Robust Inattentive Discrete Choice[*]

Lars Peter Hansen (University of Chicago)
Jianjun Miao (Boston University)
Hao Xing (Boston University)

December 18, 2024

Abstract

Rational inattention models characterize optimal decision-making in data-rich environments. In such environments, it can be costly to look carefully at all of the information. Some information is much more salient for the decision at hand and merits closer scrutiny. The inattention decision model formalizes this choice and deduces how best to navigate through the potentially vast array of data when making decisions. In the rational formulation, the decision-maker commits fully to a subjective prior distribution over the possible states of the world that could be realized. We relax this assumption and look for a robustly optimal solution to the inattention problem by allowing the decision-maker to be ambiguity averse with respect to this prior. We feature a setup that is deliberately simple by a) assuming a discrete set of choices, b) using Shannon's mutual information to quantify attention costs, and c) imposing relative entropy with respect to a baseline probability distribution to quantify prior divergence. We provide necessary and sufficient conditions for the robust solution and develop numerical methods to solve it. In comparison to the rational solution with no prior uncertainty, our decision-maker slants priors in more cautious or pessimistic directions when deducing how to allocate attention over the range of available information. This approach implements a form of robustness to prior misspecification, or equivalently, a form of ambiguity aversion. We explore some examples that show how the robust solution differs from the rational solution with a commitment to a subjective prior distribution and how it differs from imposing risk aversion.

December 18, 2024

Keywords— rational inattention | robustness | information acquisition |

[*] We gratefully acknowledge comments by Maryam Farboodi, Diana Petrova, Monika Piazzesi, and Judy Yue, and research assistance by Chun Hei Hung. We are very appreciative of feedback provided on an earlier draft by Tommaso Denti and Benjamin Hebert.

People often make choices with limited information. In so doing, they determine where to focus their attention when the available information is vast. To study such decision problems, Sims (1998, 2003) introduces a rational inattention (RI) framework by modeling information costs using the Shannon (1948) entropy-based mutual information (henceforth the Shannon model). While Sims (1998) focuses on continuous choices in dynamic settings, Caplin and Dean (2015), Matêjka and McKay (2015), and Caplin et al. (2019, 2022) adopt the RI framework to study discrete choice problems in static settings. We follow these latter authors by studying discrete choice problems.[1] Such problems have applications to such diverse fields as labor economics, industrial organization, macroeconomics, and political economy. This framework also has links to model selection problems in statistics and control theory. Those literatures study the ability to statistically discriminate among models using existing data sets. This becomes an attention allocation problem when the decision maker is unsure which statistical model is correct and explores some of a vast amount information to help with this determination.

While the Shannon model is a standard approach in the literature, it may suffer from a problem of prior misspecification. Specifically, the Shannon mutual information is defined as the difference between the entropies of the prior and the posterior. The prior is assumed to be exogenously given to the decision maker (DM), and the DM has full confidence in this prior. The perspective adopted in this paper is that of a DM with prior ambiguity who is looking to make a robustness adjustment.

Prior uncertainty can have important ramifications for the Shannon model because, as Caplin et al. (2019) show, the different specifications of the prior can generate different so called "consideration sets." Consideration sets are collections of alternatives chosen with positive probabilities. The implications of prior ambiguity carry over to these sets and change how the DM views alternatives within these sets.

Operationally, we incorporate prior robustness into a static, discrete-choice setting by integrating a version of Hansen and Sargent (2007, 2023), and Hansen and Miao (2018). As Hansen and Sargent (2007) show, there is a robust prior interpretation of the smooth ambiguity model of Klibanoff et al. (2005). While the smooth ambiguity model has multiple axiomatic defenses, Hansen and Sargent (2023) argue that the robust prior interpretation opens the door to the axiomatic formulations of Maccheroni et al. (2006) and Strzalecki (2011).[2] Intuitively, the DM does not have a single prior about the state of the world but instead has a baseline or reference prior. The DM explores the consequences of alternative priors that may be different from a baseline prior. The deviation of any prior from the baseline prior is penalized by a cost modeled by the relative entropy. The cost is scaled in utility units by a robustness parameter that dictates the degree of ambiguity aversion. Alternatively, the DM may impose a relative entropy constraint on the family of priors, giving rise to a version of Gilboa and Schmeidler (1989)'s max-min utility preferences.

1 Model

We first introduce the model setup and then provide an equivalent formulation to simplify the decision problem.

[1] An earlier version of this research was presented at the American Economic Association meetings in January of 2022 in a session on "Expectations and Macro-Finance."

[2] For other statistical motivations for smooth ambiguity, see Cerreia-Vioglio et al. (2013) and Denti and Pomatto (2022).

1.1 Setup

In what follows, we use a bold lower case letter $\mathbf{x}$ to denote a random variable and a letter x to denote its realization. The set $\Delta(Y)$ denotes the collection of probability distributions on any finite set Y. One such set is the finite state space $X = \{x_1, x_2, ..., x_M\}$ and a prior $\mu \in \Delta(X)$. Another is the set S, which denotes a finite set of signal realizations.

The DM does not observe the state, but can acquire a signal about the state. Based on the signal, the DM selects an element from a finite action set A to maximize expected utility $u : X \times A \to \mathbb{R}$. The DM also chooses an optimal information structure by paying a cost modeled by the Shannon mutual information. A strategy is a pair (d, σ) composed of

(i) an *information strategy*, d, consisting of a system of signal distributions $d(s \mid x)$, for all $s \in S$, $x \in X$;

(ii) an *action strategy*, $\sigma : S \to A$, specifying an action $a = \sigma(s)$ when observing a signal s.

Let Σ denote the set of all strategies. While the information strategy, d, is an object of choice, the prior over states, μ, is given with full commitment in the standard Shannon RI specification.

An information strategy and prior over states induce a joint distribution $d \otimes \mu$ over $X \times S$:

$$(d \otimes \mu)(x, s) = d(s|x)\,\mu(x).$$

In what follow, we let $\mathbb{E}_{d\otimes\mu}$ denote the expectation computed using $d(s|x)\mu(x)$. The implied marginal distribution over signals is:

$$\tau(s) = \sum_x \mu(x)\, d(s|x),$$

and the conditional distribution over states is

$$\mu_s(x) \overset{\text{def}}{=} \frac{d(s|x)\,\mu(x)}{\tau(s)}$$

for signals s for which $\tau(s) > 0$, and $\mu_s(x) = \mu(x)$ for signals s with $\tau(s) = 0$.

We consider Shannon entropy for both the marginal and the conditional distribution of states:

$$\mathbb{H}(\mu) = -\sum_x \mu(x) \log \mu(x)$$
$$\mathbb{H}(\mu_s) = -\sum_x \mu_s(x) \log \mu_s(x).$$

Entropy $\mathbb{H}(\mu)$ measures the amount of uncertainty embedded in the prior about the underling state. Mutual information measures the reduction of uncertainty after observing signals:

$$\mathbb{I}(d \otimes \mu) \overset{\text{def}}{=} \sum_s \left[\mathbb{H}(\mu) - \mathbb{H}(\mu_s)\right] \tau(s).$$

Equivalently,

$$\begin{aligned}&\mathbb{I}(d \otimes \mu)\\ &= \sum_{x,s} d(s \mid x)\mu(x)\left(\log\left[d(s \mid x)\mu(x)\right] - \log\left[\mu(x)\tau(s)\right]\right),\end{aligned}$$

which is the Kullback-Leibler (KL) divergence between the joint distribution of states and signals relative to the product of the marginals.[3]

As a precursor to our analysis, we first state the standard RI problem as a benchmark for comparison:

Problem 1. *(Signal RI problem)*

$$V(\mu) = \max_{(d,\sigma)\in\Sigma} \mathbb{E}_{d\otimes\mu}\left(u\left[\boldsymbol{x}, \sigma(\boldsymbol{s})\right]\right) - \lambda\mathbb{I}(d \otimes \mu).$$

In this problem, we interpret λ as the shadow cost of information expressed in utility units. It scales mutual information and hence acts as a per unit measure of cost. We refer to it as a *shadow* cost because we are not claiming that it has an actual market counterpart.

As is well known in the literature, the solution to Problem 1 is sensitive to the specification of the prior μ. Since prior misspecification is a concern, we consider a robust alternative in which in which the single prior, μ, is replaced by baseline prior, $\widehat{\mu}$, and the DM explores the adverse consequences of prior misspecification subject to a scaled version of a relative entropy or Kublack-Leibler cost:

$$\mathbb{R}\left(\mu \| \widehat{\mu}\right) \coloneqq \sum_x \mu(x) \log \frac{\mu(x)}{\widehat{\mu}(x)}.$$

Our robust alternative to Problem 1 is:

Problem 2. *(Robust signal RI problem)*

$$\begin{aligned}W(\widehat{\mu}) = \max_{(d,\sigma)\in\Sigma} \min_{\mu\in\Delta(X)} \quad & \mathbb{E}_{d\otimes\mu}\left(u\left[\boldsymbol{x}, \sigma(\boldsymbol{s})\right]\right) - \lambda\mathbb{I}(d \otimes \mu)\\ & +\xi\mathbb{R}\left(\mu \| \widehat{\mu}\right),\end{aligned} \tag{1}$$

where $\xi > 0$ denotes a robustness parameter.

Since the constraint sets are compact and the objective function is continuous, the robust RI problem has a solution. The minimization in (1) reflects the DM's aversion to prior ambiguity as parameterized by ξ.[4]

As specified, ξ governs a *smooth* tradeoff between the utility maximization and the cost of prior ambiguity as in multiplier preferences of Hansen and Sargent (2001) and the variational preferences of Maccheroni et al. (2006).[5] Large values of ξ induce large costs in deviating from the baseline prior and hence a small amount

[3] See Cover and Thomas (2012) for a textbook treatment of information theory.

[4] Mutual information depends on the prior. An alternative formulation of information cost was proposed in Woodford (2012) using channel capacity, which is the largest mutual information among all possible priors. In our Problem 2, the prior contributes to the expected utility. Notice that if we set $\xi = 0$, the minimization over priors would compute the channel capacity as the implied measure of information were we to ignore how the prior influences the expected utility.

[5] References Hansen and Sargent (2001) and Maccheroni et al. (2006) feature model misspecification, but subsequent literature has extended their approach to the case of prior misspecification. See Hansen and Sargent (2023).

of aversion to ambiguity. Alternatively, we may suppose that there is constraint on relative entropy:

$$\mathbb{R}\left(\mu || \widehat{\mu}\right) \leq \eta$$

for some $\eta > 0$. In this case η governs the ambiguity aversion as in Gilboa and Schmeidler (1989). Consider a Lagrangian specification of the problem with ξ as a multiplier as in Petersen et al. (2000) and Hansen and Sargent (2001). Using a standard duality argument:

$$\begin{aligned} W\left(\widehat{\mu}\right) = & \max_{(d,\sigma)\in\Sigma} \max_{\xi\geq 0} \min_{\mu\in\Delta(X)} \mathbb{E}_{d\otimes\mu}\left(u\left[\mathbf{x}, \sigma(\mathbf{s})\right]\right) - \lambda\mathbb{I}(d\otimes\mu) \\ & + \xi\mathbb{R}\left(\mu || \widehat{\mu}\right) - \xi\eta \\ = & \max_{\xi\geq 0} \max_{(d,\sigma)\in\Sigma} \min_{\mu\in\Delta(X)} \mathbb{E}_{d\otimes\mu}\left(u\left[\mathbf{x}, \sigma(\mathbf{s})\right]\right) - \lambda\mathbb{I}(d\otimes\mu) \\ & + \xi\mathbb{R}\left(\mu || \widehat{\mu}\right) - \xi\eta. \end{aligned}$$

The inner max-min problem of the second representation is of the same form as the smooth version posed in Problem 2.

An entirely analogous argument applies if we impose an information constraint that

$$\mathbb{I}(d\otimes\mu) \leq \kappa$$

we could add the term $\lambda\kappa$ to the objective and treat λ as a Lagrange multiplier. More generally, for each choice of (λ, ξ), we could deduce an implied information constraint and relative entropy constraint. When imposing numerical parameter values in applications, the magnitudes of the implied constraints can be informative ways to assess parameter plausibility.

Example 3. *We now consider a model selection problem familiar from statistics. Let states be models and an action be a guess of a model with utility function:*

$$u(x, a) = \begin{cases} 1 & a = x \\ 0 & a \neq x. \end{cases}$$

While we posited a symmetric utility function, the rewards for the correct identification of a model and penalties for mistakes could be distinct across partitions. This would be the case, for instance, when the alternative statistical models might have different implications for future courses of action. Robustness considerations would come into play in the choice of prior over states.

Now suppose there is a hypothetical data set that defines the upper bound of information, captured by a distribution $\bar{d}(y \mid x)$ (or a data-rich likelihood when viewed as a function of x), where y is a realization of the potentially available data. This upper bound could be formalized using Blackwell (1951)'s ordering and information quantified using the same Shannon approach applied here by comparing posteriors to the prior. In other words, a potentially interesting extension of our formulation could impose this upper bound on what can be revealed by the signals in conjunction with an information cost in a way that preserves convexity as in Blackwell's formulation. Robustness considerations could come into play, not only in the choice of prior over *states, but also in the choice of the probabilistic upper bound on the available information about the models.*[6]

1.2 Simplifying the problem solution

When solving Problem 2, there is no reason to have signals provide more information than what is necessary to incorporate into decision making since signals are costly to obtain. This suggests an alternative formulation with a probabilistic decision, $p(a \mid x)$, and an information cost that imposed directly on p. In effect, the choice p now serves as *both* an action and a signal distribution conditioned on the state. In this subsection we pose and study this alternative formulation.

Observe that a strategy $(d, \sigma) \in \Sigma$ generates a choice rule $p \in \Delta(A|X)$ where $\Delta(A|X)$ denotes the set of conditional distributions $p(\cdot|x)$ on A given by

$$p(a|x) = \Pr(\sigma(s) = a|x) = \sum_{\{s \in S : \sigma(s) = a\}} d(s|x) . \tag{2}$$

Instead of the strategy (d, σ), we now let actions play the role of signals in terms of the attention costs. We suppose that the DM specifies a choice rule that is a conditional probability of actions given the states. This conditional distribution along with the prior, μ, over states imply a joint distribution $p \otimes \mu$ over $X \times A$:

$$(p \otimes \mu)(x, a) \equiv p(a|x) \mu(x) .$$

Given this joint distribution, we define the mutual information as

$$\mathbb{I}(p \otimes \mu) \overset{\text{def}}{=} \sum_a [\mathbb{H}(\mu) - \mathbb{H}(\mu_a)] q(a),$$

where $\mu_a \in \Delta(X)$ and $q \in \Delta(A)$ denote the posterior and marginal distributions implied by the joint distribution $p \otimes \mu$:

$$\begin{aligned} q(a) &= \sum_x p(a|x) \mu(x) , & (3) \\ \mu_a(x) &= \frac{p(a|x) \mu(x)}{\sum_x p(a|x) \mu(x)}, \text{ if } q(a) > 0. & (4) \end{aligned}$$

Problem 4. *(Choice-based robust RI problem)*

$$\begin{aligned} J(\widehat{\mu}) \overset{\text{def}}{=} \max_{p \in \Delta(A|X)} \min_{\mu \in \Delta(X)} \quad & \mathbb{E}_{p \otimes \mu}[u(\mathbf{x}, \mathbf{a})] - \lambda \mathbb{I}(p \otimes \mu) \\ & + \xi \mathbb{R}(\mu || \widehat{\mu}) , \end{aligned} \tag{5}$$

where $\mathbb{E}_{p \otimes \mu}$ is an expectation operator given distribution $p \otimes \mu$.

In the online appendix, we provide a Recommendation Lemma to establish the equivalence of Problems 2 and 4. Before providing solutions to Problem 4 in the next section, we first present some special limiting

[6] The paper Brooks et al. (2004) uses a related information restriction, but applied coherently across multiple decision-makers and combined with an f-divergence bound. They deduce implications of games with imperfect information from the perspective of an outside observer.

cases. First, when $\xi = \infty$, Problem 4 is reduced to the standard RI problem. In this case the worst-case prior is the baseline prior. Second, when $\xi = 0$, Problem 4 is reduced to the following one:

$$\max_{p \in \Delta(A|X)} \min_{\mu \in \Delta(X)} \mathbb{E}_{p \otimes \mu} \left[u\left(\mathbf{x}, \mathbf{a}\right)\right] - \lambda \mathbb{I}\left(p \otimes \mu\right).$$

This is the extreme case in which the DM thinks any prior in the feasible set $\Delta\left(X\right)$ is possible and the penalty cost is zero. Third, when $\lambda = 0$, the DM can acquire signals to fully observe the state. The DM will select the highest-payoff action with probability one conditional on each state. Finally, when $\lambda = \infty$, the DM does not acquire any information about the state and selects an action to maximize expected utility given the worst-case prior.

Remark 5. *For the rational intention specification in which there is a commitment to the baseline prior, the distribution q has meaning as a statement of the ex ante probability of the different actions. Under robustness there is ambiguity about the prior. The choice of a worst-case prior is not the subjective belief of the DM, but rather it is a device to obtain a robustly optimal choice of $p^*(a \mid x)$. Since the distribution, $q^*(a)$, inherits the worst-case prior, it is not interpretable as the ex ante probability of the alternative actions. Nevertheless, it may still provide a useful summary of the robust choices $p^*(a \mid x)$ in comparison to an average using the baseline prior distribution.*

Notice that randomization among the discrete options remains a possibility, as a choice of p that is independent of the state is feasible for the DM.

2 Model Solution

We include a seemingly superfluous contribution to the maximization problem by including the marginal q over actions. At the same time, we ignore the constraint linking this marginal to the conditional $p(a \mid x)$ and the prior $\mu(x)$. That link will follow directly from the optimization that we investigate.[7] Along with the inclusion of q in the optimization, we introduce the corresponding consideration sets defined as the set of actions chosen with positive probabilities. Formally, for any $q \in \Delta\left(A\right)$, the associated consideration set is defined as $B\left(q\right) = \left\{a \in A : q\left(a\right) > 0\right\}$.

To help understand why we may include q in the optimization, a straightforward derivation shows that

$$\min_{q \in \Delta(A)} \sum_x \mu(x) \sum_a \left[\log p(a|x) - \log q(a)\right] p(a|x)$$

has

$$q^*(x) = \sum_x \mu(x)\, p(a|x)$$

as its solution, which is simply the marginal distribution over A implied by p. By adding and subtracting $\log \mu(x)$ inside the summations, observe that the minimizing objective is the information measure $\mathbb{I}(p \otimes \mu)$.

[7] This equivalent formulation utilizes special features of the Shannon entropy cost function. For general uniformly posterior-separable (UPS) cost functions, Caplin and Dean (2013) and Caplin et al. (2019) introduce a posterior-based approach. Dynamic problems with UPS cost functions are analyzed by Miao and Xing (2024) and a numerical algorithm is also proposed therein.

We solve Problem 4 by computing:

$$J\left(\widehat{\mu}\right) \stackrel{\text{def}}{=} \max_{q\in\Delta(A),p\in\Delta(A|X)} \min_{\mu\in\Delta(X)} F\left(p,q,\mu\right), \tag{6}$$

where

$$F\left(p,q,\mu\right) \stackrel{\text{def}}{=} \sum_x \mu(x)\left[G(p,q)(x) + \xi \log \frac{\mu\left(x\right)}{\widehat{\mu}\left(x\right)}\right], \tag{7}$$

and

$$G(p,q)(x) \stackrel{\text{def}}{=} \sum_a p\left(a|x\right)\left[u\left(x,a\right) - \lambda \log \frac{p\left(a|x\right)}{q\left(a\right)}\right]$$

The function F is concave in (p,q) and convex in μ. By the Minimax Theorem, we can exchange the extremization without effecting the optimized value:

$$J\left(\widehat{\mu}\right) = \min_{\mu\in\Delta(X)} \max_{q\in\Delta(A),p\in\Delta(A|X)} F\left(p,q,\mu\right). \tag{8}$$

This also gives an alternative way to compute the robust solution to the rational inattention problem. In what follows, we implement a hybrid approach.

Consider the inner minimization in the problem (6), taking $p = p^*$ and $q = q^*$ as given. This has a well known solution from robust control theory, large deviation theory and other applications of relative entropy (for example, see Donsker and Varadhan (1975), Dupuis and Ellis (1997), Petersen et al. (2000)):

$$\mu^*(x) = \frac{\exp\left[-\left(\frac{1}{\xi}\right)\upsilon(x)\right]\widehat{\mu}(x)}{\sum_y \exp\left[-\left(\frac{1}{\xi}\right)\upsilon(y)\right]\widehat{\mu}(y)} \tag{9}$$

for

$$\upsilon(x) \stackrel{\text{def}}{=} G(p^*,q^*)(x). \tag{10}$$

This displays exponential tilting of μ^* relative to $\widehat{\mu}$ towards values of x for which the values of υ are relatively low. The minimized objective is known to be:

$$J\left(\widehat{\mu}\right) = -\xi \log \sum_x \widehat{\mu}(x)\exp\left[-\left(\frac{1}{\xi}\right)\upsilon(x)\right]. \tag{11}$$

Consider the inner maximization in the problem (8). This takes $\mu = \mu^*$ as given, and as consequence the last term in the objective F can be ignored. It suffices to study:

$$\max_{q\in\Delta(A),p\in\Delta(A|X)} \sum_x \mu^*(x)G(p,q)(x) \tag{12}$$

The maximizing solution for (p,q) coincide with those of a standard rational inattention problem. Following Matêjka and McKay (2015), the first-order conditions for p imply that

$$p^*(a \mid x) = \frac{q(a)\exp\left[u(x,a)/\lambda\right]}{\sum_b q(b)\exp\left[u(x,b)/\lambda\right]} \qquad \mu^*(x) > 0. \tag{13}$$

While this formula can be deduced by ignoring a nonnegativity constraint on p, the resulting p^* will be nonnegative provided that q is.

Next we follow Caplin et al. (2019) by substituting formula (13) into the objective F for Problem 8 and maximizing the resulting objective as a function of q. Thus we form:

$$\widehat{G}(q)(x) \overset{\text{def}}{=} \lambda \log \sum_a q(a) \exp[u(x,a)/\lambda], \tag{14}$$

and q^* solves:

$$\max_{q \in \Delta(A)} \lambda \sum_x \mu^*(x) \log \sum_a q(a) \exp[u(x,a)/\lambda].$$

In particular, $\upsilon = \widehat{G}(q^*)$ where υ is defined in (10).

We now impose that q be *nonnegative*, as this constraint could well bind. Following Caplin et al. (2019), we obtain the relation:

$$\sum_x \mu^*(x) \left(\frac{\exp\left[u(x,a)/\lambda\right]}{\sum_b q^*(b) \exp\left[u(x,b)/\lambda\right]} \right) \begin{cases} \leq 1 & \forall a, \\ = 1 & a \in B(q^*) \end{cases} \tag{15}$$

as a set of first-order conditions. Since F is concave in (p,q), conditions (13) and (15) are both necessary and sufficient for the maximization given $\mu = \mu^*$.

We combine these results in the following proposition.

Proposition 6. *The triple (p^*, q^*, μ^*) is the solution to the robust RI problem if, and only if,* (9), (13), *and* (15) *are satisfied. The resulting value function J is given by*

$$J(\widehat{\mu}) = -\xi \log \sum_x \widehat{\mu}(x) \exp\left[-\left(\frac{1}{\xi}\right) \upsilon(x) \right].$$

for υ given by (10).

The robust decision is the conditional distribution $p^*(a \mid x)$, capturing the signal distribution. Recall that we "normalized" the problem so that the action coincides with the signal. While our decision maker chooses μ^* along with p^*, the worst-case prior, μ^*, is a vehicle for constructing p^* and together these determine q^*, and the worst-case posterior distribution is given by:

$$\mu_a^*(x) = \frac{p^*(a \mid x)\mu^*(x)}{q^*(a)},$$

provided that $q^*(a) > 0$.

We interpret $\upsilon(x)$ as the ex-post payoff (utility) derived from a standard RI problem for a given prior μ^*. Moreover, imposing μ^* on the maximization problem over (p,q) gives the standard RI solution. Thus one interpretation is that the robust solution replaces $\widehat{\mu}$ with μ^* in an otherwise standard RI problem where μ^* tilts the prior distribution by putting more weight on states for which ex-post payoff is relatively lower. An alternative robust preference interpretation has the analogous slanting for possible pair (p,q) in forming a ranking over such admissible pairs. Under either interpretation, the resulting value function is smaller as it reflects the cost of prior misspecification or ambiguity aversion.

Remark 7. *As Caplin et al. (2019) note, in the absence of robustness considerations, we could repose problem to be a choice of* $(\mu, q, \mu_a(\cdot))$ *subject to*

$$\sum_a \mu_a(x)q(a) = \mu(x)$$

for all states x*. The necessary and sufficient conditions are again given by (9), (13) and (15) substituting*

$$\frac{\mu_a^*(x)q^*(a)}{\mu^*(x)}$$

for $p^*(a \mid x)$*. This same insight extends to our analysis with robustness concerns with the following caveat. In the problem analyzed here, it is the twisted or worst-case probability that adjusts for prior ambiguity and limited attention that would be chosen under this strategy. In contrast, the chosen signal distributions that we feature do not require this qualification.*

3 Numerical method

An algorithm of Arimoto (1972) and Blahut (1972) finds numerical solutions to the standard RI problem.

This algorithm is an application of the general block coordinate descent algorithm, which applies to multivariate optimization problems in which the constraint set has a Cartesian product property. The key idea is that at each iteration one solves the optimization problem with respect to each of the block coordinate taken in cyclic order. Based on this idea, we propose the following generalized Arimoto-Blahut algorithm to solve the robust RI Problem 4:

1. Start with a guess $\mu^{(0)} \in \Delta(X)$ with $\mu^{(0)}(x) > 0$ for all x and a guess $p^{(0)} \in \Delta(A|X)$ with $p^{(0)}(a|x) > 0$ for all (x, a).

2. Given $\left(p^{(k)}, q^{(k)}, \mu^{(k)}\right)$ for step size $0 < \mathsf{s} \leq 1$ compute:

$$\begin{aligned}
\bar{q}^{(k+1)}(a) &= \sum_x \mu^{(k)}(x)\, p^{(k)}(a|x)\,,\\
q^{(k+1)}(a) &= q^{(k)}(a) + \mathsf{s}\left[\bar{q}^{(k+1)}(a) - q^{(k)}(a)\right]\\
v^{(k+1)}(x) &= \lambda \log \sum_a q^{(k+1)}(a) \exp\left(u(x,a)/\lambda\right)
\end{aligned}$$

3. Given $\left(q^{(k+1)}, v^{(k+1)}, p^{(k)}, \mu^{(k)}\right)$, for step size $0 < \mathsf{s} \leq 1$ construct:

$$\begin{aligned}
\bar{p}^{(k+1)}(a|x) =& \frac{q^{(k+1)}(a)\exp\left(u(x,a)/\lambda\right)}{\sum_b q^{(k+1)}(b)\exp\left(u(x,b)/\lambda\right)},\\
p^{(k+1)}(a \mid x) =& \; p^{(k)}(a \mid x)\\
&+ \mathsf{s}\left[\bar{p}^{(k+1)}(a \mid x) - p^{(k)}(a \mid x)\right]\\
\bar{\mu}^{(k+1)}(x) =& \frac{\exp\left(-v^{(k+1)}(x)/\xi\right)\hat{\mu}(x)}{\sum_y \exp\left(-v^{(k+1)}(y)/\xi\right)},\\
\mu^{(k+1)}(x) =& \; \mu^{(k)}(x) + \mathsf{s}\left[\bar{\mu}^{(k+1)}(x) - \mu^{(k)}(x)\right].
\end{aligned}$$

4. Iterate over integer $k \geq 0$ until convergence.

We applied this algorithm to compute the numerical solutions for some illustrations reported in the next two sections.[8]

4 Finding the good alternative

We begin with a consumer choice problem analyzed previously by Caplin et al. (2019). There is a discrete set of M possible consumption goods for the consumer to select among. Alternatively, the choices might be over different investment opportunities. An action is a choice of a good, and the value of the good depends on the underlying state. For simplicity, we make the action space and the state space identical in this example. Consumer preferences are captured by the utility function:

$$u(x, a) = \begin{cases} u_g & \text{if } x = a \\ u_b & \text{if } x \neq a \end{cases} \tag{16}$$

where $u_g > u_b$. For convenience, we parameterize $u_b = \lambda \log \bar{u}$ and $u_g = \lambda(\log \bar{u} + \log(1 + \delta))$. The numerical magnitude of $\bar{u}$ turns out to be inconsequential to the analysis.[9]

Let $\widehat{\mu}\,(x)$ be the baseline prior probability that option x yields the good prize. Without loss of generality, we order states according to $\widehat{\mu}\,(x_i) \geq \widehat{\mu}\,(x_{i+1}) \geq \widehat{\mu}(x_M) > 0$, for $i \in \{1, \dots, M - 1\}$.

The DM can learn about the state by paying a mutual information cost. The DM also has concerns about prior misspecification and seeks robust decision making that performs well when their prior μ may deviate from $\widehat{\mu}$. The decision problem can be formalized as Problem 4.

The solution to this decision problem will have a threshold whereby some actions may lie outside the consideration set. A central part of the solution is the characterization of this threshold. As we will show, introducing robustness into the analysis can expand the consideration set.

To construct a threshold, introduce:

$$\rho_k \stackrel{\text{def}}{=} \left[\sum_{i=1}^{k} [\widehat{\mu}(x_i)]^{\psi} \left(\frac{1}{\delta + k} \right) \right]^{\frac{1}{\psi}}$$

where

$$\psi \stackrel{\text{def}}{=} \frac{\xi}{\lambda + \xi}. \tag{17}$$

Notice that

$$\rho_1 < \widehat{\mu}(x_1).$$

Find the largest $k^* \leq M$ such that

$$\rho_k < \widehat{\mu}\,(x_k) \ \forall\, 1 \leq k \leq k^*.$$

[8] The Github repository for the computational code used in this paper can be found here. In particular, a user-friendly notebook can be accessed at here.

[9] It is straightforward to reinterpret this decision problem as one of model selection.

Only the $k \leq k^*$ are in the consideration set. Define:

$$\rho^* \overset{\text{def}}{=} \rho_{k^*}.$$

Proposition 8. *For the robust solution for the consumer choice problem, the choice rule is given by*

$$\begin{aligned}
p^*(a|x_k) =& \frac{1+\delta}{\delta}\left[\left(\frac{\widehat{\mu}(x_k)}{\rho^*}\right)^{\psi} - 1\right]\left(\frac{\rho^*}{\widehat{\mu}(x_k)}\right)^{\psi}, a = x_k, \\
p^*(a|x_k) =& \frac{1}{\delta}\left[\left(\frac{\widehat{\mu}(x_\ell)}{\rho^*}\right)^{\psi} - 1\right]\left(\frac{\rho^*}{\widehat{\mu}(x_k)}\right)^{\psi}, \forall a = x_\ell, k \neq \ell \leq k^*, \\
p^*(a|x_k) =& 0, \forall a = x_\ell, \ell > k^*
\end{aligned}$$

for $1 \leq k \leq k^*$, *and*

$$\begin{aligned}
p^*(a|x_k) =& \frac{1}{\delta}\left[\left(\frac{\widehat{\mu}(x_\ell)}{\rho^*}\right)^{\psi} - 1\right], \forall a = x_\ell, \ell \leq k^*, \\
p^*(a|x_k) =& 0, \forall a = x_\ell, \ell > k^*
\end{aligned}$$

for $M \geq k > k^*$.

The worst-case prior is given by

$$\mu^*(x_k) = \frac{(\rho^*)^{1-\psi}\,\widehat{\mu}(x_k)^{\psi}}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(x_i)^{\psi} + \sum_{i=k^*+1}^{M}\widehat{\mu}(x_i)}$$

for $1 \leq k \leq k^*$, *and*

$$\mu^*(x_k) = \frac{\widehat{\mu}(x_k)}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(x_i)^{\psi} + \sum_{i=k^*+1}^{M}\widehat{\mu}(x_i)}$$

for $M \geq k > k^*$.

Proposition 8 also implies a worst-case posterior distribution for the underlying states given the signal (action), and a worst-case marginal distribution for the actions. As is the case for the worst-case prior, these are not intended to depict the DM's actual beliefs.

As $\xi \to \infty$, a specification studied by Caplin et al. (2019), deviating from the baseline prior is increasingly costly and thus the worst-case prior approximates the baseline prior itself. In this case, Proposition 8 converges to the corresponding findings in Theorem 1 of Caplin et al. (2019).

The robust solution is qualitatively similar to that when $\xi = \infty$. Specifically, the consideration set is determined by a threshold strategy: the decision maker will consider only alternatives with a prior probability that exceeds an endogenously determined threshold ρ^*.

Quantitatively, increasing prior robustness concerns may enlarge the consideration set, leading the decision maker to consider more options. Formally, decreasing the penalty ξ increases the prior ambiguity concerns. We verify in the online appendix that it decreases ρ_k for each k, and thus k^* is larger. In the extreme, as

$\xi \downarrow 0$, ρ_k diminishes to zero, $k^* = M$, $\mu^*(x) = \frac{1}{M}$,

$$\begin{aligned} \mu_a^*(x) &\rightarrow \frac{1+\delta}{\delta+M} \text{ if } x = a, \\ \mu_a^*(x) &\rightarrow \frac{1}{\delta+M} \text{ if } x \neq a, \end{aligned}$$

for any $x \in X$.

More generally, notice that the solution depends on ξ and λ through the construction of ψ given in (17). Thus there is a one dimensional curve in the (ξ, λ) space for which the solution remains the same. For instance, reduction in ξ can be offset by reductions in λ, the attention cost, without altering the solutions for $p^*(a \mid x)$ and $\mu^*(x)$.[10]

5 Correlated options

We now consider a static investment problem with three options and two states. Options one and two are negatively correlated while option three is constant across states and hence risk-free. The two states are equally likely under the baseline prior $\widehat{\mu}(x_1) = \widehat{\mu}(x_2) = 1/2$. Let

$$u(x_i, j) = \frac{[c(x_i, j)]^{1-\alpha} - 1}{1-\alpha}$$

for $x_i \in X$, $j \in A$, and $0 \leq \alpha < 1$. Table 1 presents the payoffs $c(x_i, j)$ for each of three options as a function of the two different states.

option \ state	x_1	x_2
1	0	2 ×r
2	r	0
3	5	5

Table 1: This table displays payoffs $c(x_i, j)$ for each state $x_i \in X$ and each option $j \in A$. We will consider different values of $\mathsf{r} > 0$ in the computations that follow.

Notice that option one has the highest expected payoff and option two has the lowest under the baseline prior when $\mathsf{r} < 10$. We initially set $\alpha = 0$, implying risk neutrality. Consider first the special case in which $\lambda = \infty$, so that there is only prior ambiguity but no possibility for the DM to look for information about the underlying states. Randomization across actions independent of states is still allowed as the implied mutual information will be zero. In this case, we simply write $p(j|x_i) = p(j)$ for any $j \in A$ and $x_i \in X$.

Consider first the case in which $\mathsf{r} = 7$ as depicted in the left side of Figure 1, and $\mathsf{r} = 7.5$ on the right side of this same figure. In the absence of prior ambiguity ($\xi = \infty$), action one maximizes expected utility in both cases. For $\mathsf{r} = 7$, when ξ is twenty-five or less, the DM randomizes across actions one and three. Option two

[10] This explicit link between and ξ and λ is reminiscent of on observationally equivalent discussion in Kasa (2006) for a class of linear, Gaussian models.

is never included as part of the solution. For r=7.5, randomization only occurs at a higher levels of ambiguity aversion (smaller values of ξ).

As ξ declines to zero, the randomized choices converge to a solution to the limiting $\xi = 0$ case. In the $\xi = 0$ limit, all priors over states are entertained. In this case we recover a solution that is familiar from robust statistical decision theory. We equalize the expected payoffs conditioned on the states, which restricts $p(2) = 2p(1)$. We then maximize subject to this restriction, by choosing $p(1) \in [0, 1/3]$ and setting $p(2)$ accordingly. When $\mathsf{r} = 7.5$ all such probability choices agree. Thus there are multiple choices including the one depicted in Figure 1. When $\mathsf{r} > 7.5$, the optimized solution is $p^*(1) = 1/3$; and when $\mathsf{r} < 7.5$, $p^*(1) = p^*(2) = 0$.

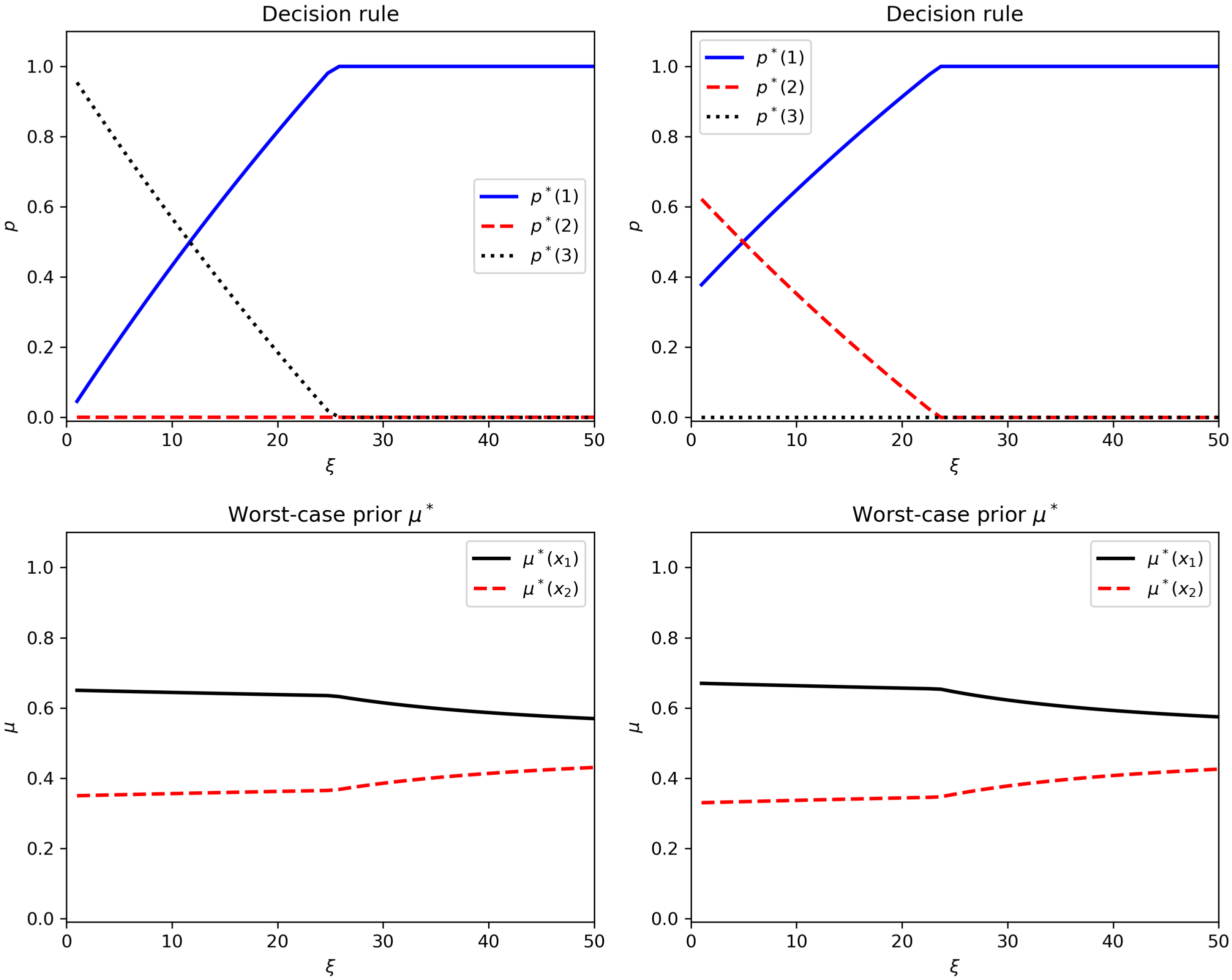

Figure 1: This figure explores the sensitivity to changes in ξ. The attention cost parameter, $\lambda = \infty$; utility curvature parameter, $\alpha = 0$; and the payoff on option two in state one is $\mathsf{r} = 7$ for the left figure and $\mathsf{r} = 7.5$ for the right.

We now explore attention allocation. We again report results for $\mathsf{r} = 7$ and set $\lambda = 10$ for illustration purposes. We report the findings in Figure 2. For lower values of ξ, option two becomes attractive. In contrast to the $\lambda = \infty$ case, this is even true for $\mathsf{r} < 7.5$. In the large ξ limiting case, the solution coincides with the rational inattention solution. In this limiting case, Caplin et al. (2019) note that option two becomes attractive because it helps the decision-maker learn about the true state of nature in contrast to option three. Robustness considerations about the prior further enhance the attractiveness of option two. This may be expected given what happens in the absence of attention considerations for $\mathsf{r} \geq 7.5$. In the presence of attention costs, increasing r, to say $\mathsf{r} = 7.5$, makes option two all the more attractive for $\lambda = 10$, as reported in the online Appendix. Option three is never chosen in this illustration.

Notice in Figure 2 that the worst-case prior is tilted towards state 1 and more so for small values of ξ (more prior ambiguity). As expected, the resulting relative entropy diminishes with ξ. Also observe that the implied mutual information is relatively stable across different values of ξ.

Remark 9. *Proposition 3 of Matêjka and McKay (2015) shows that if an action becomes more attractive according to prior beliefs, then the rationally inattentive DM will select that action with a higher unconditional probability. Prior robustness concerns provide a different perspective on this result. Without attention considerations, we see from Figure 1 that even if option one is preferred under the benchmark probabilities, this can be reversed by entertaining robustness concerns. This carries over to Figure 2 and to the* $\mathsf{r} = 7.5$ *counterpart as option one is clearly preferred under the baseline probabilities when* $\xi = \infty$*, but this gets reversed for low values of* ξ*. The Matêjka and McKay (2015) monotonicity will hold for the endogenously determined worst-case probabilities. But our DM has prior ambiguity and the worst-case probabilities are just a device to impute robust attention choices and are not intended as the actual beliefs.*

We next explore if the impact of ambiguity aversion is similar to that of risk aversion. Recall that increases in ξ are associated with less aversion to ambiguity over priors. Perhaps the most interesting comparison is between relatively large α and small ξ.

Consider the case in which $\alpha = 1$. This limit results in $u(x_i, j) = \log c(x_i, j)$. When $\mathsf{r} > 5$, we see immediately that option one is chosen only if it is known that the realized state is x_1, and option two is chosen only if it is known that the realized state is x_1. These options only happen if the attention allocation is sufficient to reveal the actual state. Otherwise, option 3 is chosen with probability one. Thus attention choice will be one of two extremes, pay no attention or pay enough attention to reveal underlying state. In what follows we explore more modest utility curvature.

Figure 3 investigates the implications for more modest values of α, while omitting prior robustness concerns. The attention cost remains at $\lambda = 10$. Option two is considerably less attractive under risk rather than aversion to prior ambiguity. Indeed, under risk aversion it is set to zero for values of α that exceed .2. At this same threshold, option three begins to be considered and it becomes more prominent for larger α. For α in excess of .5, only the risk-less option (option three) is considered. Given this, the mutual information drops to zero. This is in contrast to the risk-neutral case in Figure 2 where option three is not chosen for any of the values ξ. Overall, increasing α makes option two less attractive. Recall from Figure 2 that increases in ξ also make option two less attractive, albeit in a smoother way; but these increases imply reductions, not increases, in ambiguity aversion.

Since attention costs are depicted in utility units, changing α impacts how we should view the magnitude of the attention costs. For a different perspective on the impact of risk aversion, we also report results where attention is imposed as a constraint, $\mathbb{I} \leq .1$, in Figure 4. The quantitative magnitudes are quite different, as should be expected since the attention allocation in Figure 3 dropped to zero for the larger choices of α. Now option one remains part of the optimized solution even for larger values of α. Option two is a little bit more prominent than in the fixed λ case. Recall that probabilistic choice of signals and actions are explicitly linked in our computation of a solution as describe in Section 1.1.2.

To summarize our findings, we find important differences between the impact of risk and that of ambiguity aversion (preference for prior robustness). When the decision-maker becomes more risk averse, the riskless option three comes to dominate, as we might expect. With a fixed attention constraint, option one remains as part of the solution when we impose an information constraint in place of a cost. In contrast, when the decision-maker becomes more averse to ambiguity, depending on the payoff disparity, option two can remain attractive in contrast to options three. This outcome depends on the magnitude of the payoff disparity. By including attention choice subject to cost, option two remains even more attractive due to its role in conveying information.

We have two broad lessons to take away from this example.

i) While in the previous rational inattention literature, the impact of information revelation on choices, this example illustrates how robustness considerations amplify that role.

ii) This example captures illustrates an aspect of Frank Knight's Knight (1921) assertion that

> **Uncertainty** must be taken in a sense radically distinct from the familiar notion of **risk**, ... and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two is really present and operating.

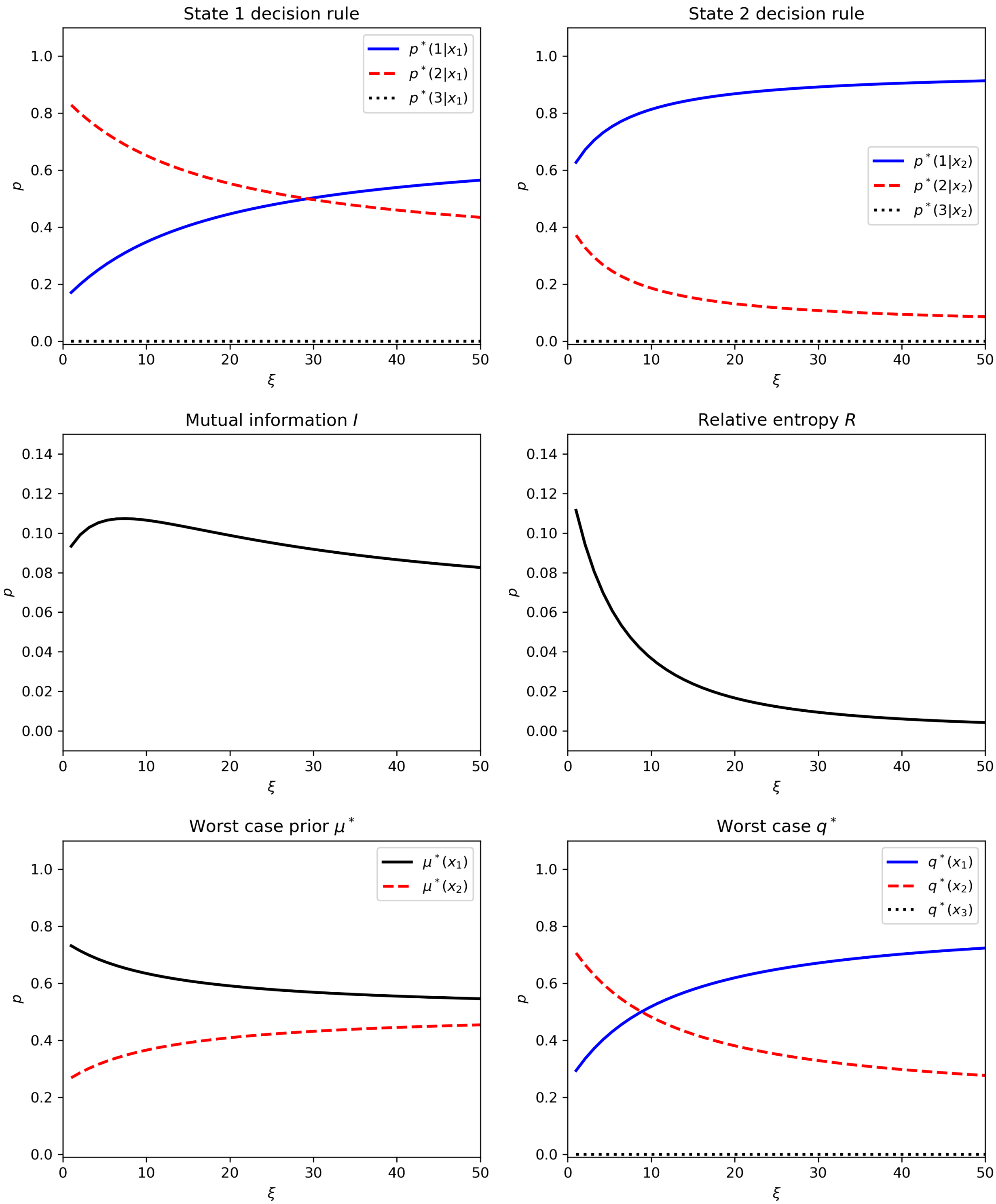

Figure 2: This figure explores the sensitivity to changes in ξ. The attention cost parameter, $\lambda = 10$; utility curvature parameter, $\alpha = 0$, and $\mathsf{r} = 7$.

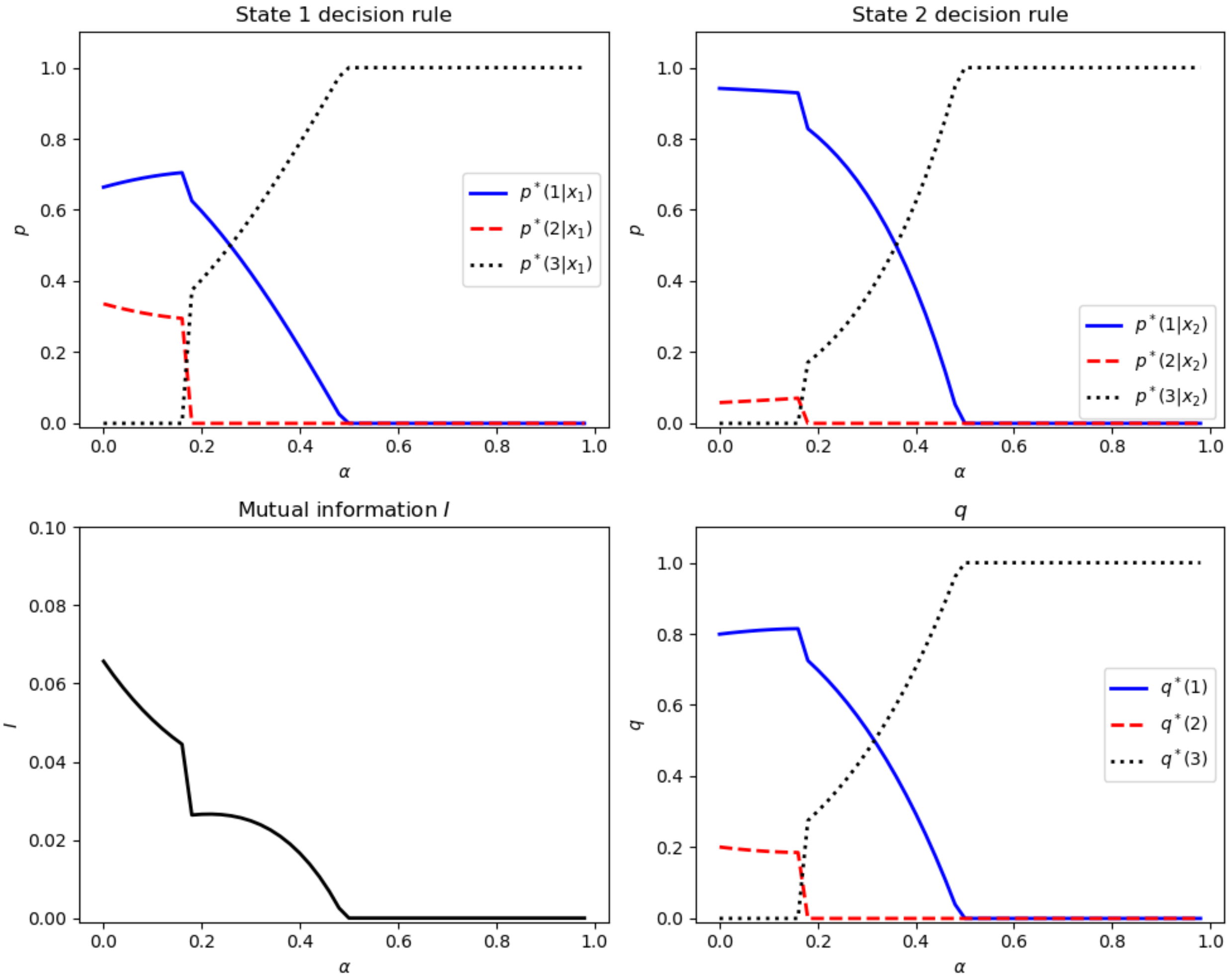

Figure 3: This figure explores the sensitivity to changes in α. The attention cost parameter, $\lambda = 10$; robustness parameter, $\xi = \infty$, and $\mathsf{r} = 7.0$

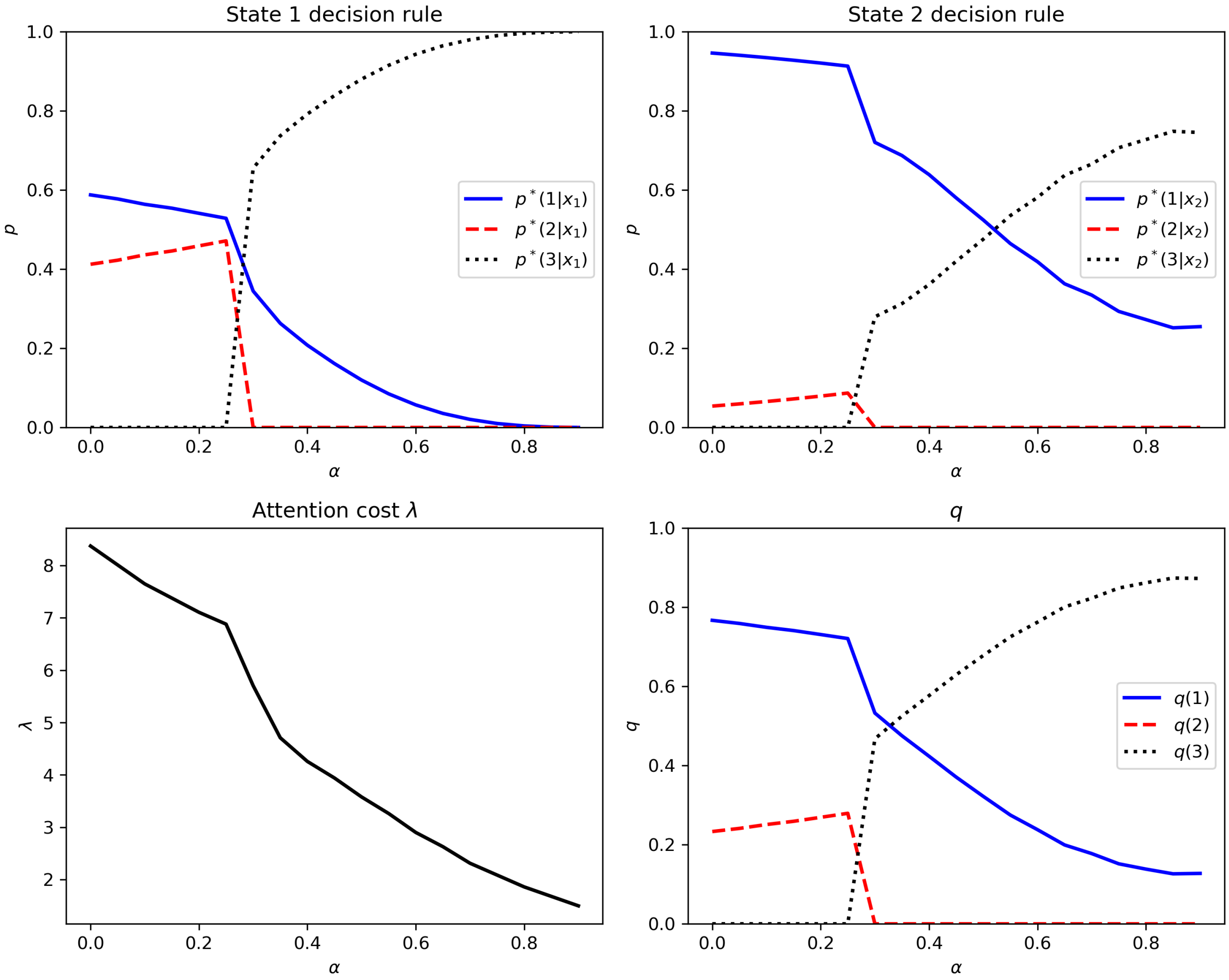

Figure 4: This figure explores the sensitivity to changes in α. The mutual information is constrained by $\mathbb{I} \leq \kappa = 0.1$; $\xi = \infty$ and $\mathsf{r} = 7.0$

6 Conclusion

Our examples are only meant to be illustrative. Many applications would seem to call for a large number of potential states. Our proposed computational algorithm could still be analyzed even with many more states. Recall that the robustly, optimal rational inattention solution will have the same number of signals as decision possibilities, which can be exploited in computation. Prior ambiguity may well be even more prominent for larger state spaces as it may be even more difficult to specify the priors over such spaces with full confidence.

While many contributions in the existing literature embrace mutual information to model attention cost, some of its behavioral implications are inconsistent with experimental evidence (see, e.g., Woodford (2012), Caplin and Dean (2013), Dean and Neligh (2020), and Dean and Neligh (2023)). Motivated by this evidence, more flexible cost functions for information acquisition are proposed by Caplin and Dean (2013), Caplin et al. (2022), Hébert and Woodford (2021), Pomatto et al. (2023), and Bloedel and Zhong (2021). These cost functions continue to depend on the DM's prior beliefs and so a robustness analysis over the priors, as we have illustrated here, continues to be a potentially important ingredient to incorporate.

In empirical or experimental studies, the observable data are state-dependent choice probabilities. To test theories of discrete choices, one has to make some additional assumptions. Caplin and Martin (2015), Caplin and Dean (2015), Caplin et al. (2022) make an important assumption that the DM and the econometrician share the same exogenously specified subjective prior over states of the world. Under these assumptions, there are unique posteriors over states which can be used when testing implications. Specifically, they construct tests based on the ratio of posteriors conditioned on different actions. In addition, Caplin and Dean (2015) point to two axioms, no improving attention cycles (NIAC) and no improving action switches (NIAS), as necessary and sufficient conditions for a costly information acquisition representation of state-dependent probabilities. In our setting, prior ambiguity implies posterior ambiguity, which brings into question many approaches to test the decision-making model. Section 2 of the Online Appendix shows that NIAC and NIAS can either hold or be violated in the prior ambiguity case. It remains an interesting challenge as to how best to test robust versions of the inattention model in the presence of ambiguity.

Here we consider only the potential misspecification of priors over states, but not of the chosen signal distributions given the state. (For instance, see Hansen and Sargent (2023) for an elaboration of potentially distinct forms of prior robustness and likelihood robustness with connections to decision theory, control theory and statistics.) When a decision maker chooses to allocate attention there may well be uncertainty about the resulting $p(a \mid x)$. This, too, would be a valuable extension of rational inattention decision theory. While we adopt a static perspective in this paper, we suspect that many applications would be better suited for a dynamic or recursive extension to the formulation we investigate in this paper. This could open an additional motivation for robust information acquisition that would operate through continuation values since "tomorrow's prior" is "today's posterior" in a dynamic setting.

In other settings, the analog to our robustness perspective results in a representation of preferences that is a special case of smooth ambiguity aversion. Within this context, we find this "special case" to be particularly interesting because of it by providing a formalization of prior robustness pertinent to robust Bayesian analysis. In this paper, our treatment of prior ambiguity extends to construction of the information cost, which would not be the case had we followed Denti et al. (2022). Denti et al. (2022) replace the attention cost used by Matêjka and McKay (2015) and others with a measure of the cost of statistical experimentation, arguing that the latter should not depend on the decision-maker prior. For instance, in the special case of information based on relative entropy, Denti et al. (2022) consider a reference probability, say uniform, that is distinct from the decision-maker prior. Both probabilities show up as part of the solution the decision-makers optimization problem, and there remains a potentially interesting scope for assessing prior robustness. For an entirely different formulation than ours of a model smooth ambiguity and limited attention see Fabbri (2024). This paper adopts a different starting point and presumes that the ambiguity adjustment in Fabbri (2024) applies to the utility contribution only.

References

Arimoto, S. 1972. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory* 18 (1):14–20.

Blackwell, D. 1951. Comparison of Experiments. *Berkley Symp on Math. Stat. and Prob* 2:93–101.

Blahut, R. 1972. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory* 18 (4):460–473.

Bloedel, Alex and Weijie Zhong. 2021. The cost of optimally-acquired information. Working Paper.

Brooks, Benjamin, Songzi Du, and Alexander Haberman. 2004. Robust Predictions with Bounded Information. Working paper.

Caplin, Andrew and Mark Dean. 2013. Behavioral Implications of Rational Inattention with Shannon Entropy. NBER working paper 19318.

———. 2015. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105:2183–2203.

Caplin, Andrew and Daniel Martin. 2015. A Testable Theory of Imperfect Perception. *Economic Journal* 125 (582):184–202.

Caplin, Andrew, Mark Dean, and John Leahy. 2019. Rational inattention, optimal consideration sets, and stochastic choice. *Review of Economic Studies* 86:1061–1094.

———. 2022. Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy. *Journal of Political Economy* 130:1676–1715.

Cerreia-Vioglio, S, F Maccheroni, M Marinacci, and L Montrucchio. 2013. Ambiguity and Robust Statistics. *Journal of Economic Theory* 148:974–1049.

Cover, Thomas M. and Joy A. Thomas. 2012. *Elements of Information Theory*. Wiley.

Dean, Mark and Nathaniel Neligh. 2020. Estimating information cost functions in models of rational inattention. *Journal of Economic Theory* 187:1–32.

———. 2023. Experimental tests of rational inattention. *Journal of Political Economy* 131 (12):3415–3461.

Denti, Tommaso and Luciano Pomatto. 2022. Model and Predictive Uncertainty: A Foundation for Smooth Ambiguity Preferences. *Econometrica* 90 (2):551–584.

Denti, Tommaso, Massimo Marinacci, and Aldo Rustichini. 2022. Experimental Cost of Information. *American Economic Review* 112:3106–3123.

Donsker, Monroe D. and S.R. Srinivasa Varadhan. 1975. Asymptotic Evaluation of Certain Markov Process Expectations for Large Time, I-IV. *Communications on Pure and Applied Mathematics* 28 (1):1–47.

Dupuis, Paul and Richard Ellis. 1997. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley and Sons, Inc.

Fabbri, Francesco. 2024. Rational Inattention with Ambiguity Aversion. Working paper.

Gilboa, Itzhak and David Schmeidler. 1989. Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics* 18:141–153.

Hansen, Lars Peter and Jianjun Miao. 2018. Aversion to ambiguity and model misspecification in dynamic stochastic environments. *Proceedings of the National Academy of Sciences* 115:9163–9168.

Hansen, Lars Peter and Thomas J. Sargent. 2001. Robust Control and Model Uncertainty. *The American Economic Review* 91:60–66.

———. 2007. Recursive Robust Estimation and Control without Commitment. *Journal of Economic Theory* 136:1–27.

———. 2023. Risk, ambiguity, and misspecification: Decision theory, robust control, and statistics. *Journal of Applied Econometrics* online (n/a):1–31.

Hébert, Benjamin and Michael Woodford. 2021. Neighborhood-based information costs. *American Economic Review* 111 (10):3225–3255.

Kasa, Kenneth. 2006. Robustness and information processing. *Review of Economic Dynamics* 9 (1):1–33.

Klibanoff, Peter, Massimo Marinacci, and Sujoy Mukerji. 2005. A Smooth Model of Decision Making Under Uncertainty. *Econometrica* 73:1849–1892.

Knight, F H. 1921. *Risk, Uncertainty, and Profit.* Houghton Mifflin.

Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini. 2006. Ambiguity Aversion, Robustness, and the Variational Representation of Preferences. *Econometrica* 74:1147–1498.

Matêjka, Filip and Alisdair McKay. 2015. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105:272–298.

Miao, Jianjun and Hao Xing. 2024. Dynamic discrete choice under rational inattention. *Economic Theory* 77:597–652.

Petersen, Ian R., Matthew R. James, and Paul Dupuis. 2000. Minimax Optimal Control of Stochastic Uncertain Systems with Relative Entropy Constraints. *IEEE Transactions on Automatic Control* 45:398–412.

Pomatto, Luciano, Philipp Strack, and Omer Tamuz. 2023. The cost of information: the case of constant marginal costs. *American Economic Review* 113 (5):1360–1393.

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

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

———. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50:665–690.

Strzalecki, Tomasz. 2011. Axiomatic Foundations of Multiplier Preferences. *Econometrica* 79:47–73.

Woodford, Michael. 2012. Inattentive valuation and reference-dependent choice. Working Paper.

A Proofs

A.1 A recommendation lemma

We first present the following recommendation lemma similar to Matêjka and McKay (2015).

Lemma 10. *Let (μ^*, d^*, σ^*) solve Problem 2. Then (d^*, σ^*) generates a choice rule p such that (μ^*, p) solves Problem 4. Conversely, let (μ^*, p^*) solve Problem 4. Then p^* induces a strategy (d, σ) such that (μ^*, d, σ) solves Problem 2. Moreover, $W(\widehat{\mu}) = J(\widehat{\mu})$.*

Proof. Use the Minimax Theorem to exchange the extremization in [1] and [5]. Because the relative entropy penalty term $R(\mu||\widehat{\mu})$ is the same in these two problems, it suffices to prove

$$\max_{p\in\Delta(A|X)} \mathbb{E}_{p\otimes\mu}\left[u(\mathbf{x},\mathbf{a})\right] - \lambda\mathbb{I}(p\otimes\mu) = \max_{(d,\sigma)\in\Sigma} \mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] - \lambda\mathbb{I}(d\otimes\mu), \tag{18}$$

for any fixed prior $\mu \in \Delta(X)$.

Given any strategy $(d, \sigma) \in \Sigma$, we can construct a choice rule p as in [2] and hence define $\mathbb{I}(p\otimes\mu)$. We will prove that $\mathbb{I}(p\otimes\mu) \leq \mathbb{I}(d\otimes\mu)$. This statement is equivalent to

$$\sum_a q(a)\,\mathbb{H}(\mu_a) \geq \sum_s \nu(s)\,\mathbb{H}(\mu_s),$$

where $\nu(s)$ is the marginal distribution over signals. Since $a = \sigma(s)$, we have

$$\mu_a(x) = \sum_s \mu_s(x)\Pr(s|a),\ \ x \in X.$$

Since Shannon entropy $\mathbb{H} : \Delta(X) \to \mathbb{R}$ is a concave function, it follows from Jensen's inequality that

$$\mathbb{H}(\mu_a) \geq \sum_s \Pr(s|a)\,\mathbb{H}(\mu_s).$$

Multiplying both sides by $q(a)$ and summing over a, we obtain

$$\sum_a q(a)\,\mathbb{H}(\mu_a) \geq \sum_s\sum_a \Pr(s|a)q(a)\,\mathbb{H}(\mu_s) = \sum_s \nu(s)\,\mathbb{H}(\mu_s),$$

as desired. Since $\mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] = \mathbb{E}_{p\otimes\mu}\left[u\left(\mathbf{x},\mathbf{a}\right)\right]$ by construction, we have

$$\mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] - \lambda\mathbb{I}(d\otimes\mu) \leq\ \mathbb{E}_{p\otimes\mu}\left[u\left(\mathbf{x},\mathbf{a}\right)\right] - \lambda\mathbb{I}(p\otimes\mu).$$

Let the strategy (d^*, σ^*) achieve the maximum of the problem on the right-hand side of (18). Let p be the induced choice rule. We then have

$$\max_{(d,\sigma)\in\Sigma} \mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] - \lambda\mathbb{I}(d\otimes\mu) \leq\ \mathbb{E}_{p\otimes\mu}\left[u\left(\mathbf{x},\mathbf{a}\right)\right] - \lambda\mathbb{I}(p\otimes\mu). \tag{19}$$

Conversely, given any choice rule p, we can construct a strategy (d, σ). Specifically, let S be any finite set with $|S| = |A|$ and fix any bijection $\phi : A \to S$. Define

$$d(s|x) = p(a|x),\, \sigma(s) = a, \qquad \text{for } s = \phi(a).$$

This construction implies

$$\mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] - \lambda\mathbb{I}(d\otimes\mu) = \mathbb{E}_{p\otimes\mu}\left[u\left(\mathbf{x},\mathbf{a}\right)\right] - \lambda\mathbb{I}(p\otimes\mu).$$

Let p^* achieve the maximum of the problem on the left-hand side of (18) and (d,σ) be the induced strategy. Then we have

$$\max_{p\in\Delta(A|X)} \mathbb{E}_{p\otimes\mu}\left[u\left(\mathbf{x},\mathbf{a}\right)\right] - \lambda\mathbb{I}(p\otimes\mu) = \mathbb{E}_{d\otimes\mu}\left[u\left(\mathbf{x},\sigma\left(\mathbf{s}\right)\right)\right] - \lambda\mathbb{I}(d\otimes\mu) \tag{20}$$

Combining (19) and (20), we obtain the desired result. □

A.2 Proof of Proposition 6

It is straightforward to check that the objective function F is convex in μ and jointly concave in p and q. We first solve the inner maximization problem of [8] for a fixed μ. This inner maximization problem is a standard rational inattention problem. The optimal choice probabilities

$$p^*(a \mid x) = \frac{q(a)\exp\left[u(x,a)/\lambda\right]}{\sum_b q(b)\exp\left[u(x,b)/\lambda\right]}, \qquad \mu(x) > 0, \tag{21}$$

and the resulting optimal value in [14] are obtained by Matêjka and McKay (2015). Caplin and Dean (2013) and Caplin et al. (2019) characterize the necessary and sufficient conditions which consistent of (21) and

$$\sum_x \mu(x)\left(\frac{\exp\left[u(x,a)/\lambda\right]}{\sum_b q^*(b)\exp\left[u(x,b)/\lambda\right]}\right)\begin{cases}\leq 1 & \forall a,\\ = 1 & a\in B(q^*).\end{cases} \tag{22}$$

They argue that these conditions are important for identifying the consideration set. They also show that the value function of the inner rational inattention problem is

$$V(\mu) = \sum_x \mu(x)v(x), \quad v(x) = \lambda\log\sum_a q^*(a)\exp\big(u(x,a)/\lambda\big), \tag{23}$$

and V is convex and satisfies

$$\frac{\partial V(\mu)}{\partial\mu(x)} = v(x) - v(x_M),\;\; x = x_1, x_2, ..., x_{M-1}. \tag{24}$$

Now we consider the outer minimization problem of [8], which can be written as

$$\min_{\mu\in\Delta(X)} V\left(\mu\right) + \xi\sum_x \mu\left(x\right)\log\frac{\mu\left(x\right)}{\widehat{\mu}\left(x\right)}. \tag{25}$$

Since V and the relative entropy are convex in μ, this is a convex optimization problem. Replace $\mu(x_M)$ by

$$\mu(x_M) = 1 - \sum_{i=1}^{M-1} \mu(x_i). \tag{26}$$

Using (24) and taking first-order conditions with respect to $\mu(x)$ yields

$$v(x) - v(x_M) + \xi \left[\log \frac{\mu^*(x)}{\widehat{\mu}(x)} - \log \frac{\mu^*(x_M)}{\widehat{\mu}(x_M)} \right] = 0, \tag{27}$$

for $x = x_1, ..., x_{M-1}$. Solving for $\mu(x)$, summing over $x = x_1, x_2, ..., x_{M-1}$, and using (26), we can derive

$$\mu^*(x_M) = \frac{\exp(-v(x_M)/\xi)\,\widehat{\mu}(x_M)}{\sum_{x'} \exp(-v(x')/\xi)\,\widehat{\mu}(x')}.$$

Plugging this expression into (27), we can derive

$$\mu^*(x) = \frac{\exp(-v(x)/\xi)\,\widehat{\mu}(x)}{\sum_{x'} \exp(-v(x')/\xi)\,\widehat{\mu}(x')}, \quad x = x_1, ..., x_{M-1}.$$

We then obtain [9]. Replacing μ with μ^* in (21) and (22) yields [13] and [15]. Plugging μ^* into (25) and using (23), we can derive the value function $J(\widehat{\mu})$ in the proposition. □

A.3 Proof of Proposition 8:

Notice that we set $A = X = \{x_1, x_2, ..., x_M\}$. Choosing an action $a \in A$ is the same as choosing some state x_k.

It follows from (23) and [16] that

$$v(x) = \lambda \log \sum_a q^*(a) \exp(u(x,a)/\lambda) = \lambda \log(\bar{u}(1 + \delta q^*(x))).$$

Plugging the above expression into [9], we obtain

$$\mu^*(x) = \frac{(1 + \delta q^*(x))^{-\frac{\lambda}{\xi}} \widehat{\mu}(x)}{\sum_y (1 + \delta q^*(y))^{-\frac{\lambda}{\xi}} \widehat{\mu}(y)}. \tag{28}$$

Let B be the consideration set. For any chosen action $a \in B$, the equality in [15] implies

$$\begin{aligned}
1 =& \sum_x \frac{\mu^*(x) \exp[u(x,a)/\lambda]}{\sum_b q^*(b) \exp[u(x,b)/\lambda]} \\
=& \frac{(1+\delta)\mu^*(a)}{1 + \delta q^*(a)} + \sum_{a' \in B \setminus \{a\}} \frac{\mu^*(a')}{1 + \delta q^*(a')} + \sum_{b \in A \setminus B} \mu^*(b) \\
=& \frac{\delta \mu^*(a)}{1 + \delta q^*(a)} + \sum_{a' \in B} \frac{\mu^*(a')}{1 + \delta q^*(a')} + \sum_{b \in A \setminus B} \mu^*(b),
\end{aligned} \tag{29}$$

where the second equality follows from $\sum_b q^*(b) \exp[u(x,b)/\lambda] = \sum_{b \neq x} q^*(b)\bar{u} + q^*(x)\bar{u}(1+\delta) = \bar{u}[1 + \delta q^*(x)]$

for $x \in B$; and $\sum_b q^*(b)\exp[u(x,b)/\lambda] = \sum_{b\in B} q^*(b)\bar{u} = \bar{u}$ for $x \notin B$. Observe that the last two terms on the right-hand side of (29) are independent of a. Therefore the first term $\frac{\delta\mu^*(a)}{1+\delta q^*(a)}$ must be identical for any chosen action a. Using (28) to replace $\mu^*(a)$ and noticing that the denominator of the term on the right-hand side of (28) is independent of a, we deduce that $(1+\delta q^*(a))^{-\frac{\lambda}{\xi}-1}\widehat{\mu}(a)$ is identical for any $a \in B$. Therefore, we denote

$$\rho := [1+\delta q^*(a)]^{-\frac{\lambda}{\xi}-1}\,\widehat{\mu}(a), \qquad \text{for any } a \in B. \tag{30}$$

Recall $\psi = \frac{\xi}{\lambda+\xi}$ from [17]. It then follows from (30) and (28) that

$$1+\delta q^*(a) = \left[\frac{\rho}{\widehat{\mu}(a)}\right]^{-\psi} \quad \text{and} \quad \mu^*(a) = \frac{\rho^{1-\psi}\widehat{\mu}(a)^{\psi}}{\sum_x (1+\delta q^*(x))^{-\frac{\lambda}{\xi}}\widehat{\mu}(x)}. \tag{31}$$

Combining (28), (29), and (30), we obtain

$$\frac{\rho(\delta+|B|)}{\sum_x (1+\delta q^*(x))^{-\frac{\lambda}{\xi}}\widehat{\mu}(x)} = 1 - \sum_{b\in A\setminus B}\mu^*(b) = \sum_{a\in B}\mu^*(a). \tag{32}$$

Then combining (31) and (32) yields

$$(\delta+|B|)\rho^{\psi} = \sum_{a\in B}\widehat{\mu}(a)^{\psi}. \tag{33}$$

Equation (30) implies $\rho < \widehat{\mu}(a)$ for any chosen action $a \in B$ with $q^*(a) > 0$. It follows from (33) that

$$\widehat{\mu}(a) > \rho = \left[\frac{\sum_{a\in B}\widehat{\mu}(a)^{\psi}}{\delta+|B|}\right]^{\frac{1}{\psi}}, \qquad \text{for any } a \in B. \tag{34}$$

For $a \notin B$, the inequality in [15] yields

$$\begin{aligned} 1 &\geq \sum_x \frac{\mu^*(x)\exp(u(x,a)/\lambda)}{\sum_b q^*(b)\exp(u(x,b)/\lambda)} \\ &= \sum_{a'\in B}\frac{\mu^*(a')}{1+\delta q^*(a')} + \sum_{b\in A\setminus B}\mu^*(b) + \delta\mu^*(a), \end{aligned}$$

which is equivalent to

$$\delta\mu^*(a) \leq \sum_{a'\in B}\left[\mu^*(a') - \frac{\mu^*(a')}{1+\delta q^*(a')}\right] = \sum_{a'\in B}\frac{\delta q^*(a')}{1+\delta q^*(a')}\mu^*(a'). \tag{35}$$

Because a is not chosen, $q^*(a) = 0$. Then (28) implies that

$$\mu^*(a) = \frac{\widehat{\mu}(a)}{\sum_x (1+\delta q^*(x))^{-\frac{\lambda}{\xi}}\widehat{\mu}(x)}, \qquad \text{for any } a \notin B. \tag{36}$$

Plugging the previous equation and (28) into (35), we obtain

$$\widehat{\mu}(a) \leq \sum_{a'\in B} q^*(a')\,[1+\delta q^*(a')]^{-\frac{\lambda}{\xi}-1}\,\widehat{\mu}(a') = \sum_{a'\in B} q^*(a')\,\rho = \rho, \quad \text{for any } a \notin B, \tag{37}$$

where the last equality follows from $\sum_{a' \in B} q^*(a') = 1$.

Combination of (34) and (37) identifies the consideration set B by the following relation

$$\widehat{\mu}(a) > \rho, \text{ for any } a \in B; \quad \widehat{\mu}(a) \leq \rho, \text{ for any } a \notin B, \tag{38}$$

where

$$\rho = \left[\frac{\sum_{a \in B} \widehat{\mu}(a)^{\psi}}{\delta + |B|} \right]^{\frac{1}{\psi}}.$$

The relationship (38) and the monotonicity of $\widehat{\mu}(x_i)$ imply that the consideration set must be a threshold type: if $a = x_k \in B$, then $a = x_{k'} \in B$ for any $k' \leq k$.

We now identify the consideration set B. To this end, define

$$\rho_k = \left[\frac{\sum_{i=1}^{k} \widehat{\mu}(x_i)^{\psi}}{\delta + k} \right]^{\frac{1}{\psi}}.$$

Notice that $\rho_1 < \widehat{\mu}(x_1)$. Find the largest $k^* \leq M$ such that

$$\rho_k < \widehat{\mu}(x_k), \quad \forall 1 \leq k \leq k^*. \tag{39}$$

Then (38) is satisfied when $B = \{x_1, \ldots, x_{k^*}\}$ and $\rho = \rho_{k^*}$. To verify this claim, we use the monotonicity of $\widehat{\mu}(x_i)$ and (39) to obtain

$$\rho_{k^*} < \widehat{\mu}(x_{k^*}) \leq \widehat{\mu}(x_k), \quad \forall 1 \leq k \leq k^*. \tag{40}$$

Meanwhile, being the largest k^* satisfying (39) implies

$$\rho_{k^*} \geq \widehat{\mu}(x_k), \quad \forall k^* < k \leq M. \tag{41}$$

Assume otherwise, if $\rho_{k^*} < \widehat{\mu}(x_k)$ for some k satisfying $k^* < k \leq M$, then $\rho_{k^*} < \widehat{\mu}(x_{k^*+1})$, which is equivalent to

$$\left[\frac{\sum_{i=1}^{k^*} \widehat{\mu}(x_i)^{\psi}}{\delta + k^*} \right]^{\frac{1}{\psi}} < \widehat{\mu}(x_{k^*+1}).$$

Raising to the ψ-th power, multiplying by $\delta + k^*$, and adding $\widehat{\mu}(x_{k^*+1})^{\psi}$ on both sides, we obtain

$$\sum_{i=1}^{k^*+1} \widehat{\mu}(x_i)^{\psi} < (\delta + k^* + 1)\widehat{\mu}(x_{k^*+1})^{\psi},$$

which is equivalent to

$$\rho_{k^*+1} < \widehat{\mu}(x_{k^*+1}).$$

This contradicts the choice of k^*.

Next we prove that k^* is the unique index satisfying both (40) and (41). To this end, we first claim that

$$\rho_k \geq \widehat{\mu}(x_k), \quad \forall k^* < k \leq M. \tag{42}$$

To prove (42), introduce

$$\nu_k := \sum_{i=1}^{k} \widehat{\mu}(x_i)^{\psi} - (\delta + k)\widehat{\mu}(x_k)^{\psi} = \nu_k^1 + \nu_k^2$$

where

$$\nu_k^1 := \sum_{i=1}^{k} \left[\widehat{\mu}(x_i)^{\psi} - \widehat{\mu}(x_k)^{\psi}\right] \quad \text{and} \quad \nu_k^2 := -\delta\widehat{\mu}(x_k)^{\psi}.$$

Both ν_k^1 and ν_k^2 are increasing in k, implying that ν_k is as well. Notice that $\nu_1^1 = 0$ and $\nu_1^2 < 0$, and thus $\nu_1 < 0$. By the definition of k^*, k^* is the largest positive integer less than M for which ν_{k^*} is strictly negative. By the monotonicity, ν_k is necessarily negative for smaller values of k and nonnegative for larger values of k. This implies that

$$\begin{aligned} \rho_k &< \widehat{\mu}(x_k), \quad 1 \le k \le k^* \\ \rho_k &\ge \widehat{\mu}(x_k), \quad k^* < k \le M. \end{aligned}$$

Using (42), we can now show that k^* is the unique index satisfying both (40) and (41). Suppose that ℓ is another case where (40) and (41) are satisfied with k^* replaced by ℓ therein. If $\ell > k^*$, it cannot satisfy

$$\rho_\ell < \widehat{\mu}(x_\ell),$$

because it contradicts (42). If $\ell < k^*$, (39) implies $\rho_{\ell+1} < \widehat{\mu}(x_{\ell+1})$, which is equivalent to

$$\left[\frac{\sum_{i=1}^{\ell+1} \widehat{\mu}(x_i)^{\psi}}{\delta + \ell + 1}\right]^{\frac{1}{\psi}} < \widehat{\mu}(x_{\ell+1}).$$

Raising to the ψ-th power, multiplying by $\delta + \ell + 1$, and subtracting $\widehat{\mu}(x_{\ell+1})^{\psi}$ on both sides, we obtain

$$\sum_{i=1}^{\ell} \widehat{\mu}(x_i)^{\psi} < (\delta + \ell)\widehat{\mu}(x_{\ell+1})^{\psi},$$

which is equivalent to $\rho_\ell < \widehat{\mu}(x_{\ell+1})$. Therefore (41), where k^* replaced by ℓ therein, is violated.

Having shown that k^* is the unique index such that both (40) and (41) are satisfied, we use the characterization of the consideration set in (38) to conclude that $B = \{x_1, \ldots, x_{k^*}\}$ and $\rho = \rho_{k^*}$.

Using the first equation of (31), we identify q^* as

$$\begin{aligned} q^*(a) =& \frac{1}{\delta}\left[\left(\frac{\widehat{\mu}(x_k)}{\rho^*}\right)^{\psi} - 1\right], \text{ if } a = x_k, 1 \le k \le k^*, \\ q^*(a) =& 0, \qquad\qquad\qquad\quad \text{if } a = x_k, k^* < k \le M. \end{aligned} \tag{43}$$

Using (31) and (36), we identify the worst-case prior μ^* as

$$\begin{aligned} \mu^*(x_k) =& \frac{(\rho^*)^{1-\psi}\,\widehat{\mu}(x_k)^\psi}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(x_i)^\psi + \sum_{i=k^*+1}^M \widehat{\mu}(x_i)},\ 1 \le k \le k^*, \\ \mu^*(x_k) =& \frac{\widehat{\mu}(x_k)}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(x_i)^\psi + \sum_{i=k^*+1}^M \widehat{\mu}(x_i)},\ k^* < k \le M. \end{aligned} \tag{44}$$

For the optimal choice rule, let us first consider the case conditioning on x_k with $1 \le k \le k^*$. When $a = x_k$,

$$p^*(a|x_k) = \frac{q^*(a)\exp(u(x_k,a)/\lambda)}{\sum_b q^*(b)\exp(u(x_k,b)/\lambda)} = \frac{1+\delta}{\delta}\left[\Big(\frac{\widehat{\mu}(x_k)}{\rho^*}\Big)^\psi - 1\right]\Big(\frac{\rho^*}{\widehat{\mu}(x_k)}\Big)^\psi,$$

where the second equality follows from (43) and

$$\sum_b q^*(b)\exp(u(x_k,b)/\lambda) = \bar{u}\big(1+\delta q^*(x_k)\big) = \bar{u}\Big(\frac{\widehat{\mu}(x_k)}{\rho^*}\Big)^\psi. \tag{45}$$

When $a = x_\ell$ and $k \ne \ell \le k^*$, using (43) and (45), we obtain

$$p^*(a|x_k) = \frac{1}{\delta}\left[\Big(\frac{\widehat{\mu}(x_\ell)}{\rho^*}\Big)^\psi - 1\right]\Big(\frac{\rho^*}{\widehat{\mu}(x_k)}\Big)^\psi.$$

When $a = x_\ell$ and $\ell > k^*$, it follows from $q^*(a) = 0$ that $p^*(a|x_k) = 0$.

Now consider the case conditioning on x_k with $k^* < k \le M$. When $a = x_\ell$ and $\ell \le k^*$,

$$p^*(a|x_k) = \frac{q^*(a)\exp(u(x_k,a)/\lambda)}{\sum_b q^*(b)\exp(u(x_k,b)/\lambda)} = \frac{1}{\delta}\left[\Big(\frac{\widehat{\mu}(x_k)}{\rho^*}\Big)^\psi - 1\right],$$

where the second equality follows from (43) and $\sum_b q^*(b)\exp(u(x_k,b)/\lambda) = \bar{u}$. When $a = x_\ell$ and $\ell > k^*$, $p^*(a|x_k) = 0$.

The optimal posterior can also be determined. When $a = x_k$ and $k \le k^*$,

$$\mu_a^*(x_k) = \frac{\exp(u(x_k,a)/\lambda)\mu^*(x_k)}{\sum_b q^*(b)\exp(u(x,b)/\lambda)} = \frac{(1+\delta)\mu^*(x_k)}{1+\delta q^*(a)} = \frac{(1+\delta)\rho^*}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(x_i)^\psi + \sum_{i=k^*+1}^M \widehat{\mu}(x_i)}.$$

When $a = x_\ell$ and $k \ne \ell \le k^*$,

$$\mu_a^*(x_k) = \frac{\mu^*(x_k)}{1+\delta q^*(a)} = \frac{\rho^*}{(\rho^*)^{1-\psi}\sum_{i=1}^{k^*}\widehat{\mu}(i)^\psi + \sum_{i=k^*+1}^M \widehat{\mu}(x_i)}.$$

When $k > k^*$, $a = x_\ell$, and $\ell \le k^*$, $\mu^*(x_k \,|\, a) = \mu^*(x_k)$.

When $\xi \downarrow 0$, then $\psi \downarrow 0$ and ρ_k converges to zero for any k. Equation (44) implies that $\mu^*(x_k)$ converges to $1/M$ for any $1 \le k \le M$. Moreover, $q^*(a)$ converges to $1/M$ and $\mu_a^*(x)$ converges to $\frac{1+\delta}{\delta+M}$ when $x = a$, or $\frac{1}{\delta+M}$ when $x \ne a$. The proof is completed. □

Introduce

$$\rho_k(\psi) = \left[\frac{\sum_{i=1}^k \widehat{\mu}(x_i)^\psi}{\delta+k}\right]^{\frac{1}{\psi}}.$$

Because ψ increases with ξ, the following result implies that $\rho_k(\psi)$ increases with ξ.

Lemma 11. *For any fixed k, $\rho_k(\psi)$ increases with ψ.*

Proof. Consider $\psi < \psi'$,

$$\rho_k(\psi) = \left[\left(\frac{\sum_{i=1}^k \widehat{\mu}(x_i)^\psi}{\delta + k}\right)^{\frac{\psi'}{\psi}}\right]^{\frac{1}{\psi'}} = \left[\left(\frac{k}{\delta + k}\right)^{\frac{\psi'}{\psi}} \left(\frac{1}{k}\sum_{i=1}^k \widehat{\mu}(x_i)^\psi\right)^{\frac{\psi'}{\psi}}\right]^{\frac{1}{\psi'}}$$
$$< \left[\frac{k}{\delta + k}\left(\frac{1}{k}\sum_{i=1}^k \widehat{\mu}(x_i)^\psi\right)^{\frac{\psi'}{\psi}}\right]^{\frac{1}{\psi'}} < \left[\frac{k}{\delta + k}\left(\frac{1}{k}\sum_{i=1}^k \widehat{\mu}(x_i)^{\psi'}\right)\right]^{\frac{1}{\psi'}} = \rho_k(\psi'),$$

where the first inequality follows from $\frac{k}{\delta+k} < 1$ and $\frac{\psi'}{\psi} > 1$, the second inequality follows from the Jensen's inequality. □

B NIAC and NIAS

When DM has no prior robustness concern, Caplin and Dean (2015) show that for given prior $\widehat{\mu}$ and utility function u, a dataset of state dependent choices has a costly information acquisition representation if and only if it satisfies *no improving attention cycles* (NIAC) and *no improving action switches* (NIAS). In this section, we will construct examples to show that both NIAC and NIAS can hold or be violated in the prior robustness setting. The following examples show that both NIAC and NIAS can be violated when the payoff disparity among state-action pairs are sufficiently severe. The payoff disparity is amplified by the exponential tilting in the worst-case prior, which leads to violation of NIAC and NIAS.

For two decision problems with two states $X = \{x_1, x_2\}$ and two actions $A^i = \{a^i, b^i\}$ for $i = 1, 2$ with $u(a^i, x_1) > u(b^i, x_1)$ and $u(b^i, x_2) > u(a^i, x_2)$, the NIAC condition is

$$\begin{aligned} &\widehat{\mu}(x_1)\Big[p(a^1|x_1) - p(a^2|x_1)\Big]\Big[\big(u(a^1, x_1) - u(b^1, x_1)\big) - \big(u(a^2, x_1) - u(b^2, x_1)\big)\Big] \\ &+\widehat{\mu}(x_2)\Big[p(b^1|x_2) - p(b^2|x_2)\Big]\Big[\big(u(b^1, x_2) - u(a^1, x_2)\big) - \big(u(b^2, x_2) - u(a^2, x_2)\big)\Big] \geq 0. \end{aligned} \tag{46}$$

For any decision problem with two states $X = \{x_1, x_2\}$ and two actions $A = \{a, b\}$, the NIAS condition is

$$\begin{aligned} \widehat{\mu}(x_1)p(a|x_1)\big(u(a, x_1) - u(b, x_1)\big) + \widehat{\mu}(x_2)p(a|x_2)\big(u(a, x_2) - u(b, x_2)\big) \geq 0, \\ \widehat{\mu}(x_1)p(b|x_1)\big(u(b, x_1) - u(a, x_1)\big) + \widehat{\mu}(x_2)p(b|x_2)\big(u(b, x_2) - u(a, x_2)\big) \geq 0. \end{aligned} \tag{47}$$

We now consider the following examples. Given $\epsilon, \delta > 0$, consider the risk neutral case and the payoff of state-dependent actions are given by

$$u(a^\delta, x_1) = \delta, \quad u(b^\delta, x_2) = \epsilon\delta, \quad u(a^\delta, x_2) = u(b^\delta, x_1) = 0.$$

For two decision problems indexed by δ and δ_0 respectively, the NIAC condition (46) becomes

$$\widehat{\mu}(x_1)\Big[p(a^\delta|x_1) - p(a^{\delta_0}|x_1)\Big](\delta - \delta_0) + \widehat{\mu}(x_2)\Big[p(b^\delta|x_2) - p(b^{\delta_0}|x_2)\Big]\epsilon(\delta - \delta_0) \geq 0. \tag{48}$$

In the following, we consider two examples with two different values of ϵ. For each case, we fix $\delta_0 = 1$ and consider the left-hand side of (48) as a function of δ. We will plot this function to check NIAC.

For decision problem indexed by δ, the NIAS condition (47) becomes

$$\begin{aligned} \widehat{\mu}(x_1)p(a^\delta|x_1)\delta - \widehat{\mu}(x_2)p(a^\delta|x_2)\epsilon\delta \geq 0 \\ -\widehat{\mu}(x_1)p(b^\delta|x_1)\delta + \widehat{\mu}(x_2)p(b^\delta|x_2)\epsilon\delta \geq 0. \end{aligned} \tag{49}$$

We consider the left-hand side of (49) as a function of δ. We will plot this function to check NIAS.

We first set $\epsilon = 1$, which corresponds to the problem in Section 4 of the paper. Other parameter values are listed in Figure 5. There are cases where $p(a^\delta|x_1)$ is non-monotone in δ, as illustrated in the left panel in the first row of Figure 5. The reason why $p(a^\delta|x_1)$ is non-monotone in δ is the following. When δ is high, choosing a^δ in the state x_1 and b^δ in the state x_2 generate high payoff, hence $p(a^\delta|x_1)$ and $p(b^\delta|x_2)$ are high. When δ decreases, payoff of a^δ in the state x_1 get closer to the payoff of b^δ in the same state. Hence $p(a^\delta|x_1)$ reduces, so does $p(b^\delta|x_2)$. When δ is close to zero, payoff of a^δ and b^δ are almost the same in both states. However, information is costly. In this case, DM does not acquire information to distinguish x_1 and x_2, and chooses the action is most likely in the worst-case belief. Meanwhile, as δ decreases, the exponential tilting to x_2 becomes less in the worst-case belief. DM is more likely to choose action which generates high payoff in the state x_1, and this action is a^δ.

However, NIAC holds in this case. This is becuase when $\epsilon = 1$, the sign of left-hand side of (48) depends on the monotonicity of $p(a^\delta|x_1)+p(b^\delta|x_2)$. In Figure 5, $p(a^\delta|x_1)+p(b^\delta|x_2)$ increases in δ. Even though the worst-case prior μ^* also changes with δ, NIAC holds in this case. NIAS holds as well, because $p(a^\delta|x_1) \geq p(a^\delta|x_2)$ and $p(b^\delta|x_2) \geq p(b^\delta|x_1)$.

Now we set $\epsilon = 0.1$ and other parameter values in Figure 6. The contribution from the second term on the left-hand side of (48) is much smaller. As a result, the first term dominates, and NIAC is violated in this case. NIAS is also violated for action b^δ when δ is larger than one, because the growth of $\epsilon p(b^\delta|x_2)$ is less than the growth of $p(b^\delta|x_1)$.

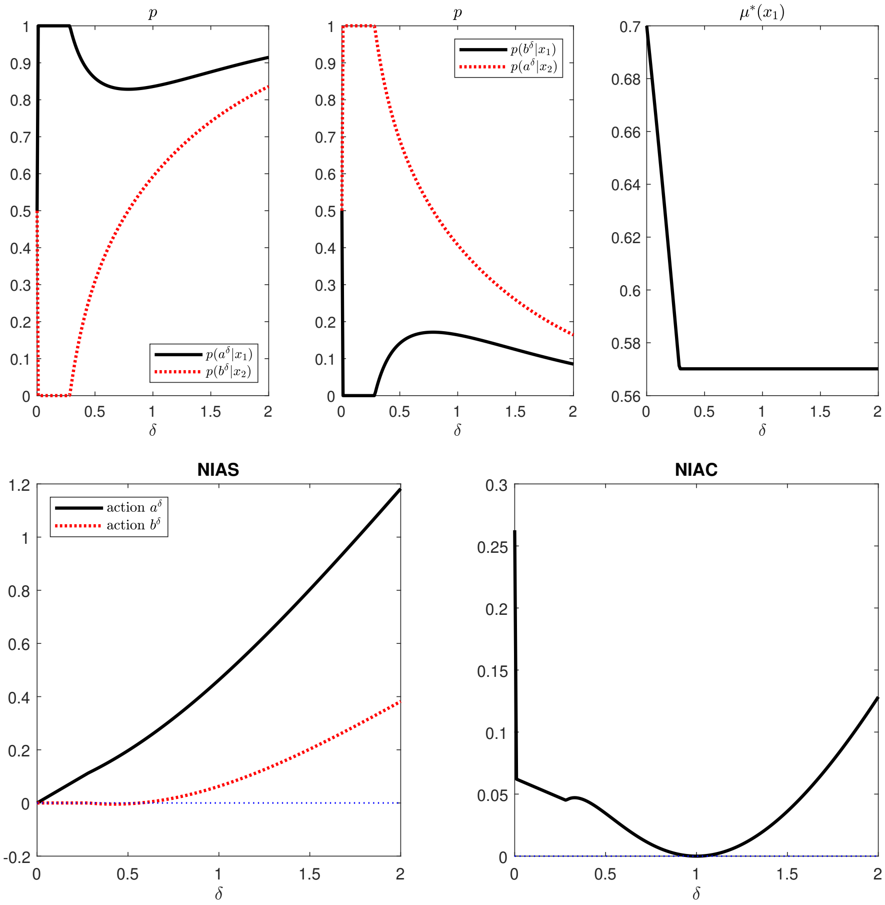

Figure 5: NIAC & NIAS hold

Parameters: $\epsilon = 1$, $\widehat{\mu}(x_1) = 0.7, \widehat{\mu}(x_2) = 0.3$, $\lambda = 1$, and $\xi = 0.5$. The left lower panel plots the left-hand side of (49) as a function of δ, the right lower panel plots the left-hand side of (48) as a function of δ.

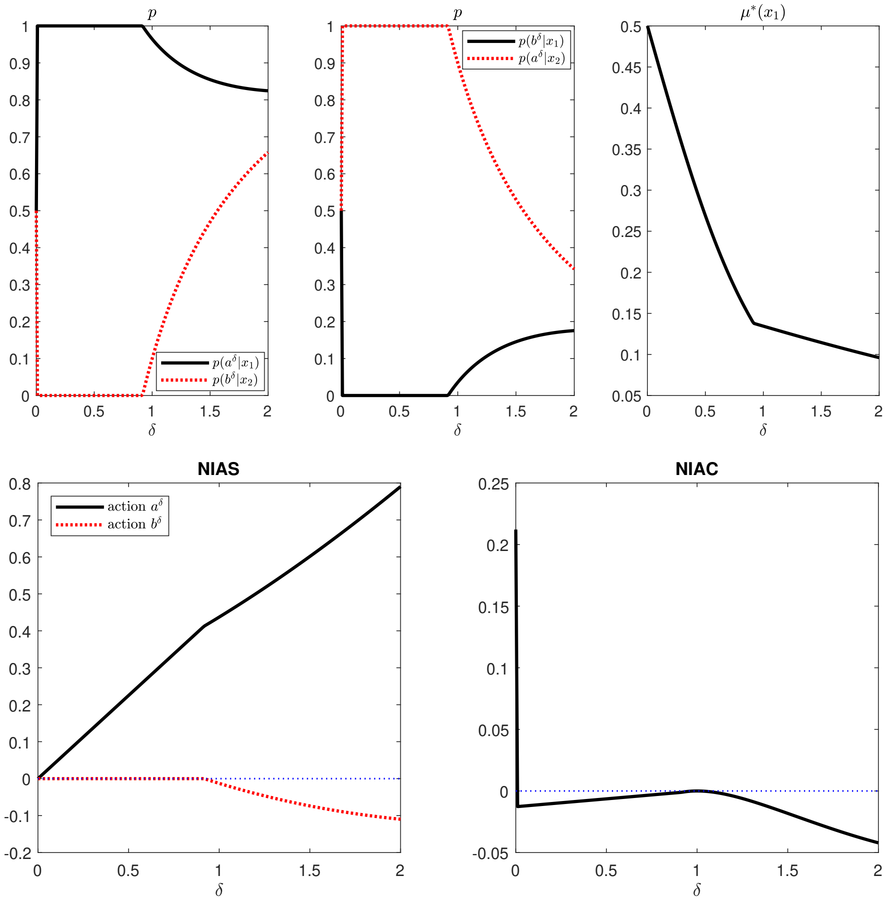

Figure 6: NIAC & NIAS are violated

Parameters: $\epsilon = 0.1$, $\widehat{\mu}(x_1) = \widehat{\mu}(x_2) = 0.5$, $\lambda = 1$, and $\xi = 0.5$. The left lower panel plots the left-hand side of (49) as a function of δ, the right lower panel plots the left-hand side of (48) as a function of δ.

C Additional Results

C.1 Comparative statics for Section 4

We use a numerical example to illustrate Proposition 8. Let $M = 3$, $\lambda = 1$, $\widehat{\mu}(x_1) = 0.5$, $\widehat{\mu}(x_2) = 0.35$, $\widehat{\mu}(x_3) = 0.15$, $u_G = 1$, and $u_B = 0$. Figure 7 plots the solutions for different values of ξ. When ξ approaches infinite, the worst-case prior approaches the baseline prior and the robust solution approaches the standard solution for which only action 1 and 2 are considered. When ξ is sufficiently small, action 3 also enters the consideration set. The worst-case prior puts more weight to x_3. As ξ approaches zero, the worst-case prior approaches a uniform distribution over all states.

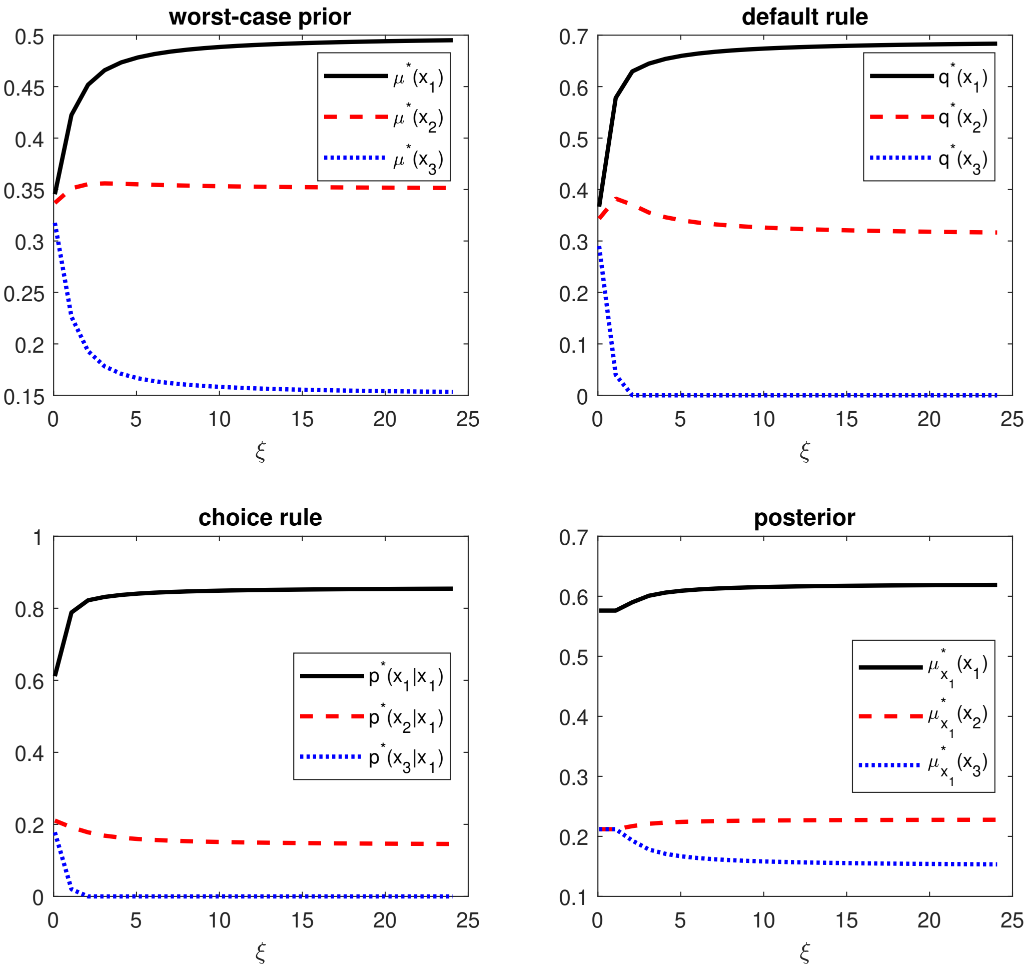

Figure 7: Solutions for the consumer choice problem in Section 4 with different degrees of ambiguity aversion.

D Posterior-Based Approach

In this section, we present an equivalent posterior-based approach for the robust RI problem. In contrast to the rational RI problem, the DM in the robust problem does not have a single posterior that is of interest. The baseline probabilities imply one posterior and the worst-case prior implies another one. While both are interesting constructs, neither is intended to capture the unique posterior beliefs of the DM. The analysis here features the worst-case posterior as a device to provide additional insights into our solution.

Let us first rewrite Problem 3 as

$$J\left(\widehat{\mu}\right)=\min_{\mu\in\Delta(X)}V\left(\mu\right)+\xi R\left(\mu||\widehat{\mu}\right),$$

where $V\left(\mu\right)$ is the value function for the standard RI problem with the prior μ.

Caplin and Dean (2013) and Caplin et al. (2019) propose a posterior-based approach to solve the standard Shannon model for any given prior μ. In the posterior-based approach, the value function is defined as

$$V\left(\mu\right)=\max_{q\in\Delta(A),\mu_a\in\Delta(X),a\in A}\sum_a q\left(a\right)N^a\left(\mu_a\right)-\lambda H\left(\mu\right) \tag{50}$$

subject to

$$\mu\left(x\right)=\sum_a \mu_a\left(x\right)q\left(a\right),\ x\in X, \tag{51}$$

where N^a denotes the net utility function associated with action a and it is defined as

$$N^a\left(\mu_a\right):=\sum_x \mu_a\left(x\right)u\left(x,a\right)+\lambda\mathbb{H}\left(\mu_a\right).$$

We can then reformulate the choice-based robust RI problem as an equivalent posterior-based problem of choosing a worst-case prior $\mu\in\Delta\left(X\right)$, a default rule $q\in\Delta\left(A\right)$, and a posterior probability $\mu_a\in\Delta\left(X\right)$, $a\in A$. Once q and $\{\mu_a\}_{a\in A}$ are determined, we can recover the choice rule as

$$p\left(a|x\right)=\frac{q\left(a\right)\mu_a\left(x\right)}{\mu\left(x\right)},\ \text{if }\mu\left(x\right)>0.$$

Notice that the net utility function $N^a\left(\mu_a\right)$ is concave in μ_a due to the concavity of entropy $\mathbb{H}\left(\mu_a\right)$. However, the objective function in (50) is not jointly concave in q and $\{\mu_a\}_{a\in A}$. Thus one has to solve a concavification problem. This posterior-based approach has a nice geometric interpretation of the solution, which helps understand economic intuition. Specifically, for any given prior μ, one graphs the net utilities associated with all actions and finds the point on the convex hull directly above the prior μ; the optimal posteriors are the points of tangency of the supporting hyperplane at this point and the net utility functions (see Figure 8). This generates the value function $V\left(\mu\right)$ for the standard RI problem without prior ambiguity. On top of this problem, the robust strategy determines the worst-case prior μ^* that minimizes $V\left(\mu\right)$ plus a penalty cost measured by the entropy relative to the baseline prior $\widehat{\mu}$. The robust posteriors and default rule are optimal relative to μ^*.

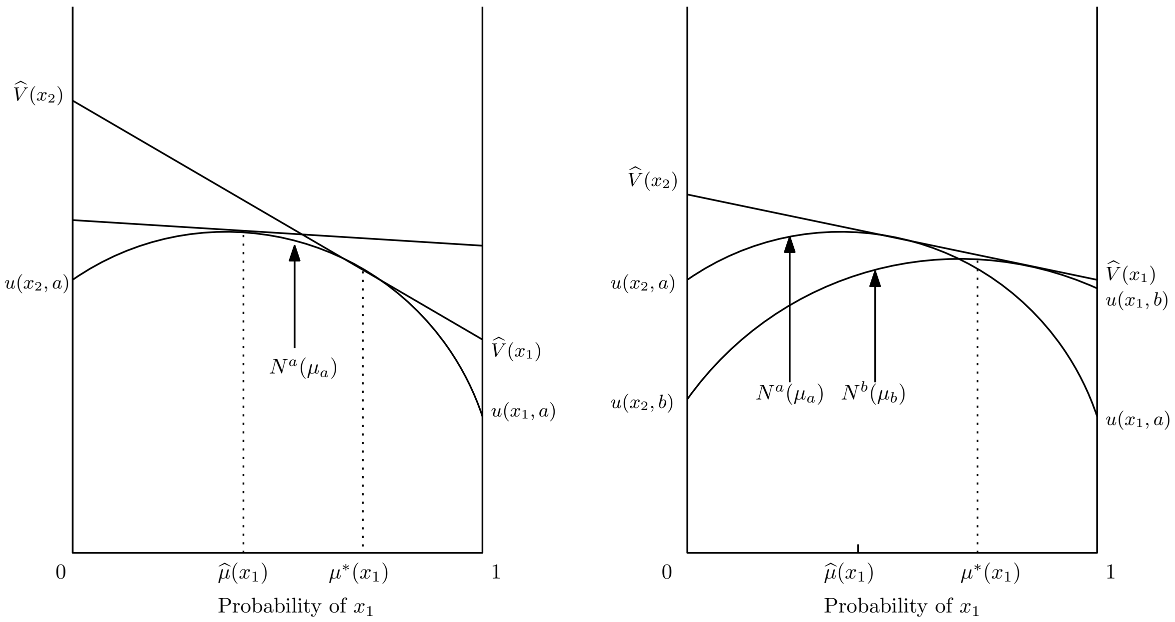

Figure 8: Geometric illustration of standard solutions and robust solutions. The left panel shows the case with two states x_1 and x_2 and one action a. The right panel shows the case with two states x_1 and x_2 and two actions a and b.

By (23), we have

$$V(\mu^*) = \sum_x \mu^*(x)\, v(x) = \sum_x \mu^*(x)\, \widehat{V}(x) + \lambda \sum_x \mu^*(x) \log \mu^*(x),$$

where $\widehat{V}(x) := v(x) - \lambda \log \mu^*(x)$. Miao and Xing (2024) show that $\widehat{V}(x)$ is the height of the supporting hyperplane at the point with $\mu^*(x) = 1$ and $\mu^*(x') = 0$ for all $x' \neq x$. Replacing $v(x)$ in the equation above by the expression from (9), we can show that

$$\mu^*(x) = \frac{\widehat{\mu}(x)^{\frac{\xi}{\lambda+\xi}} \exp\left(-\frac{1}{\lambda+\xi}\widehat{V}(x)\right)}{\sum_{x'} \widehat{\mu}(x')^{\frac{\xi}{\lambda+\xi}} \exp\left(-\frac{1}{\lambda+\xi}\widehat{V}(x')\right)}.$$

Thus the worst-case prior μ^* puts more weight on a lower payoff $\widehat{V}(x)$. The payoff $\widehat{V}(x)$ is related to $v(x)$ and has a better geometric interpretation. Notice that $\widehat{V}(x)$ or $v(x)$ tends to be low when $u(x,a)$ is low for a chosen action a.

The left panel of Figure 8 illustrates the case with two states x_1 and x_2 and one action a. For the standard Shannon model, the tangency point of the supporting plane lies directly above the prior $\widehat{\mu}(x_1)$ so that the optimal posterior is the prior. For the robust Shannon model, the optimal posterior is the same as the worst-case prior $\mu^*(x_1)$ given only one action a. Since $q^*(a) = 1$, we have $v(x) = u(x, a)$ and

$$\mu^*(x) = \frac{\exp(-u(x,a)/\xi)\,\widehat{\mu}(x)}{\sum_x \exp(-u(x,a)/\xi)\,\widehat{\mu}(x)}, \quad x = x_1, x_2.$$

Since $u(x_1, a) < u(x_2, a)$ in the left panel of Figure 8, we have $\mu^*(x_1) > \widehat{\mu}(x_1)$ and the tangent hyperplane tilts down toward state x_1.

The right panel of Figure 8 illustrates the case with two states and two actions. The worst-case prior $\mu^*(x_1)$ lies in the convex hull of the tangency points, which give the optimal posteriors $\mu_a^*(x_1)$ and $\mu_b^*(x_1)$. Then both actions a and b are in the consideration set.

As in Caplin et al. (2019), for an action to be in the consideration set, its net utility must touch the supporting hyperplane. Except in cases of indifference, this means that the net utility function associated with this action would pierce the hyperplane associated with a problem that did not include this action. This is more likely if the net utility associated with this action is higher (i.e. the payoffs are higher) or the hyperplane is lower (i.e. the payoffs to the other actions are lower). The latter case reflects a hedging motive: an action is more likely to be considered when it pays off more in states in which other actions pay off less.

For the standard Shannon model, Caplin and Dean (2015) establish a locally invariant posteriors (LIP) property; that is, optimal posterior distributions are locally invariant to changes in priors in the convex hull of the optimal posteriors. This result is intuitive using the geometric interpretation discussed earlier: local changes of priors in the convex hull of the tangency points of the hyperplane and net utility functions do not change these tangency points. In our robust RI model, these priors are not the primitives of the model. They are the worst-case priors endogenously derived from a robust control problem given a baseline prior $\widehat{\mu}$ and a robustness parameter ξ. Similar to Caplin and Dean (2015), we have the following invariance result. The proof is similar to that of Corollary 1 in Caplin and Dean (2013) and hence is omitted here.

Corollary 12. *Let B and $\{\mu_a\}_{a\in A}$ be the optimal consideration set and posterior distribution for the robust RI problem with the baseline prior $\widehat{\mu}$ and the penalty parameter ξ. If $B' \subset B$ and $\mu_a = \mu'_a$ for any $a \in B'$, then B' and $\{\mu'_a\}_{a\in B'}$ are optimal for a robust RI problem with some baseline prior $\widehat{\mu}'$ and penalty parameter ξ'.*

It merits emphasis that the LIP of Caplin and Dean (2013) does not hold in our robust RI model. More specifically, our invariance result only holds for a combination of $\widehat{\mu}'$ and ξ'. Local changes of baseline priors $\widehat{\mu}$ in the convex hull of the tangency points may affect the optimal posteriors. But a local change of $\widehat{\mu}$ such that the associated worst-case prior μ^* remains in that convex hull will not change the optimal posteriors. See the right panel of Figure 1 for the intuition.

Caplin et al. (2019) establish a converse result that finds exogenous priors associated with any given consideration set. We are unable to provide a similar result. In our model, a baseline prior and a robustness parameter are primitives. Any given consideration set must be generated by both some baseline prior and some robustness parameter.

E Additional Figures for Section 5

In this section, we present the counterpart of Figures 2, 3, and 4 when $r = 7.5$. In contrast to Figure 1, Figures 9, 10, 11 present quantitatively similar results to Figures 2, 3, and 4, when the information acquisition cost is finite.

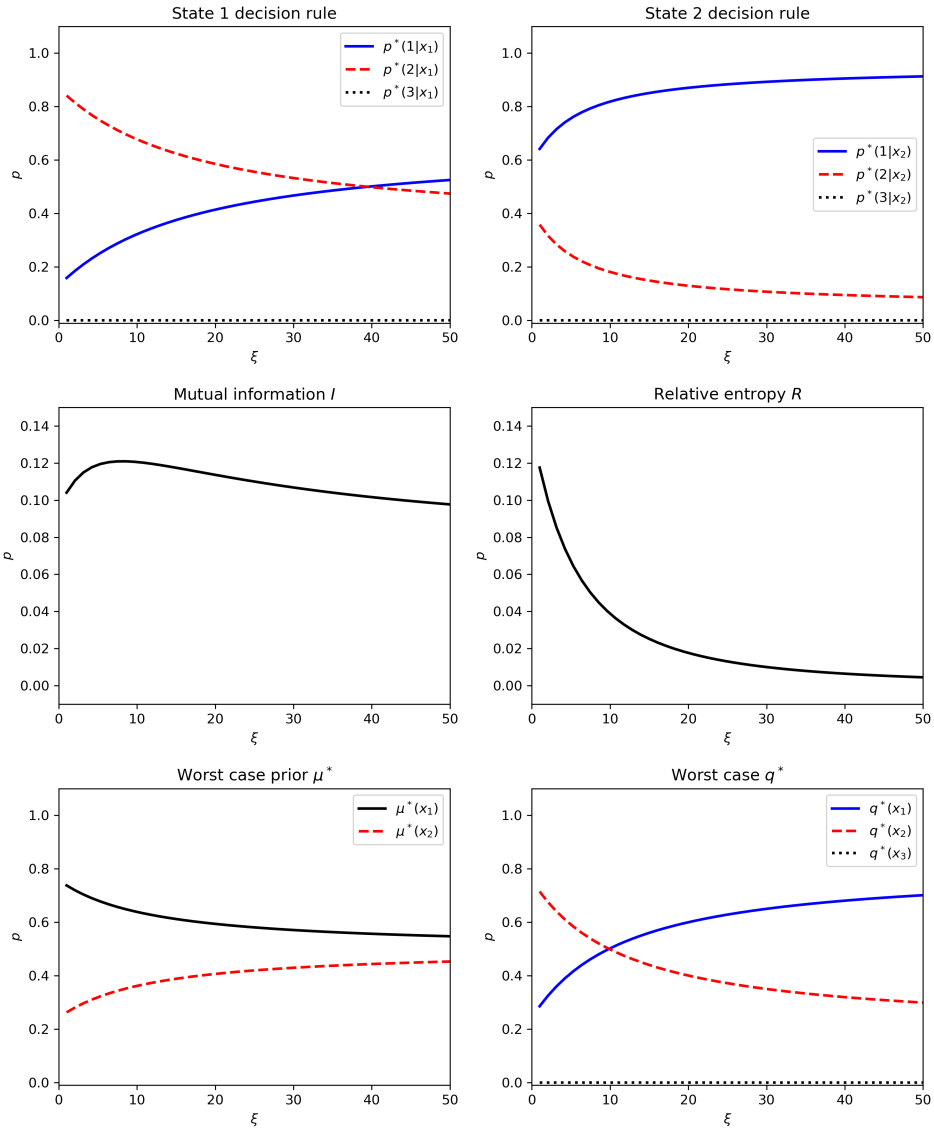

Figure 9: This figure explores the sensitivity to changes in ξ. The attention cost parameter, $\lambda = 10$; utility curvature parameter, $\alpha = 0$, and $\mathsf{r} = 7.5$.

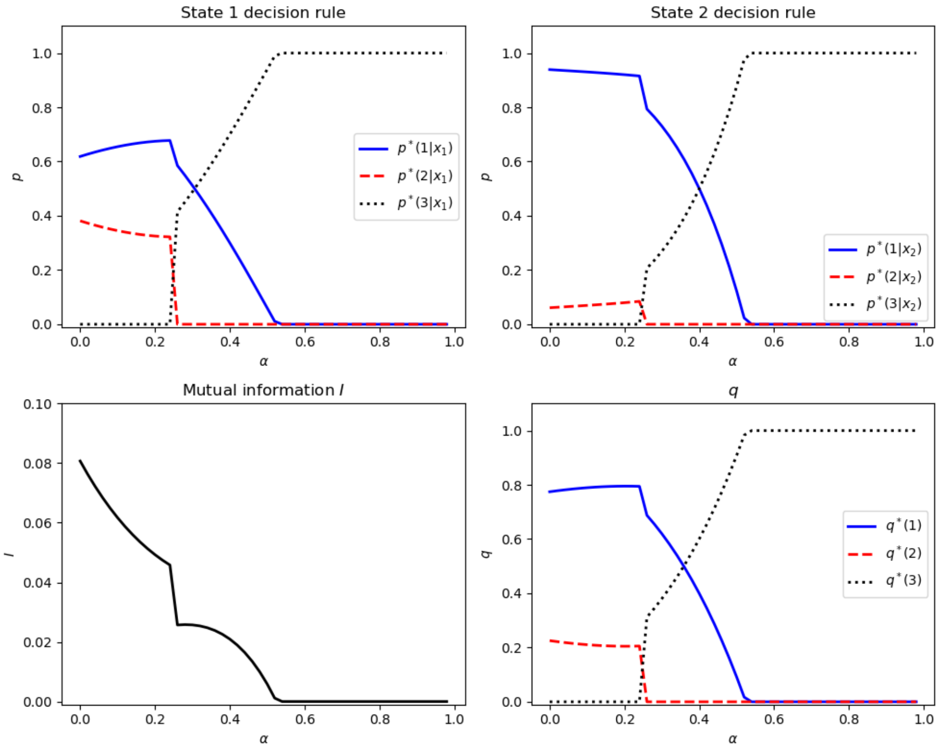

Figure 10: This figure explores the sensitivity to changes in α. The attention cost parameter, $\lambda = 10$; robustness parameter, $\xi = \infty$, and $\mathsf{r} = 7.5$.

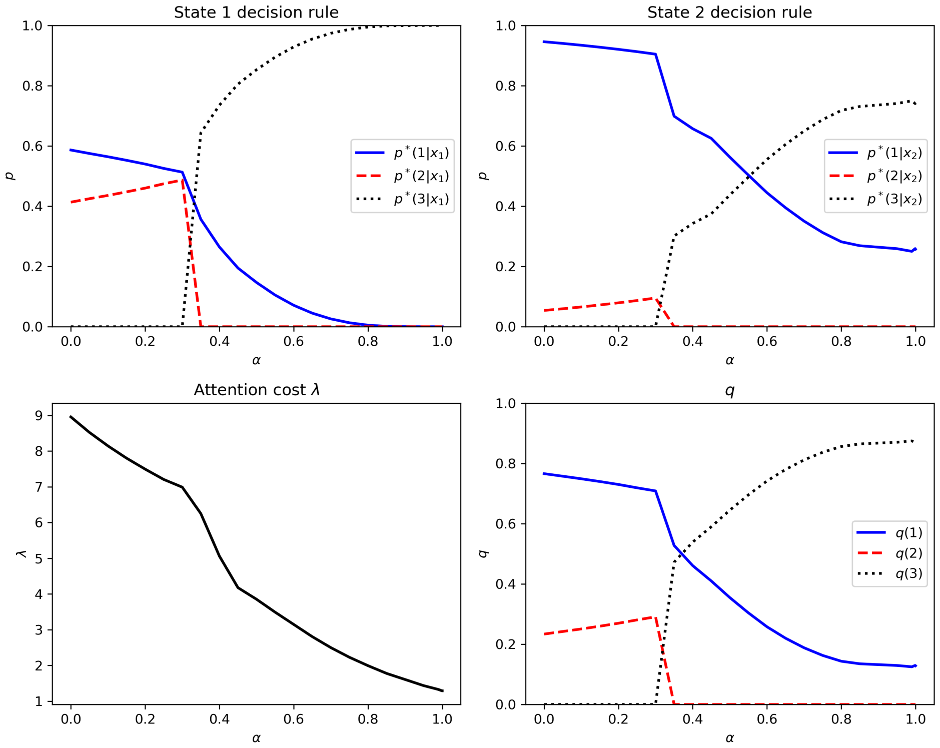

Figure 11: This figure explores the sensitivity to changes in α. The mutual information is constrained by $I \leq \kappa = 0.1$; $\xi = \infty$ and $\mathsf{r} = 7.5$.