(z, x(n)) \\
&= \boldsymbol{G}^n(\gamma(i)) - \boldsymbol{G}^n(\gamma(j))
\end{aligned}$$

We now compare $\boldsymbol{G}^n(\gamma(i))$ to $\boldsymbol{G}^{n+1}(\gamma(i))$. Consider $\gamma(i) \in I(n) \subset I(n+1)$,

$$\begin{aligned}
\boldsymbol{G}^{n+1}(\gamma(i)) &= \bar{\boldsymbol{B}}(\gamma(i), x(n+1)) - \bar{\boldsymbol{B}}(z, x(n+1)) \\
&= \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(x(n+1), x(n)) - \bar{\boldsymbol{B}}(z, x(n)) + \bar{\boldsymbol{B}}(x(n+1), x(n)) \\
&= \bar{\boldsymbol{B}}(\gamma(i), x(n)) - \bar{\boldsymbol{B}}(z, x(n)) \\
&= \boldsymbol{G}^n(\gamma(i))
\end{aligned}$$

where the second equality follows from Lemma 5.25 applied to $\tilde{\gamma}$ with $\tilde{\gamma}(i) = x(n+1)$, $\tilde{\gamma}(j) = x(n)$ and $\tilde{\gamma}(k) = \gamma(i)$:

$$\tilde{T}_{(ji)}(\tilde{\gamma}) = \bar{\boldsymbol{B}}(x(n+1), x(n)) = \bar{\boldsymbol{B}}(x(n+1), \gamma(i)) - \bar{\boldsymbol{B}}(x(n), \gamma(i)) = -\bar{\boldsymbol{B}}(\gamma(i), x(n+1)) + \bar{\boldsymbol{B}}(\gamma(i), x(n))$$

and similarly for z in place of $\gamma(i)$.

Hence we can define a limit function $f : \cup_{n=1}^{\infty} I(n) \to \mathbb{R}$ unambiguously by taking any $x \in \cup_{n=1}^{\infty} I(n)$, selecting a particular $\bar{n}$ such that $x \in I(\bar{n})$, and defining,

$$f(x) = \boldsymbol{G}^{\bar{n}}(x).$$

By Lemma 5.27, we know that with $\gamma(i), \gamma(j) \in I(n)$, a dense subset of $(0, 1 - \epsilon(n))$, (128) holds,

$$\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j). \tag{129}$$

Since $\lim_{n\to\infty} \epsilon(n) = \lim_{n\to\infty} \epsilon(n) = 0$, note that

$$\bar{I} \equiv \cup_{n=1}^{\infty} I(n)$$

is a dense subset of $(0, 1)$, establishing the Lemma. ■

Lemma 5.29: The function $f : \bar{I} \longrightarrow \mathbb{R}$ defined in Lemma 5.28 for which (128) holds is non-decreasing, and can be extended to a function $f : (0, 1) \longrightarrow \mathbb{R}$ that is non-decreasing.

Proof. We pick arbitrary $x, x + \epsilon \in \cup_{n=1}^{\infty} I(n) = \bar{I}$ with $\epsilon > 0$ and show that $f(x + \epsilon) \geq f(x)$. Consider $\gamma \in \tilde{\Gamma}$ with $\gamma(i) = x$ and $\gamma(j) = x + \epsilon$, by Lemma 5.28

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

If we now define $\gamma' \in \tilde{\Gamma}$ as,

$$\gamma' = \gamma + \epsilon(e_i - e_j),$$

Lemma 5.28 implies that,

$$\tilde{T}_{(ji)}(\gamma') = f(\gamma'(i)) - f(\gamma'(j)) = f(x+\epsilon) - f(x).$$

By the monotonicity lemma 5.14, $\gamma' = \gamma + \epsilon(e_i - e_j)$ for $\epsilon > 0$ implies $\tilde{T}_{(ji)}(\gamma') \geq \tilde{T}_{(ji)}(\gamma)$, which translates to,

$$f(x+\epsilon) - f(x) \geq f(x) - f(x+\epsilon),$$

which directly implies $f(x+\epsilon) \geq f(x)$, completing the proof that f is non-decreasing on $\bar{I}$.

To complete the proof, pick $x \in \bar{I} \backslash (0,1)$. Since $\bar{I} \subset (0,1)$ is of full measure in $(0,1)$, we can find sequence $\{x(n)\}_{n=1}^{\infty} > x$ with $x(n+1) < x(n)$ and $\lim_{n\to\infty} x(n) = x$ such that $x(n) \in \bar{I}$. We define $f(x)$ as the corresponding limit,

$$f(x) = \lim_{n\to\infty} f(x(n)$$

Since we have just shown f to be non-decreasing, the limit is well-defined, and also non-decreasing.
■

We now show a connection between continuity properties of f and existence of two-sided directional derivatives.

Lemma 5.30: Given $\eta \in \tilde{\Gamma}$, $\tilde{T}_{(ji)}(\eta)$ exists if and only if $f(\gamma(i)) - f(\gamma(j))$ is continuous at η.

Proof. Consider $\eta \in \tilde{\Gamma}$. Pick distinct states $1 \leq i, j \leq J$ and define

$$Y(\eta, i, j) = \{\gamma \in \mathbb{R}^J | \gamma(k) = \eta(k), k \neq i, j\}.$$

Note $Y(\eta, i, j)$ is convex set. Per the general prescription, define the restricted function $\tilde{T}^{\;Y(\eta,i,j)}$. By Lemma 5.16 this function is differentiable almost everywhere in the relative interior of the restricted domain $\tilde{\Gamma}(Y(\eta, i, j))$. Hence the directional derivative in the only relevant direction,

$$\tilde{T}^{Y(k,\epsilon)}_{(ji)}(\gamma) = \tilde{T}_{(ji)}(\gamma),$$

exists almost everywhere. Hence we can find sequences approaching from η both corresponding directions. We now select $\{\epsilon(n)\}_{n=1}^{\infty} > 0$ with $\lim_{n\to\infty} \epsilon(n) = 0$ such that given $\gamma_n = \eta + \epsilon(n)(e_i - e_j)$, $\tilde{T}_{(ji)}(\gamma_n)$ exists. We select also $\{\epsilon'(n)\}_{n=1}^{\infty} < 0$ with $\lim_{n\to\infty} \epsilon'(n) = 0$ such that, defining $\gamma'_n = \eta + \epsilon'(n)(e_i - e_j)$, $\tilde{T}_{(ji)}(\gamma'_n)$ exists.

Since $\tilde{T}$ is convex we know that the one-sided directional derivatives are monotonically increasing,

$$\lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma'_n) \leq -\tilde{T}_{\overrightarrow{ji}}(\eta) \leq \tilde{T}_{\overrightarrow{ji}}(\eta) \leq \lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma_n). \tag{130}$$

We now use Lemma 5.28 to substitute in (130) at all γ_n and γ'_n since $\tilde{T}_{(ji)}(\circ)$ is well-defined at these points, to arrive at,

$$\begin{aligned} \tilde{T}_{(ji)}(\gamma_n) &= f(\eta(i)+\epsilon_n) - f(\eta(j)-\epsilon_n); \\ \tilde{T}_{(ji)}(\gamma'_n) &= f(\eta(i)+\epsilon'_n) - f(\eta(j)-\epsilon'_n); \end{aligned}$$

Now suppose that $f(\gamma(i)) - f(\gamma(j))$ is continuous at η. In this case,

$$\lim_{n\to\infty} \left[f(\eta(i)+\epsilon'_n) - f(\eta(j)-\epsilon'_n) \right] = \lim_{n\to\infty} \left[f(\eta(i)+\epsilon_n) - f(\eta(j)-\epsilon_n) \right],$$

so that correspondingly,

$$\lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma'_n) = \lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma_n),$$

hence by (130),

$$-\tilde{T}_{\overrightarrow{ji}}(\eta) = \tilde{T}_{\overrightarrow{ji}}(\eta),$$

establishing through Lemma 5.9 that $\tilde{T}_{(ji)}(\eta)$ exists.

Suppose conversely that $\tilde{T}_{(ji)}(\eta)$ does not exist. In this case we know by Lemma 5.9 that $\tilde{T}_{\overrightarrow{ji}}(\eta) < \tilde{T}_{\overrightarrow{ji}}(\eta)$, so that by (130),

$$\lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma_n) = \lim_{n\to\infty} f(\eta(i)+\epsilon'_n) - f(\eta(j)-\epsilon'_n) < \lim_{n\to\infty} \tilde{T}_{(ji)}(\gamma_n) = \lim_{n\to\infty} f(\eta(i)+\epsilon_n) - f(\eta(j)-\epsilon_n),$$

establishing that $f(\gamma(i)) - f(\gamma(j))$ is discontinuous at η, and completing the proof. ■

A5.2.13: Existence of Directional Derivatives

Lemma 5.31: Given η and given $\alpha = \frac{\eta(k)}{\eta(l)}$, if $\tilde{T}_{(kl)}(\eta)$ exists then $\tilde{T}$ is differentiable for almost all $\gamma \in \Gamma_{kl}(\alpha)$.

Proof. Consider $\eta \in \tilde{\Gamma}$ such that $\tilde{T}_{(kl)}(\eta)$ exists and set $\alpha = \frac{\eta(k)}{\eta(l)}$. Since $\Gamma_{kl}(\alpha)$ is convex, $\tilde{T}^{\Gamma_{kl}(\alpha)}$ almost everywhere differentiable on the relative interior of $\Gamma_{kl}(\alpha)$ by Lemma 5.16. At points of differentiability, we know from Lemma 5.9 that $\tilde{T}_{(ji)}(\gamma) = \tilde{T}^{\Gamma_{kl}(\alpha)}_{(ji)}(\gamma)$ exists provided $\Gamma_{kl}(\alpha)$ contains a line segment through γ in direction $(e_i - e_j)$, By definition of $\Gamma_{kl}(\alpha)$, this holds for all directions except that defined by the pair of states (lk) whose posterior belief ratio is held fixed through the set.

Consider $\gamma \in \Gamma_{kl}(\alpha)$ at which $\tilde{T}^{\Gamma_{kl}(\alpha)}$ is differentiable. As $\tilde{T}_{(kl)}(\eta)$ exists, Lemma 5.18 implies that $\tilde{T}_{(kl)}(\gamma)$ also exists. Hence at all such γ, we know that all 2-sided directional derivatives exist. Following precisely the steps in Lemma 5.25, we can remove an arbitrary state $k \neq i, j$ from the domain and construct set X as in (118), then define $\eta(x) \in \tilde{\Gamma}$ on $x \in X$ as in (119) and function $H(x)$ on X by (120), whose partial derivatives are precisely the directional derivatives $\tilde{T}_{(km)}(\gamma)$,

$$\tilde{T}_{(km)}(\gamma) = H_1(\gamma(m), \gamma(i)),$$

all $m \neq k$.

Since all partials of this function therefore exist, we note from Rockafellar theorem 25.2 that $H(\gamma)$ is differentiable at γ and that the directional derivative function $H'(\gamma|y)$ is linear in direction $y \in \mathbb{R}^2$. Re-application of Rockafellar theorem 25.2 implies that $\tilde{T}$ is differentiable at γ, completing the proof. ■

Lemma 5.32: $\tilde{T}_{(ji)}(\gamma)$ exists for all i, j and $\gamma \in \tilde{\Gamma}$.

Proof. The proof is by contradiction. Consider a posterior η at which $\tilde{T}_{(ji)}(\eta)$ does not exist. It follows from Lemma 5.30 that $f(\eta(i)) - f(\eta(j))$ is discontinuous at this point:

$$\lim_{\epsilon \uparrow 0} f(\eta(i)+\epsilon) - f(\eta(j)+\epsilon) \leq \lim_{\epsilon \downarrow 0} f(\eta(i)+\epsilon) - f(\eta(j)+\epsilon)$$

Without loss of generality, suppose that it is $f(\eta(i))$ that is discontinuous.

Since f is monotonic, $f(\gamma(j))$ is continuous for almost all $\gamma(j) \in (0, 1-\eta(i))$ (Rudin (1976) Theorem 4.30). The discontinuity of f at $\eta(i)$ and the continuity of f almost everywhere else implies that $f(\eta(i)) - f(\gamma(j))$ is discontinuous in the direction (ji) for almost all $\gamma(j) \in (0, 1-\eta(i))$. Hence by Lemma 5.30, $\tilde{T}_{(ji)}(\gamma)$ does not exist for almost all γ such that $\gamma(i) = \eta(i)$ and $\gamma(j) \in (0, 1-\eta(i))$. It follows that for almost all $\alpha \in \left(\frac{\eta(i)}{1-\eta(i)}, \infty\right)$, there exists $\gamma \in \Gamma^{\alpha}_{ji}$ such that $\tilde{T}_{(ji)}(\gamma)$ does not exist. But by Lemma 5.18, if $\tilde{T}_{(ji)}(\gamma)$ exists for any $\eta \in \Gamma^{\alpha}_{ji}$, then $\tilde{T}_{(ji)}(\gamma)$ exists for all $\gamma \in \Gamma^{\alpha}_{ji}$. Hence for almost all $\alpha \in \left(\frac{\eta(i)}{1-\eta(i)}, \infty\right)$, $\tilde{T}_{(ji)}(\gamma)$ does not exist for any $\gamma \in \Gamma^{\alpha}_{ji}$. But $\left\{\gamma | \gamma \in \Gamma_{ji}(\alpha), \alpha \in \left(\frac{\eta(i)}{1-\eta(i)}, \infty\right)\right\}$ is a set of positive measure and $\tilde{T}$ is differentiable almost everywhere. This contradiction establishes the result. ■

Lemma 5.33: $\tilde{T}$ is continuously differentiable on $\gamma \in \tilde{\Gamma}$ and $f(\gamma(j))$ is continuous on $\tilde{\Gamma}$

Proof. Lemma 5.32 establishes that the directional derivatives $\tilde{T}_{(ji)}(\gamma)$ exist for all (ji) and all $\gamma \in \tilde{\Gamma}$. It follows from Rockafellar (1970) Theorem 25.2 that $\tilde{T}$ is differentiable and from Rockafellar (1970) Corollary 25.5.1 that $\tilde{T}$ is continuously differentiable on $\gamma \in \tilde{\Gamma}$. Since $\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j))$ continuity of f follows. ■

A5.2.14: Existence of Cross-Directional Derivatives

The twice differentiability of a convex function such as $\tilde{T}$ normally a subtle object as convexity alone is not sufficient to establish differentiability on any open set. In our case, however, we know from Lemma 5.33 that $\tilde{T}$ is continuously differentiable on $\gamma \in \tilde{\Gamma}$ and hence standard notions of twice differentiability apply. We now introduce the cross derivatives of $\tilde{T}$, which are of the essence in the proof of Theorem 1.

Definition 14 *Given $\gamma \in \tilde{\Gamma}$ and any two pair of states $1 \leq i \neq j \leq J$ and $1 \leq k \neq l \leq J$ we define the corresponding **cross derivative in direction** lk of $\tilde{T}_{(ji)}$, as the corresponding directional derivative of $\tilde{T}_{(ji)}$, should it exist:*

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

The next results show that the cross-directional derivatives exist almost everywhere.

Lemma 5.34: $\tilde{T}_{(ji)(il)}$ exists almost everywhere in $\tilde{\Gamma}$

Proof. Rockafellar (1999) Theorem 2.3 states that for almost all $\gamma \in \tilde{\Gamma}$ the gradient of $\tilde{T}(\gamma)$, $\nabla\tilde{T}(\gamma)$, exists and

$$\nabla\tilde{T}(\gamma) = \nabla\tilde{T}(\gamma + \omega) + A\omega + o(|\omega|)$$

holds with respect to $\{\omega | \gamma + \omega \in \text{dom}\nabla\tilde{T}\}$ for some $J-1$ by $J-1$ matrix A. By Lemma 5.33, $\tilde{T}$ is differentiable everywhere on $\tilde{\Gamma}$, and so $\text{dom}\nabla\tilde{T} = \tilde{\Gamma}$. It follows that $\tilde{T}_{(ji)(il)}$ exists almost everywhere in $\tilde{\Gamma}$, and where $\tilde{T}_{(ji)(il)}$ exists it is equal to:

$$\begin{aligned}\tilde{T}_{(ji)(il)} &= \lim_{\epsilon \to 0} \frac{\tilde{T}_{(ji)}(\eta + \epsilon(e_l - e_i)) - \tilde{T}_{(ji)}(\eta)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{A\epsilon(e_l - e_i) + o(|\epsilon(e_l - e_i)|)}{\epsilon} \\ &= A(e_l - e_i).\end{aligned}$$

■

Lemma 5.35: Given $\eta \in \tilde{\Gamma}$ at which $\tilde{T}_{(ij)(lk)}(\eta)$ exists and given $\alpha = \frac{\eta(k)}{\eta(l)}$, $\tilde{T}_{(ij)(lk)}(\nu)$ exists for all $\nu \in \Gamma_{kl}(\alpha)$.

Proof. Choose η at which $\tilde{T}_{(ij)(lk)}$ exists and set $\alpha = \frac{\eta(k)}{\eta(l)}$. Consider $\nu \in \Gamma_{kl}(\alpha)$. Lemma 5.12 establishes the existence of a parameterized set of problems (μ_t, A) indexed by $t \in [0,1]$ where:

$$\mu_t(j) = \begin{cases} t\,[\mu(k) + \mu(l)] & \text{for } j = k; \\ (1-t)\,[\mu(k) + \mu(l)] & \text{for } j = l; \\ \mu(j) & \text{otherwise;} \end{cases}$$

and η_t is the revealed posterior for action a and ν_t is the revealed posterior for action b where

$$\eta_t(j) = \frac{\eta(j)}{\mu(j)}\mu_t(j) \qquad \text{and} \qquad \nu_t(j) = \frac{\nu(j)}{\mu(j)}\mu_t(j).$$

By Lemma 5.31, $\tilde{T}$ is differentiable everywhere in $\tilde{\Gamma}$. It follows from Lemma 5.9 that $\tilde{T}_{(ij)}(\eta_t)$ and $\tilde{T}_{(ij)}(\nu_t)$ exist for all t. Lemma 5.11 implies therefore that,

$$\tilde{N}^a_{(ij)}(\eta_t) = \tilde{N}^b_{(ij)}(\nu_t)$$

for all t. Substituting the definition of net utility yields

$$\tilde{T}_{(ij)}(\nu_t) = \tilde{T}_{(ij)}(\eta_t) - u(a,k) + u(a,l) + u(b,k) - u(b,l) \tag{131}$$

Differencing (131) evaluated at t and $\bar{t}$ and taking $\mu_t = \mu + \epsilon_t(e_k - e_l)$ implies $\eta_t = \eta + \epsilon_t \frac{\eta(k)}{\mu(k)}(e_k - e_l)$ and $\nu_t = \nu + \epsilon_t \frac{\nu(k)}{\mu(k)}(e_k - e_l)$ yields

$$\lim_{\epsilon_t \to 0} \frac{\tilde{T}_{(ij)}\left(\nu + \epsilon_t \frac{\nu(k)}{\mu(k)}(e_k - e_l)\right) - \tilde{T}_{(lk)}(\eta)}{\epsilon_t \frac{\nu(k)}{\mu(k)}} = \lim_{\epsilon_t \to 0} \frac{\tilde{T}_{(ij)}\left(\eta + \epsilon_t \frac{\eta(k)}{\mu(k)}(e_k - e_l)\right) - \tilde{T}_{(lk)}(\eta)}{\epsilon_t \frac{\eta(k)}{\mu(k)}} \tag{132}$$

Since $\tilde{T}_{(ij)(lk)}(\eta)$ exists, the right-hand side limit exists, and so the left-hand side limit exists, establishing the result. ■

Lemma 5.36: Given $\gamma \in \tilde{\Gamma}$, $1 \leq i \neq j \leq J$ and $1 \leq k \neq l \leq J$, all cross-derivatives $\tilde{T}_{(ji)(lk)}(\gamma)$ exist.

Proof. Note first that for all $\gamma \in \tilde{\Gamma}$ for any distinct states i, j, k, and l, $\tilde{T}_{(ji)(lk)}(\gamma) = 0$, since Lemma 5.29 and 4.33 imply that $\tilde{T}_{(ji)}(\gamma) = f(\gamma(i)) - f(\gamma(j))$ which is independent of $\gamma(k)$ and $\gamma(l)$.

Now consider cases with overlap. Consider first the case in which $k = i$ and $l \neq j$. The proof is by contradiction. Consider a posterior $\eta \in \tilde{\Gamma}$. By Lemma 5.32, $\tilde{T}$ is differentiable at η, and hence by Lemma 5.29, $\tilde{T}_{(ji)}(\eta) = f(\eta(i)) - f(\eta(j))$. It follows that

$$\frac{\tilde{T}_{(ji)}(\eta + \epsilon(e_i - e_l)) - \tilde{T}_{(ji)}(\eta)}{\epsilon} = \frac{f(\eta(i) + \epsilon) - f(\eta(i))}{\epsilon}$$

so that $\tilde{T}_{(ji)}$ is differentiable in the direction (li) if and only if f is differentiable at $\eta(i)$. Suppose now that f is not differentiable at $\eta(i)$. Consider the set of posteriors ν such that $\nu(i) = \eta(i)$ and $v(j) \in (1, 1 - \eta(i))$. Since

$$\frac{\tilde{T}_{(ji)}(\nu + \epsilon(e_i - e_l)) - \tilde{T}_{(ji)}(\nu)}{\epsilon} = \frac{f(\eta(i) + \epsilon) - f(\eta(i))}{\epsilon}$$

$\tilde{T}_{(ji)(il)}(\gamma)$ does not exist for all such ν. It follows that for each $\alpha \in (\frac{\eta(i)}{1-\eta(i)}, \infty)$, there exists $\gamma \in \Gamma^{\alpha}_{ji}$ such that $\tilde{T}_{(ji)(il)}(\gamma)$ does not exist, namely any γ such that $\gamma(i) = \eta(i)$ and $\gamma(j) = \eta(j)/\alpha$. But by Lemma 5.35, if $\tilde{T}_{(ji)(il)}(\gamma)$ exists at any $\gamma \in \Gamma_{ji}(\alpha)$ then $\tilde{T}_{(ji)(il)}$ exists for all $\gamma \in \Gamma_{ji}(\alpha)$. Hence $\tilde{T}_{(ji)(il)}(\gamma)$ does not exist for any $\gamma \in \Gamma_{ji}(\alpha)$ such that $\alpha \in \left(\frac{\eta(i)}{1-\eta(i)}, \infty\right)$. But this is a set of positive measure in $\tilde{\Gamma}$ which contradicts the result of Lemma 5.34. This contradiction establishes that $\tilde{T}_{(ji)(il)}(\gamma)$ exists for $k = i$ and $l \neq j$.

Finally, note that the above proves the differentiability of f which establishes that $\tilde{T}_{(ji)(il)}(\gamma)$ exists in the case that $k = i$ and $l = j$. ■

A5.3: Theorem 1 (Sufficiency)

Theorem 1: *If data set $C \in \mathcal{C}$ with a UPS representation satisfies IUC, it has a Shannon representation.*

Proof. We are looking to show that if $C \in \mathcal{C}$ *has a UPS representation* $K \in \mathcal{K}^{UPS}$ *and satisfies IUC*, then there exists $\kappa > 0$ such that, given $(\mu, Q) \in \mathcal{F}$ such that $Q \in \hat{\mathcal{Q}}(\mu|K)$,

$$K(\mu, Q) = \Sigma_{\gamma \in \Gamma(Q)} Q(\gamma) T(\gamma),$$

where,

$$T(\gamma) = \kappa \sum_{\omega \in \Gamma(\gamma)} \gamma(\omega) \ln(\gamma(\omega)). \tag{133}$$

The proof has three parts. The first establishes that, given any fixed set of states $\bar{\Omega}$ of cardinality $J \geq 4$, there exists $\kappa^J > 0$ such that, for all corresponding interior posteriors $\gamma \in \tilde{\Gamma}$ with $\bar{\Omega}$, the corresponding strictly convex function $\tilde{T} : \tilde{\Gamma} \to \mathbb{R}$ in the UPS representation has the form,

$$\tilde{T}(\gamma) = \kappa^J \sum_{j=1}^{J} \gamma(j) \ln \gamma(j). \tag{134}$$

The second part of the proof shows all optimal strategies are precisely as if κ^J applied to all posteriors $\gamma \in \Gamma(\mu)$ with $|\Omega(\gamma)| = L \leq J$. This implies that one can identify all optimal strategies, and hence the full data in this CIR, using the necessary and sufficient conditions for optimality when this function to all feasible posteriors. The final step is to apply IUC (A1) not only to show that $\kappa^{J+1} = \kappa^J$ for $J \geq 4$, but also that the same Shannon functional form and multiplier applies for all $J \geq 2$, which completes the proof.

To prove the first part, we pick any fixed set of states $\bar{\Omega}$ of cardinality $J \geq 4$, and define the corresponding interior posteriors $\tilde{\Gamma}$. We choose $\eta \in \tilde{\Gamma}$, set $\alpha = \frac{\eta(k)}{\eta(l)}$, and consider $\nu \in \Gamma_{kl}(\alpha)$, so that,

$$\frac{\eta(k)}{\eta(l)} = \frac{\nu(k)}{\nu(l)} > 0.$$

By Lemma 5.7 (Symmetric Costs), we can order the arguments such that $l = J$. Define the mean belief $\bar{\mu} = \frac{\eta+\nu}{2}$, and μ_t for $t \in [0, 1]$ by:

$$\mu_t(j) = \begin{cases} t[\bar{\mu}(k) + \bar{\mu}(l)] & \text{for } j = k; \\ (1-t)[\bar{\mu}(k) + \bar{\mu}(l)] & \text{for } j = l; \\ \bar{\mu}(j) & \text{otherwise.} \end{cases}$$

By Lemma 5.12 we know that there exists $a, b \in \mathcal{A}$ with $u(a, k) = u(a, l)$ and $u(b, k) = u(b, l)$ such that $C(\mu_t, \{a, b\}) = \{P_t\}$ and for $t \in (0, 1)$, η_t is the revealed posterior for action a and ν_t is the revealed posterior for action b where,

$$\eta_t(j) = \left[\frac{\eta(j)}{\bar{\mu}(j)}\right] \mu_t(j) \qquad \text{and} \qquad \nu_t(j) = \left[\frac{\nu(j)}{\bar{\mu}(j)}\right] \mu_t(j)$$

for $1 \leq j \leq J$. Since $\mu_t(j) = \bar{\mu}(j)$ for $j \neq k, l$ and $\mu_t(k) = t[\bar{\mu}(k) + \bar{\mu}(l)]$, note that $\mu_{\bar{t}} = \bar{\mu}$ if and only it,

$$\bar{t} = \frac{\bar{\mu}(k)}{\bar{\mu}(k) + \bar{\mu}(l)},$$

in which case,

$$\eta_{\bar{t}}(j) = \eta(j) \text{ and } \nu_{\bar{t}}(j) = \nu(j).$$

By Lemma 5.33, $\tilde{T}$ is differentiable for all $\gamma \in \tilde{\Gamma}$. By Lemma 5.11, we know that (81) holds for all $t \in (0,1)$,

$$u(a,i) - u(a,j) - \tilde{T}_{(ji)}(\eta_t) \equiv \tilde{N}^a_{(ji)}(\eta_t) = \tilde{N}^b_{(ji)}(\nu_t) \equiv u(b,i) - u(b,j) - \tilde{T}_{(ji)}(\nu_t).$$

By Lemma 5.36, $\tilde{T}_{(ji)}$ is differentiable in the direction (lk). Since this equation holds for all $t \in (0,1)$, we can differentiate this identity with respect to t at $\bar{t}$,

$$\frac{d\tilde{T}_{(ji)}(\eta_{\bar{t}})}{dt} = \frac{d\tilde{T}_{(ji)}(\nu_{\bar{t}})}{dt}. \tag{135}$$

A change in t at $\bar{t}$ raises $\eta_t(k)$ and lowers $\eta_t(l)$ each by $\eta(k) + \eta(l)$ and leaves $\eta_t(j)$ unchanged for all $j \neq k, l$. The chain rule then implies,

$$\frac{dT_{(ji)}(\eta_{\bar{t}})}{dt} = \tilde{T}_{(ji)(lk)}(\eta)\left[\eta(k) + \eta(l)\right] \tag{136}$$

Combining (135) and (136) and setting $t = \bar{t}$ yields,

$$\tilde{T}_{(ji)(lk)}(\eta)\left[\eta(k) + \eta(l)\right] = \tilde{T}_{(ji)(lk)}(\nu)\left[\nu(k) + \nu(l)\right].$$

Since,

$$\frac{\nu(k) + \nu(l)}{\eta(k) + \eta(l)} = \frac{\bar{t}\left[\nu(k) + \nu(l)\right]}{\bar{t}\left[\eta(k) + \eta(l)\right]} = \frac{\nu(k)}{\eta(k)}$$

we have

$$\eta(k)\tilde{T}_{(ji)(lk)}(\eta) = \nu(k)\tilde{T}_{(ji)(lk)}(\nu). \tag{137}$$

Equation (137) must hold for all η and ν in $\Gamma_{kl}(\alpha)$, therefore $\gamma(k)\tilde{T}_{(ji)(lk)}(\gamma)$ is constant across $\gamma \in \tilde{\Gamma}$. Note that by Lemma 5.7, this constant is independent of the states i, j, k, and l, although at this point it may depend on the dimension J. Since the additive separability of $\tilde{T}$ implies $\tilde{T}_{(ji)(lk)} = 0$ for distinct states i, j, k, and l. The interesting cases involve overlap. Taking $i = k$.

$$\gamma(i)\tilde{T}_{(ji)(li)}(\gamma) = \kappa^J. \tag{138}$$

for some constant κ^J.

We look for the general form of $\tilde{T}_{(ji)}$ that solves (138). We know from Lemma 5.28 that,

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

Hence

$$\gamma(i)f'(\gamma(i)) = \kappa^J$$

The solution to this differential equation is

$$f(\gamma(i)) = \kappa^J \ln\gamma(i) + \varsigma$$

for some constant of integration ς. It follows that

$$\tilde{T}_{(ji)}(\gamma) = \kappa^J \ln\gamma(i) - \kappa^J \ln\gamma(j). \tag{139}$$

Note that we can rule out the dependence of ς on $\gamma(m)$ for $m \neq i$ as $f(\gamma(i))$ depends only on $\gamma(i)$.

The general solution to (139) is

$$\tilde{T}(\gamma) = \kappa^J \gamma(i)[\ln(\gamma(i)) - 1] + \kappa^J \gamma(j)[\ln(\gamma(j)) - 1] + G(\gamma(i) + \gamma(j), \{\gamma(k)\}_{k \neq i,j}).$$

Here G combines constants of integration with the potential for a shift between $\gamma(i)$ and $\gamma(j)$ to offset each other. Lemma 5.7 states that $\tilde{T}$ is symmetric. Hence,

$$\tilde{T}(\gamma) = \sum_j \kappa^J \gamma(j)[\ln(\gamma(j)) - 1] + G\left(\sum_j \gamma(j)\right).$$

As $\sum_j \gamma(j) = 1$ and $\tilde{T}$ is defined only to an affine transformation,

$$\tilde{T}(\gamma) = \kappa^J \sum_j \gamma(j) \ln(\gamma(j)). \tag{140}$$

This completes the proof of (134).

For the second part of the proof, the first key observation is that, given $J \geq 4$, all optimal strategies are precisely as if κ^J applied to all posteriors $\gamma \in \Gamma$ with $|\Omega(\gamma)| = L \leq J$. Note that we replace $\tilde{T}$ with T from this point forward in the proof since this function is designed to apply to boundary as well as to interior posteriors. What we require then is that cost cannot be strictly lower than this formula implies,

$$T(\gamma) \geq \kappa^J \sum_{l=1}^{L} \gamma(l) \ln \gamma(l). \tag{141}$$

To see that this is sufficient, suppose that this has been established and replace the costs of all such posteriors with precisely the lower bound. Note that in this case one can apply the standard necessary and sufficient conditions for optimal choices to conclude that, even with this lower bound imposed it is never optimal to select any such posteriors. The equal likelihood ratio necessary conditions for $\lambda \in \Lambda(\mu, A)$ to be optimal in the Shannon model with general cost parameter $\kappa^J > 0$ in Caplin *et al.* [2016]) which asserts that, for $\lambda \in \Lambda(\mu, A)$ to be an optimal strategy requires, that for all chosen actions $a, b \in \mathcal{A}(\lambda)$,

$$\frac{\gamma_\lambda^a(j)}{\exp(u(a,j)/\kappa^{J+1})} = \frac{\gamma_\lambda^b(j)}{\exp(u(b,j)/\kappa^J + 1)} \text{ all } j. \tag{142}$$

Note that this is inconsistent with there being any posterior with $\gamma_\lambda^a(j) = 0$, since we know that $\mu(j) > 0$, so that by Bayes' rule there must be some strictly positive values $\gamma_\lambda^b(j) > 0$ for some chosen action, hence all must be strictly positive for (142) to hold. Hence replacement of $\kappa^J \sum_{l=1}^{L} \gamma(l) \ln \gamma(l)$ when some of the ex ante possible states are ruled out cannot make strategies with such posteriors optimal: hence there is no loss from the view point of optimization and hence data observed in the CIR in applying this cost function.

We now demonstrate the validity of (141). Suppose to the contrary that there exists some $\gamma \in \Gamma$ with $|\Omega(\gamma)| = L < J$ such that the opposite holds,

$$T(\gamma) = \kappa^J \sum_{l=1}^{L} \gamma(l) \ln \gamma(l) - \delta, \tag{143}$$

for $\delta > 0$. Note that there is no loss of generality in supposing that $L = (J-1)$, since there must be some highest number of states $\bar{L} < J$ for which this is true, which implies that (141) held for all $\bar{L}+1$ yet not for $\bar{L}$, so that the comparison below works at least as well for $\bar{L}+1$ relative to $\bar{L}$, as it does for J relative to $\bar{L}$. It then simplifies to set $\bar{L} = J-1$. We show that this contradicts completeness, since it is cheaper and cannot lower expected utility to rule out an ex ante possible state state than to leave minimal ignorance.

Consider a particular posterior $\bar{\gamma}$ of this form such that (143) holds with difference $\bar{\delta} > 0$. Without loss of generality, suppose that the prior possible state that is impossible state in this posterior is state 1, $\bar{\gamma}(1) = 0$. In this case we can express costs in simple form,

$$T(\bar{\gamma}) = \kappa^J \sum_{j=2}^{J} \bar{\gamma}(j) \ln \bar{\gamma}(j) - \bar{\delta}. \tag{144}$$

We now define a second posterior $\bar{\eta}$ that permutes $\bar{\gamma}$ by reversing the posteriors associated with states 1 and 2,

$$\bar{\eta}(j) = \begin{cases} \bar{\gamma}(2) \text{ if } j = 1; \\ 0 \text{ if } j = 2; \\ \bar{\gamma}(j) \text{ if } j \geq 3. \end{cases}$$

By the Symmetry of Costs Lemma, we know that $T(\bar{\gamma}) = T\left(\bar{\eta}\right)$.

We now define $\bar{\mu}$ to be the average posterior,

$$\bar{\mu} = \frac{\bar{\gamma} + \bar{\eta}}{2} = \begin{cases} \frac{\bar{\gamma}(2)}{2} \text{ if } j = 1, 2; \\ \bar{\gamma}(j) \text{ if } j \geq 3. \end{cases}.$$

and consider the parameterized families of posteriors $\bar{\gamma}^\alpha, \bar{\eta}^\alpha \in \Gamma(\bar{e})$ on $\alpha \in [0,1]$ by:

$$\bar{\gamma}^\alpha(j) = \begin{cases} \alpha\bar{\gamma}(2) \text{ if } j = 1; \\ (1-\alpha)\bar{\gamma}(2) \text{ if } j = 2 \\ \bar{\gamma}(j) \text{ if } j \geq 3. \end{cases} \text{ and } \bar{\eta}^\alpha(j) = \begin{cases} (1-\alpha)\bar{\gamma}(2) \text{ if } j = 1; \\ \alpha\bar{\gamma}(2) \text{ if } j = 2 \\ \bar{\gamma}(j) \text{ if } j \geq 3. \end{cases} \tag{145}$$

By construction, for all $\alpha \in [0,1]$ the simple average of the posteriors $\bar{\gamma}^\alpha, \bar{\eta}^\alpha$ is precisely $\bar{\mu}$,

$$\frac{1}{2}\left[\bar{\gamma}^\alpha(j) + \bar{\eta}^\alpha(j)\right] = \begin{cases} \frac{\bar{\gamma}(2)}{2} \text{ if } j = 1, 2 \\ \bar{\gamma}(j) \text{ if } j \geq 3. \end{cases}$$

Given that the data has a PS representation, theorem 3 implies that Axiom A4 (Completeness) holds. Hence we know that, for any $\alpha > 0$, there exists action set $\{a^\alpha, b^\alpha\}$ such that decision problem $(\bar{\mu}, \{a^\alpha, b^\alpha\}) \in \mathcal{D}$ has data $P \in C\left(\bar{\mu}, \{a^\alpha, b^\alpha\}\right)$ that assigns equal likelihood $\frac{1}{2}$ to each of $\bar{\gamma}^\alpha, \bar{\eta}^\alpha$. Since this is a CIR, there exists an optimal strategy $\lambda^\alpha \in \hat{\Lambda}\left(\bar{\mu}, \{a^\alpha, b^\alpha\}\right)$ with the corresponding property,

$$Q_{\lambda^\alpha}(\bar{\gamma}^\alpha) = Q_{\lambda^\alpha}(\bar{\eta}^\alpha) = \frac{1}{2}.$$

We show now that if $\kappa^J < \kappa^{J+1}$, this cannot hold for small enough $\alpha > 0$. The proof involves demonstrating that alternative strategy λ' that assigns equal likelihood to posteriors $\bar{\gamma}$ and $\bar{\eta}$

$$Q_{\lambda^\alpha}(\bar{\gamma}) = Q_{\lambda^\alpha}(\bar{\eta}) = \frac{1}{2},$$

strictly dominates in this limit.

That such a strategy produces no lower expected utility follows directly from the fact that the posterior distribution in λ^α is a garbling of λ', so that λ' is Blackwell more informative than λ^α. With regard to the costs, note that, for any $\alpha \in (0,1)$, the relevant parameter in the Shannon function is $\kappa^J > 0$, since both possible posteriors have all prior possible states still possible,

$$|\Omega(\bar{\gamma}^\alpha)| = |\Omega(\bar{\eta}^\alpha)| = J.$$

Given that the posteriors are permutations of one another, the corresponding Shannon costs are simple to compute as,

$$K(\bar{\mu}, Q_{\lambda^\alpha}) = \kappa^J \left[\sum_j \bar{\gamma}(j)\ln\bar{\gamma}(j) + \alpha\bar{\gamma}(2)\ln(\alpha\bar{\gamma}(2)) + (1-\alpha)\bar{\gamma}(2)\ln[(1-\alpha)\bar{\gamma}(2)]\right] \tag{146}$$

By construction, equation (144) and the symmetry of costs shows that the corresponding computation for λ' has cost strictly $\bar{\delta} > 0$ below what it would be according to the corresponding Shannon cost,

$$K(\bar{\mu}, Q_{\lambda'}) = \kappa^J \left[\sum_j \bar{\gamma}(j)\ln\bar{\gamma}(j) + \bar{\gamma}(2)\ln(\bar{\gamma}(2))\right] - \bar{\delta}. \tag{147}$$

Subtraction of (147) from (146) reveals that the costs of λ^α are strictly higher than those of λ' provided,

$$\bar{\delta} > \alpha\bar{\gamma}(2)\ln(\alpha\bar{\gamma}(2)) + (1-\alpha)\bar{\gamma}(2)\ln[(1-\alpha)\bar{\gamma}(2)] - \bar{\gamma}(2)\ln(\bar{\gamma}(2)). \tag{148}$$

Note that the LHS of (148) is a fixed strictly positive constant independent of α. With regard to the RHS, note that in the limit as $\alpha \downarrow 0$ it approaches zero, since,

$$\begin{aligned}\lim_{\alpha\downarrow 0}(1-\alpha)\bar{\gamma}(2)\ln[(1-\alpha)\bar{\gamma}(2)] &= \bar{\gamma}(2)(2)\ln[\bar{\gamma}(2)]; \\ \lim_{\alpha\downarrow 0}[\alpha\bar{\gamma}(2)\ln(\alpha\bar{\gamma}(2))] &= 0.\end{aligned}$$

For $\alpha > 0$ but sufficiently small we conclude that,

$$K(\bar{\mu}, Q_{\lambda'}) < K(\bar{\mu}, Q_{\lambda^\alpha}),$$

contradicting optimality of strategy $\lambda(\alpha)$ and thereby establishing (141).

We know now that the applicable cost function for working out all optimal strategies and hence all observed data for priors involving $J \geq 4$ possible states the Shannon cost function with parameter κ^J as defined for posteriors γ with $|\Omega(\gamma)| = J$ in equation (140). The final part of the proof uses IUC (A1) to iterate down in dimension. To be precise, define K^J to be the Shannon cost function with parameter κ^J for $J \geq 4$ as defined on all posteriors with that state space or below,

$$K^J(\gamma) \equiv \kappa^J \sum_{j \in \Omega(\gamma)} \gamma(j) \ln \gamma(j), \text{ all } \gamma \in \Gamma \text{ with } |\Omega(\gamma)| \leq J. \tag{149}$$

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

$$P \in C(\mu, A) \text{ iff } \exists \lambda \in \hat{\Lambda}(\mu, A|K^J) \text{ such that } \mathbf{P}_\lambda = P. \tag{150}$$

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

Consider first $P \in C(\mu, A)$ where $|\Omega(\mu)| = J - 1$ and label the states in $\Omega(\mu)$ by $2 \leq j \leq J$. Consider now an additional state, $j = 1$, that in payoff terms is a replica of state $j = 2$,

$$u(a, 1) = u(a, 2) \text{ all } a \in A;$$

and the prior μ' that divides up the $\mu(1)$ equally between states $j = 1, 2$,

$$\mu'(j) = \begin{cases} \frac{\mu(1)}{2} \text{ if } j = 1, 2; \\ \mu(j) \text{ if } j \geq 3. \end{cases}$$

By construction (μ, A) is a basic version of (μ', A), $(\mu, A) \in \mathcal{B}(\mu', A)$. Hence by IUC we know that data set P' that agrees with P in all states and repeats in state $j = 1$ what P specifies in state $j = 1$,

$$P'(a|j) = \begin{cases} P(a|2) \text{ if } j = 1; \\ P(a|j) \text{ if } j \geq 2; \end{cases}$$

satisfies $P' \in C(\mu, A)$. Consider any chosen action $a \in \mathcal{A}(P')$ and define the corresponding revealed posterior $\bar{\gamma}^a_{P'}$. Since this is a CIR, we know that $\exists \lambda' = (Q', q') \in \hat{\Lambda}(\mu', A|K^J)$ such that $\mathbf{P}_\lambda = P'$ and with this posterior possible, and action a chosen deterministically at this posterior (by Lemma 2.7, FIO),

$$Q'(\bar{\gamma}^a_{P'}) > 0 \text{ and } q'(a|\bar{\gamma}^a_{P'}) = 1.$$

Applying this to all chosen actions and noting that the strategy is optimal, we know that it satisfies the full bank of necessary and sufficient conditions for an optimal strategy. Specifically, given $a, b \in \mathcal{A}(P')$ we therefore know that the ILR equality holds:

$$\frac{\bar{\gamma}^a_{P'}(j)}{\exp(u(a, j)/\kappa^J)} = \frac{\bar{\gamma}^b_{P'}(j)}{\exp(u(b, j)/\kappa^J)} \text{ all } j; \tag{151}$$

together with the corresponding inequality: given $a \in \mathcal{A}(P')$ and $c \in A$,

$$\sum_{j=1}^{J} \left[\frac{\bar{\gamma}^a_{P'}(j)}{\exp(u(a, j)/\kappa^J)} \right] \exp(u(c, j)/\kappa^J) \leq 1. \tag{152}$$

Note that the revealed posteriors defined by data set P on chosen actions $a \in \mathcal{A}(P') = \mathcal{A}(P)$ are readily derived from those associated with data set P' by application of Bayes' rule,

$$\bar{\gamma}_P^a = \begin{cases} 0 & \text{if } j = 1; \\ \bar{\gamma}_{P'}^a(1) + \bar{\gamma}_{P'}^a(2) & \text{if } j = 2; \\ \gamma_{\lambda^{J-1}}^a(j) & \text{if } j \geq 3. \end{cases}$$

Now consider the strategy $\lambda = (Q_\lambda, q_\lambda) \in \Lambda(\mu, A)$ designed to precisely mirror $\lambda' = (Q', q') \in \hat{\Lambda}(\mu', A|K^J)$,

$$Q_\lambda(\bar{\gamma}_P^a) = Q'(\bar{\gamma}_{P'}^a) \text{ and } q_\lambda(a|\bar{\gamma}_P^a) = q'(a|\bar{\gamma}_{P'}^a).$$

Note by construction that $\mathbf{P}_\lambda = P$. Note also that validity of (151) and (152) survives this amalgamation,

$$\begin{aligned} a, b &\in \mathcal{A}(P) \Longrightarrow \frac{\bar{\gamma}_P^a(j)}{\exp(u(a,j)/\kappa^J)} = \frac{\bar{\gamma}_P^b(j)}{\exp(u(b,j)/\kappa^J)} \text{ all } j; \\ a &\in \mathcal{A}(P), c \in A \Longrightarrow \sum_{j=1}^{J} \left[\frac{\bar{\gamma}_P^a(j)}{\exp(u(a,j)/\kappa^J)} \right] \exp(u(c,j)/\kappa^J) \leq 1. \end{aligned}$$

Hence this strategy satisfies the necessary and sufficient conditions for $\lambda \in \hat{\Lambda}(\mu, A|K^J)$ as required.

The final step in the proof is to show the converse implication: given $\lambda = (Q_\lambda, q_\lambda) \in \hat{\Lambda}(\mu, A|K^J)$, the corresponding data satisfies $\mathbf{P}_\lambda \in C(\mu, A)$. We work again with the necessary and sufficient conditions (151) and (152) that characterize such optimal strategies: given $a, b \in \mathcal{A}(\lambda)$:

$$\begin{aligned} a, b &\in \mathcal{A}(\lambda) \Longrightarrow \frac{\gamma_\lambda^a(j)}{\exp(u(a,j)/\kappa^J)} = \frac{\gamma_\lambda^b(j)}{\exp(u(b,j)/\kappa^J)} \text{ all } j \geq 2; \\ a &\in \mathcal{A}(\lambda), c \in A \Longrightarrow \sum_{j=2}^{J} \left[\frac{\gamma_\lambda^a(j)}{\exp(u(a,j)/\kappa^J)} \right] \exp(u(c,j)/\kappa^J) \leq 1. \end{aligned}$$

We define the additional state $j = 1$ and the corresponding prior μ' precisely as above. We then derive strategy $\lambda' = (Q', q') \in \Lambda(\mu', A)$ from $\lambda = (Q_\lambda, q_\lambda) \in \Lambda(\mu, A)$ by reversing the process above. Given $a \in \mathcal{A}(\lambda)$ we first define corresponding posteriors,

$$\gamma_{\lambda'}^a = \begin{cases} \frac{\gamma_\lambda^a(2)}{2} & \text{if } j = 1, 2; \\ \gamma_\lambda^a(j) & \text{if } j \geq 3. \end{cases}$$

We then define the strategy $\lambda' = (Q', q') \in \Lambda(\mu', A)$ by

$$Q'(\gamma_{\lambda'}^a) = Q_\lambda(\gamma_\lambda^a) \text{ and } q_\lambda(a|\gamma_{\lambda'}^a) = q'(a|\gamma_\lambda^a).$$

To round out the proof, we note first that this strategy satisfies the full conditions (151) and (152) characterizing optimal strategies for the Shannon model. Hence we conclude that $\lambda' = (Q', q') \in \hat{\Lambda}(\mu', A|K^J)$. Hence, since this is a CIR, we conclude that $\mathbf{P}_{\lambda'} \in C(\mu, A')$. Finally, we apply IUC, which shows that the corresponding data satisfies $\mathbf{P}_\lambda \in C(\mu, A)$. This completes the proof. ■

Corollary 3: Data set $C \in \mathcal{C}$ has a Shannon representation if and only if it satisfies A1 through A9.

Proof. Suppose first that $C \in \mathcal{C}$ satisfies A1 through A9. By theorem 4 sufficiency we know that, since $\mathcal{C}$ satisfies A2 through A9, it has a UPS representation. At this point theorem 1 sufficiency applies, whereby, since $C \in \mathcal{C}$ has a UPS representation and satisfies A1, it has a Shannon representation.

To complete the proof, we show that having a Shannon representation implies satisfaction of A1 through A9. Theorem 1 necessity shows directly that if $C \in \mathcal{C}$ has a Shannon representation it satisfies A1. To prove that A2 through A9 are implied, we show that any $C \in \mathcal{C}$ that has Shannon representation is regular, $C \in \mathcal{C}^R$. This will complete the proof, since by theorem 4 necessity, we know that, since $C \in \mathcal{C}^R$ has a Shannon representation, it also has a UPS representation, hence satisfies A2 through A9.

To establish that any $C \in \mathcal{C}$ that has Shannon representation is regular, $C \in \mathcal{C}^R$, consider $\mu_1 \in \Gamma$ and $Q \in \Delta(\Gamma(\mu_1))$ with $\Gamma(Q) \subset \Gamma^C(\mu_1)$ and $\sum\limits_{\gamma \in \Gamma(\mu_2)} \gamma Q(\gamma) = \mu_2$. It is implied directly from the invariant likelihood ratio characterization of optimal strategies that, given prior $\mu_1 \in \Gamma$, the set of observed posteriors associated with a Shannon representation is precisely the interior set of posteiors, $\Gamma^C(\mu_1) = \tilde{\Gamma}(\mu_1)$. Hence given $Q \in \Delta(\Gamma(\mu_1))$ with $\Gamma(Q) \subset \Gamma^C(\mu_1)$, we know that $\Omega(\gamma) = \Omega(\mu_1)$ all $\gamma \in \Gamma(Q)$. Hence the same applies to their weighted average,

$$\sum_{\gamma \in \Gamma(\mu_2)} \gamma Q(\gamma) = \mu_2.$$

Given that $\Omega(\mu_2) = \Omega(\mu_1)$, we know that $\Gamma^C(\mu_2) = \tilde{\Gamma}(\mu_2)$. Overall,

$$\Gamma(Q) \subset \Gamma^C(\mu_1) = \tilde{\Gamma}(\mu_1) = \tilde{\Gamma}(\mu_2),$$

establishing regularity, $C \in \mathcal{C}^R$, and completing the proof of the corollary. ■

References

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

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

Andrew Caplin, Mark Dean, and John Leahy. Rational inattention, optimal consideration sets and stochastic choice. 2016.

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

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

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